

**Universidade do Minho**  
Escola de Engenharia

António João Oliveira da Silva

*Text Mining* e Processamento de Linguagem Natural para  
Interpretação de Notas Clínicas

Dissertação de Mestrado  
Mestrado Integrado em Engenharia e Gestão de Sistemas de  
Informação

Trabalho realizado sob a orientação de:

**Professor Doutor Manuel Filipe Vieira Torres dos Santos**  
(Professor Associado do DSI)

Professor Doutor Carlos Filipe Portela (Professor Auxiliar  
Convidado DSI)

Outubro, 2016

## DECLARAÇÃO REPOSITÓRIUM: DISSERTAÇÃO MESTRADO

Nome: António João Oliveira da Silva

Nº Cartão Cidadão /BI: 14176466/OZZ3

Tel./Telem.: 916 050 039

Correio eletrónico: joaosilvagmr92@gmail.com / a64895@alunos.uminho.pt

Curso: Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação

Ano de conclusão da dissertação: 2016

Escola de Engenharia, Departamento: Departamento de Sistemas de Informação

### TÍTULO DISSERTAÇÃO:

Título em PT : *Text Mining* e Processamento de Linguagem Natural para Interpretação de Notas Clínicas

Título em EN : *Text Mining and Natural Language Processing for Clinical Notes Interpretation*

Orientador: Prof. Dr. Manuel Filipe Santos

Co-orientador: Prof. Dr. Carlos Filipe Portela

Nº ECTS da Dissertação: 45

Classificação em valores (0-20):

Classificação ECTS com base no percentil (A a F):

Declaro sob compromisso de honra que a dissertação agora entregue corresponde à que foi aprovada pelo júri constituído pela Universidade do Minho.

Declaro que concedo à Universidade do Minho e aos seus agentes uma licença não-exclusiva para arquivar e tornar acessível, nomeadamente através do seu repositório institucional, nas condições abaixo indicadas, a minha dissertação/trabalho de projeto, em suporte digital.

Concordo que a minha dissertação seja colocada no repositório da Universidade do Minho com o seguinte estatuto (assinale um):

1.  Disponibilização imediata do trabalho para acesso universal;
2.  Disponibilização do trabalho para acesso exclusivo na Universidade do Minho durante o período de  1 ano,  2 anos ou  3 anos, sendo que após o tempo assinalado autorizo o acesso universal.
3.  Disponibilização do trabalho de acordo com o **Despacho RT-98/2010 c)** (embargo\_\_\_\_\_anos)

Guimarães, 31 /10/2016

Assinatura: \_\_\_\_\_

## Agradecimentos

O desenvolvimento desta dissertação de mestrado assinala o encerramento de um capítulo marcante, tanto na minha vida académica como pessoal. Porém, esta dissertação de mestrado só foi concretizável devido ao apoio de várias pessoas.

Em primeiro lugar, quero agradecer ao meu coorientador, o Professor Doutor Carlos Filipe Portela, por toda a ajuda que me prestou e pela total disponibilidade no desenvolvimento desta tese com conselhos que irei levar após esta dissertação.

Ao Professor Doutor Manuel Filipe Santos, meu orientador, por me ter disponibilizado este tema para a realização da minha dissertação de mestrado.

Ao Centro Hospitalar do Porto – Hospital de Santo António, pelo fornecimento dos dados, pois sem os mesmos não seria possível a realização da dissertação.

À Universidade do Minho e aos docentes, que me deram condições e apoio na realização da dissertação.

Aos meus amigos e colegas de curso, que estiveram presentes nestes cinco anos, por todos os momentos que passamos, e pelas amizades criadas que irão prevalecer após o final deste capítulo da minha vida.

Um agradecimento especial para a Raquel Neves, por todo o apoio, paciência, carinho e motivação que me prestou ao longo destes anos em especial na realização desta dissertação onde me motivou sempre a fazer mais e melhor.

Por fim, um agradecimento especial aos meus Pais por tudo o que fizeram na minha vida, onde sempre me incentivaram a estudar e a lutar por um futuro melhor para mim.

## Resumo

Esta dissertação enquadra-se no âmbito da conclusão do Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação na Universidade do Minho, sendo o seu tema “*Text Mining* e Linguagem Natural Para Processamento de Notas Clínicas”. O tema resultou de uma relação contínua entre um Grupo de Investigação da Universidade do Minho e o Centro Hospitalar do Porto (CHP) – Hospital de Santo António. O CHP tem vários doentes internados nos seus serviços, que estão a receber tratamento para a doença que lhes foi diagnosticada. Para fazer o registo das doenças, dos seus tratamentos e respetivos resultados, os médicos utilizam as Notas Clínicas. Estas armazenam toda a informação relativa ao doente e, frequentemente, a sua análise torna-se complexa porque não existe um padrão de escrita seguido por todos os médicos. O maior desafio deste projeto centrou-se na análise do texto que constitui as Notas Clínicas para, posteriormente, criar um sistema de suporte à decisão e ajudar os médicos a deliberar o melhor para o doente. Para fazer esta análise, foram utilizadas duas técnicas na área da análise e da compressão de texto, o *Text Mining* e o Processamento de Linguagem Natural (PLN). Desenvolveram-se modelos focados na descoberta de conhecimento no texto presente nas Notas Clínicas, com o objetivo de criar padrões de informação. Além destes, também foram levados a cabo outros modelos de previsão de acontecimentos com base na análise de texto, que se revelaram extremamente positivos. Por fim, foi elaborado um dicionário com termos médicos com base na análise das Notas Clínicas fornecidas pelo CHP. Os resultados obtidos foram bastante positivos, onde por exemplo com a criação de modelos de interpretação de texto, foi possível identificar diferentes tipos de diagnóstico efetuados a doentes que sofreram morte cerebral, bem como, a que tipo de doentes isto ocorreu. Estes modelos foram criados com as palavras que constituíam os relatórios médicos, assim como, pelas palavras do dicionário criado. Os modelos de previsão criados também obtiveram resultados bastante satisfatórios, atingindo os 88% de acuidade, o que torna viável a aplicação dos mesmos num ambiente real, como por exemplo o CHP. Todo o trabalho desenvolvido no projeto seguiu na sua vertente prática a metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM) e a metodologia de investigação *Design Science Research* (DSR).

**Palavras-chave:** Sistema de Apoio à Decisão, *Text Mining*, Processamento de Linguagem Natural, Notas Clínicas, CRISP-DM

## **Abstract**

*This dissertation lies framed in the conclusion of the integrated Masters in Engineering and Management of Information Systems at the University of Minho, and the dissertation topic is "Text Mining and Natural Language for Clinical Notes Processing". The theme is the result of an ongoing relationship between the School of the University of Minho, and the Hospital of Porto (CHP) – Hospital de Santo António. The CHP has several patients admitted to their hospital service, and when they are admitted, patients are receiving treatment for their disease. To make the registration of diseases, their treatments and outcomes, doctors use the Clinical Notes. The clinical notes store all information related to the patient, and often, their analysis becomes complex because the doctors do not have a writing standardization. The biggest challenge of this project focused on the analysis of the text that is within clinical notes, to subsequently create a decision support system to help the physicians make the best decision for the patient's treatment. To do the analysis of clinical notes there were used two techniques in the field of analysis and text compression, which are the Text Mining, and Natural Language Processing. There were developed design models which focused on the discovery of knowledge in the text present in the Clinical Notes in order to create patterns of information. In addition to these models, predictive models have also been developed based on the analysis of these texts, and these extremely positive results. Finally, it was developed a dictionary of medical terms based on analysis of Clinical Notes provided by the CHP. The results obtained in the projects were very positive, which for example, with the creation of text interpretation models that were able to identify types of diagnoses made in patients who had suffered brain death, and identify types of patients who had suffered brain death. These models have been created with the words which were in the medical reports as well the words of the dictionary that was developed. The prediction models created also achieved very positive results, with the best models achieving the 88% of accuracy, what makes feasible the application of the models in a real environment. This project followed the Cross Industry Standard Process methodology for Data Mining (CRISP-DM) as a practical methodology, and the research methodology Design Science Research (DSR).*

**Keywords:** *Decision Support System, Text Mining, Natural Language Processing, Clinical Notes, CRISP-DM*

## Índice

1.	Introdução .....	1
1.1.	Enquadramento e Motivação .....	1
1.2.	Objetivos do Projeto.....	2
1.3.	Estrutura do Documento.....	3
2.	Materiais e Métodos.....	5
2.1.	Abordagem Metodológica .....	5
2.1.1.	<i>Design Science Research (DSR)</i> .....	5
2.1.2.	<i>Cross Industry Standard Process for Data Mining (CRISP-DM)</i> .....	7
2.2.	Ferramentas Utilizadas.....	11
3.	Revisão da Literatura .....	13
3.1.	Processo de Pesquisa .....	13
3.2.	Inteligência Artificial.....	14
3.3.	<i>Text Mining</i> .....	16
3.3.1.	<i>Contextualização</i> .....	16
3.3.2.	<i>Contexto Histórico do Text Mining</i> .....	17
3.3.3.	<i>Crescimento do Text Mining</i> .....	18
3.3.4.	<i>Questões legais do DM e Text Mining</i> .....	19
3.3.5.	<i>Knowledge Discovery in Text</i> .....	19
3.3.6.	<i>Modelo de Processamento de Texto Conceptual de Perrin e Petry</i> .....	21
3.3.7.	<i>Técnicas Mais Utilizadas No Text Mining</i> .....	23
3.3.8.	<i>Áreas de aplicação do Text Mining</i> .....	24
3.3.9.	<i>Text Mining na Medicina</i> .....	25
3.3.10.	<i>Fases do Text Mining</i> .....	28
3.4.	Linguagem Natural .....	30
3.4.1.	<i>Contexto Histórico</i> .....	31

3.4.2. Ontologias.....	33
3.4.3. Diferentes Níveis da Linguagem Natural.....	36
3.4.4. Abordagens da Linguagem Natural.....	39
3.4.5. Aplicações do PLN.....	40
3.4.6. PLN na Medicina.....	40
3.5. Notas Clínicas.....	42
3.5.1. Tipos de Notas Clínicas.....	44
3.5.2. Regras das Notas Clínicas (Diários Clínicos).....	44
3.5.3. Guidelines para os Diários Clínicos.....	45
3.5.4. Problemas das Notas Clínicas.....	46
3.6. Sistemas de Apoio à Decisão Clínica.....	47
4. Soluções envolvendo o <i>Text Mining</i> o Processamento de Linguagem Natural e as Notas Clínicas.....	49
4.1. Usar o TM e o PLN para o Processamento de Seguros de Saúde.....	49
4.2. Extração automática em texto livre de micro-organismos e os seus <i>habitats</i> usando <i>workflows</i> de TM.....	51
5. Estudo dos Dados.....	56
5.1. Contexto do Problema.....	56
5.1.1. Morte Cerebral.....	56
5.1.2. Análise de Texto das Notas Clínicas.....	57
5.1.3. Previsão de Morte Cerebral após a Realização de raio-x.....	58
5.2. Análise do Estudo.....	58
5.3. Estudo dos Dados.....	59
5.3.1. Descrição dos dados e extração dos dados Iniciais.....	59
5.3.2. Exploração e Compreensão dos dados.....	60
5.3.3. Qualidade dos Dados.....	67
5.4. Preparação dos dados.....	68

5.5.	Criação do dicionário.....	71
5.5.1.	<i>Processo de Criação do dicionário</i> .....	72
5.5.2.	<i>Dicionários Criados</i> .....	72
A)	<i>Dicionário para o KH Coder em Português e Inglês</i> .....	72
B)	<i>Dicionário para o KNIME em Português</i> .....	73
6.	Criação de Modelos de Análise de Texto utilizando o <i>KH Coder</i> .....	74
6.1.	Seleção dos Dados.....	74
6.2.	Selecionar as Técnicas de Análise .....	75
6.3.	Resultados da Análise.....	76
6.3.1.	<i>Análise sem Dicionário</i> .....	76
A)	<i>Frequência de Palavras</i> .....	76
B)	<i>Análise Hierárquica de Clusters</i> .....	78
C)	<i>Mapa Auto Organizacional</i> .....	82
D)	<i>Coocorrência de Rede sem dicionário</i> .....	84
E)	<i>Análise de Correspondência</i> .....	88
F)	<i>Escala Multidimensional de Termos</i> .....	90
6.3.2.	<i>Análise com Dicionário</i> .....	94
A)	<i>Frequência de Palavras com a Utilização do Dicionário</i> .....	94
B)	<i>Análise Hierárquica de Clusters</i> .....	95
C)	<i>Mapa Auto Organizacional com Dicionário</i> .....	100
D)	<i>Coocorrência de rede com Dicionário</i> .....	101
E)	<i>Análise de Correspondência</i> .....	102
F)	<i>Escala multidimensional de Códigos</i> .....	103
7.	Criação de uma Ontologia .....	104
7.1.	Contexto do Problema .....	104
7.2.	Desenvolvimento da Ontologia .....	104



7.3.	Discussão da Ontologia .....	106
8.	Criação de Modelos de Previsão com o KNIME.....	107
8.1.	Compreensão do Negócio.....	107
8.2.	Compreensão dos Dados.....	107
8.3.	Preparação dos Dados .....	108
8.4.	Modelação .....	109
8.5.	Avaliação .....	114
8.6.	Discussão dos Resultados .....	116
9.	Discussão de Resultados.....	117
9.1.	Resultados Obtidos nos Modelos de Análise criados pelo <i>KH Coder</i> .....	117
A)	<i>Frequência de Palavras</i> .....	117
B)	<i>Análise Hierárquica de Clusters</i> .....	117
C)	<i>Mapa Auto Organizacional</i> .....	120
D)	<i>Coocorrência de Rede</i> .....	122
E)	<i>Análise de Correspondência</i> .....	124
F)	<i>Escala Multidimensional de Termos</i> .....	125
G)	<i>Frequência de Palavras com Dicionário</i> .....	128
H)	<i>Análise Hierárquica de Clusters com Dicionário</i> .....	128
I)	<i>Mapa Auto Organizacional com Dicionário</i> .....	129
J)	<i>Coocorrência de Rede com Dicionário</i> .....	131
K)	<i>Análise de Correspondência com Dicionário</i> .....	131
L)	<i>Escala Multidimensional de Códigos</i> .....	132
9.2.	Resultados Obtidos pelos Modelos de Previsão do KNIME .....	133
A)	<i>Resultados obtidos com a utilização do algoritmo Decision Tree</i> .....	133
B)	<i>Resultados obtidos com a utilização do algoritmo K-Nearest Neighbor</i> .....	134

<i>C) Resultados obtidos com a utilização do algoritmo Decision Tree com cross validation</i>	134
<i>D) Resultados obtidos com a utilização do algoritmo K-Nearest Neighbor com cross validation</i>	135
10. Conclusão.....	136
10.1. Considerações finais .....	136
10.2. Dificuldades Encontradas .....	138
10.3. Trabalho Futuro.....	139
11. Referências.....	140
<b>ANEXOS</b> .....	152

## Índice de Figuras

Figura 1– Ciclos do <i>Design Science Research</i> .....	5
Figura 2 – Fases do CRISP-DM .....	8
Figura 3 – Número de publicações existentes no PubMed com a palavra “ <i>Text Mining</i> ” ou “ <i>Literature Mining</i> ” como palavra-chave ( <i>keyword</i> ) ou estando presente no resumo ( <i>abstract</i> )..	18
Figura 4 – Exemplo de uma Nota Clínica Eletrônica .....	43
Figura 5 – Abordagem Utilizada Pelos Autores .....	53
Figura 6 – Diagrama Entidade-Relação dos Datasets fornecidos .....	66
Figura 7 – Exemplo do Dicionário Estruturado para o <i>KH Coder</i> .....	73
Figura 8 – Exemplo do Dicionário Estruturado para o <i>KNIME</i> .....	73
Figura 9 – Análise Hierárquica de Palavras com termos que aparecem pelo menos vinte vezes no documento .....	79
Figura 10 – Análise Hierárquica de Palavras com termos que aparecem pelo menos quinze vezes no documento .....	80
Figura 11 – Análise Hierárquica de Palavras com termos que aparecem pelo menos dez vezes no documento .....	81
Figura 12 – Mapa Auto Organizacional com termos que aparecem pelo menos vinte vezes no documento .....	82
Figura 13 – Mapa Auto Organizacional com termos que aparecem pelo menos quinze vezes no documento .....	83
Figura 14 – Mapa Auto Organizacional com termos que aparecem pelo menos dez vezes no documento .....	84
Figura 15 – Coocorrência de Rede de Palavras com termos que aparecem pelo menos vinte vezes no documento .....	85
Figura 16 – Coocorrência de Rede de Palavras com termos que aparecem pelo menos vinte vezes no documento .....	86
Figura 17 – Coocorrência de Rede de Palavras com termos que aparecem pelo menos dez vezes no documento .....	87
Figura 18 – Análise de Correspondência com Termos que aparecem pelo menos vinte vezes no documento .....	88

Figura 19 – Análise de Correspondência com Termos que aparecem pelo menos quinze vezes no documento.....	89
Figura 20 – Análise de Correspondência com Termos que aparecem pelo menos dez vezes no documento.....	90
Figura 21 – Escala multidimensional de termos que aparecem pelo menos vinte vezes no documento.....	91
Figura 22 – Escala multidimensional de termos que aparecem pelo menos quinze vezes no documento.....	92
Figura 23 – Escala multidimensional de termos que aparecem pelo menos dez vezes no documento.....	93
Figura 24 – Análise Hierárquica de Clusters dos Temas do Dicionário .....	99
Figura 25 – Mapa Auto Organizacional dos Temas presentes no dicionário.....	100
Figura 26 – Coocorrência de Rede de Palavras do Dicionário .....	101
Figura 27 – Análise de Correspondência de Palavras presentes no Dicionário.....	102
Figura 28 – Escala multidimensional de temas presentes no dicionário .....	103
Figura 29 - Classes da Ontologia.....	105
Figura 30 - Diagrama das Classes da Ontologia.....	106
Figura 31 – Representação Gráfica da diferença de doentes que morreram com e sem oversampling.....	109
Figura 32 – Primeiro conjunto do workflow aplicado no KNIME .....	111
Figura 33 – Segundo conjunto do workflow aplicado no KNIME.....	111
Figura 34 – Terceiro conjunto do workflow utilizado no KNIME .....	112

## Índice de Tabelas

Tabela 1 – Matriz Confusão .....	10
Tabela 2 – Descrição e a identificação das ferramentas de suporte à dissertação .....	11
Tabela 3 – Performance das Abordagens de Dicionário e CRF .....	54
Tabela 4 – Performance da Abordagem utilizada pelos Autores .....	55
Tabela 5 – Performance da Abordagem extração-relação.....	55
Tabela 6 – Composição e descrição dos dados do <i>Dataset</i> REPORT_RX_DATA_TABLE .....	60
Tabela 7 – Composição e descrição dos dados do <i>Dataset</i> OBITOS_DATA_TABLE .....	63
Tabela 8 – Composição e descrição dos dados do <i>Dataset</i> ReportRX_NEW .....	64
Tabela 9 – Estrutura da tabela <i>Report</i> no <i>Microsoft SQL Server</i> .....	68
Tabela 10 – Estrutura da tabela <i>Obitos</i> no <i>Microsoft SQL Server</i> .....	69
Tabela 11 – Estrutura da tabela <i>ReportRXNEW</i> no <i>Microsoft SQL Server</i> .....	69
Tabela 12 – Estrutura da <i>view</i> <i>ReportsObitosInicial</i> .....	70
Tabela 13 – Estrutura da <i>view</i> <i>ReportsObitos</i> .....	70
Tabela 14 – Estrutura da <i>view</i> <i>ReportsVivos</i> .....	71
Tabela 15 – Exemplo de <i>stopwords</i> utilizadas .....	76
Tabela 16 – Extrato da análise de Frequência de Palavras.....	77
Tabela 17 – Extrato do resultado de Frequência de Palavras do Dicionário .....	94
Tabela 18 – Tabela com informação sobre doentes que morreram com e sem <i>oversampling</i> . 108	
Tabela 19 – Algoritmos utilizados no teste e os seus detalhes .....	110
Tabela 20 – Métodos do <i>Workflow</i> .....	112
Tabela 21 – Tabela com os melhores resultados por Algoritmo .....	115

## Índice de Abreviações, Siglas e Acrónimos

ACIS – *Axonwave Content Intelligence System*

ALPAC – *Automatic Language Processing Advisory Committee of National Academy of Science – National Research Council*

CHP – *Centro Hospitalar do Porto*

CIS – *Content Intelligence System*

CRF – *Conditional Random Field*

CRISP-DM – *Cross Industry Standard Process for Data Mining*

CSL – *Concept Specification Language*

cTAKES – *clinical Text Analysis and Knowledge Extraction System*

DM – *Data Mining*

DT – *Decision Tree*

DTCV – *Decision Tree com cross validation*

DSR – *Design Science Research*

FRS – *Framingham Risk Score*

GATE – *General Architecture for Text Engineering*

HITEx – *Health Information Text Extraction*

I2B2 – *National Center for Biomedical Computing, Informatics for Integrating Biology & the Bedside*

IA – *Inteligência Artificial*

IBM – *International Business Machines*

ICF – *Internacional Classification of Functioning, Disability, and Health*

KNN – *K-Nearest Neighbor*

KNNCV – *K-Nearest Neighbor com cross validation*

KDD – *Knowledge Discovery in Data*

KDT – *Knowledge Discovery in Text*

LSP – *Linguistic String Project*

MALLET – *Machine Learning for Language Toolkit*

MedLEE – *Medical Language Extraction and Encoding System*

MEDLINE – *Medical Literature Analysis and Retrieval System Online*

medKAT/P – *Medical Knowledge Analysis Tool*

MedTAS/P – *Medical Text Analysis System*

MLP – *Medical Language Processor*

NLG – *Natural Language Generation*

PLN – *Processamento de Linguagem Natural*

TC – *Tomografia Computadorizada*

TM – *Text Mining*

VSD – *Vaccine Safety Datalink Project*

# 1. Introdução

Este capítulo contém uma introdução ao tema desta dissertação, a sua contextualização e as suas motivações. Posteriormente, é apresentada a questão de investigação e os objetivos da dissertação. Por fim, é descrita a estrutura do documento com o objetivo de facilitar a leitura do mesmo.

## 1.1. Enquadramento e Motivação

A medicina é uma vertente da área da saúde que trabalha com o objetivo de prevenir e descobrir cura para as doenças que afetam os seres humanos. As Notas Clínicas são tradicionalmente registadas pelos médicos, que procedem ao seu registo durante o período em que os doentes estão internados (notas diárias), no momento em que são internados ou saem do hospital (notas de admissão/notas de alta), bem como, durante as suas consultas.

Atualmente os médicos do Centro Hospitalar do Porto (CHP) fazem o registo das Notas Clínicas em regime de texto livre, o que dificulta a análise e interpretação das mesmas. Assim, é necessário interpretá-las e criar um dicionário clínico que permita, de forma automática, compreender o que é escrito sobre o doente e ajudar os médicos a tomar uma decisão rápida e eficaz.

Para interpretar as Notas Clínicas, utilizou-se o *Text Mining* (TM) e a Linguagem Natural. O TM refere a extração de informação de dados estruturados, neste caso, um conjunto de texto num espaço vetorial com o objetivo de descobrir um novo conhecimento (Ishikiryama, Miro, Francisco, & Gomes, 2015). A Linguagem Natural é uma área que se define pela utilização de conhecimentos sobre a língua e a comunicação humana, tanto para a comunicação com sistemas computacionais, como para melhorar a comunicação entre os seres humanos (D. Santos, 2001). A Linguagem Natural, ou o Processamento de Linguagem Natural (PLN), é uma área que provém da Inteligência Artificial (IA) e tem como objetivo converter os dados armazenados eletronicamente e não organizados/padronizados em dados que sejam compreensíveis de modo a que seja possível fazer uma análise e interpretação fiáveis dos dados recolhidos.



Este trabalho foi realizado no Departamento de Sistemas de Informação da Universidade do Minho em conjunto com o CHP que, por sua vez, é um hospital central e universitário pela sua associação ao Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto, que visa a excelência em todas as suas atividades numa perspetiva global e integrada da saúde. A sua missão é a prestação de cuidados de saúde humanizados, competitivos e de referência, promovendo a articulação com os outros parceiros do sistema, a valorização do ensino pré e pós-graduado e da formação profissional, a dinamização e incentivo à investigação e desenvolvimento na área de Saúde.

O tema desta dissertação surgiu no sentido de tentar encontrar tratamentos padronizados para os doentes, e para introduzir o TM e o PLN na área da saúde de forma a que ambos tenham um papel fulcral na mesma. A importância de um tratamento padronizado para os doentes é enorme pois, além de poder reduzir custos nos hospitais, poderá sobretudo tornar mais eficazes os tratamentos aos doentes, de maneira a reduzir, por exemplo, o seu tempo de internamento. Além disso, a criação de uma ferramenta responsável pela análise de texto livre poderá ser um importante passo para o estudo de outras questões noutras áreas que não sejam do ramo da saúde, solucionando outros eventuais problemas que podem acontecer na área da Saúde.

## 1.2. Objetivos do Projeto

No âmbito deste estudo, podemos identificar a seguinte questão de investigação como aquela que esta dissertação procurou dar resposta:

*De que forma o processo de decisão clínico poderá beneficiar da introdução do Text Mining e Linguagem natural para a análise de texto livre?*

O principal objetivo desta dissertação foi a interpretação das Notas Clínicas e a criação de um dicionário clínico, que permitisse de forma automática interpretar o que é escrito sobre um doente, e ajudar os médicos a tomar uma decisão rápida e eficaz. Os modelos foram produzidos e testados através de modelos de TM, utilizando dados reais provenientes do CHP. Como objetivos específicos desta dissertação apresentaram-se a:

- Criação de um sistema que apoie a decisão na Saúde;
- Ajuda na padronização dos tratamentos aos doentes;
- Introdução do TM e do PLN na área da Saúde;
- Criação de um sistema que utilize TM;

- Criação de um sistema que faça a análise de texto livre.

Para esta dissertação foi projetado como resultado criar uma ferramenta capaz de analisar e interpretar Notas Clínicas, criar novos algoritmos de interpretação de informação clínica, elaborar um dicionário clínico, e explorar um novo conhecimento na área dos Sistemas de Informação aplicados à saúde.

### 1.3. Estrutura do Documento

O documento contém um estudo realizado sobre o TM e a Linguagem Natural para interpretação das Notas Clínicas. Para uma melhor compreensão do mesmo, este foi organizado pelos seguintes capítulos:

- **Capítulo 1 – Introdução** – Onde é apresentado ao leitor um enquadramento sobre o tema a ser estudado, o seu ambiente e o porquê do estudo do tema, a motivação do mesmo, os objetivos, e os resultados esperados.
- **Capítulo 2 – Metodologias** – Onde são descritas as metodologias práticas e de investigação que foram utilizadas no desenvolvimento desta dissertação. Por fim, é realizada uma síntese das ferramentas utilizadas na mesma.
- **Capítulos 3 e 4 – Revisão de Literatura** – Onde está presente o estado da arte de todos os temas e áreas relacionados com este projeto. Primeiramente, é abordada a IA, a sua história e as suas bases. No TM, é realizada uma contextualização histórica do mesmo, o seu conceito, as suas técnicas mais utilizadas, o *Knowledge Discovery in Text* (KDT), bem como, o seu estado na Medicina. Também no PLN foi abordada a contextualização histórica, o conceito, o seu estado na medicina, os seus diferentes níveis e ontologias existentes. Nas notas, foi efetuada uma revisão sobre a definição de Notas Clínicas, identificação dos seus diferentes tipos, bem como, os seus padrões de escrita e problemas da interpretação. Dentro do capítulo da Revisão da Literatura, elaborou-se a descrição sucinta de soluções existentes que eram parecidas com o projeto em si e que se revelaram úteis na realização do mesmo.
- **Capítulo 5 – Estudo dos dados** – neste capítulo foi efetuado um estudo dos dados fornecidos, com o objetivo de verificar a viabilidade dos mesmos e de tratar as Notas Clínicas, de forma a manter as que contêm a informação relevante para o projeto. Foi ainda abordada a criação de um dicionário com termos médicos, com base na análise

de um conjunto de Notas Clínicas, cujo objetivo foi uma melhor categorização da informação das mesmas, o que se traduziu na criação de padrões de informação, bem como, na previsão de acontecimentos.

- **Capítulo 6 – Criação de Modelos de Análise de Texto Utilizando do *KH Coder*** – Este presente capítulo contém o estudo que se focou na análise e obtenção de informação existente nas notas clínicas com o objetivo de criar padrões de informação sobre as mesmas. Foram efetuadas diversas análises aos dados estudados com o objetivo de encontrar os modelos que ofereciam informações mais objetivas sobre as notas clínicas.
- **Capítulo 7 – Criação de Modelos de Previsão com o *KNIME*** – Neste capítulo foi desenvolvido um estudo cujo objetivo era a previsão de morte cerebral com base nas análises feitas as notas clínicas. Neste capítulo estão descritos todos os passos do estudo, como a compreensão do negócio e dos dados, a preparação dos dados, a modelação e uma avaliação breve aos resultados obtidos
- **Capítulo 8 – Discussão de Resultados** – Este capítulo contém uma discussão aprofundada aos resultados obtidos nos estudos anteriores, onde todos os resultados considerados para esta dissertação foram analisados, e no fim foi feita uma análise crítica sobre os resultados obtidos nesta dissertação.
- **Capítulo 9 – Conclusão** – Por fim, são apresentadas as considerações finais da dissertação, as dificuldades encontradas na sua realização e é efetuada a reflexão sobre o trabalho futuro que esta dissertação poderá ter.
- **Referências** – Este capítulo contém todas as referências bibliográficas que se revelaram úteis para a realização desta dissertação.

## 2. Materiais e Métodos

Neste capítulo são descritas as metodologias utilizadas nesta dissertação. Primeiramente, detalha-se a metodologia de investigação e, posteriormente, é realizada a descrição da metodologia prática. No final, estão discriminadas as ferramentas utilizadas para o desenvolvimento desta dissertação.

### 2.1. Abordagem Metodológica

No desenvolvimento da dissertação foram utilizadas duas metodologias, uma de investigação, e outra orientada para projetos em *Text Mining* (TM). A metodologia de Investigação adotada foi a *Design Science Research* (DSR) e a metodologia aplicada ao estudo de TM foi o *Cross Industry Standard Process for Data Mining* (CRISP-DM). Para uma utilização correta das metodologias, estas foram respeitadas em todas as fases do projeto.

#### 2.1.1. *Design Science Research* (DSR)

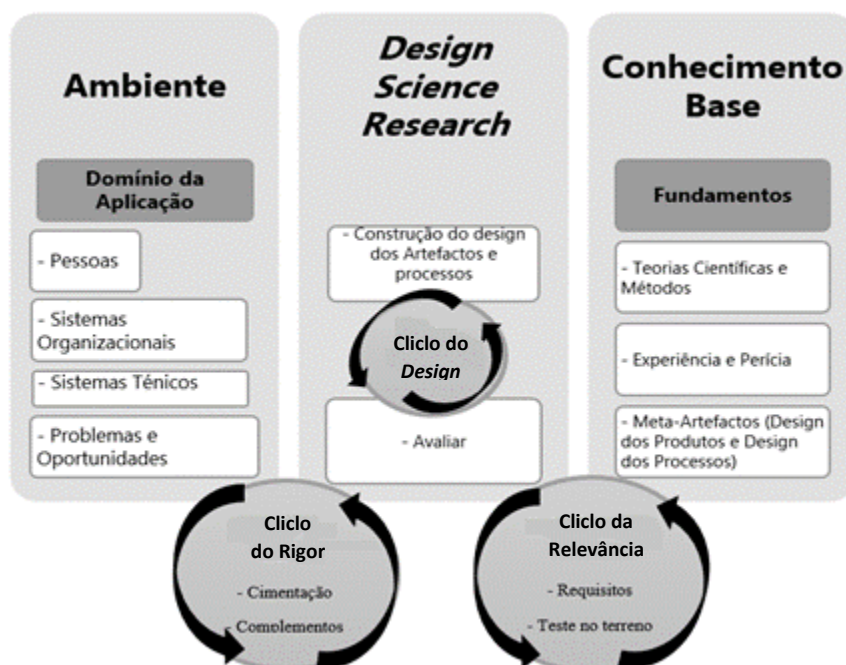


Figura 1– Ciclos do Design Science Research

A metodologia de investigação utilizada no projeto foi o DSR. Compreender e comunicar o processo de DSR é essencial, não só para incentivar a aceitação entre os profissionais de Sistemas de Informação, mas também para estabelecer a credibilidade do DSR nos Sistemas de informação como metodologia *major* entre os investigadores de *Design Project* nos diversos campos da engenharia, arquitetura, as artes, e outras comunidades orientadas para o *design* (Hevner, 2007).

Como se pode observar na Figura 1, o DSR divide-se em três ciclos importantes: o Ciclo da Relevância, o Ciclo do Rigor e o Ciclo do *Design*. O Ciclo da Relevância preenche o ambiente contextual do projeto de pesquisa com as atividades da ciência de *design*. O Ciclo do Rigor liga as atividades da ciência de *design* com a base de conhecimento dos fundamentos científicos, experiência, e conhecimentos que informam o projeto de pesquisa. O Ciclo do *Design* opera entre as atividades do núcleo de construção e avaliação dos artefactos de *design* e nos processos da pesquisa (Hevner, 2007).

Este trabalho, numa alusão à aplicação do DSR, teve como principal problema o facto de o tratamento médico aos doentes não ser padronizado, e estes precisarem de um sistema que os apoie a tomar a decisão mais correta. Para isso acontecer, foi necessário construir um artefacto (utilizando TM e Linguagem Natural) que, com base nos recursos existentes (essencialmente texto livre), fosse possível determinar os melhores procedimentos, criando assim uma possível solução para o problema inicialmente formulado.

No Paradigma de *Design Science*, o conhecimento e o entendimento de um determinado problema e a sua solução são alcançados através da construção da aplicação definida pelo artefacto desenhado (Hevner, March, & Park, 2004). Hevner et al. (2004) definiu sete *guidelines* para o *Design Science in Information Systems Research*, e elas são, por esta ordem:

- **Design como um Artefacto** – O DSR deve produzir um artefacto viável na forma de um modelo, método ou instanciação. – Os artefactos criados no âmbito desta dissertação foram modelos de análise e de descoberta de informação em documentos de texto, e a criação de modelos de previsão usando as técnicas de TM.
- **A Relevância do Problema** – O Objetivo do DSR é desenvolver soluções tecnológicas para problemas de negócios relevantes e importantes. – Nesta fase foi efetuado o levantamento e interpretação do problema que tinha por base a descoberta de conhecimento em ficheiros de texto.

- **Avaliação do *Design*** – A utilidade, qualidade, e eficácia do artefacto devem ser demonstradas de forma rigorosa por via de métodos de avaliação bem executados. – Este passo focou-se em verificar a viabilidade das técnicas utilizadas bem como a confirmação dos resultados obtidos.
- **Contribuições da Pesquisa** – Um projeto eficaz em DSR deve disponibilizar contribuições claras e verificáveis nas áreas do *design* do artefacto, fundamentos e/ou metodologias. – Neste ponto foi realizado um levantamento dos conceitos e das metodologias que foram utilizadas para a realização da dissertação.
- **Rigor da Pesquisa** – No DSR tem de se verificar a aplicação de métodos rigorosos, tanto na construção como na avaliação do artefacto. – Nesta dissertação, este passo descreveu o processo de pesquisa utilizado para a realização do estado da arte.
- ***Design* como um Processo de Pesquisa** – a pesquisa por um artefacto eficaz requer a utilização de meios disponíveis para alcançar os fins desejados, ao mesmo tempo que tem que satisfazer as leis do ambiente do problema. – Nesta etapa foi efetuado um levantamento do estado da arte existente nas áreas de aplicação desta dissertação, com o objetivo de criar modelos eficazes para a análise e previsão de acontecimentos, tendo como base as Notas Clínicas.
- **A Comunicação da Pesquisa** – Os resultados do DSR devem ser apresentados para as pessoas mais orientadas ao meio tecnológico, como as pessoas mais orientadas à gestão (ex. artigos, relatórios, apresentações). – Por fim, neste passo foram escritos artigos científicos [(Silva, Portela, Santos, Abelha, & Machado, 2016), (Silva, Portela, Santos, Machado, & Abelha, 2016)] que permitiram disseminar as soluções desenvolvidas e os resultados atingidos com esta dissertação.

### 2.1.2. *Cross Industry Standard Process for Data Mining* (CRISP-DM)

A metodologia CRISP-DM é apresentada na Figura 2 e descreve o processo de DM em seis fases: o estudo do negócio, o estudo dos dados, a preparação dos dados, a modelação, a avaliação e a implementação. Esta metodologia tem inúmeras vantagens quando aplicada a projetos de DM, tais como: uma maior celeridade, custos de execução menores, maior segurança e maior exequibilidade e viabilidade dos projetos (Santos & Azevedo, 2005).

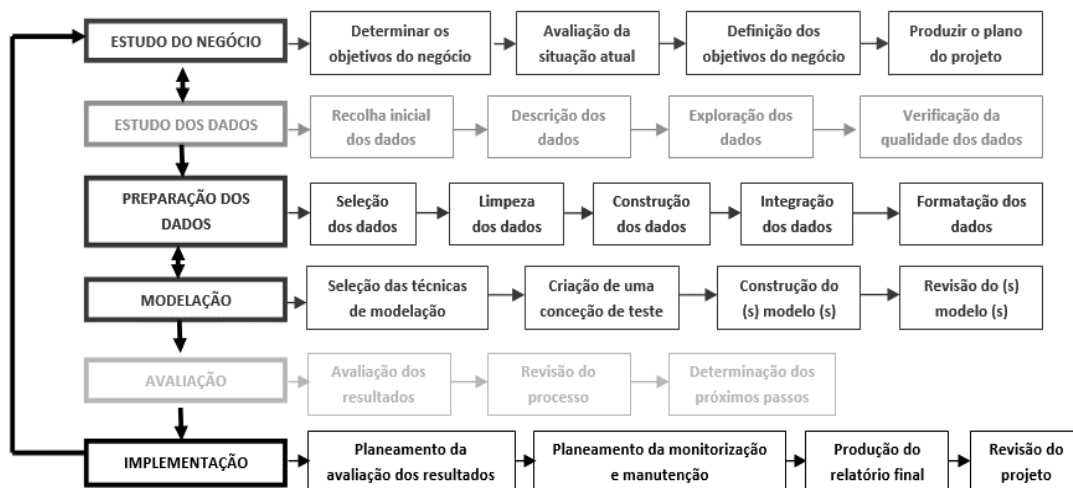


Figura 2 – Fases do CRISP-DM (Adaptado de Santos & Azevedo, 2005)

A metodologia CRISP-DM é constituída por seis fases principais que irão ser descritas nos seguintes pontos abaixo. As descrições são realizadas com base no livro de Santos & Azevedo (Santos & Azevedo, 2005):

- **Estudo do negócio** – A primeira etapa do modelo CRISP-DM, como vimos anteriormente, é o denominado “Estudo do negócio”. É neste ponto que se avalia a real necessidade de empreender o projeto de *Data Mining* (DM), que se compreende o problema, se definem os objetivos a atingir e se avaliam os meios necessários e disponíveis. – Nesta dissertação, o estudo do negócio focou-se em avaliar o problema que envolvia as Notas Clínicas e a morte cerebral. Em termos de DM, foi estudada a viabilidade da criação de modelos de previsão de morte cerebral com base em Notas Clínicas (raio-x).
- **Estudo dos Dados** – A etapa referente ao Estudo dos dados comporta quatro tarefas principais que são a: Recolha dos Dados, a Descrição dos Dados, a Exploração dos Dados e a Verificação da Qualidade dos Dados. – Nesta fase foi realizado um estudo dos mesmos, onde inicialmente foi efetuada uma recolha dos dados fornecidos pelo Centro Hospitalar do Porto (CHP) – Hospital de Santo António. Após isto, foi realizada a sua análise, onde estes foram explorados de modo a conhecer o que significava cada campo e, posteriormente, foi realizado o seu tratamento cujo objetivo era eliminar dados nulos e dados cuja integridade não estivesse assegurada (sem relação entre os *datasets*).

- **Preparação dos dados** – Esta fase abrange todas as atividades relativas à construção do conjunto final de dados, aquele que será usado na ferramenta de modelação, que inevitavelmente será objeto de várias otimizações. Estão incluídas a seleção de tabelas, registos e atributos, bem como, a transformação e limpeza dos dados a usar na ferramenta de modelação. – Nesta etapa os dados foram preparados de modo a serem utilizados pelo KNIME, e também de modo a que a máquina onde foram realizados os testes pudesse aguentar com a carga de dados. Os dados foram reduzidos para uma amostra de dados ontem só continham os raio-x realizados entre 2009 e 2010.
- **Modelação** – É nesta fase que se selecionam as várias técnicas de modelação e em que há um ajuste dos parâmetros, de forma a otimizar os resultados. Nesta seleção há que atender não só à adequação da técnica ao problema de DM, mas também aos requisitos específicos que algumas destas técnicas têm, e só depois é possível submeter os dados previamente preparados na fase anterior para a modelação, aplicando o modelo ao conjunto de dados. – Na modelação foram descritos os modelos de TM desenvolvidos para realizar a previsão de morte cerebral. Durante esta fase, foram estabelecidos os requisitos mínimos para a viabilidade dos modelos. Ao todo foram desenvolvidos cinquenta e seis modelos de TM com base na seguinte fórmula (quatro técnicas de *Mining* x um alvo x dois métodos de amostragem x sete cenários).
- **Avaliação** – Nesta fase avalia-se a utilidade dos modelos, reveem-se os passos executados e verifica-se se permitem atingir os objetivos do negócio. Os três passos desta fase são a Avaliação dos Resultados, a Revisão do Processo e a Determinação dos Próximos Passos. – Na avaliação, foram explicitados os resultados obtidos pelos modelos criados, onde só foram considerados os valores que cumpriam os requisitos estabelecidos na preparação.
- **Implementação** – Nesta fase planeia-se a avaliação dos resultados, onde se define a estratégia para a implementação dos resultados de DM, incluindo os passos e a forma de os executar; planeia-se a monitorização e manutenção; produz-se o relatório final e efetua-se a revisão do projeto, que foi precisamente o que fizemos ao elaborar este relatório final. – Por fim, a implementação refere-se à implementação dos dados num ambiente real, neste caso, aplicar os modelos que cumpriram todos os requisitos mínimos no CHP. Este passo não irá ser executado neste projeto, mas os modelos criados estão prontos para serem introduzidos no sistema INTCare.



Os resultados dos modelos de previsão criados, irão ter a forma de uma Matriz Confusão, como está representada na tabela 1 e, de modo a obter conhecimento desta, foi necessário fazer um estudo da mesma.

Tabela 1 – Matriz Confusão

	n' (Previsão)	n' (Previsão)
p (Atual)	Verdadeiros Positivos	Falsos Negativos
n (Atual)	Falsos Positivos	Verdadeiros Negativos

Witten, Frank & Hall, (2011) definem os Verdadeiros Positivos (VP) e os Verdadeiros Negativos (VN) como as classificações corretas. O Falso Positivo (FP) é quando o resultado é previsto incorretamente como Positivo (p) quando o seu valor atual é negativo (n). Um Falso Negativo (FN) é quando o resultado obtido é Negativo (n), quando o seu valor atual é Positivo (p).

Para fazer uma avaliação mais precisa dos modelos, foram escolhidas quatro análises que demonstram com maior detalhe a precisão dos modelos. Essas avaliações são a **Acuidade**, o **Erro**, a **Sensibilidade** e a **Especificidade**.

A **Acuidade** é a divisão ente valores Verdadeiros das amostras e os valores Falsos, da mesma, e o resultado é a percentagem de acertos do modelo, a expressão da Acuidade é a seguinte:

$$Acuidade = 100 \times \frac{VP + VN}{(VP + VN + FP + FN)}$$

O **Erro** é a percentagem de acertos errados do modelo, e é representado pela seguinte expressão:

$$Erro = 100 - Acuidade$$

A **Sensibilidade** avalia a eficácia do classificador ao reconhecer as amostras positivas, e é definido na expressão abaixo (Halkidi & Vazirgiannis, 2005):

$$Sensibilidade = 100 \times \frac{VP}{VP + FN}$$

A **Especificidade** mede a eficácia do classificador em reconhecer amostras negativas e é definida como (Halkidi & Vazirgiannis, 2005):

$$Especificidade = 100 \times \frac{VN}{VN + FP}$$

## 2.2. Ferramentas Utilizadas

No desenvolvimento desta dissertação, foram utilizadas algumas ferramentas para alcançar os objetivos propostos. Assim, a tabela 2 contém a descrição e a identificação das ferramentas que deram suporte a esta dissertação.

Tabela 2 – Descrição e a identificação das ferramentas de suporte à dissertação

Máquina	Especificações
<i>ASUS N55SL</i>	Processador: <i>Intel Core-i7 2670QM CPU @ 2.20Ghz</i> RAM: 4GB DDR3 Disco Rígido: 500 GB HDD Sistema Operativo: <i>Microsoft Windows 10 Home (64 Bits)</i>
Software	Funcionalidade
<i>Microsoft Excel 2016</i>	Está integrado no pacote de ferramentas do <i>Microsoft Office</i> . Esta ferramenta permite a leitura e exploração de folhas de cálculo. Neste projeto, este <i>software</i> foi utilizado para explorar e estudar os dados fornecidos pelo CHP. Serviu também para ler as folhas de cálculo antes de serem analisadas pelo KNIME.
Bloco de Notas	Serve para escrever texto que pode ser de todo o tipo, como anotações rápidas, texto corrido, ou parágrafos inteiros. Neste projeto serviu para a criação dos dicionários, para preparar os dados para o <i>KH Coder</i> , e também para escrever pequenas notas sobre a organização do trabalho.
<i>SQL Server 2014 Express (Management Studio)</i>	É disponibilizado pela <i>Microsoft</i> como um sistema de gestão de base de dados. Para esta dissertação este <i>software</i> foi usado para armazenar as bases de dados extraídas do <i>Oracle SQL Developer</i> e criar as <i>views</i> desejadas.

Oracle SQL Developer	É uma ferramenta utilizada para a criação e gestão de bases de dados de uma maneira simplificada. Neste projeto, este software foi utilizado para extrair os dados recolhidos do CHP – Hospital de Santo António.
KH Coder 3.Alpha.07b	É uma ferramenta <i>open source</i> para análises quantitativas de texto, e de TM e também é utilizada para linguística computacional. Tem uma grande vantagem de poder analisar texto em várias línguas como Japonês, Inglês, Francês, Alemão, Italiano, Português e Espanhol (KH Coder Index Page, 2016).
KNIME 3.1.2	O KNIME é uma ferramenta <i>open source</i> que permite fazer várias análises a dados e texto, bem como também pode integrar sistemas de <i>Business Intelligence</i> , <i>Big Data</i> , e muitas mais aplicações através de <i>plugins</i> que são fornecidos pela Equipa que desenvolve a Ferramenta.
Protégé 5.1.0	O <i>Protégé</i> é um <i>software</i> desenvolvido pelo <i>Stanford Center for Biomedical Informatics Research</i> e é um recurso para a construção de ontologias e bases de conhecimento principalmente na área da medicina.

## 3. Revisão da Literatura

Este capítulo tem como objetivo descrever os conceitos teóricos e científicos relacionados com o desenvolvimento deste projeto, bem como, a revisão da literatura que envolve os mesmos. Em primeiro lugar, é descrito o processo de pesquisa que envolveu a revisão de literatura, bem como, a descrição dos métodos utilizados para a pesquisa. Posteriormente, são abordados os temas que envolvem a dissertação, como a Inteligência Artificial (IA), Sistemas de Apoio à Decisão Clínica (SADCs) e as Notas Clínicas. Além de abordar os temas, são descritos os processos relacionados com a descoberta de conhecimento como o Text Mining (TM) e o Processamento de Linguagem Natural (PLN).

### 3.1. Processo de Pesquisa

Para o desenvolvimento da revisão da literatura, foram utilizados bastantes motores de pesquisa e bases de dados informacionais, que continham artigos científicos publicados, livros, capítulos de livros etc. Nos diversos *sites* utilizados, estes foram os que foram utilizados com maior frequência:

- *ScienceDirect;*
- *Google Scholar;*
- *Google;*
- *Springer;*
- *Scopus.*

As palavras e termos para a pesquisa foram variando conforme o que se encontrava com base em cada pesquisa. No entanto, dentro do grande leque de palavras-chave utilizadas, destacam-se as seguintes:

- *Clinical Notes;*
- *Text Mining;*
- *Natural Language Processing;*
- *Health Records Documentation;*
- *Text Mining in Medicine;*
- *Natural Language in Medicine.*

Também foram realizadas pesquisas em língua portuguesa com estes termos acima devidamente traduzidos, mas não produziram grandes resultados ao desenvolvimento do documento.

Os critérios de seleção dos artigos e livros tinham como ponto de partida a leitura do *abstract* destes pois, ao ler o mesmo, já se criava uma ideia sobre o que o artigo abordava e o que continha, e por isso foi inicialmente por aí que se realizava a seleção. Quando o *abstract* não era conclusivo, foi necessária uma leitura mais aprofundada do artigo, para ver se este se enquadrava no trabalho que se estava a desenvolver.

Durante a pesquisa foram utilizados alguns artigos com mais de dez anos de existência, mas como os conceitos abordados dos mesmos ainda eram atuais nos dias de hoje, por isso foram escolhidos para integrarem esta revisão de literatura.

### 3.2. Inteligência Artificial

Existem várias abordagens de IA e estas definições variam ao longo de duas dimensões principais. Existem abordagens que estão focadas no pensamento e raciocínio e abordagens que estão focadas no comportamento (Russel & Norvig, 2003).

As abordagens de IA podem ser descritas da seguinte forma:

- **Sistemas que pensam como humanos** – O novo esforço entusiasmante para fazer os computadores pensar... máquinas com mentes, no seu sentido literal” (Haugeland, 1985); [A automatização] atividades que associamos como pensamento humano, atividades como a criação de decisão, resolução do problema, aprendizagem (Bellman, 1978).
- **Sistemas que agem como humanos** – a arte de criar máquinas que executam funções que requerem inteligência quando são executadas por pessoas (Kurzweil, 1990); o estudo de como fazer os computadores fazerem coisas que, neste momento as pessoas fazem melhor (Rich & Knight, 1991).
- **Sistemas que pensam racionalmente** – O estudo das faculdades mentais através de modelos computacionais (Charniak & McDermott, 1985); O estudo de computações que fazem possível o entendimento, a razão e a ação (Winston, 1992).

- **Sistemas que agem racionalmente** –Inteligência computacional é o estudo do *design* dos agentes inteligentes (Poole, Mackworth, & Goebel, 1998); preocupa-se com o comportamento inteligente em artefactos (Nilsson, 1998).

Historicamente, as quatro abordagens para a IA têm sido seguidas. Como se poderia esperar, existe uma tensão entre as abordagens centradas em torno de seres humanos e abordagens centradas em torno de racionalidade. Uma abordagem centrada no ser humano deve ser uma ciência empírica, envolvendo hipótese e confirmação experimental. A abordagem racionalista envolve uma combinação de matemática e engenharia. Cada abordagem desacredita outras abordagens, mas também já se ajudaram na sua viabilidade (Russel & Norvig, 2003).

Existem várias bases/ áreas de aplicação da IA, como por exemplo (Russel & Norvig, 2003):

- Filosofia;
- Matemática;
- Económica;
- Neurociência;
- Psicologia;
- Engenharia Computacional;
- Teoria do Controlo e Cibernética;
- Linguística ou Linguagem Natural.

A Linguística moderna e a IA "nasceram" praticamente ao mesmo tempo, e cresceram juntas, cruzando-se num campo híbrido chamado linguística computacional ou PLN. O problema da compreensão da linguagem rapidamente acabou por ser consideravelmente mais complexo do que parecia em 1957. O entendimento da linguagem requer uma compreensão do assunto e contexto, não apenas uma compreensão da estrutura das frases. Isso pode parecer óbvio, mas não foi apreciado amplamente até os anos de 1960. Grande parte do trabalho no início da representação do conhecimento (o estudo de como colocar o conhecimento numa forma que pode ser entendida computacionalmente) foi associado à linguagem e informado pela pesquisa em linguística, que foi ligado através vez de décadas de trabalho na análise filosófica da linguagem (Russel & Norvig, 2003).

### 3.3. Text Mining

O TM é um conjunto de técnicas cujo objetivo é a descoberta de conhecimento em documentos de texto. Neste capítulo foi realizado um estudo sobre esta ferramenta, onde contém a sua contextualização, o seu contexto histórico, bem como as técnicas mais utilizadas do TM, e a sua aplicação na medicina.

#### 3.3.1. Contextualização

O TM ou análise de texto como é comum designar-se, são termos que descrevem um variado leque de técnicas, que têm como objetivo extrair informação útil para análise de um documento ou de coleções de documentos (Feldman & Sanger, 2007). Nos últimos tempos o TM tem crescido muito, tendo sido alvo de bastante discussão, já que, a quantidade de dados cresce de dia para dia, e são necessárias ferramentas e técnicas para poder analisar os dados em forma de texto. Este está relacionado também com o *Data Mining* (DM), no entanto, diferem bastante (Truyens & Van Eecke, 2014).

O TM é conhecido pela descoberta de conhecimento em bases de dados textuais (Hearst, 1999). Este refere-se ao processo de extração de padrões interessantes ou a extração de conhecimento de documentos de texto não estruturados. Como a forma mais natural de armazenamento de informação é realizada em texto, o TM tem maior potencial comercial do que o DM. Em relação ao DM, o TM é uma técnica mais complexa porque envolve lidar com dados de texto que estão inerentemente não estruturados e difusos (Tan, 1999). O TM deriva muito da inspiração e direção da pesquisa elaborada no DM. Existem também outras áreas práticas na mesma região do TM como as Estatísticas Gerais, *Machine Learning*, Gestão de Base de Dados, IA e linguagem computacional (Truyens & Van Eecke, 2014).

Este processo é realizado através da identificação e exploração de padrões de interesse em conjuntos de dados não estruturados como livros, páginas *web*, *e-mails*, relatórios ou até mesmo descrições de produtos. Pode ainda ser definido formalmente como a criação de uma nova informação não-óbvia (como padrões, relações) de uma coleção de documentos textuais (Hearst, 1999).

Normalmente, as tarefas que estão incluídas na técnica de TM incluem atividades como ferramentas de procura (Truyens & Van Eecke, 2014):

- *Text Categorization* – associar um texto a uma ou várias categorias;
- *Clustering* – agrupar e juntar textos similares;
- *Concept/entity extraction* – encontrar o assunto das discussões;
- *Sentiment analysis* – encontrar o “tom” do texto;
- *Entity Relation Modelling* – sumarizar os documentos e encontrar relações entre as entidades descritas no texto.

O TM tem diferentes técnicas e ferramentas, sendo uma delas a *PubMed*. O *PubMed* é um motor de busca que tem como referência, a Base de Dados Literária da *Medical Literature Analysis and Retrieval System Online* (MEDLINE). A pesquisa é efetuada por via de seleção das áreas de interesse ou pela inserção de autores. Assim que a informação é obtida usando o TM, o próximo passo é a *curation*. Esta fase pode ter um cariz automático, semiautomático ou manual. Na vertente automática, os requerimentos adicionais podem ser adicionados de maneira binária ou por categorização. Por outro lado, a vertente manual requer experiência do indivíduo na área, já que envolve limpar as listas de dados ou informações desnecessárias com base nos seus critérios e vivência de outras situações semelhantes. No próximo passo, são estabelecidas conexões sobre o texto selecionado e os restantes dados. Este processo chama-se integração e é essencial para criar novo conhecimento (Piedra & Ferrer, 2014).

### 3.3.2. Contexto Histórico do Text Mining

Hearst (Hearst, 1999) escreveu que o domínio primordial do TDM trouxe grande entusiasmo, mas ainda assim não existe quase ninguém a praticar. O nome também não parecia ser claro, já que, com base num motor de busca bastante utilizado, a frase “*Text Mining*” aparece dezassete vezes mais do que a palavra “*Text Data Mining*” na *Web*, e o “*Data Mining*” aparece 500 vezes mais do que o “*Text Data Mining*”. Além disso, o significado do nome não é descrito de forma clara. Hearst define como DM o acesso à informação e à linguística computacional baseada em *corpus* e discute a relação destes como TDM, mas não define esse termo (Hearst, 1999). A literatura acerca do DM é muito mais extensa, e também mais focada. Há inúmeros livros didáticos e revisões críticas que traçam o seu desenvolvimento desde o início em *Machine Learning* e aplicações estatísticas.

O DM conseguiu acompanhar o aumento da alta tecnologia na década de 1990, e estabeleceu-se como uma prática amplamente utilizada na área da tecnologia (Franklin, 2002).



Ao invés do DM, o TM surgiu pouco antes do *crash* do mercado, os primeiros *Workshops* foram realizados na *International Machine Learning Conference* em julho de 1999 e na *International Joint Conference on Artificial Intelligence* em agosto de 1999 – sendo aí que perdeu a oportunidade de ganhar uma posição sólida nos anos de impulso.

### 3.3.3. Crescimento do Text Mining

O crescimento do TM tem crescido como se pode ver na Figura 3, onde se verifica o crescimento do mesmo na área biomédica. Como se pode constatar, as publicações que contêm a palavra “*Text Mining*” ou “*Literature Mining*” como palavra-chave (*keyword*) ou no resumo (*abstract*) cresceram de uma forma aproximadamente exponencial desde o ano 2000 até ao ano de 2011.

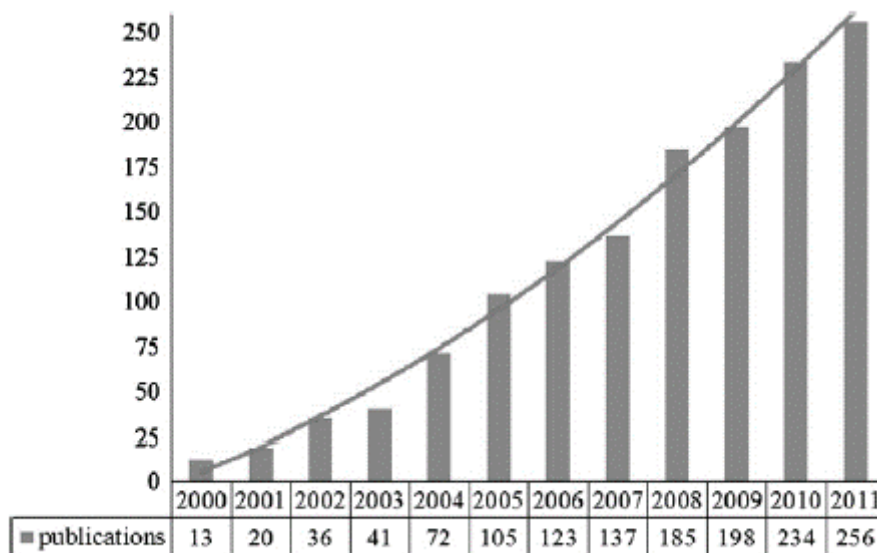


Figura 3 – Número de publicações existentes no PubMed com a palavra “*Text Mining*” ou “*Literature Mining*” como palavra-chave (*keyword*) ou estando presente no resumo (*abstract*) (Zhu, Patumcharoenpol, Zhang, Yang, & Chan, 2013)

Uma das razões para este aumento pode passar pelas vantagens do TM em descobrir um conhecimento capaz de melhorar o desenvolvimento e a pesquisa biomédica, especialmente em doenças malignas como o cancro, que em 2008 causou cerca de 7,4 milhões de mortes (World Health Organization, 2009). Com base nas preocupações relativas a esta doença tem-se verificado um aumento do número de publicações sobre pesquisas na área da medicina. Só em 2011 foram

realizadas mais de 73000 publicações que continham a palavra “cancro” no título, ou no resumo ou nas palavras-chave (Zhu, Patumcharoenpol, Zhang, Yang, & Chan, 2013).

#### *3.3.4. Questões legais do DM e Text Mining*

De um ponto de vista legal, o DM sempre foi alvo de uma grande discussão no campo da proteção de dados, devido ao facto de estar fortemente relacionado com tópicos de caracterização comportamentais. Assim, este tem assim recebido atenção significativa dos agentes legais devido ao seu impacto direto na privacidade.

O TM, pelo contrário, tem recebido muito menos atenção dos agentes legais, já que tem um impacto muito menor nas questões de privacidade. Porém, a Comissão Europeia reconheceu recentemente a importância do TM e do DM e quer promover o seu uso para efeitos de pesquisas e trabalhos científicos (European Commission, 2012).

Apesar das atenções relativas ao TM estarem a crescer lentamente, as questões legais acerca do mesmo mantêm-se escassas. Dentro da União Europeia não existe qualquer conhecimento ou qualquer indício de decisões de tribunais envolvendo-o diretamente. No entanto, o TM tem relações com outras ferramentas que já estiveram relacionadas com processos e ações legais, como por exemplo motores de busca, *screen scraping* e extração de base de dados.

Com o aumento da utilização do *Big Data*, esta situação pode inverter-se num futuro próximo, pois este tem uma área de aplicação mais abrangente. Neste passo, a conexão e a integração os dados obtidos das áreas específicas nem sempre se fazem espontaneamente usando os métodos tradicionais. Assim, os métodos automáticos que selecionam e põem informação significativa à disposição do profissional que gere o sistema são necessários, independentemente, da área de conhecimento na qual o texto foi gerado (Truyens & Van Eecke, 2014).

#### *3.3.5. Knowledge Discovery in Text*

A descoberta de conhecimento é definida como a extração não trivial de informação implícita, previamente desconhecida e potencialmente útil a partir de determinados dados (Frawley, Piatetsky-shapiro, & Matheus, 1991). A heterogeneidade e o número de fontes podem levar a um problema de sobre carregamento da informação, que acontece quando se chega ao ponto de ter demasiada informação para lidar. Para minimizar o sobre carregamento da

informação e para ajudar a extrair informação dos textos surgiu a Descoberta da Informação em Texto (Feldman & Dagan, 1995). A descoberta de conhecimento em texto, como o nome sugere, aplica as técnicas da descoberta de conhecimento em bases de dados em textos (Loh, Wives, & de Oliveira, 2000). A Descoberta de Conhecimento em Texto refere-se a *Knowledge Discovery in Text* (KDT) e a Descoberta de Conhecimento em Base de Dados refere-se a *Knowledge Discovery in Data* (KDD).

No entanto, Feldman e os seus parceiros, [(Feldman & Dagan, 1995) (Feldman & Hirsh, 1997) (Feldman, Dagan, & Hirsh, 1998)], reconheciam que existia o problema de aplicar as ferramentas da Descoberta de conhecimento em bases de dados em palavras-chave que são atribuídas a grupos de texto. Estas técnicas de *Mining* utilizam análises estatísticas para descobrir regras de associação e padrões de interesse, em vez das palavras-chave e associações/relações. Para utilizar o processo de Descoberta de Informação em Texto, as palavras-chave devem de ser atribuídas previamente aos textos. Esta associação pode ser feita manualmente, ou automaticamente com *software* especializado.

Por outro lado, Lin et al. (1998) usaram termos extraídos automaticamente dos textos para os categorizar e para encontrar associação. Os termos mais frequentes do texto são assinalados como palavra-chave (atributos). No entanto, ao analisar os termos podem existir problemas de vocabulário, como por exemplo existirem erros semânticos por causa dos sinónimos e da polissemia. Se existe apenas análise dos termos, atribuídos ou extraídos do texto, o processo de descoberta de conhecimento pode ser induzido em erro por falhas semânticas (Loh et al., 2000).

Outra abordagem passa por aplicar as técnicas de Descoberta de conhecimento de base de dados após o uso das técnicas de Extração da Informação que transforma a informação dos textos numa base de dados estruturada. Quando a informação textual é estruturada numa base de dados, pode-se fazer análises úteis com a ajuda de Sistemas de Gestão de Base de Dados (Cowie, Lehnert, & Wilks, 1996). Por exemplo, usando técnicas de associação, podem ser descobertas relações entre itens a examinar transações numa base de dados. Em alguns casos, a Extração de Informação tem-se mostrado viável em explorar o conteúdo textual. Soderland (1997) extraiu informações sobre previsão do tempo em textos da Web. Etzioni (1996) cita alguns casos de sucesso de aplicações ou seja, usando *wrappers* (Extratores de informação *Web*).

Apesar dos resultados promissores na Extração de Informação, infelizmente, a maioria dos sistemas de Extração de Informação atualmente dependem de um número reduzido de *wrappers* codificados para aceder a um conjunto fixo de recursos da *Web* (Garofalakis & Rastogi, 1999). Isto significa que os sistemas de Extração de Informação são muito dependentes do domínio, sendo úteis apenas para aplicações específicas ou para trabalhar apenas com uma classe especial de tipos de documentos. Além disso, para criar tais sistemas, é necessário um grande conhecimento sobre o domínio (engenharia do conhecimento), examinar estilos de texto e saber como a informação é codificada em frases em linguagem natural (Chinchor, Hirschman, & Lewis, 1993). Mattox, Seligman & Smith, (1999) concluem que o conhecimento semântico sobre o domínio (como as ontologias) é essencial para Extração de Informação e que há um esforço não-trivial para gerar invólucros, mesmo com ferramentas.

### *3.3.6. Modelo de Processamento de Texto Conceptual de Perrin e Petry*

Para abordar o KDT existem investigadores que criaram modelos que permitem uma melhor eficácia do KDT, nesta revisão vamos verificar o modelo utilizado por Patrick Perrin & Fred Petry (1998) que criaram um modelo para o processamento de texto contextual. O seu objetivo primeiramente era encontrar discurso implícito na estrutura do texto identificando os temas mais importantes do texto. O segundo objetivo era extrair automaticamente e selecionar expressões colocacionais em cada tema.

#### *I. Primeira fase: representar o conteúdo do texto*

Nesta fase os autores utilizaram cláusulas programa-definido, um subconjunto de lógica de primeira ordem para representar a estrutura do tópico do discurso e as afirmações relevantes. Cada peça de informação no texto passou a ser uma expressão lógica na forma de Tema (expressão\_colocacional, número\_de\_texto), o que denota um conceito particular que é indicado por uma expressão colocacional que ocorre no texto relevante para o número de casos afirmados, e tem sido observado para definir o tema indicado (Perrin & Petry, 1998).

#### *II. Segunda Fase: descobrir o discurso na estrutura do tópico*

Nesta fase, os autores querem responder a duas perguntas. A primeira, qual o tipo de estrutura que é inerente no discurso. A segunda, quais os mecanismos e os aspetos da linguagem que podem ser utilizados para detetar isso. Existem certos aspetos na estrutura do discurso que podem por exemplo ser revelados pela via dos padrões de distribuição lexical. O objetivo é extrair

uma estrutura global do tópicos para partir o texto em unidades contextuais não fixas, coesas no seu conteúdo, e, portanto, atribuir uma etiqueta contextual para quaisquer fatos que serão extraídos dessas unidades contextuais. Este é um problema de rotulagem e a abordagem proposta é simples, pode ser sistemática e não requer um *oracle*, que são todas qualidades desejadas de um paradigma KDT, o que faz com que a abordagem seja robusta.

Para a classe de textos técnica, pode-se notar que o discurso é estruturado e, muitas vezes há uma estrutura explícita, sob a forma de títulos principais (por exemplo, *history\_of\_present\_illness, past\_medical\_history*). Esta estrutura é bastante fácil de extrair, não requer uma codificação extensa e específica do conhecimento, e mostra-se muito útil para explorar. No caso de não existir tal estrutura explícita, Hearst propõe um método direto baseado num particionamento contextual do documento (Hearst, 1994). O objetivo dos autores em revelar a estrutura do tópicos é ter a capacidade de rotular facilmente qualquer coisa que é extraída das secções identificadas. Esta é uma informação contextual local suplementar que será útil para a descoberta de relações causais. No caso de existir uma estrutura explícita, como nos relatórios médicos psiquiátricos, qualquer facto que seja extraído das sessões mencionada como *history\_of\_present\_illness*, será rotulado como *history\_of\_present\_illness* (o tema predicado) (Perrin & Petry, 1998).

### *III. Terceira Fase: descobrir afirmações de texto significativas*

Os autores desenvolveram um algoritmo que identifica automaticamente segmentos de prosódia (factos) no texto e avalia-os para a sua adequação. Esta abordagem é tentadora, porque fornece um meio para tomar sistematicamente medidas mínimas a partir de qualquer texto sem restrições, garantindo a representação de todos os temas do texto. Acaba por se tornar num problema de análise de características não supervisionadas, onde as expressões colocacionais são as medidas. A medida de entropia serve para reduzir a dimensionalidade do espaço de características para tornar a descoberta do problema tratável.

Os autores (Perrin & Petry, 1998) aumentaram a definição tradicional do *collocates* pares de palavras para uma definição mais geral, devido ao facto de que a concorrência de palavras de conteúdo pode não ser adjacente. Uma característica textual é um conceito típico numa unidade de texto, e pode ser representado por um grupo “a” palavras consecutivas, que aparecem juntos dentro “b” palavras uma unidade de texto. Os recursos são identificados pela sua frequência relativa de ocorrência (significância) no texto (dependência de contexto). Como a definição de um

recurso textual é geral o suficiente para incluir qualquer sequência de “a” palavras consecutivas, usamos uma medida de Informação Mútua para distinguir os mais significativos, ou seja, aqueles que carregam informações mais propensas a ajudar a extrair conhecimento a partir do texto.

Por exemplo, com um “a” =2 e “b”=5, todas as combinações de duas palavras consecutivas na seguinte frase (respeitando a ordem das palavras na frase) são características:

*“... durante o exame, a doente negou qualquer febre ...”*

"Febre negada", " doente de exame " ou " o doente ", são características, mas " doente negado " e " durante a febre " não são. Uma relação colocacional é uma característica relevante textual, isto é, um recurso com elevado grau de coesão inter-relação. Como o ser humano descarta informações irrelevantes nas tarefas cognitivas, os autores modelaram este aspeto pelo *ranking* dos temas extraídos por seu grau de relevância no texto que descrevem.

Verificou-se que a relevância é relativa ao contexto, o que torna o problema difícil, porque não há nenhuma forma de saber precisamente que contexto alguém tinha em mente a qualquer momento. Tem sido sugerido que uma suposição é relevante num contexto, na medida em que os seus efeitos de contexto, neste contexto, são grandes; e não é importante na medida em que o esforço necessário para processá-lo neste contexto é grande (Akman & Surav, 1996). Ordenação de expressões colocacionais pelo seu conteúdo de informação revela quais são as expressões mais significativas, o que acaba por ser aquelas expressões que contêm palavras de conteúdo (Perrin & Petry, 1998).

### *3.3.7. Técnicas Mais Utilizadas No Text Mining*

O TM tem bastantes técnicas nas suas áreas de aplicação, Chauhan Shrihari R (Chauhan Shrihari R, 2015), define algumas delas como as mais utilizadas.

#### *1. Extração de Informação*

A extração de informação é o primeiro passo para fazer a análise do texto não estruturado e a sua relação. Este processo é elaborado por correspondência de padrões, e é usado para procurar e pré-definir a sequência do texto. Existem duas técnicas, a inclusão da verificação e a segmentação das frases, que são muito importantes para documentos com textos bastante densos. Muitos desafios na informação eletrónica estão na forma do PLN, e a Extração de Informação resolve este problema transformando o documento de texto num formato estruturado.

## *II. Clustering*

O *Clustering* é um método não supervisionado. A técnica de *clustering* é usada para agrupar documentos similares, mas difere na categorização, onde estes documentos são *clusterizados*. Este método é baseado no conceito da divisão de texto similar no mesmo *cluster*, e cada *cluster* contém um número similar de documentos.

## *III. Sumarização*

Devido à grande quantidade de dados existente nos dias de hoje, é preciso sumarizar os dados a partir do número do documento onde se resumem os mesmos, sem mudar o sentido do documento e o tamanho dos dados. A sumarização produz um resumo de um grupo de documentos. Após essa sumarização ser realizada o documento inteiro (ou o grupo de documentos) é substituído pelo resumo. A sumarização é útil para o utilizador ler um pequeno resumo do documento em vez de o ler na sua totalidade.

## *IV. Visualização*

No TM, a visualização reduz a dificuldade de descobrir informação. Um grupo do documento ou simplesmente um documento de texto marcado é utilizado para mostrar o documento e a cor usada. Este método fornece informação compreensível e em massa, o que ajuda a descobrir ou a criar o *mining* padrão da coleção dos documentos. Utiliza cores diferentes, distâncias relacionais etc.

## *V. Categorização*

A Categorização é similar à classificação de texto. A categorização é uma técnica supervisionada porque se baseia em exemplos e *input* *output* para a classificação. O identificador de texto é usado para a categorização do documento do texto para pré-definir a classe do documento. A marca da predefinição da classe é baseada no conteúdo do documento textual. Uma forma comum no processo de categorização de texto consiste no pré-processamento, indexação, redução das dimensões e classificação. O objetivo da categorização é treinar o classificador com uma base de conhecimento, onde os exemplos desconhecidos são categorizados automaticamente.

### *3.3.8. Áreas de aplicação do Text Mining*

O TM pode ser aplicado em diversas áreas, sendo a mais promissora a área Biomédica. Nesta área, este pode ser utilizado para descobrir previamente relações ocultas num

conhecimento existente. Por exemplo, um gene específico pode ser mencionado em alguns artigos de pesquisa, mas pode não se diferenciar por si mesmo. Após análise de diversos artigos de pesquisa, o TM poderá ser capaz de filtrar esse mesmo gene, e identifica-lo para que ele possa proporcionar uma interessante via numa pesquisa futura (Truyens & Van Eecke, 2014).

Além de poder ser aplicado na área Biomédica como referido, o TM poderá ser utilizado na área clínica para extrair informação relevante não estruturada e converter essa mesma informação em informação estruturada. Posteriormente, esta poderá ser utilizada para calcular o *Framingham Risk Score* (FRS). O FRS foi desenvolvido como parte do Estudo Cardíaco de *Framingham* (Mahmooda, Levy, Vasan, & Wang, 2014). Um dos objetivos do estudo foi desenvolver modelos de previsão, que estimam a probabilidade de desenvolver várias doenças cardiovasculares e cerebrovasculares (Jonagaddala et al., 2015).

Outra das variantes do TM é o *Agile Text Mining*. Este último é definido como um desenvolvimento interativo de regras semânticas e léxico-sintáticas (*queries*). O desenvolvimento de *queries* interativas permite aos utilizadores generalizar ou especializar as mesmas e comparar os resultados devolvidos. Isto permite e facilita a extração *one-off*, bem como a extração normal de tarefas, como acontece na relação gene-doença ou o estágio do cancro. Ao construir estas regras numa forma *data-driven*, aumenta-se a precisão e reduz-se a experiência requerida no domínio (Cormack, Nath, Milward, Raja, & Jonnalagadda, 2015).

### 3.3.9. Text Mining na Medicina

Apesar do TM já existir há algum tempo, só nos tempos mais recentes foi aumentando a necessidade de processar automaticamente grandes quantidades de informação. Já há mais de 50 anos que são utilizadas técnicas semelhantes ao TM, para a divulgação de análises, e para a estudos da estrutura linguística.

Na área da medicina foram desenvolvidos alguns trabalhos que estabelecem relações aparentemente sem ligações ao fenómeno, assim como enxaquecas e a deficiência de magnésio. Weeber et al. (2013), descobriram o potencial benéfico de curcumina e talidomida em doentes com a doença de *Crohn*. Devido a estes sucessos recentes, este tipo de pesquisa é neste momento usada pela biologia e pela tecnologia, de modo a estabelecer relações entre ambos. Por exemplo, os genes e as doenças ou avaliar interações entre várias proteínas. Um instrumento interessante



para esta proposta é o DisGeNET, desenhado para relatar a informação derivada de vários *omics* com dados gerados em doenças diferentes (Piedra & Ferrer, 2014).

Vários estudos têm-se centrado no tratamento de informação textual disponível em conjuntos de dados de saúde. Uma breve visão geral dos estudos que destacam a importância de dados textuais e a sua adequação em ambientes de pesquisa é apresentado neste ponto (Raja, Mitchell, Day, & Hardin, 2008).

Uma iniciativa notável foi a pesquisa realizada na *Vanderbilt Clinic, New York* (Kukafka, Bales, Burkhardt, & Friedman, 2006). O objetivo foi determinar se um programa de PLN podia codificar automaticamente informações de estado funcional de acordo com os requisitos da *Internacional Classification of Functioning, Disability, and Health* (ICF). A codificação automática é uma escolha óbvia para este tipo de iniciativas. Esta é extremamente importante para efeitos de reembolso e de manutenção dos registos, no entanto é processo muito tedioso e demorado. Se a codificação for realizada com precisão pode evitar que as instalações médicas gastem uma quantidade substancial de recursos.

Os investigadores estenderam o código de PLN – *Medical Language Extraction and Encoding System* (MedLEE) existente para codificar notas de alta. Dez códigos ICD-9 foram pré-selecionados pela sua relação conhecida nas mudanças do estado funcional. As avaliações foram realizadas pelo sistema PLN, com codificadores experientes, e com codificadores não-especializados. Estes descobriram que o sistema PLN codificado teve resultados semelhantes com resultados dos codificadores humanos, o que se revelou numa descoberta promissora para a investigação de codificação automática de códigos ICD-9, que são a principal base para o reembolso, em maioria dos serviços de saúde (Raja et al., 2008).

Um estudo realizado na Universidade de *Utah* (Penz, Wilcox, & Hurdle, 2007), utiliza uma versão modificada do MedLEE, bem como um algoritmo *phrasematching* para extrair dados para iniciativas de investigação. A maioria dos registos eletrónicos são ditados numa forma narrativa e recuperam manualmente dados específicos para a pesquisa, o que pode ser demorado e caro. O objetivo deste estudo foi extrair os dados relacionados com eventos adversos ligados à colocação do cateter venoso central. Os eventos adversos podem ser infeções, complicações decorrentes de extravio e pneumotórax (um colapso pulmonar). Os testes foram conduzidos usando cada método

individualmente e, em seguida, utilizando os métodos todos juntos com uma amostra de registos que tinha anteriormente tido sido avaliada de forma manual.

Os ensaios que utilizaram os métodos individuais foram infrutíferos. O algoritmo *phrasematching* não era suficientemente específico e o sistema PLN não foi sensível o suficiente. Estes produziram valores de previsão positivos de 6,4 e 6,2%, respetivamente. No entanto, quando usados em conjunto, os resultados foram promissores: produziram uma sensibilidade de 72,0% e uma especificidade de 80,1%, o que são valores aceitáveis. Este estudo mostra potencial para a utilização de sistemas de PLN para automatizar a extração de dados de pesquisa.

A deteção de eventos é outra área importante de pesquisa. Hazlehurst, Mullooly, Naleway, & Crane (2005) realizaram um estudo para identificar reações vacinais para o Vaccine Safety Datalink Project (VSD). A VSD é uma parceria entre o Centro de Prevenção e Controlo de Doenças e oito grandes Organizações de Manutenção da Saúde, para investigar eventos adversos após a imunização, através da análise de bases de dados de assistência médica e de Notas Clínicas. Neste estudo, foi utilizada uma versão modificada do sistema de PLN MediClass que tinham sido treinados com o conhecimento necessário para detetar possíveis reações à vacinação. O sistema alcançou uma elevada taxa de sensibilidade e especificidade. Em comparação com os métodos que são usados por médicos, este sistema melhora de forma significativa o valor positivo previsto. Estudos como estes são especialmente importantes porque o objetivo final é a migração para um sistema que pode prever tais ocorrências no futuro.

Recentemente, as ferramentas de TM têm sido utilizadas em pesquisas na área da saúde, por exemplo, Cerrito e Cerrito (Cerrito & Cerrito, 2011) analisaram os registos médicos eletrónicos do departamento de urgência de um hospital ao longo de um período de seis meses, usando TM. Estes descobriram que doentes que tinham queixas semelhantes foram tratados de forma diferente, dependendo do médico de serviço. Tais diferenças podem afetar a qualidade de atendimento e custos. Portanto, o TM de tratamento especializado pode fornecer aos médicos de serviço um plano de tratamento otimizado para os doentes, podendo levar ao desenvolvimento de protocolos para aliviar a disparidade de tratamento (Raja et al., 2008).

### 3.3.10. Fases do Text Mining

O TM tem várias fases, nos pontos abaixo é realizada uma breve descrição dos passos mais importantes que são utilizados no mesmo:

#### I. Recuperação de Informação

Além dos sistemas convencionais de recuperação de informação, existem também sistemas de recuperação de informação de conhecimento avançado. Estes integram os dados de diferentes recursos dentro de um contexto simples para facilitar a entender os sistemas biomédicos complexos (Zhu et al., 2013).

Por exemplo, para aceder aos resultados do TM e de outros dados, Maier et al. (2011) gerou uma base de conhecimento sobre a doença crónica de obstrução pulmonar e desenvolveu um sistema de conhecimento integrado. *Salivaomics Knowledge Base* definiu a *Saliva Ontology* como termos e relações de vocabulário, para facilitar a recuperação de informação e integração da mesma através de múltiplos campos de pesquisa juntamente com a análise dos dados e DM. QuExT, um documento retirado da PubMed sobre sistemas de recuperação de informação, seguiu um conceito *query* orientado à expansão da metodologia para encontrar documentos que contenham conceitos das *query words* (Matos, Arrais, Maia-Rodrigues, & Oliveira, 2010).

Na era do genoma, com avanços na biotecnologia e métodos para a análise de genes, irá haver um crescimento contínuo na necessidade de TM e na recuperação da Informação para ajudar os investigadores a encontrar artigos relevantes para os seus estudos (Zhu et al., 2013).

#### II. Reconhecimento das entidades mencionadas e a sua relação-extração

O reconhecimento das entidades nomeadas é o passo mais importante na extração do conhecimento, no qual o objetivo principal é identificar os termos específicos, como o gene a proteína a doença e o medicamento (Leser & Hakenberg, 2005). No entanto, na prática ainda existem muitos obstáculos para a identificação automática dos termos biomédicos.

Por exemplo, um termo biomédico pode ter várias formas diferentes de se escrever. Por exemplo, a epilepsia e a síndrome vertiginoso referem-se à mesma doença, que consiste num distúrbio no sistema nervoso central caracterizado pela perda de consciência e convulsões. Além disso, uma entidade pode ser representada de maneira diferente, como cancro, que tanto pode ser referenciado como uma doença, bem como um signo astrónomo (termo brasileiro por exemplo). As abreviações também podem ter problemas de ambiguidade, por exemplo, PC, pode

significar *Personal Computer*, ou também pode ser uma abreviação para cancro da Próstata (Zhu et al., 2013).

As técnicas de reconhecimento das entidades mencionadas estão divididas em três categorias: abordagens *dictionary-based*, abordagens *rule-based* e abordagem ao *Machine Learning* (Rebholz-Schuhmann et al., 2011). Um termo biomédico pode aparecer na forma de abreviação e também pode conter múltiplos sinónimos no texto. O reconhecimento da abreviação e o reconhecimento do sinónimo são necessários para unir e normalizar os termos biomédicos no reconhecimento das entidades nomeadas (Zhu et al., 2013).

Neste momento, existem cada vez mais pesquisas e investigadores interessados nos termos de identificação e normalização. Um dos exemplos envolve a *BioCreative III* (Arighi et al., 2011), focada na normalização do gene, que define as menções do gene e liga os genes aos identificadores padronizados, por exemplo, uma base de dados.

Na era genómica, muitos investigadores estão interessados em interações “*mining gene-gene*”, “*protein-protein*” e outras no âmbito do genoma, que fornecem bases para mais análises interativas de expressões de genes e de anotação de base de dados (Cohen & Hersh, 2005), bem como outras relações extensivas (Arighi et al., 2011). Para além disso, os investigadores estão focados nas relações entre genes e outras entidades biomédicas, como as relações gene-doença e relações de proteínas sub-celulares (Zhu et al., 2013).

### III. *Descoberta de Conhecimento*

O Conhecimento incluindo factos, informação ou descrições (implícitas ou explícitas) refere-se à compreensão prática ou teórica de um domínio ou de um assunto, a descoberta de conhecimento e a criação de grandes quantidades de dados não estruturados e estruturados. Com base nisso, o conhecimento obtido pode tornar-se em informação adicional que poderá ser usada para descobrimento futuro. Na área do TM e do DM a descoberta de conhecimento é uma parte muito importante. A descoberta de conhecimento permite integrar texto biomédico com múltiplas fontes de dados para gerar um novo contexto interpretativo (Zhu et al., 2013).

### IV. *Geração de Hipótese*

Quando não existem factos e/ou informação explanada de uma maneira satisfatória com o conhecimento disponível, surge a hipótese científica. Esta é uma tentativa de solução para o problema em vez de ser só uma teoria, podendo ser sugerida ou proposta para investigação futura.

As experiências podem ser usadas para avaliar as hipóteses propostas antes de resolver o problema. A hipótese científica é um pouco como a imaginação científica, que se baseia em evidências e conhecimento existente.

A geração de hipótese é um passo fulcral no TM, porque é muito importante para os investigadores que querem inferir fatos desconhecidos que podem ser utilizados para explicar os resultados experimentais ou para orientar a conceção de novas experiências. Esta tarefa tem recebido muito mais atenção dos investigadores (Zhu et al., 2013).

### 3.4. Linguagem Natural

O PLN é um cruzamento de disciplinas baseado em linguística, lógica, fisiologia, psicologia, ciência computacional, matemática e outras teorias relacionadas. O PLN pode processar linguagem falada e linguagem textual. Até ao momento, o PLN é usado em máquinas de tradução automáticas, sistemas de diálogos, *Machine Learning*, computadores inteligentes, computadores de multimédia, sistemas experientes, DM, robótica, recolha de informação, e várias fronteiras entre a ciência e a tecnologia (Liddy, 2001; Yue, Di, Yu, Wang, & Shi, 2012).

O PLN pode ser conhecido também por *Natural Language Understand*, ou *Computacional Linguistics* num campo académico linguístico, embora alguns experientes dizem que o PLN contém o *Natural Language Understand* e a Geração de Linguagem Natural [(*Natural Language Generation* (NLG))]. Os diferentes nomes que são associados ao PLN são uma descrição especial do mesmo, porque envolvem várias disciplinas (Yue et al., 2012).

A aplicação do PLN está a ser investigado por várias universidades, escolas e institutos. A maior parte das pessoas utiliza os resultados da pesquisa do PLN, sendo esta muito importante para muitos países do mundo pois estes estão a aplicar muitos recursos humanos, financeiros e materiais na investigação da mesma (Yue et al., 2012).

O objetivo do PLN é realizar o processamento da linguagem semelhante à linguagem humana. Nos primeiros tempos da IA, o PLN era referido como sendo NLG. Um Sistema completo de Processamento de Linguagem Natural deve ser capaz de (Liddy, 2001) :

- Parafrasear um texto inserido;
- Traduzir o texto noutra idioma;
- Questionar sobre o conteúdo do texto;

- Inferir sobre o texto.

Apesar dos três primeiros pontos terem sofrido avanços consideráveis nos últimos anos, o PLN ainda encontra dificuldades em resolver o último ponto.

Existem objetivos mais práticos do PLN, como por exemplo, desenvolver um sistema baseado em recolha de informação com o objetivo de fornecer um conjunto de informação mais precisa e completa, consoante as necessidades reais do utilizador. Mais especificamente o objetivo do sistema é representar o sentido verdadeiro e a intenção da *query* do utilizador, que poderá ser expressa naturalmente na linguagem corrente como se estivessem a falar para um bibliotecário de referência (Liddy, 2001).

#### 3.4.1. Contexto Histórico

A pesquisa no âmbito do PLN tem acontecido por várias décadas começando no final dos anos 40 do século passado. O primeiro computador com aplicações baseadas na Linguagem Natural foi o *Machine Translation*. Em 1946, Weaver e Booth (1949) desenvolveram um computador de *Machine Translation* que estava especializado em descobrir os códigos dos inimigos durante a Segunda Guerra Mundial. Esse projeto inspirou vários projetos posteriores a este. *Weaver* sugeriu usar ideias da criptografia e da teoria de informação para a tradução da linguagem, a ideia foi aceite e as pesquisas iniciaram-se em várias instituições dos Estados Unidos durante os anos seguintes (Weaver, 1949).

Os primeiros trabalhos realizados em *Machine Translation* enveredaram por um ponto de vista mais simples, onde apenas se via as diferenças entre os vários idiomas, que residiam nos seus vocabulários e nas ordenações de palavras. Os sistemas desenvolvidos por esta perspetiva usavam apenas um dicionário-*lookup* para as palavras apropriadas para a tradução e guardavam as palavras depois de traduzidas para as colocar nas regras de ordem de palavras do seu idioma, sem ter em conta a ambiguidade do campo léxico da linguagem natural. Esta prática teve resultados insatisfatórios, e para a resolver os investigadores tiveram uma tarefa bastante mais complicada do que a prevista, já que, precisavam de uma teoria mais adequada da linguagem (Liddy, 2001).

Em 1957, Chumsky (Chumsky, 1957) introduziu a ideia da gramática generativa, o que fez obter uma maior visão, se ou como, a linguística dominante podia ajudar o *Machine*

*Translation*. Durante este período, começaram a emergir outras áreas do PLN como o reconhecimento da fala. A comunidade do processamento de linguagem e a comunidade da fala foram divididos em dois campos com o processamento de linguagem dominado pela perspectiva teórica da gramática generativa e métodos estatísticos *hostis*, e a comunidade da fala dominada pela teoria da informação estatística e teorias linguísticas *hostis*.

Em 1950, existiu um grande entusiasmo porque as pessoas acreditavam que os sistemas automáticos de grande qualidade na tradução de idiomas seriam capazes de reproduzir resultados indistinguíveis dos tradutores humanos, e que esses sistemas iriam estar operacionais dentro de alguns anos. Isso não era realístico pois na altura não existiam sistemas de conhecimento linguístico computacionais disponíveis (Liddy, 2001).

Devido às inadequações dos sistemas existentes na época e para travar ao entusiasmo existente, foi emitido um comunicado pela *Automatic Language Processing Advisory Committee of National Academy of Science – National Research Council* (ALPAC) a esclarecer que o *Machine Translation* não era imediatamente alcançável e recomendável, bem como, não era consolidado. Este comunicado levou à suspensão de grande parte dos trabalhos em PLN e de *Machine Translation* nos Estados Unidos (ALPAC, 1966).

Apesar de grande parte dos trabalhos em PLN terem sido suspensos nos anos seguintes ao comunicado feito pela ALPAC, surgiram desenvolvimentos significativos, tanto nos problemas teóricos bem como na construção de sistemas protótipos. O trabalho teórico desenvolvido na década de 60 e 70 focou-se no problema de como representar o sentido e desenvolver soluções computacionalmente tratáveis, que as teorias existentes à data não foram capazes de desenvolver, como por exemplo, em 1970 a Força Aérea dos Estados Unidos da América começou a utilizar o *Systran*. Em 1976 foi a vez da Comissão das Comunidades Europeias implementá-lo. O *Systran* é um sistema de *Machine Translation* (Hutchins, 2005).

A par do desenvolvimento teórico, bastantes sistemas protótipos foram desenvolvidos para demonstrar a efetividade de princípios particulares. Por exemplo, foram criados sistemas que replicavam a conversa entre um psicólogo e o seu doente onde só seria necessário permutar ou ecoar o *user input*. Também existiram tentativas de encarnar uma teoria da paranoia num sistema, em que em vez de palavras-chave individuais, foram utilizados grupos de palavras-chaves, e sinónimos usados se as palavras-chaves não fossem encontradas (Liddy, 2001).

Neste mesmo período também se verificaram trabalhos relevantes na *Natural Language Generation*. O Planeador de Discurso de McKeown (TEXT) (McKeown, 1985) e o gerador de resposta de McDonald (MUMBLE) (McDonald & Pustejovsky, 1985) usavam predicados retóricos para produzir descrições declarativas em forma de pequenos textos, normalmente parágrafos. A habilidade da ferramenta de McKeown onde era possível gerar respostas coerentes online foi considerada um grande feito na área.

Já no início dos anos 80, motivado pelo aumento na disponibilidade de recursos computacionais críticos, verificou-se uma crescente consciência de cada comunidade das limitações de soluções isoladas de problemas de PLN. Esses dois fatores em conjunto originaram um impulso geral para o desenvolvimento das aplicações que trabalham com uma linguagem ampla no contexto do mundo real. Posto isto, os investigadores voltaram a analisar abordagens não-simbólicas que tinham perdido popularidade nos primeiros tempos da Linguagem Natural (Liddy, 2001).

Na década de 90, o campo do PLN cresceu rapidamente e isso deveu-se a certos fatores, como o aumento da disponibilidade de grandes quantidades de texto eletrónico, a disponibilidade de computadores com maior memória e velocidade de processamento e da chegada da Internet. Várias abordagens estatísticas têm surgido para lidar com vários problemas genéricos nas linguísticas computacionais como a identificação do *part-of-speech*, por exemplo. Os investigadores de PLN têm desenvolvido uma geração de sistemas que lidam razoavelmente bem com texto generalizado e contam com uma boa porção da variabilidade e da ambiguidade da linguagem (Liddy, 2001).

#### 3.4.2. Ontologias

Brank, Grobelnik, & Mladeni, (2007) afirmam que se pode observar que o foco dos sistemas de informação modernos está a mover-se de processamento de dados no sentido de processamento de conceito, o que significa que a unidade básica de processamento é cada vez menos um pedaço atómico de dados e está a tornar-se cada vez mais num conceito semântico que carrega uma interpretação e que existe num contexto com outros conceitos. Uma ontologia é normalmente usada como uma estrutura de captura de conhecimento sobre uma determinada área, fornecendo os conceitos e as relações relevantes entre eles.



A análise textual de dados desempenha um papel importante na construção e no uso das ontologias, especialmente com a crescente popularidade de construção semiautomática de ontologias. Existem diferentes métodos de descoberta de conhecimento que têm sido adotadas para o problema da construção semiautomática de ontologias (Brank et al., 2007), incluindo a visualização semi-supervisionada e supervisionada. Esta consiste em aprender sobre uma coleção de documentos de texto, usando PLN para obter um gráfico semântico de um documento, visualização de documentos, extração de informações para encontrar conceitos relevantes, visualização de contexto de entidades nomeadas num grupo de documentos (Brank et al., 2007).

Um fator chave que torna uma disciplina específica ou abordagem científica é a capacidade de avaliar e comparar as ideias dentro da sua área. As ontologias são uma estrutura de dados fundamental para fazer uma conceptualização do conhecimento que na maioria dos casos práticos é suave e não é unicamente expressado. Como consequência somos, em geral, capazes de construir muitas ontologias diferentes de concetualização do mesmo corpo de conhecimento e devemos ser capazes de dizer quais dessas ontologias servem melhor alguns dos critérios pré-definidos anteriormente.

Assim, a avaliação de ontologias é uma questão importante que deve ser abordada para verificar se as ontologias estão a ser amplamente utilizadas na área da web semântica e outras aplicações de reconhecimento da semântica. Os utilizadores que enfrentam uma multiplicidade de ontologias precisam de os avaliar e decidir qual a melhor ontologia que se adapta às suas necessidades. Da mesma forma, as pessoas que constroem uma ontologia precisam de ter uma forma de avaliar a ontologia desenvolvida, para orientar o processo de construção e quaisquer passos de refinamento. As técnicas automáticas ou semiautomáticas de aprendizagem de ontologias também exigem medidas de avaliação eficazes, que podem ser usadas para seleccionar a melhor ontologia de um grupo, para seleccionar valores de parâmetros ajustáveis do algoritmo de aprendizagem, ou para dirigir o próprio processo de aprendizagem se este for formulado como encontrar um caminho através de um espaço de procura (Brank et al., 2007).

Uma ontologia é uma estrutura bastante complexa e muitas vezes é mais prático concentrar-se na avaliação dos diferentes níveis da ontologia separadamente, em vez de tentar avaliar diretamente como um todo. Isto é particularmente certo se a ênfase se focar no facto da avaliação se processar automaticamente ao invés de ser totalmente realizada por utilizadores / peritos humanos. Outra razão para ser feita uma abordagem baseada em nível é porque quando

são utilizadas técnicas de *Machine Learning* para a construção de uma ontologia, as técnicas envolvidas na construção são substancialmente diferentes para os diferentes níveis (Brank et al., 2007). Os níveis individuais foram definidos diferentemente por diferentes autores [por exemplo, (A Gómez-Pérez, 1994), (Asunción Gómez-Pérez, 1996), (Burton-Jones, Storey, Sugumaran, & Ahluwalia, 2004), (Porzel & Malaka, 2004), (Ehrig & Haase, 2005)], mas estas definições diferentes tendem a ser bastante semelhantes e envolvem geralmente os seguintes níveis:

- **Lexical, vocabulário, ou camada de dados** – O foco é sobre os conceitos, exemplos, fatos, etc. que foram incluídos na ontologia, e o vocabulário usado para representar ou identificar esses conceitos. A avaliação neste nível tende a envolver comparações com várias fontes de dados relativos ao domínio do problema (por exemplo, domínio *corpus* de texto específico), bem como técnicas, tais como medidas sequência de similaridade (por exemplo, editar distância).
- **Hierarquia ou taxonomia** – Uma ontologia inclui tipicamente uma hierarquia “é-um” ou subsunção relação entre conceitos. Embora também podem ser definidas várias outras relações entre os conceitos, a relação “é-um” é muitas vezes particularmente importante e pode ser o foco dos esforços de avaliação específicos.
- **Outras relações semânticas** – A ontologia pode conter outras relações além “é-um”, e estas relações podem ser avaliados separadamente. Isso normalmente inclui medidas como a precisão e *recall*.
- **Nível de contexto** – 1) Uma ontologia pode ser parte de uma coleção maior de ontologias, e pode fazer referência ou ser referenciada por várias definições nessas outras ontologias. Neste caso, pode ser importante ter em consideração este contexto quando a avaliação é feita [(Supekar, 2005), (Burton-Jones et al., 2004), (Patel, Supekar, Lee, & Park, 2004)]. 2) Uma outra forma de contexto é a aplicação onde a ontologia vai ser utilizada. Basicamente, em vez de avaliar a ontologia per se, pode ser mais prático para avalia-la dentro do contexto de uma aplicação particular para ver como os resultados da aplicação são afetados pela utilização de uma ontologia em questão. Em vez de se concentrar numa aplicação individual, a pessoa também pode concentrar-se na avaliação do ponto de vista dos utilizadores individuais ou da organização (por exemplo, empresas) que irá utilizar a ontologia (Fox, Barbuceanu, Gruninger, & Lin, 1998).

- **Nível Sintático** – A avaliação neste nível pode ser de particular interesse para as ontologias que tenham sido construídas manualmente. A ontologia é geralmente descrita numa linguagem formal particular e deve corresponder aos requisitos sintáticos da linguagem (o uso das palavras-chave corretas, etc.). Várias outras considerações sintáticas, como a presença de documentação de linguagem natural, evitando *loops* de entre as definições, etc., também podem ser considerados (A Gómez-Pérez, 1994). De todos os aspetos da avaliação ontologia, este é provavelmente mais próximo um processamento automatizado.
- **Estrutura, arquitetura, *design*** – Ao contrário dos três primeiros níveis desta lista, que incidem sobre os conjuntos de conceitos atuais, instâncias, relações, etc. envolvidos na ontologia, este nível centra-se em decisões de *design* de alto nível que foram usados durante o desenvolvimento da ontologia. Este nível tem maior interesse nas ontologias construídas manualmente. Partindo do princípio de que algum tipo de princípios de *design* ou critérios foram acordados antes da construção da ontologia, a avaliação a este nível significa verificar em que medida a ontologia resultante corresponde a esses critérios. Preocupações estruturais envolvem a organização da ontologia e sua adequação para o desenvolvimento (por exemplo, a adição de novos conceitos, alteração ou remoção de antigos) (A Gómez-Pérez, 1994)(Asunción Gómez-Pérez, 1996). Para algumas aplicações, também é importante que as definições formais e declarações da ontologia sejam acompanhadas por uma documentação adequada de linguagem natural, que deve ser significativa e coerente, atualizada e consistente com as definições formais, suficientemente detalhada, etc. A avaliação destas qualidades a este nível deve normalmente ser feito em grande parte ou até inteiramente manualmente por pessoas tais como engenheiros na área da ontologia e especialistas na área (Brank et al., 2007).

### 3.4.3. *Diferentes Níveis da Linguagem Natural*

A maneira mais fácil de explicar o que está a acontecer com um sistema de PLN é fazer uma abordagem aos níveis da linguagem que é utilizada. Isto também é referido como o modelo de linguagem assíncrona e distingue-se dos modelos sequenciais anteriores, em que levanta hipótese de que os níveis do processamento de linguagem humana seguem um ao outro numa forma estritamente sequencial. Existem pesquisas psicolinguísticas que sugerem que o

processamento de linguagem é muito mais dinâmico, assim como os níveis podem interagir com a variação de ordens.

A introspeção revela a utilização frequente de informação que ganhamos pelo que pensamos, ou seja, utiliza-se um nível elevado de processamento para ajudar a uma análise de baixo nível. Por exemplo, o conhecimento pragmático que se lê de um documento é biológico. Quando se encontra uma palavra que tem diferentes significados para quem lê, a palavra irá ser interpretada como tendo um sentido biológico.

Os diferentes níveis de PLN são (Liddy, 2001):

- **Fonologia** – Este nível lida com a interpretação dos sons da fala, dentro e através de palavras. Existem três tipos de regras utilizadas numa análise fonológica:
  - Regras Fonéticas, para sons dentro de palavras;
  - Regras Fonémicas, para variações de pronúncia quando as palavras são faladas em conjunto;
  - Regras prosódicas, para flutuação em *stress* e entoação através de uma frase.
- **Morfologia** – Este nível lida com a natureza da composição das palavras no qual são compostos por morfemas (as menores unidades de significado). Um sistema de PLN pode reconhecer o significado transportado por cada morfema com o fim de obter e representar o significado. Por exemplo, ao juntar o sufixo –am a um verbo, o sistema reconhece que a ação do verbo teve lugar no passado.
- **Lexical** – A este nível, os humanos bem como os sistemas de PLN interpretam o significado de palavras individuais. Existem vários tipos de processamento que contribuem para a compressão de nível de palavra, como por exemplo a atribuição de uma única *tag part-of-speech* para cada palavra. Neste processamento as palavras que podem ter mais funcionalidades do que um *part-of-speech* são atribuídas ao mais provável *tag part-of-speech* com base no contexto em que ocorrem.
- **Sintático** – O nível sintático foca em analisar as palavras de uma frase de modo a descobrir a estrutura gramatical da frase. Isto requer uma gramática (dicionário) e um analisador.

O *output* deste nível de processamento é uma representação da frase que revela as relações de dependência estrutural entre as palavras. Existem várias gramáticas que podem ser

utilizadas e que por sua vez irão influenciar a escolha de um analisador. No nível sintático do PLN nem todas as aplicações requerem uma análise integral das frases, portanto, as fases restantes da análise do anexo da frase preposicional e o conjunto de *scoping* não param as aplicações para quais as suas dependências das frases e dependências de cláusulas são suficientes. A sintaxe demonstra o que significa as palavras na maioria das línguas porque a ordem e a dependência contribuem para o significado da frase. Por exemplo estas duas frases “o cão perseguiu o gato” e o “o gato perseguiu o cão” diferem apenas em termos de sintaxe, mas conseguem transmitir significados bastante diferentes:

- **Semântico** – A semântica determina o processamento de possíveis significados de uma frase ao focar-se nas interações dentro dos significados do nível de palavra numa frase. Este nível de processamento inclui uma desambiguação semântica de palavras com múltiplos sentidos. A desambiguação semântica permite um e apenas um sentido de palavras polissémicas que sejam selecionadas e incluídas numa representação semântica da frase. Por exemplo, a palavra “banco” pode referir-se a uma entidade financeira ou simplesmente a um sítio para uma pessoa se sentar. Se a infirmação do resto da frase for necessária para a desambiguação, o nível semântico, e não o nível léxico, fará a desambiguação. Existe um leque abrangente de métodos que podem ser implementados para a fazer a desambiguação, alguns precisam de informação como a frequência com que cada sentido ocorre num particular *corpus* de interesse, ou numa maneira geral, alguns podem precisar da consideração do contexto local, e outros que utilizam conhecimento pragmático do assunto do documento.
- **Discurso** – Enquanto a sintaxe e a semântica trabalham com as unidades do cumprimento da frase, o nível de discurso do PLN trabalha com unidades de texto maiores do que uma frase. Isto faz com que não interprete textos como muitas frases concatenadas em que cada uma poderá ser individualmente interpretada. Em vez disso, o discurso foca as propriedades do texto como um todo e transmite o significado fazendo conexões entre as frases que compõe o texto. Existem vários tipos de processamento de discurso que podem ocorrer neste nível, dois dos mais comuns são a resolução anáfora e o reconhecimento da estrutura do texto. A resolução anáfora e a substituição de palavras como pronomes, com a entidade que o texto quer referir. O Reconhecimento da estrutura de texto determina as funções das frases no texto,

em que podem contribuir para representação significativa do texto. Por exemplo, os artigos dos jornais podem ser desconstruídos em componentes de texto/discurso como: História principal, eventos anteriores, avaliação.

- **Pragmático** – Este nível preocupa-se com o uso correto da linguagem e utiliza o contexto sobre e acima do conteúdo do texto para o entendimento do mesmo. O objetivo é explicar como o significado extra é lido nos textos sem que ele seja codificado neles. Isto requer muito conhecimento incluindo o significado as intenções, planos e objetivos. Algumas aplicações em PLN podem utilizar bases de conhecimento e módulos de inferência.

Os sistemas de PLN tendem a implementar módulos para alcançar maioritariamente os níveis mais baixos de processamento. Isto acontece por várias razões, tais como a aplicação não requerer a interpretação a níveis mais altos. Como os níveis mais baixos têm sido mais investigados e implementados, os mesmos lidam com unidades mais pequenas de análise, por exemplo, morfemas, palavras, frases, nas quais tem as suas regras *rule-governed*, enquanto os níveis mais altos de processamento de linguagem são os que lidam com textos e conhecimento do mundo os quais são *regularity-governed* (Liddy, 2001).

#### 3.4.4. Abordagens da Linguagem Natural

As abordagens do PLN são divididas em quatro categorias:

- **Abordagem Simbólica** – Atua nas análises profundas de fenómenos linguísticos e é baseada numa representação explícita dos factos sobre a linguagem através de esquemas de representação de conhecimento de fácil compreensão e de algoritmos associados.
- **Abordagem Estatística** – Utiliza várias técnicas matemáticas e normalmente usa grandes *corpus* de texto para desenvolver modelos generalizados de fenómenos linguísticos. Estes modelos baseiam-se em exemplos de atuações desses fenómenos fornecidos pelo *corpus* de texto sem adicionar linguísticas significantes ou *world knowledge*. As abordagens estatísticas usam os dados observáveis como a primeira fonte de evidência.
- **Abordagem Conectora** – Assim como a abordagem estatística, também desenvolve modelos generalizados através de fenómenos linguísticos. O que diferencia a abordagem conectora dos métodos estatísticos é que os modelos conetores

combinam a aprendizagem estatística com várias teorias de representação, assim as representações conectoras permitem a transformação, inferência, e manipulação da fórmula lógica. Para além disso, nos sistemas conectores, os modelos linguísticos são difíceis de observar devido ao facto das arquiteturas conectoras serem menos restritas do que as arquiteturas estatísticas.

- **Abordagem Híbrida** – Ainda em desenvolvimento, estas juntam os pontos fortes de cada abordagem para resolver os problemas dos Sistemas de PLN de uma maneira mais efetiva e flexível (Liddy, 2001).

#### 3.4.5. Aplicações do PLN

O PLN dispõe de bastantes implementações para um leque variado de aplicações. As aplicações que mais frequentemente utilizam o PLN são (Weikum, 2002):

- Aplicações de Recolha de Informação;
- Aplicações de Extração de Informação;
- Aplicações de Pergunta-Resposta;
- Aplicações de Sumarização;
- Aplicações de Tradução Máquina;
- Aplicações em Sistemas de diálogo;
- Aplicações em *Clusterização*.

#### 3.4.6. PLN na Medicina

Na medicina existem vários sistemas de PLN, tanto parciais como integrais. Muitos dos sistemas parciais operam ao nível da palavra, isto é, da análise morfológica; e no nível conceptual, mais concretamente, análise semântica, e também geralmente em ferramentas de codificação automáticas. Os sistemas integrais de PLN tem como alvo a representação de conhecimento expresso nas frases do documento. Um objetivo importante é a recolha de informação pelo meio definido pelo utilizador, com *queries* (Spyns, 1996).

A visão inicial de um sistema de PLN médico foi implementada no sistema no *Linguistic String Project* (LSP) que desenvolveu os componentes básicos e a representação formal da narrativa clínica, e implementou a transformação de documentos clínicos de texto livre numa representação formal (Sager, Lyman, Bucknall, Nhan, & Tick, 1994). O sistema LSP evoluiu para

o *Medical Language Processor* (MLP), que inclui o campo léxico-sintático Inglês sobre cuidados de saúde e o campo léxico médico *tagged*. O analisador MLP analisa a gramática de Inglês médico, seleciona padrões de coocorrência médica, transforma o Inglês, realiza a regularização sintática, o mapeamento de estruturas de formato da informação médica, e tem um conjunto de ferramentas XML para navegação e amostra.

O MedLEE é um sistema de PLN que extrai informações de narrativas Clínicas e apresenta essas informações num formato estruturado usando um vocabulário controlado. Este usa um campo léxico para mapear termos em classes semânticas e uma gramática semântica para gerar uma representação formal das frases. Ele está a ser usado *Columbia University Medical Center*, e é um dos poucos sistemas de PLN integrados com sistemas de informação clínica. O MedLEE foi usado com sucesso para processar relatórios de radiologia, Notas de alta, Notas *regist-out*, relatórios de patologias, dos relatórios de eletrocardiograma, e relatórios de ecocardiograma (Chen, Hripcsak, Xu, Markatou, & Friedman, 2008). Friedman (Friedman, 2005) faz uma visão geral em profundidade do sistema com um cenário de caso.

A arquitetura *Text Analytics* desenvolvida em colaboração entre a *Mayo Clinic* e a *International Business Machines* (IBM) usou a *Unstructured Information Architecture Management* para identificar entidades clinicamente relevantes nas Notas Clínicas. As entidades estão a ser subsequentemente usadas para Recuperação de Informação (*Information Retrieval*) e DM (Pakhomov, Buntrock, & Duffy, 2005). O desenvolvimento contínuo desta arquitetura resultou em duas vias especializadas: *Medical Knowledge Analysis Tool* (medKAT/P), que extrai características de cancros de relatórios de patologia, e o *clinical Text Analysis and Knowledge Extraction System* (cTAKES), que identifica distúrbios, drogas, sítios anatómicos e procedimentos em Notas Clínicas. Avaliadas num conjunto de relatórios de patologia do cancro do cólon anotados manualmente, o *Medical Text Analysis System* (MedTAS/P) alcançou pontuação de *F1* na zona dos 90% na extração de histologia, entidades anatómicas e tumores primários (Codon et al., 2009). A pontuação mais baixa alcançada para tumores metastáticos foi atribuído ao pequeno número de casos nos conjuntos de treinamento e teste (Codon et al., 2009). O cTAKES e *Health Information Text Extraction* (HITEx), descritos abaixo, são os primeiros sistemas de PLN em clínica geral, a serem disponibilizados ao público.

Desenvolvido no *National Center for Biomedical Computing, Informatics for Integrating Biology & the Bedside* (I2B2), o HITEx é uma ferramenta baseada em *General Architecture for Text*



*Engineering* (GATE) e é um sistema modular que reúne uma via diferente para extrair conclusões específicas da narrativa clínica. Por exemplo, uma via para extrair diagnósticos é formada aplicando sequencialmente um divisor de secção, um filtro de secção, um divisor de frases, um *Tokenizer* de frases, POStagger, um substantivo localizador de frase, um conceito mapeador de UMLS, e o localizador de negação (Zeng et al., 2006). Uma via para a extração da história da família de Notas de alta e Notas ambulatorio avaliadas em 350 frases alcançou 85% de precisão e 87% de *recall* na identificação de diagnósticos; 96% de precisão e 93% de *recall* na diferenciação da história familiar da história do doente; e 92% de precisão e *recall* assumindo exatamente os diagnósticos a atribuir aos membros da família (Goryachev, Kim, & Zeng-Treitler, 2008).

O sistema de classificação médico, MediClass, foi projetado para detetar automaticamente eventos clínicos em qualquer registo médico eletrónico, analisando as partes codificadas e sem texto de registo. Este foi avaliado na deteção de prestação de cuidados para a cessação do tabagismo; eventos adversos de imunização; e subtipos de retinopatia diabética. Embora a arquitetura do sistema se tenha mantido constante para cada tarefa clínica de deteção de eventos, foram definidas novas regras de classificação e terminologia para cada tarefa (Hazlehurst, Frost, Sittig, & Stevens, 2005). Por exemplo, para detetar possíveis reações vacinais nas Notas Clínicas, os desenvolvedores do MediClass identificaram os conceitos relevantes e as estruturas linguísticas usadas em Notas Clínicas para gravar e atribuir um evento adverso para uma imunização ou vacina (Hazlehurst, Mullooly, et al., 2005).

Os termos e estruturas identificadas foram codificados em regras de um módulo de conhecimento MediClass que define o esquema de classificação para a deteção automática de possíveis reações vacinais. O esquema requer a deteção de uma menção explícita de um evento imunização e deteta ou infere pelo menos um achado de um evento adverso (Hazlehurst, Mullooly, et al., 2005). Em 227 dos 248 casos (92%), o MediClass detetou corretamente uma possível reação vacinal (Hazlehurst, Mullooly, et al., 2005).

### 3.5. Notas Clínicas

Uma nota clínica eletrónica é computadorizada numa organização de cuidados de saúde, como um hospital, consultório, centro de saúde, etc. Esta tende a ser parte de um sistema isolado

de sistema de informação na saúde que permite o armazenamento, a recolha e a manipulação de dados médicos para reduzir o erro médico (-Alanazi, Jalab, Alam, Zaidan, & Zaidan, 2010).


<ul style="list-style-type: none"> <li>• Help</li> <li>• Logout</li> </ul>	<b>Patient Details</b>  <p><b>GME0000 Smith, Caroline</b></p> <p>Sex: Female DOB: 1940/01/01 Next of kin: John Smith</p> <p>Phone: 365-565-9090 Address: 19 Provincial Rd. Edmonton AB T6M 1R7</p>		<b>GP Details</b> <b>Name:</b> Jones, Evans <b>Phone:</b> 333-465-5545 <b>Address:</b> 11 Terrence Ave., Edmonton, AB T4Y 8U9																																																																																	
	<b>Patient Record</b> <ul style="list-style-type: none"> <li>• Summary</li> <li>• Lab Results</li> <li>• Diagnostic Images</li> <li>• Details</li> <li>• Notes or Comments</li> </ul>		<b>Other Healthcare Providers</b> <table border="1"> <thead> <tr> <th>Name</th> <th>Disp.</th> <th>Last Encounter</th> <th>Next encounter</th> <th>Right of Access</th> </tr> </thead> <tbody> <tr> <td>Diaz, Ellen</td> <td>Cardiology</td> <td>01/2006</td> <td>07/2006</td> <td>Y</td> </tr> <tr> <td>Fournier, Janice</td> <td>RN</td> <td>08/2005</td> <td>-</td> <td>N</td> </tr> <tr> <td>Cohen, Richard</td> <td>Dermatology</td> <td>07/2005</td> <td>-</td> <td>N</td> </tr> </tbody> </table>		Name	Disp.	Last Encounter	Next encounter	Right of Access	Diaz, Ellen	Cardiology	01/2006	07/2006	Y	Fournier, Janice	RN	08/2005	-	N	Cohen, Richard	Dermatology	07/2005	-	N																																																												
Name	Disp.	Last Encounter	Next encounter	Right of Access																																																																																
Diaz, Ellen	Cardiology	01/2006	07/2006	Y																																																																																
Fournier, Janice	RN	08/2005	-	N																																																																																
Cohen, Richard	Dermatology	07/2005	-	N																																																																																
<b>Alerts</b> Allergies – Sulfa Drugs • Pap smear due • Td due • A1C above target		<b>Medications</b> <table border="1"> <thead> <tr> <th>Date</th> <th>Medications</th> <th>Prescriptions</th> <th>Last Filled</th> </tr> </thead> <tbody> <tr> <td>11/1989</td> <td>Hydrochlorothiazide 25 mg</td> <td>One tab at breakfast</td> <td>12/2005</td> </tr> <tr> <td>03/1999</td> <td>Glyburide 5 mg</td> <td>One tab twice daily</td> <td>12/2005</td> </tr> <tr> <td>01/2001</td> <td>Metformin 500 mg</td> <td>Two tabs twice daily</td> <td>12/2005</td> </tr> <tr> <td>03/2001</td> <td>Atorvastatin 20 mg</td> <td>One tab at supper</td> <td>12/2005</td> </tr> <tr> <td>02/2002</td> <td>Atenolol 50 mg</td> <td>One tab at breakfast</td> <td>12/2005</td> </tr> <tr> <td>02/2002</td> <td>ECASA 325 mg</td> <td>One tab at breakfast</td> <td>12/2005</td> </tr> <tr> <td>02/2006</td> <td>Ramipril 10mg</td> <td>One tab at supper</td> <td>02/2006</td> </tr> <tr> <td>06/2005</td> <td>Cloxacillin 500 mg</td> <td>Discontinued</td> <td>-</td> </tr> <tr> <td>05/2004</td> <td>Beclothemason Cream</td> <td>Discontinued</td> <td>-</td> </tr> </tbody> </table>		Date	Medications	Prescriptions	Last Filled	11/1989	Hydrochlorothiazide 25 mg	One tab at breakfast	12/2005	03/1999	Glyburide 5 mg	One tab twice daily	12/2005	01/2001	Metformin 500 mg	Two tabs twice daily	12/2005	03/2001	Atorvastatin 20 mg	One tab at supper	12/2005	02/2002	Atenolol 50 mg	One tab at breakfast	12/2005	02/2002	ECASA 325 mg	One tab at breakfast	12/2005	02/2006	Ramipril 10mg	One tab at supper	02/2006	06/2005	Cloxacillin 500 mg	Discontinued	-	05/2004	Beclothemason Cream	Discontinued	-																																									
Date	Medications	Prescriptions	Last Filled																																																																																	
11/1989	Hydrochlorothiazide 25 mg	One tab at breakfast	12/2005																																																																																	
03/1999	Glyburide 5 mg	One tab twice daily	12/2005																																																																																	
01/2001	Metformin 500 mg	Two tabs twice daily	12/2005																																																																																	
03/2001	Atorvastatin 20 mg	One tab at supper	12/2005																																																																																	
02/2002	Atenolol 50 mg	One tab at breakfast	12/2005																																																																																	
02/2002	ECASA 325 mg	One tab at breakfast	12/2005																																																																																	
02/2006	Ramipril 10mg	One tab at supper	02/2006																																																																																	
06/2005	Cloxacillin 500 mg	Discontinued	-																																																																																	
05/2004	Beclothemason Cream	Discontinued	-																																																																																	
<b>Diagnosis</b> <table border="1"> <thead> <tr> <th>Diagnosis</th> <th>Date</th> <th>Status</th> </tr> </thead> <tbody> <tr> <td>Hypertension</td> <td>11/1989</td> <td>Ongoing</td> </tr> <tr> <td>Diabetes</td> <td>05/1996</td> <td>Ongoing</td> </tr> <tr> <td>Coronary Artery Disease</td> <td>02/2002</td> <td>Ongoing</td> </tr> <tr> <td>Fasting lipids</td> <td>12/2005</td> <td>-</td> </tr> <tr> <td>Exercise stress test</td> <td>1/2005</td> <td>-</td> </tr> <tr> <td>Coronary angiogram / Cellulitis</td> <td>02/2005</td> <td>Resolved</td> </tr> <tr> <td>Cholecystectomy</td> <td>05/1981</td> <td>Resolved</td> </tr> <tr> <td>Cesarian section</td> <td>01/1967</td> <td>Resolved</td> </tr> </tbody> </table>		Diagnosis	Date	Status	Hypertension	11/1989	Ongoing	Diabetes	05/1996	Ongoing	Coronary Artery Disease	02/2002	Ongoing	Fasting lipids	12/2005	-	Exercise stress test	1/2005	-	Coronary angiogram / Cellulitis	02/2005	Resolved	Cholecystectomy	05/1981	Resolved	Cesarian section	01/1967	Resolved	<b>Encounter History</b> <table border="1"> <thead> <tr> <th>Date</th> <th>Facility</th> <th>Speciality</th> <th>Clinician</th> <th>Reason</th> <th>Type</th> </tr> </thead> <tbody> <tr> <td>02/2006</td> <td>GP</td> <td>-</td> <td>-</td> <td>Hypertension</td> <td>-</td> </tr> <tr> <td>01/2006</td> <td>Cardio Assoc</td> <td>Cardiology</td> <td>Diaz, E.</td> <td>CAD</td> <td>Outpatient</td> </tr> <tr> <td>12/2005</td> <td>GP</td> <td>-</td> <td>-</td> <td>Diabetes</td> <td>-</td> </tr> <tr> <td>10/2005</td> <td>General Hosp</td> <td>Dietician</td> <td>Johnson, H.</td> <td>Diabetes teaching</td> <td>Outpatient</td> </tr> <tr> <td>08/2005</td> <td>GP</td> <td>-</td> <td>-</td> <td>Diabetes</td> <td>-</td> </tr> <tr> <td>08/2005</td> <td>GP</td> <td>-</td> <td>-</td> <td>Cellulitis</td> <td>-</td> </tr> <tr> <td>08/2005</td> <td>Home Visit</td> <td>RN</td> <td>Fournier, J.</td> <td>Cellulitis</td> <td>-</td> </tr> <tr> <td>07/2005</td> <td>Polyclinic</td> <td>Dermatology</td> <td>Cohen, R.</td> <td>Stasis dermatitis</td> <td>Outpatient</td> </tr> </tbody> </table>		Date	Facility	Speciality	Clinician	Reason	Type	02/2006	GP	-	-	Hypertension	-	01/2006	Cardio Assoc	Cardiology	Diaz, E.	CAD	Outpatient	12/2005	GP	-	-	Diabetes	-	10/2005	General Hosp	Dietician	Johnson, H.	Diabetes teaching	Outpatient	08/2005	GP	-	-	Diabetes	-	08/2005	GP	-	-	Cellulitis	-	08/2005	Home Visit	RN	Fournier, J.	Cellulitis	-	07/2005	Polyclinic	Dermatology	Cohen, R.	Stasis dermatitis	Outpatient
Diagnosis	Date	Status																																																																																		
Hypertension	11/1989	Ongoing																																																																																		
Diabetes	05/1996	Ongoing																																																																																		
Coronary Artery Disease	02/2002	Ongoing																																																																																		
Fasting lipids	12/2005	-																																																																																		
Exercise stress test	1/2005	-																																																																																		
Coronary angiogram / Cellulitis	02/2005	Resolved																																																																																		
Cholecystectomy	05/1981	Resolved																																																																																		
Cesarian section	01/1967	Resolved																																																																																		
Date	Facility	Speciality	Clinician	Reason	Type																																																																															
02/2006	GP	-	-	Hypertension	-																																																																															
01/2006	Cardio Assoc	Cardiology	Diaz, E.	CAD	Outpatient																																																																															
12/2005	GP	-	-	Diabetes	-																																																																															
10/2005	General Hosp	Dietician	Johnson, H.	Diabetes teaching	Outpatient																																																																															
08/2005	GP	-	-	Diabetes	-																																																																															
08/2005	GP	-	-	Cellulitis	-																																																																															
08/2005	Home Visit	RN	Fournier, J.	Cellulitis	-																																																																															
07/2005	Polyclinic	Dermatology	Cohen, R.	Stasis dermatitis	Outpatient																																																																															
		<b>Immunizations</b> <table border="1"> <thead> <tr> <th>Type</th> <th>Most Recent</th> <th>Number Received</th> </tr> </thead> <tbody> <tr> <td>Influenza</td> <td>11/2005</td> <td>7</td> </tr> <tr> <td>Pneumovax</td> <td>03/2005</td> <td>1</td> </tr> <tr> <td>Twinrix</td> <td>08/2002</td> <td>3</td> </tr> <tr> <td>Td</td> <td>04/1996</td> <td>1</td> </tr> </tbody> </table>		Type	Most Recent	Number Received	Influenza	11/2005	7	Pneumovax	03/2005	1	Twinrix	08/2002	3	Td	04/1996	1																																																																		
Type	Most Recent	Number Received																																																																																		
Influenza	11/2005	7																																																																																		
Pneumovax	03/2005	1																																																																																		
Twinrix	08/2002	3																																																																																		
Td	04/1996	1																																																																																		
		<b>Diabetic Indices</b> <table border="1"> <thead> <tr> <th>Type</th> <th>Value</th> <th>Most Recent</th> </tr> </thead> <tbody> <tr> <td>A1C</td> <td>0.071</td> <td>12/2005</td> </tr> <tr> <td>LDL</td> <td>2.41</td> <td>12/2005</td> </tr> <tr> <td>BP</td> <td>135/75</td> <td>02/2006</td> </tr> <tr> <td>Urine Microalb</td> <td>0.02</td> <td>08/2005</td> </tr> <tr> <td>Eye Exam</td> <td>-</td> <td>05/2005</td> </tr> <tr> <td>Home Gluc (average)</td> <td>7.4</td> <td>01/2006</td> </tr> </tbody> </table>		Type	Value	Most Recent	A1C	0.071	12/2005	LDL	2.41	12/2005	BP	135/75	02/2006	Urine Microalb	0.02	08/2005	Eye Exam	-	05/2005	Home Gluc (average)	7.4	01/2006																																																												
Type	Value	Most Recent																																																																																		
A1C	0.071	12/2005																																																																																		
LDL	2.41	12/2005																																																																																		
BP	135/75	02/2006																																																																																		
Urine Microalb	0.02	08/2005																																																																																		
Eye Exam	-	05/2005																																																																																		
Home Gluc (average)	7.4	01/2006																																																																																		

Figura 4 – Exemplo de uma Nota Clínica Eletrônica (retirado de Office of the Auditor General of Canada (“Office of the Auditor General of Canada,” 2010))

As Notas Clínicas eletrônicas são um conjunto longitudinal de Notas Clínicas com a informação do utente resultante das suas idas às organizações que prestam cuidados de saúde. Como podemos ver na Figura 4, uma Nota Clínica inclui todos os dados do utente e permite ver esses dados ao consultar a mesma.

As Notas Clínicas eletrônicas contêm todo o histórico de doenças do utente, assim como a sua demografia, problemas atuais de saúde, medicação habitual, sinais vitais, imunizações, dados de laboratório e relatórios de raio-x. As Notas Clínicas eletrônicas automatizam e criam linhas cronológicas do doente. Um dos pontos fortes das mesmas é criar um registo completo da passagem do doente pela organização de cuidados de saúde, bem como, servir de apoio a outras atividades relacionadas com os serviços de saúde, diretamente ou indiretamente, através da interface. Esta inclui o suporte à decisão de eventos ocorridos, a gestão de qualidade e obtenção dos resultados de relatórios médicos. Um relatório médico poderá ser criado por cada serviço médico que o doente usufrua, como a radiologia, laboratório, a farmácia ou como resultado de uma ação administrativa. Existem também sistema médicos que permitem a captura de sinais vitais, anotações de enfermagem, e ordens do médico (-Alanazi et al., 2010).

Uma nota clínica eletrônica pode fornecer uma infraestrutura eletrônica para oito tipos de atividades Clínicas e administrativas executadas fisicamente. Além disso, as Notas Clínicas eletrônicas têm um vasto leque de capacidade e tem bastante potencial para aumentar a qualidade do serviço (R. H. Miller & Sim, 2004).

### *3.5.1. Tipos de Notas Clínicas*

Existem bastantes tipos de Notas Clínicas, como por exemplo (Voorhees & Hersh, 2012):

- Relatórios de Radiologia;
- Relatórios do Histórico do Utente;
- Relatórios de Exames Médicos;
- Relatórios de Consulta;
- Relatórios da Urgência;
- Relatórios de Progresso;
- Relatórios / Diários por serviço / especialidade;
- Notas de Alta;
- Notas de Admissão;
- Diários Clínicos:
- Relatórios Operacionais;
- Relatórios de Cirurgia Patológica;
- Relatórios de Cardiologia.

### *3.5.2. Regras das Notas Clínicas (Diários Clínicos)*

Regras Básicas na documentação de Diários Clínicos (Columbia University, 2006):

- As Notas Clínicas têm de ter data e hora;
- Os médicos têm de incluir um título breve para todas as entradas de Notas Clínicas; identificar-se e seu papel;
- Os médicos são aconselhados a evitar abreviações;
- Não é permitido copiar e colar a partir de Notas anteriores sem edição e atualização; usando observações e avaliações de outro provedor é antiético e não profissional;
- Além do médico assinar as Notas, ele terá que imprimir o seu nome de forma legível e incluir o seu número de *pager*.

### 3.5.3. Guidelines para os Diários Clínicos

A *Columbia University* (Columbia University, 2006), definiu as seguintes *Guidelines* para as Notas Clínicas Diárias:

1. O objetivo dos diários clínicos é fornecer uma informação diária dos doentes e as suas doenças, e anotar a evolução do seu diagnóstico e tratamento.
2. Os Diários Clínicos devem ser escritos no formato de SOAP (subjetivo, objetivo, avaliação e plano) orientado para o problema, como por exemplo:
  - Subjetivo – deve incluir as informações do doente sobre seus sintomas e desejos, relatos da família e de outros profissionais de saúde (por exemplo. "Enfermeira relata que o doente teve uma noite sem dormir.")
  - Objetivo – só deve incluir novas informações. A informação que está disponível para outras pessoas pode ser resumida, ou incluir apenas os resultados anormais ou de mudança. Esta seção deve incluir os seguintes tipos de informações:
    - I. Sinais vitais e exame físico;
    - II. Dados de laboratório;
    - III. Dados de imagem.
  - Avaliação e planos devem ser resumidos por problema. Os problemas podem ser doenças diagnosticadas ou síndromes, ou sintomas, complexos de sintomas ou anormalidades dos exames, laboratórios ou imagem. Embora às vezes um problema pode ser melhor expresso por referência a um sistema de órgãos, por exemplo: "anormalidades pulmonares", geralmente a lista de problemas não é simplesmente uma lista de sistemas de órgãos (algumas doenças ou problemas que envolvem vários sistemas de órgãos, alguns sistemas de órgãos podem ter mais do que um problema discreto).
    - I. A avaliação é a parte mais importante. A avaliação pode ser breve, mas deve incluir o diagnóstico diferencial de trabalho ou diagnóstico estabelecido, a gravidade ou o prognóstico quando for o caso, e o *status* do problema, ou seja, se o doente melhorou, piorou, ou desenvolveu problemas adicionais. É onde se interpreta as mudanças subjetivas no estado do doente, os novos resultados do teste de diagnóstico, resume o *input* de consultores, e articula uma opinião "re": o desdobramento de diagnóstico e tratamento do doente.

II. O plano é o que está a seguir; que pode ser dividido entre os planos de diagnóstico (ou monitorização) e planos terapêuticos.

- O último problema necessário (análogo à conclusão de listas de problemas ambulatoriais com "manutenção da saúde") deve ser "disposição" ou "planeamento de alta" e deve sempre incluir planeamento ou necessidades de alta.
  - A Formulação ou impressão não é necessária para cada nota de progresso, e não deve ser simplesmente uma recapitulação de diagnósticos ou conclusões passadas ou presentes do doente. Uma avaliação global breve pode ser prestada quando é útil ter uma avaliação resumida do estado do doente ou o progresso, ou a sua morbilidade cumulativa, ou a interação de vários problemas diferentes.
3. Notas Clínicas diárias podem e devem ser relativamente breves, com foco na evolução desde a nota anterior, e recapitulando apenas os problemas ativos em curso. Não é permitido cortar e colar a partir de Notas anteriores, sem edição ou atualização e informação desatualizada e redundante devem ser eliminados a partir de Notas.

#### *3.5.4. Problemas das Notas Clínicas*

O Texto Livre nas Notas Clínicas coloca desafios técnicos significativos para PLN. Além disso, há exigências éticas e sociais quando se trabalha com esses dados, que são destinados ao uso por profissionais e médicos treinados que apreciam as restrições que impõe a confidencialidade do doente (Pestian et al., 2007).

O estado da arte nos Sistemas de PLN lida melhor e com mais cuidado o texto editado do que Notas fragmentárias, e a linguagem clínica é conhecida por apresentar características únicas de sub-idiomas (Hirschman & Sager, 1982)(Friedman, Kra, & Rzhetsky, 2002) (Stetson, Johnson, Scotch, & Hripcsak, 2002) (por exemplo, frases sem verbo, semântica de pontuação de domínio específico, e metonímias incomuns) que podem limitar o desempenho das ferramentas de PLN.

Mais importante ainda, os requisitos de confidencialidade levam tempo e esforço para se desenvolver, por isso não é de estranhar que muito trabalho na área biomédica tenha sido focado em artigos editados em revistas (e no domínio genómico) em vez de texto livre nas Notas Clínicas. A verdade, no entanto, é que a automação de fluxos de trabalho na área da Saúde pode trazer benefícios importantes ao tratamento (Hurtado, Swift, & Corrigan, 2001) e reduzir a carga

administrativa. Nestes casos o texto livre é um componente crítico de esses fluxos de trabalho (Pestian et al., 2007).

Há vários desafios para traduzir frases de texto livre em texto padronizado. A pesquisa por códigos-conceito deve ser precisa e rápida o suficiente para que os anotadores não percam a paciência. Os anotadores também são propensos a cometer erros ortográficos e muitas vezes usam abreviaturas que podem ter mais de um significado. Além disso, o conceito de *Unified Medical Language System* (UMLS) pode ser descrito de várias maneiras diferentes, ou os anotadores podem desejar codificar um conceito de que simplesmente não existe nas UMLS. Às vezes, o anotador pode não estar satisfeito com o nível de especificidade dos códigos retornados e pode querer olhar para conceitos relacionados. Estas questões são abordadas e comparações de precisão e de pesquisa são elaboradas para uma variedade de frases médicas (Shu, 2005).

### 3.6. Sistemas de Apoio à Decisão Clínica

Os Sistemas de Apoio à Decisão Clínica (SADCs) são programas interativos computadorizados que têm como objetivo ajudar os médicos e outros profissionais de saúde (Gamberger et al., 2008). Estes ajudam na prescrição de medicamentos, diagnóstico e gestão de doenças, para melhorar os serviços e reduzir custos, riscos e erros (El-Sappagh & El-Masri, 2014). Podem ser utilizados para verificar se há alergias a medicamentos, comparar custos dos mesmos e diferentes laboratórios, avaliar o potencial de interações entre estes, sugerir alternativas, bloquear pedidos duplicados, sugerir posologias, e fornecer recomendações. Além disso, os SADCs podem fornecer conhecimentos clínicos e padrões de melhores práticas e diretrizes para médicos não-especialistas (El-Sappagh & El-Masri, 2014).

O SADC deve ser integrado com sistemas de Notas Clínicas eletrônicas e sistemas computadorizados, que estão ligados a outros Sistemas de Informação (por exemplo, de laboratório, radiologia, faturação). Os componentes básicos de um SADC incluem uma base de conhecimento médico e um mecanismo de inferência (geralmente um conjunto de regras que são derivados de especialistas e medicina baseada em evidências) e são executados através de módulos lógicos médicos com base em uma linguagem como a sintaxe *Arden*, ou utilizando uma rede neural artificial (El-Sappagh & El-Masri, 2014).

O INTCare é um SADC baseado no KDD, e em paradigmas *Agent-Based* com o objetivo de ajudar a tomada de decisão em questões médicas. O INTCare é um sistema que ajuda os médicos a tomar decisões através da deteção das condições do doente através de atualizações contínuas sobre o seu estado de saúde e aplicando o modelo de previsão para prever as possíveis falhas que possam ocorrer no próximo dia. O INTCare também executa uma manutenção *up-to-date* sobre a probabilidade de morte usadas num processo de decisão *end-of-life*. Além disto o INTCare também avalia os cenários de evolução da condição do doente, permitindo aos médicos comparar as consequências de diferentes procedimentos médicos (Gago et al., 2002).

## 4. Soluções envolvendo o *Text Mining* o Processamento de Linguagem Natural e as Notas Clínicas

Este capítulo aborda o levantamento do estado da arte relacionado com esta dissertação. São descritos dois estudos que utilizam as ferramentas de descoberta de conhecimento que foram utilizadas nesta dissertação, e que foram úteis para a realização da mesma. Os resultados encontrados foram relativamente poucos e não estão relacionados diretamente com a área da saúde, pois ainda não existe no mercado muitos estudos que envolvam o *Text Mining* (TM) e o Processamento de Linguagem Natural (PLN) na descoberta de conhecimento em Notas Clínicas.

### 4.1. Usar o TM e o PLN para o Processamento de Seguros de Saúde

Popowich (Popowich, 2005) criou um sistema que junta o TM e o PLN para o processamento de Seguros na Saúde. Quando um doente que possui um seguro de saúde precisa de cuidados médicos, a seguradora tem de os pagar ao doente, mas existem casos de fraude à seguradora e casos onde existe outra seguradora que terá de cobrir os custos de saúde. Na utilização do PLN, Popowich usou o *Content Intelligence System* e o *Concept Specification Language*.

O *Axonwave Content Intelligence System* (ACIS) contém sistemas de PLN centrais que realizam PLN *rule-based* baseado em estatística. Este é capaz de alavancar fontes de conhecimento existentes, além de fornecer a capacidade para utilizadores comuns para adaptar ou personalizar a base de conhecimento com conceitos que são de interesse para eles (Popowich, 2005).

O sistema faz uso de uma *tagger* estatística, baseada num analisador parcial de *rule-based*, juntamente com recursos externos, incluindo *Wordnet* (G. A. Miller, Beckwith, Fellbaum, Gross, & Miller, 1993). O *tagger* e o analisador parcial são robustos e capazes de lidar com texto



sem gramática encontrado em registos de chamadas, que contém numerosos casos de abreviaturas e acrónimos, bem como texto mais elaborado encontrado em documentos do plano de serviços médicos. O analisador parcial fornece mais informações do que apenas palavras com o *tag*. Ele a fornece identificação (nome próprio), também determina os argumentos e modificadores de relacionamentos e entidades encontradas em um documento (conforme o caso) (Popowich, 2005).

A tecnologia de núcleo diz respeito à adequação dos conceitos que são representados num *Concept Specification Language* (CSL). CSL é usado para especificar padrões linguísticos ricos que incorporam fundamentalmente a noção de recursão (incorporação) de padrões e vários predicados linguísticos (Popowich, 2005).

CSL e o conceito de correspondência são incorporados no CIS, que analisa a estrutura de palavras, frases (fazendo uso das regras linguísticas de uso geral e dicionários). A primeira fase da análise de expansão consiste na abreviatura e correção ortográfica, que é então seguido do *tagging*, e em seguida, a análise parcial (Abney, 1996). As informações específicas podem então ser extraídas de acordo com regras e conceitos formulados com o CSL, que é organizado dentro de várias taxonomias. CSL permite a definição de conceitos ou termos-chave; e a especificação da inter-relação entre os conceitos na forma de vários operadores, tais como *OR*, *NOT*, *“Immediately Precedes”*, *“Is Related”*, ou *“Causes”*; e também a formulação de categorias avançadas para conceitos, como se um conceito é uma palavra, tem sinónimos, ou é um termo geral ou específico, etc. (Popowich, 2005).

No documento foram identificados alguns indicadores que são importantes para descobrir quais as principais partes que devem de ser responsáveis para cobrir os custos dos seguros médicos. Esses indicadores recaíram nestas categorias (Popowich, 2005):

- *Commercial Coordination of Benefits;*
- *Medicare Coordination of Benefits;*
- *No-fault Recovery;*
- *Subrogation Recovery;*
- *Workers Compensation.*

Embora seja possível criar especificações muito complexas e precisas utilizando o CSL, esta pode ser uma tarefa muito demorada, além disso, pode exigir bastante conhecimento

linguístico, e experiência na área. Para facilitar esta tarefa, pode-se alavancar a perícia linguística e domínio contido dentro das regras linguísticas e da base de conhecimento de um sistema de PLN para ajudar na criação de novos CSL. Assim, podemos arrancar a partir de um sistema existente para criar um novo sistema que tem uma base de conhecimento mais rica usando o texto de criação de conceito, permitindo que um utilizador consiga criar um sistema CSL sem qualquer conhecimento da CSL (Popowich, 2005).

O algoritmo de criação do conceito baseado em texto consiste nos seguintes oito etapas (Popowich, 2005):

- *Input* de fragmentos de texto;
- Fragmentos separados em palavras;
- Seleção de palavras relevantes;
- Relação de conceito;
- Remoção de Relação de Conceito;
- Contruir relações de conceito;
- Relações escritas como um conceito CSL.

O resultado deste processo de avaliação é uma lista priorizada de reivindicações médicas, onde a pontuação é o valor calculado a partir dos diferentes indicadores. A pontuação é um número inteiro entre 1 e 1000. Finalmente, uma vez que um sistema de reivindicações de cuidados de saúde a auditoria é um sistema que envolve um ser humano na investigação dos créditos resultantes e casos, é possível, ao longo do tempo, construir um *corpus* rico que, juntamente com uma vasta variedade de indicadores associados. Com estes dados, é possível mudar automaticamente os pesos associados com os diferentes indicadores, ou mesmo introduzir novos indicadores na equação (Popowich, 2005).

#### 4.2. Extração automática em texto livre de micro-organismos e os seus *habitats* usando *workflows* de TM

Kolluru, Nakjang, Hirt, Wipat, & Ananiadou, (2011) fizeram um teste usando técnicas de TM para extrair automaticamente instâncias de micro-organismos e os seus *habitats* no texto livre. Estas entradas podem ser curadas e adicionadas em bases de dados diferentes. Para este fim, os autores usaram o um classificador *Conditional Random Field* (CRF) como parte dos seus *workflows*

para extrair a menção dos micro-organismos, habitats e a sua inter-relação entre organismos e os seus *habitats*.

O objetivo desta investigação foram os seguintes:

- Até que ponto as aproximações estatísticas são eficazes para extrair micro-organismos e os seus *habitats*?
- Desenvolver *workflows* para combinar o processamento de texto, e o reconhecimento da entidade nomeada e a relação-*mining*.

Como esta investigação é um novo passo na aplicação do TM, não existiam *corpus* padronizados para treinar os dados para o algoritmo de *Machine Learning* ou para servir uma avaliação. Então, os autores tiveram de criar um *corpus* novo, que é o *corpus* “Organismo-habitat”. Nesse *corpus*, duas classes de entidades foram identificadas: Micro-organismos e *habitats* (Kolluru et al., 2011).

- Micro-organismos: os nomes científicos de micro-organismos, incluindo bactérias, *archaea* e eucariotas microbianas, são anotadas:
  1. Se eles forem especificados, pelo menos, ao nível de género de precisão, espécie, estirpe. Os exemplos típicos de microrganismos são *Campylobacter spp.*, de *Escherichia coli K12* e *Trichomonas vaginalis*.
  2. Se eles estão em frases que contêm habitat ou fonte de isolamento informações para o organismo.
- *Habitats*: *habitats* ou de isolamento de fontes de organismos são identificadas:
  1. Se eles estão relacionados com o contexto ou pode ser referido como um habitat ou o isolamento de um organismo fonte. Por exemplo, se eles se referem a um organismo hospedeiro (humano, vaca), uma parte do corpo ou órgão de um organismo hospedeiro (pulmão, intestino, abscesso pulmonar), referem-se a um habitat ambiental (águas residuais), ou empregar o adjetivo formas de *habitats* listados acima (bovina, pulmonar).
  2. Se eles não estão associados diretamente com a doença, por exemplo, diarreia, infeção do trato respiratório.
  3. Se forem em frases, que contêm o organismo associado com o habitat.

Uma frase totalmente anotada com organismo e habitat informação é mostrada aqui, sendo a abordagem seguida pelos autores a presente na Figura 5, *Bacteroides salyersae. sp. nov.* isolado a partir de amostras clínicas humanas de origem intestinal.



Figura 5 – Abordagem Utilizada Pelos Autores (retirado de (Kolluru et al., 2011))

O *workflow* do reconhecedor da entidade nomeada emprega campos aleatórios condicionais, CRF's (Lafferty, McCallum, & Pereira, 2001) usando uma combinação de dicionário (NCBI / Lista de habitat) recursos, recursos lexicais, características ortográficas e características contextuais. Os CRF são um tipo de probabilidades discriminativos e modelos gráficos não direcionados usados frequentemente para marcação de dados sequenciais e no reconhecimento de entidades mencionadas em domínios de processamento e de linguagem natural biológica [(McDonald & Pereira, 2005), (Finkel & Manning, 2009)]. Nesta implementação, os autores usaram o *Machine Learning for Language Toolkit* (MALLET) ("MALLET," 2013) na implementação de CRF com recursos lexicais e ortográficas para treinar o modelo CRF. Também aplicaram dois dicionários que foram adaptados para a tarefa a partir de uma combinação de ontologias de domínio estabelecidos e com coração e listas de termos fornecidas pelos especialistas do domínio (Kolluru et al., 2011).

Recursos para o modelo CRF incorporam três conjuntos principais de recursos de base, inspiradas em pesquisas anteriores no reconhecimento de nome-entidade biomédico (Sasaki, Tsuruoka, McNaught, & Ananiadou, 2008):

- Recursos lexicais são a palavra atual, a forma raiz da palavra atual, e a parte da *tag* de discurso da palavra atual, calculado pelo identificador *Genia* (Tsuruoka & Tsujii, 2005).

- Características ortográficas são feitas de *substring* e palavras características de forma. Nos recursos de forma de palavra, todas as letras maiúsculas serão convertidas para 'A', minúsculas para 'a', e todos os números para '0'. Os dois e quatro primeiros caracteres e os dois e quatro últimos caracteres da palavra original e a forma da palavra são escolhidos como características (Kolluru et al., 2011).
- Critérios de dicionário são características binárias para indicar a presença da palavra no dicionário e a posição da palavra dentro de quaisquer entradas do dicionário (Kolluru et al., 2011).

Para cada uma das características de base, as características correspondentes para palavras dentro de uma janela de contexto são adicionadas para a representação. A janela varia de 1-3 palavras anteriores e posteriores à palavra atual.

- **Abordagem Híbrida do Dicionário-*Machine Learning* (CRF Híbrido)**

Para o reconhecimento de entidades, foi utilizada a abordagem baseada CRF. A atual abordagem emprega um classificador de sequência, treinada num *corpus* anotado à mão, que foi convertido em um formato BIO padrão [Começo (*Begin*) de uma sequência, dentro (*Inside*) de uma sequência, Fora (*Outside*) de uma sequência]. Este *corpus* composto por trinta e dois trabalhos completos de várias revistas e foi especificamente anotado para micro-organismos e *habitats*. Os CRF's foram empregados com um modelo de cadeia linear (Kolluru et al., 2011).

Tabela 3 – Performance das Abordagens de Dicionário e CRF

Resultados	Micro-Organismos			<i>Habitats</i>		
	<i>P</i> (precisão)	<i>R</i> (recall)	<i>F</i> (F-score)	<i>P</i>	<i>R</i>	<i>F</i>
Dicionário	54	75	63	58	55	56
CRF	84	79	81	68	50	57

Na Abordagem baseada pelo dicionário, os resultados encontram-se nos 63% e os 56%, e pelos métodos baseados no *Machine Learning* (CRF) os valores vão para os 57% e os 81%. Na identificação dos Micro-Organismos a abordagem CRF é melhor que o dicionário, enquanto o *Recall* mantém-se num bom nível para a abordagem do dicionário, a precisão desde para os 54%.

Por outras palavras a abordagem do dicionário teve um elevado número de falsos-positivos (Kolluru et al., 2011).

Tabela 4 – Performance da Abordagem utilizada pelos Autores

Classe de Entidades	Precisão (%)	Recall (%)	F-Score (%)
Micro-Organismos	84	79	81
Habitats	68	50	57

Tabela 5 – Performance da Abordagem extração-relação

	Precisão (%)	Recall (%)	F-Score (%)
Relações	85	49	57

Na abordagem do CRF Híbrido presente nas tabelas 4 e 5, o classificador obteve um F-Score perto de 80% para os micro-organismos e 57% para os *habitats* numa *Cross Validation 9-fold* (Kolluru et al., 2011).

## 5. Estudo dos Dados

Neste capítulo são apresentados os estudos realizados sobre a utilização do *Text Mining* (TM) e o Processamento de Linguagem Natural (PLN) na análise e previsão de mortes cerebrais. Este capítulo é composto pela Contextualização do Problema, onde vai ser exposto o problema em causa. Na descrição do estudo é elaborado um resumo dos estudos que irão ser realizados. No Estudo dos Dados, contém a descrição dos dados fornecidos. Na preparação dos dados, realizou-se o tratamento dos dados para que eles pudessem ser analisados pelas ferramentas de análise.

### 5.1. Contexto do Problema

Os doentes têm a sua história clínica registada em Notas Clínicas. Estas Notas estão atualmente armazenadas em bases de dados que pertencem à organização de saúde que as armazena. As Notas Clínicas têm tipos diferentes. As Notas Clínicas podem ser Notas Clínicas de admissão, alta e progresso, por exemplo.

O maior foco das análises a realizar é a Morte Cerebral, onde foi feita uma análise onde procurou-se descobrir padrões de doentes que faleceram após o raio-x, e foram criados modelos de previsão com o objetivo de prever se os doentes iriam falecer com base no seu diagnóstico de raio-x.

O tipo de raio-x que são usados neste projeto são as tomografias computadorizadas (TC) cranioencefálicas.

#### *5.1.1. Morte Cerebral*

A morte cerebral é declarada quando os reflexos da unidade respiratória e motoras do tronco encefálico estão ausentes em uma norma térmica num doente sem medicação em coma, lesão cerebral irreversível, com uma massa conhecida e não contribuem para distúrbios metabólicos verificados. A determinação de morte cerebral em adultos tornou-se parte integrante da prática do campo neurológico e neurocirúrgico, mas pode incluir qualquer especialidade médica (Eelco & Wijdicks, 2007).

Existe uma clara diferença entre danos cerebrais graves e morte cerebral. O médico deve entender essa diferença, porque a morte encefálica significa o suporte básico de vida é inútil, e é o principal requisito para a doação de órgãos para transplante. Em adultos, as principais causas de morte cerebral são as lesões cerebrais traumáticas e hemorrágicas (Wijdicks, 1995). As considerações éticas, religiosas e filosóficas sobre a definição de morte foram abordadas em uma monografia recente (Brock, 1999).

A interpretação da TC é essencial para determinar a causa da morte cerebral. Normalmente, a TC relata o doente com uma hérnia cerebral em massa, múltiplas lesões hemisféricas com edema ou inchaço isolado. No entanto, tal declaração na TC não elimina a necessidade de uma cuidadosa pesquisa por fatores de confusão. Além disso, a TC pode ser normal no período inicial após paragem cardíaca e em doentes com meningite fulminante ou encefalite. O exame do líquido cefalorraquidiano deve revelar resultados de diagnóstico em infecção das condições do sistema nervoso central (Greer, Varelas, Haque, Eelco, & Wijdicks, 2008).

Outra necessidade de confirmação para confirmar a morte cerebral é o eletroencefalograma. O eletroencefalograma é um registo dos potenciais elétricos cerebrais por elétrodos no couro cabeludo. A atividade elétrica cerebral inclui os potenciais de ação que são breves e produzem campos elétricos circunscritos, mais lentos, mais generalizados, e com potenciais pós-sinápticos. A magnitude do sinal é gravada a partir de um gerador de neural e depende do ângulo sólido subtendido no elétrodo. Posteriormente, a atividade de um único neurónio pode ser gravada por um microelétrodo adjacente, mas não a um elétrodo de escalpe distante. A atividade síncrona num agregado laminar horizontal de neurónios com orientação paralela pode, no entanto, constituir um gerador de extensão suficiente para ser detetável no couro cabeludo (Binnie & Prior, 1994).

Existem outros testes confirmatórios para determinar se há uma morte cerebral, como a Angiografia Cerebral, o Doppler transcraniano e a cintilografia cerebral (Eelco & Wijdicks, 2007).

### *5.1.2. Análise de Texto das Notas Clínicas*

Dentro das Notas Clínicas, os raio-x são um tipo de Notas Clínicas, porque o médico tem de analisar e descrever os raio-x usando texto livre e por isso, texto não estruturado. Este tipo de escrita dificulta o processo de recuperação de informação, devido ao fato de a informação não estar estruturada. A análise de texto livre usando algoritmos inteligentes é algo relativamente novo,



por isso há poucas ferramentas capazes de fazer análise de texto em várias línguas. Uma das vantagens é que análise de texto livre ou não estruturada pode dar novas perspectivas sobre o que está escrito sobre uma determinada coisa.

Por exemplo, na área da medicina, podemos encontrar o evento que acontece mais frequentemente, em seguida, estabelecer normas baseadas nestes eventos. O objetivo desta análise é analisar texto não estruturado, e tentar encontrar padrões, a fim de fazer uma análise mais concreta do texto não estruturado em um tipo específico de eventos clínicos: Morte Cerebral. A Morte Cerebral é um dos piores eventos clínicos e alguns relatórios de TC foram usadas para realizar este estudo. análise de texto não estruturado e descoberta de conhecimento em textos é algo novo nos sistemas de informação aplicados a esta área da medicina.

### *5.1.3. Previsão de Morte Cerebral após a Realização de raio-x*

Utilizando também os raio-x, esta análise diferencia-se da anterior pois esta tem o objetivo de criar modelos de previsão que consigam prever com fiabilidade a Morte Cerebral após a Realização de um raio-x(TC). A previsão de acontecimentos é algo que começa a ser comum nos dias de hoje, mas os tipos de dados utilizados para a previsão dos acontecimentos são numéricos. Nesta análise o foco vira-se para previsão de Morte Cerebral com base no diagnóstico do raio-x.

Este diagnóstico, como já foi mencionado anteriormente, está em formato de Texto não estruturado, o que poderá dificultar a criação de modelos fiáveis, mas também poderá abrir portas para novos tratamentos na medicina para prever doenças e acontecimentos através de Notas Clínicas.

## 5.2. Análise do Estudo

O estudo realizado é composto por três partes, a análise de texto não estruturado, e recolha dessa informação utilizando ferramenta *KH Coder*, onde o objetivo foi a criação de padrões de doentes que tenham falecido após a realização do raio-x.

A criação de modelos de previsão utilizando o KNIME, tendo como objeto de estudo os relatórios de raio-x de doentes que faleceram e Sobrevieram. Este processo foi realizado tendo como base o *Cross Industry Standard Process for Data-Mining* (CRISP-DM).

Um último estudo foi realizado, utilizando o KNIME, com o objetivo de retirar as palavras mais utilizadas dos Relatórios de raio-x de doentes que faleceram e sobreviveram. Este estudo foi útil para ver quais são as palavras mais utilizadas em cada tipo de doente, e ajudou na criação de padrões dos doentes que realizaram o raio-x.

### 5.3. Estudo dos Dados

Neste ponto é descrito o estudo efetuado aos dados fornecidos pelo CHP com o objetivo de obter as informações sobre os dados para posteriormente aplicar as técnicas de TM aos mesmos.

#### *5.3.1. Descrição dos dados e extração dos dados Iniciais*

Um dos objetivos desta dissertação contemplou a análise de texto não estruturado na área da medicina. Para alcançar esse objetivo foi necessário o fornecimento de dados reais para realizar a mesma. Como a dissertação foi elaborada em colaboração com o CHP, a entidade disponibilizou os dados necessários.

Os dados fornecidos são registos de acontecimentos reais que aconteceram no CHP, nomeadamente os óbitos registados naquele local, e os relatórios de raio-x (dentro dos quais, o TC) feitos aos doentes. Apesar destes dados serem reais e conterem informação sobre os doentes, a confidencialidade destes últimos encontra-se salvaguardada.

Em termos da formação dos dados, estes dividem-se em três *Datasets*:

- *Dataset* com Relatórios de raio-x (Inicial) – **REPORT\_RX\_DATA\_TABLE** – Contém a informação dos raio-x realizados desde 27 de maio de 2008 até ao dia 20 de agosto de 2008, com um total de 867 registos.
- *Dataset* com a informação dos óbitos – **OBITOS\_DATA\_TABLE** – Contém o registo dos óbitos verificados no Hospital desde o dia 6 de dezembro de 2005 até ao dia 1 de março de 2016, com um total de 14528 registos.
- *Dataset* com Relatórios de raio-x (final) – **ReportRX\_NEW** – Contém a informação de todos os raio-x realizados no Hospital desde o dia 12 de outubro de 2009 até ao dia 17 de maio de 2016 contendo um total de 147973 registos.

Os *datasets* e os dados utilizados são descritos com mais detalhe no ponto abaixo. Após a recolha dos dados do Hospital, estes foram alojados numa base de dados *Oracle SQL*, e foram extraídos com o programa *Oracle SQL Developer* para, posteriormente, fazer a análise dos mesmos. Os ficheiros foram extraídos para uma folha de cálculo.

### 5.3.2. Exploração e Compreensão dos dados

Após a extração dos dados, foi necessário explorar e compreender os mesmos, como está descrito no CRISP-DM. Assim, realizou-se uma análise aos *dataset*, para identificar quais os campos mais relevantes e não relevantes para a análise, e ter uma ideia geral de como estão estruturados.

Os dados que compõe a tabela **REPORT\_RX\_DATA\_TABLE** são descritos detalhadamente na tabela abaixo:

*Tabela 6 – Composição e descrição dos dados do Dataset REPORT\_RX\_DATA\_TABLE*

Nome da Coluna	Descrição	Tipo de Dados	Valores Possíveis
<b>NUMEXAME</b>	Identifica o exame	Texto	Ex. RXC.511.2008.2639
<b>ESTADO</b>	Código do Estado do doente	Número	0 - 3
<b>SERVICO</b>	Código do serviço onde o doente vai ser submetido a exames e tratamento	Texto	Ex. RXU
<b>SERVICO_DESC</b>	Descrição do Serviço onde o doente vai ser submetido a exames e tratamento	Texto	Ex. RXU - Radiologia Urgência
<b>PROCESSO</b>	Código do Processo que identifica o Doente	Número	0 - 1201878
<b>TITULO</b>	Tipo de Exame a que o doente vai ser submetido	Texto	Ex. Tc Cranio Encefálica
<b>DATA</b>	Dia da realização do exame	Texto	5/26/2008 – 11/26/2008
<b>HORA</b>	Hora da Realização do Exame	Texto	1:00:08 PM – 9:59:11 AM

<b>IDENTIFICACAO</b>	Identificação do doente	Texto	Ex. Informação clínica: ;Despiste hidrocefalia tardia pós-TCE. Alt. comportamento com tendência a !maior lentificação psicomotora (+/- 6M pós-TCE).;
<b>DESCRICAO</b>	Descrição dos resultados do exame	Texto	<p>Ex. TÉCNICA:</p> <p>Foram efectuados cortes axiais de 2,5mm de espessura dirigidos à fossa posterior e de 5mm para o compartimento supratentorial. As imagens foram processadas em algoritmos adequados a tecidos moles e a detalhe ósseo.</p> <p>RELATÓRIO:</p> <p>O sistema ventricular apresenta morfologia e dimensões sobreponíveis às observadas no exame prévio (TAC de 07/11/2007), não existindo imagens de hidrocefalia.</p> <p>Reencontra-se a extensa hipodensidade cortico-subcortical temporal</p>

			<p>esquerda, estendendo-se ao polo temporal, com alargamento dos sulcos corticais regionais e dilatação ex-vácuo do corno temporal, correspondente a sequela de lesão encefaloclástica (provável contusão).</p> <p>Excluem-se colecções pericerebrais.</p> <p>As cisternas da base estão patentes.</p> <p>A charneira nervosa occipito-vertebral é normal.</p> <p>Referência para focos de espessamento da mucosa nos seios maxilares e câmara esquerda do seio esfenoidal, sugerindo processo inflamatório.</p>
<b>CONCLUSAO</b>	Código da Conclusão do Relatório	Texto	<p>Ex. Em Conclusão, as imagens são sobreponíveis ao exame anterior, nomeadamente quanto às dimensões do hematoma e seu efeito de massa e relativamente às dimensões do sistema ventricular.</p>

<b>CONCLUSAO_TEXTO</b>	Descrição da Conclusão	Texto	Ex. CONCLUSÃO: Resolução de hematoma subdural crónico.
------------------------	------------------------	-------	---

O *dataset* **OBITOS\_DATA\_TABLE**, que contém os dados sobre os doentes que faleceram. Após uma primeira análise ao *dataset*, verificou-se que este tinha bastantes dados que não eram relevantes para a análise e não foram considerados para este projeto como por exemplo as tabelas FSMS, FEMAIL, DSMS, DEMAIL, SMSID. Com a exclusão destes dados a tabela 7 descreve os dados que compõem o *dataset*.

Tabela 7 – Composição e descrição dos dados do Dataset **OBITOS\_DATA\_TABLE**

Nome da Coluna	Descrição	Tipo de Dados	Valores Possíveis
<b>URG_OBITO</b>	Identifica o óbito ocorrido	Número	10000001 - 16000407
<b>NUM_SEQUENCIAL</b>	Numero atribuído a cada doente (Automático)	Número	18 - 1630592
<b>NUM_PROCESSO</b>	Identifica o processo do doente que faleceu	Número	0 - 1676161
<b>DTA_OBITO</b>	Data do Óbito	Texto	06.12.200 – 01.03.2016
<b>HORA_OBITO</b>	Hora do Óbito	Número	0 - 86340
<b>FALECEU_HOSP</b>	Refere se o local da morte foi o hospital	Texto	Ex. N
<b>COD_ESPECIALIDADE</b>	Código da especialidade onde foi registado o óbito do doente	Número	0 - 49514
<b>HSA</b>	Se o óbito ocorreu no Hospital de Santo António	Número	0 - 1
<b>NUM_EPISODIO</b>	Número do Episódio que se relaciona com o Óbito	Número	0 - 16025161
<b>COD_MODULO</b>	Código do Modulo de Internamento onde estava o doente	Texto	Ex. INT

<b>DTA_REGISTO</b>	Data do registo do óbito	Texto	10.01.01 – 16.03.01
<b>DES_ESPECIALIDADE</b>	Descrição da especialidade onde o doente faleceu	Texto	Ex. CE PED-DOENTES VENTILADOS/HSA

O *dataset* **ReportRX\_NEW** contém informação dos raio-x, mas este *dataset* ao contrário do *dataset* **REPORT\_RX\_DATA\_TABLE**, contém todos os raio-x efetuados no Hospital.

Os dados do *dataset* **ReportRX\_NEW** são descritos detalhadamente na tabela 8.

*Tabela 8 – Composição e descrição dos dados do Dataset ReportRX\_NEW*

<b>Nome da Coluna</b>	<b>Descrição</b>	<b>Tipo de Dados</b>	<b>Valores Possíveis</b>
<b>NUMEXAME</b>	Identifica o exame	Texto	Ex. RXU.511.2009.7769
<b>ESTADO</b>	Código do Estado do doente	Número	0-3
<b>SERVICO</b>	Código do serviço onde o doente vai ser submetido a exames e tratamento	Texto	Ex. RXU
<b>SERVICO_DESC</b>	Descrição do Serviço onde o doente vai ser submetido a exames e tratamento	Texto	Ex. RXU - Radiologia Urgência

<b>PROCESSO</b>	Código do Processo que identifica o Doente	Número	0 - 1682080
<b>TITULO</b>	Tipo de Exame a que o doente vai ser submetido	Texto	Ex. Tc Cranioencefálica
<b>DATA</b>	Dia da realização do exame	Texto	12-10-2009 – 17-05-2016
<b>HORA</b>	Hora da Realização do Exame	Texto	00:00:00 – 23:59:59
<b>IDENTIFICACAO</b>	Identificação do doente	Texto	Ex. quot;Doente vitima de queda com perda de consciencia (segundo acompanhantes do local do trabalho por volta de 20 minutos) no local de trabalho- com nauseas e vomitos actualmente com cefaleias occipital- Agradezia avaliação de despiste de lesões traumaticas agudas._quot;
<b>DESCRICA0</b>	Descrição dos resultados do exame	Texto	Ex. Foram observados cortes axiais com 2.5 e 5mm de espessura, respectivamente centrados na fossa posterior e no compartimento supratentorial, sem contraste._#xD;_#xA;_#xD;_#xA;Não são visíveis anomalias



			densitométricas ou morfológicas encefálicas. Não há evidência de lesões extra-parenquimatosas. As vias de circulação de LCR conservam correctas dimensões e configuração. As amígdalas cerebelosas têm topografia normal. Não se observam traços de fractura.
CONCLUSAO	Código da Conclusão do Relatório	Texto	Ex. Não contém dados disponíveis.
CONCLUSAO_TEXTO	Descrição da Conclusão	Texto	Ex. Exame globalmente sobreponível face ao realizado dia 24.01.2010.

Pela análise realizada aos três *datasets* verificou-se que existem dois elementos que fazem a ligação entre ambos, que é a coluna **PROCESSO** (no *dataset* **REPORT\_RX\_DATA\_TABLE** e **ReportRX\_NEW**) e a coluna **NUM\_PROCESSO** (no *dataset* **OBITOS\_DATA\_TABLE**), o que poderá fazer a verificação se o doente que fez um exame de raio-x terá falecido.

Na Figura 6, encontra-se representado um Diagrama Entidade-Relação onde se pode visualizar a ligação e a estrutura dos dados:



Figura 6 – Diagrama Entidade-Relação dos Datasets fornecidos

### 5.3.3. Qualidade dos Dados

A nível da Qualidade, os dados apresentavam muitas incoerências e verificou-se imediatamente que estes não eram dados tratados, bem como, o texto que estava inserido nas suas tabelas se tratava de texto não estruturado. O facto de o texto não estar estruturado é um indicador positivo, visto que o objetivo era retirar informação de texto não estruturado, mas para fazer a análise do mesmo tem que se utilizar dados viáveis e que tenham os seus campos principais preenchidos. Com base na análise dos dados, foi criado um critério de exclusão dos dados que continham as seguintes características:

No *dataset* **REPORT\_RX\_DATA\_TABLE** foram definidos os seguintes critérios de exclusão:

- Linhas com valor nulo na coluna **PROCESSO**;
- Linhas com valor 0 na coluna **NUM\_PROCESSO**;
- Linhas com valor nulo na coluna **DESCRICA0**;
- Linhas com valor diferente de “TC Crânio Encefálica” na coluna **TITULO**.

No *dataset* **OBITOS\_DATA\_TABLE** foram definidos os seguintes critérios de exclusão:

- Linhas com valor nulo na coluna **NUM\_PROCESSO**;
- Linhas com o número 0 na coluna **NUM\_PROCESSO**.

No *dataset* **ReportRX\_NEW** foram definidos os seguintes critérios de exclusão:

- Linhas com valor nulo na coluna **PROCESSO**;
- Linhas com valor 0 na coluna **NUM\_PROCESSO**;
- Linhas com valor nulo na coluna **DESCRICA0**;
- Linhas com valor diferente de “TC Cranioencefálica” na coluna **TITULO**.

O critério foi elaborado com base na ligação entre os *datasets*, pois tinha que existir uma relação fiável para com os mesmos, e as colunas com o **PROCESSO** e o **NUM\_PROCESSO** com valor 0 ou nulo não permitiam fazer uma relação entre os dois *datasets*. Sobre a coluna **DESCRICA0**, como esta está identificada como a coluna que contém a informação do raio-x, isto é, a descrição do exame realizado e os seus resultados, se esta tivesse um valor nulo, a linha não teria qualquer tipo de utilidade para a análise dos dados pois não fornecia informação relevante.

Na coluna **TITULO**, que identifica qual é o tipo de raio-x feito aos doentes, o que interessa para o estudo, como já foi referido anteriormente, são as radiografias feitas ao cérebro (TC cranioencefálica), logo todos os outros raio-x efetuados não são para o interesse desta dissertação.

Em relação aos restantes dados, não é necessária mais nenhuma alteração pois estes não são importantes para a análise do texto, nem para verificar se o doente morreu ou não, e também para manter a integridade dos dados, pois é preferencial analisá-los no seu estado bruto.

#### 5.4. Preparação dos dados

Este ponto trata a seleção e o tratamento dos dados, de modo a que estes sejam viáveis para serem introduzidos nas ferramentas de TM. Após a análise aos dados e depois da definição de critérios da sua exclusão, estes são carregados em bruto para uma base de dados do *Microsoft SQL Server*, usando o *Microsoft SQL Server 2014 Management Studio*. O carregamento dos dados para a base de dados dividiu-se em três Tabelas (Tabelas 9, 10 e 11) correspondendo aos três *datasets* selecionados. Foram então criadas as seguintes tabelas:

- **Report** – a tabela 9 contém os dados do *dataset* **REPORT\_RX\_DATA\_TABLE** contendo no total 866 linhas.

Tabela 9 – Estrutura da tabela Report no Microsoft SQL Server

NUMEXAME	ESTA...	SERVICO	SERVICO_DESC	PROCESSO	TITULO	DATA	HORA	IDENTIFICACAO	DESCRICAO
1	RXC.511.2008.2580	3	RXC	RXC - Radiologia Central - Neuroradiologia	886358	Tc Cranio Encefálica	2008-06-01 1:28:25 PM	MARIA ALICE CUNHA RIBEIRO PROCESSO 886358 DN 2...	RELATÓRIO:
2	RXC.511.2008.2584	3	RXC	RXC - Radiologia Central - Neuroradiologia	517979	Tc Cranio Encefálica	2008-06-02 1:31:55 PM	Informação clínica: &apos;Sexo feminino 48 anos com sens...	RELATÓRIO:
3	RXC.511.2008.2588	3	RXC	RXC - Radiologia Central - Neuroradiologia	1197857	Tc Cranio Encefálica	2008-06-02 11:44:19 AM	NULL	NULL
4	RXC.511.2008.2589	3	RXC	RXC - Radiologia Central - Neuroradiologia	1190084	Tc Cranio Encefálica	2008-06-17 12:24:30 PM	TC CEREBRAL Informação clínica: TCE em 02/08, acident...	Relatório: Co
5	RXC.511.2008.2590	3	RXC	RXC - Radiologia Central - Neuroradiologia	737153	Tc Cranio Encefálica	2008-06-06 2:31:14 PM	Informação clínica: &apos;16 anos. TCE em Julho/07 (bicicl...	NULL
6	RXC.511.2008.2591	3	RXC	RXC - Radiologia Central - Neuroradiologia	1198007	Tc Cranio Encefálica	2008-06-02 11:29:48 AM	NULL	TAC CEREBR
7	RXC.511.2008.2592	3	RXC	RXC - Radiologia Central - Neuroradiologia	703548	Tc Cranio Encefálica	2008-06-02 1:21:14 PM	Informação clínica: &apos;Pós-op HSD crónico - F/P esquerdo 25...	RELATÓRIO:
8	RXC.511.2008.2593	3	RXC	RXC - Radiologia Central - Neuroradiologia	1198037	Tc Cranio Encefálica	2008-06-02 11:42:19 AM	NULL	TAC CEREBR
9	RXC.511.2008.2594	3	RXC	RXC - Radiologia Central - Neuroradiologia	683371	Tc Cranio Encefálica	2008-06-02 1:11:35 PM	Informação clínica: &apos;Cefaleias tipo enxaqueca, paradi...	RELATÓRIO:
10	RXC.511.2008.2599	3	RXC	RXC - Radiologia Central - Neuroradiologia	1198022	Tc Cranio Encefálica	2008-06-02 15:46:52	TAC CEREBRAL Data: 02-06-2008 Serviço: T. C. E. -INT...	RELATÓRIO:
11	RXC.511.2008.2600	3	RXC	RXC - Radiologia Central - Neuroradiologia	1198025	Tc Cranio Encefálica	2008-06-02 18:47:14	TAC CRANIO-ENCEFALICO Data: 02-06-2008 Serviço: TC...	NULL
12	RXC.511.2008.2601	3	RXC	RXC - Radiologia Central - Neuroradiologia	1175046	Tc Cranio Encefálica	2008-06-02 18:36:23	TAC CRANIO-ENCEFALICO Data: 02-06-2008 Serviço: M...	NULL
13	RXC.511.2008.2603	3	RXC	RXC - Radiologia Central - Neuroradiologia	1197857	Tc Cranio Encefálica	2008-06-04 3:46:49 PM	TAC CEREBRAL Data: 02-06-2008 Serviço: NEUROCIRU...	RELATÓRIO:
14	RXC.511.2008.2605	3	RXC	RXC - Radiologia Central - Neuroradiologia	1198101	Tc Cranio Encefálica	2008-06-03 1:28:00 AM	NULL	TC CRANIOE

- **Obitos** – a tabela 10 contém os dados do *dataset* **OBITOS\_DATA\_TABLE** contendo no total 14527 linhas.

Tabela 10 – Estrutura da tabela *Obitos* no Microsoft SQL Server

URQ_OBI...	NUM_SEQUENCI...	NUM_PROCES...	Datapt	HORA_OBI...	FALECEU_HO...	COD_ESPECIALIDADE	HSA	NUM_EPISOD...	COD_MODU...	FSMS	FEMAIL	DSMS	DEMAIL	DTA_REGIS...
1	15000336	343929	845963	2015-02-09	10680	S	11400	1	15015586	URG	1	1	2015-02-09	2015-02-10
2	15001048	1556130	1636472	2015-05-18	46200	S	31406	1	15013246	INT	1	1	2015-05-18	2015-05-19
3	15001175	115546	615918	2015-06-08	82800	S	31406	1	15016065	INT	1	1	2015-06-09	2015-06-10
4	15001197	481112	1088733	2015-06-13	7200	S	31100	1	15016359	INT	1	1	2015-06-13	2015-06-14
5	15001266	633765	1124818	2015-06-26	9000	S	31403	1	15017089	INT	1	1	2015-06-26	2015-06-27
6	15001377	1519472	1617529	2015-01-02	43177	N	0	0	0	OBI	9	9	1999-11-30	1999-11-30
7	15001378	1469932	1563222	2015-06-20	43281	N	0	0	0	OBI	9	9	1999-11-30	1999-11-30
8	15001600	1602968	1658142	2015-08-24	73200	S	39601	1	15022322	INT	1	1	2015-08-24	2015-08-25
9	15001676	70702	570967	2015-09-07	29100	S	0	1	15102665	URG	1	1	2015-09-07	2015-09-08
10	15001785	1230612	1394166	2015-09-23	58800	S	31406	1	15025182	INT	1	1	2015-09-23	2015-09-24
11	15001886	1569402	1638893	2015-10-11	58800	S	31405	1	15026179	INT	1	1	2015-10-11	2015-10-12
12	15002061	926773	1541796	2015-11-10	3600	S	11400	1	15129971	URG	1	1	2015-11-10	2015-11-11
13	16000126	1533069	1622415	2016-01-19	33600	S	31100	1	16000290	INT	1	1	2016-01-19	2016-01-20
14	16000186	132929	633350	2016-01-28	25800	S	31403	1	16002746	INT	1	1	2016-01-28	2016-01-29

- **ReportRXNEW** – a tabela 11 contém os dados do *dataset* **ReportRX\_NEW** contendo no total 149836 linhas.

Tabela 11 – Estrutura da tabela *ReportRXNEW* no Microsoft SQL Server

NUMEXAME	ESTA...	SERVICIO	SERVICIO_DESC	PROCESSO	TITULO	DATA	HORA	IDENTIFICACAO	DESCRICAO
1	RXU.511.2009.7654	3	RXU	RXU - Radiologia Urgência	600367	Tc Cranioencefálica	04-11-2009	14:32:47	..._quotDoente vítima de queda com perda de consciencia (segund...
2	RXU.511.2009.7637	3	RXU	RXU - Radiologia Urgência	1204096	Tc Cranioencefálica	03-11-2009	17:18:58	Antecedentes de macroadenoma de hipófise: operado e recidivad...
3	RXU.501.2009.1402	3	RXU	RXU - Radiologia Urgência	0	Ecografia renal e supra-renal	15-10-2009	19:26:36	1) 27 anos - Cólica renal esquerda não complicada desde há cer...
4	RXU.502.2009.6473	3	RXU	RXU - Radiologia Urgência	545593	Ecografia Abdominal Superior	24-10-2009	14:51:00	abd-pelvica - suspeita de colecistite aguda_#KD_#fA;
5	RXU.511.2009.7209	3	RXU	RXU - Radiologia Urgência	0	Tc Cranioencefálica	18-10-2009	17:47:37	52 anos- Hidrocefalo compensado conhecido em contexto de malf...
6	RXU.502.2009.6435	3	RXU	RXU - Radiologia Urgência	1195395	Ecografia renal e supra-renal	21-10-2009	19:48:58	1) Colica renal bilateral. Eco anterior com calculo coraliforme esq...
7	RXU.511.2009.7267	3	RXU	RXU - Radiologia Urgência	1372880	Tc do Tórax	21-10-2009	20:01:33	1) POLITRAUMATISMO COM TRAUMATISMO TORACICO DIREIT...
8	RXU.501.2009.1501	3	RXU	RXU - Radiologia Urgência	0	Ecografia renal e supra-renal	23-10-2009	10:31:30	ITU_#KD_#fA;
9	RXU.501.2009.1502	3	RXU	RXU - Radiologia Urgência	1121389	Ecografia Escrotal	23-10-2009	10:37:43	33anos- antecedentes de cx a varicocele (palomo) am set- vem pa...
10	RXU.502.2009.6606	3	RXU	RXU - Radiologia Urgência	1099226	Ecografia Abdominal Superior	28-10-2009	21:54:40	Dor hálto e epigastrelgio- vomitos bilares e hipertermia. BLG duod...
11	RXU.502.2009.6838	3	RXU	RXU - Radiologia Urgência	1106410	Ecografia Abdominal Superior	05-11-2009	11:30:24	QUEDA COM TRAUMATISMO TORAX ABDOMINAL- DESCARTAR...
12	RXU.502.2009.6461	3	RXU	RXU - Radiologia Urgência	982467	Ecografia Abdominal Superior	23-10-2009	15:29:56	dor na FID; doente com DÇA de Crohn;- Agudização_#KD_#fA;
13	RXU.511.2009.7656	3	RXU	RXU - Radiologia Urgência	0	Tc Cranioencefálica	04-11-2009	14:57:39	..._quotHomem de 62 anos. Instabilidade da marcha pref. para e es...
14	RXU.511.2009.7769	3	RXU	RXU - Radiologia Urgência	1042100	Tc Cranioencefálica	08-11-2009	0:04:44	..._quotMulher de 73 anos - Já há um ano teve episódio de vertige...

Após o carregamento dos dados, foi realizado o tratamento dos mesmos. Este foi possível com base nos critérios de exclusão definidos anteriormente. Para alcançar os dados desejados, foi criada uma *view*. O propósito de ser uma *view* consiste na manutenção da integridade dos dados, de forma a que estes sejam analisados no seu estado mais bruto pois a *view* não altera, apenas exclui os dados que não são relevantes para a análise. Posto isto, foram criadas três *views*:

A *view* **ReportsObitosInicial** – esta *view* foi realizada com base nas tabelas **Report** e **Obitos**, e o objetivo da mesma era conter os raio-x de pessoas que faleceram após o exame, para os dados serem analisados pela ferramenta **KH Coder**. A *view* **ReportsObitosInicial** foi realizada com uma *Query SQL*.

A *Query SQL* cumpriu os requisitos de exclusão definidos para o *dataset* que foi carregado para a Tabela **Report**, onde esta excluiu todas as Linhas com valores 0 na coluna **PROCESSO**, valores nulos na coluna **DESCRICAO**, valores diferentes a “TC Crânio Encefálica”, e para conhecer

se o doente faleceu, a *query* apenas seleciona os doentes cujo **PROCESSO** aparece na tabela **Obitos** (na tabela **Obitos** a coluna é **NUM\_PROCESSO**).

O resultado da *query* dá a seguinte *view* (tabela 12) com 48 registos totais:

Tabela 12 – Estrutura da *view ReportsObitosInicial*

NUMEXAME	ESTA...	SERVICO	SERVICO_DESC	PROCESSO	TITULO	DATA	HORA	IDENTIFICACAO	DESCRICA
1	RXC.511.2008.2730	3	RXC	RXC - Radiologia Central - Neuroradiologia	1165743	Tc Cranio Encefálica	2008-06-11	5:01:35 PM	.Transplantedo, como induzido. Alteração da coagulação. alter...
2	RXC.511.2008.2646	3	RXC	RXC - Radiologia Central - Neuroradiologia	702251	Tc Cranio Encefálica	2008-06-04	3:18:41 PM	Técnica: Realizaram-se cortes axiais de 2,5 e 5mm de espessu...
3	RXC.511.2008.2650	3	RXC	RXC - Radiologia Central - Neuroradiologia	638839	Tc Cranio Encefálica	2008-06-05	9:35:22 AM	Informação clínica: 67 anos, múltiplos FRCV. No dia 28/05/20...
4	RXC.511.2008.2661	3	RXC	RXC - Radiologia Central - Neuroradiologia	1196785	Tc Cranio Encefálica	2008-06-16	1:28:04 PM	TAC CEREBRAL. Data: 05.06.2008 Serviço: MEDICINA 1A e...
5	RXC.511.2008.2847	3	RXC	RXC - Radiologia Central - Neuroradiologia	706816	Tc Cranio Encefálica	2008-06-18	8:42:07 PM	NULL
6	RXC.511.2008.2767	3	RXC	RXC - Radiologia Central - Neuroradiologia	1198679	Tc Cranio Encefálica	2008-06-13	11:57:20 AM	.HipoNa+ sintomática. Dte previamente autónoma. Obj - despis...
7	RXC.511.2008.2939	3	RXC	RXC - Radiologia Central - Neuroradiologia	1152669	Tc Cranio Encefálica	2008-06-28	12:12:59 AM	Múltiplos factores de risco cardio-vasculares. Hoje fez cateteris...
8	RXC.511.2008.2969	3	RXC	RXC - Radiologia Central - Neuroradiologia	1199019	Tc Cranio Encefálica	2008-06-26	5:56:22 PM	TAC CEREBRAL. Data: 26-06-2008 Serviço: NEUROCIURGI...
9	RXC.511.2008.3054	3	RXC	RXC - Radiologia Central - Neuroradiologia	1158288	Tc Cranio Encefálica	2008-07-02	1:01:44 PM	ISAJURA ROCHA COUTINHO RODRIGUES FERREIRA TAC C...
10	RXC.511.2008.3029	3	RXC	RXC - Radiologia Central - Neuroradiologia	508623	Tc Cranio Encefálica	2008-06-30	1:44:42 PM	TC CEREBRAL. Informação clínica: Antecedentes de adenocar...
11	RXC.511.2008.3125	3	RXC	RXC - Radiologia Central - Neuroradiologia	848117	Tc Cranio Encefálica	2008-07-07	3:36:10 PM	Informação Clínica: .Síndrome... Alterações cognitivas. Alzhei...
12	RXC.511.2008.3155	3	RXC	RXC - Radiologia Central - Neuroradiologia	1195267	Tc Cranio Encefálica	2008-07-08	11:55:13 AM	TAC CEREBRAL. Serviço: Cons. Ext. de Neurologia Informa...
13	RXC.511.2008.3102	3	RXC	RXC - Radiologia Central - Neuroradiologia	1055831	Tc Cranio Encefálica	2008-07-03	1:13:18 PM	NULL
14	RXC.511.2008.3114	3	RXC	RXC - Radiologia Central - Neuroradiologia	1197816	Tc Cranio Encefálica	2008-07-08	1:05:16 PM	TAC CEREBRAL. 03/07/08 Serviço: Cons. Ext. de Neurologia ...

A *view ReportsObitos* – esta *view* é elaborada com base nas tabelas **ReportRXNEW** e **Obitos**, e o objetivo da mesma era conter os raio-x de pessoas que faleceram após o exame, para os dados serem analisados pela ferramenta KNIME. A *view ReportsObitos* foi feita com uma *Query SQL*.

A *Query SQL* cumpre os requisitos de exclusão definidos para o *dataset* que foi carregado para a Tabela **ReportRXNEW**, onde esta exclui todas as Linhas com a valores 0 na coluna **PROCESSO**, valores nulos na coluna **DESCRICA**, valores diferentes a “TC Crânio Encefálica”, e para saber se o doente faleceu, a *query* apenas seleciona os doentes cujo **PROCESSO** aparece na tabela **Obitos** (na tabela **Obitos** a coluna é **NUM\_PROCESSO**).

O resultado da *query* dá a seguinte *view* com 4899 registos totais:

Tabela 13 – Estrutura da *view ReportsObitos*

NUMEXAME	ESTADO	SERVICO	SERVICO_DESC	PROCESSO	TITULO	DATA	HORA	IDENTIFICACAO	DESCRICA	CONCLUSAO	CONCLUSAO_TEXTO	
1	RXU.511.2009.7769	3	RXU	RXU - Radiologia Urgência	1042100	Tc Cranioencefálica	08-11-2009	0:04:44	„quot;Mulher de 73 a...	Técnica: Realiza...	NULL	NULL
2	RXU.511.2009.7767	3	RXU	RXU - Radiologia Urgência	908384	Tc Cranioencefálica	08-11-2009	0:18:52	toe lmaior70 anos; de...	Técnica: Realiza...	NULL	NULL
3	RXU.511.2009.7804	3	RXU	RXU - Radiologia Urgência	1161766	Tc Cranioencefálica	09-11-2009	16:27:33	COM CONTRASTE- m...	Cortes axiais ent...	NULL	NULL
4	RXU.511.2009.7713	3	RXU	RXU - Radiologia Urgência	387552	Tc Cranioencefálica	08-11-2009	1:00:42	F- 71 a Hoje durante ...	TC CEREBRAL...	NULL	NULL
5	RXU.511.2009.7676	3	RXU	RXU - Radiologia Urgência	864484	Tc Cranioencefálica	05-11-2009	1:50:54	„quot;Doente com HL...	_#d_ #sA.TAC...	NULL	NULL
6	RXU.511.2009.7513	3	RXU	RXU - Radiologia Urgência	1083562	Tc Cranioencefálica	30-10-2009	11:48:31	queda TCE perdida c...	Técnica: Realiza...	NULL	NULL
7	RXU.511.2009.7098	3	RXU	RXU - Radiologia Urgência	605037	Tc Cranioencefálica	15-10-2009	1:07:32	Mulher de 79 anos an...	Duas áreas de ...	NULL	NULL
8	RXU.511.2009.7632	3	RXU	RXU - Radiologia Urgência	556170	Tc Cranioencefálica	03-11-2009	14:25:37	Hoje- instalação egud...	Cortes axiais de ...	NULL	NULL
9	RXU.511.2009.7174	3	RXU	RXU - Radiologia Urgência	711675	Tc Cranioencefálica	17-10-2009	12:45:35	„quot;TCE há 11 dias...	Forum observad...	NULL	NULL
10	RXU.511.2009.7093	3	RXU	RXU - Radiologia Urgência	610795	Tc Cranioencefálica	14-10-2009	20:32:49	Homem 74 anos- HTA...	Técnica: Cortes t...	NULL	NULL
11	RXU.511.2009.7477	3	RXU	RXU - Radiologia Urgência	1041073	Tc Cranioencefálica	29-10-2009	0:52:33	„quot;Homem de 80 a...	Forum observad...	NULL	NULL
12	RXU.511.2009.7132	3	RXU	RXU - Radiologia Urgência	583094	Tc Cranioencefálica	16-10-2009	10:40:56	Angio-TC torex. suspe...	Observe-se defe...	NULL	NULL
13	RXU.511.2009.7520	3	RXU	RXU - Radiologia Urgência	723293	Tc Cranioencefálica	30-10-2009	17:12:35	„quot;59 anos- valvul...	Técnica: Realiza...	NULL	NULL
14	RXU.511.2009.7696	3	RXU	RXU - Radiologia Urgência	502152	Tc Cranioencefálica	05-11-2009	17:41:39	M- 74a. HTA. Vertige...	Sem contraste_#...	NULL	NULL
15	RXU.511.2009.7188	3	RXU	RXU - Radiologia Urgência	913053	Tc Cranioencefálica	17-10-2009	23:24:39	Mulher de 92 anos- d...	M CÂNDIDA ARA...	NULL	NULL

A *view* **ReportsVivos** – esta *view* realizada com base nas tabelas **ReportRXNEW** e **Obitos**, e o objetivo da mesma era conter os raio-x de pessoas que não faleceram, para os dados serem analisados pela ferramenta KNIME. A *view* **Reportsvivos** foi realizada com uma *Query SQL*.

A *query SQL* cumpriu os requisitos de exclusão definidos para o *dataset* que foi carregado para a Tabela **ReportRXNEW**, onde esta excluiu todas as Linhas com valores 0 na coluna **PROCESSO**, valores nulos na coluna **DESCRICAO**, valores diferentes de “TC Cranioencefálica”, e conhecer se o doente não faleceu, a *query* compara os valores das colunas **PROCESSO(ReportRXNEW)** e **NUM\_PROCESSO(Obitos)** e seleciona apenas as linhas que não aparecem na tabela **Obitos**.

O resultado da *query* traduziu-se na seguinte *view* com 30904 registos totais:

Tabela 14 – Estrutura da *view* ReportsVivos

NUMEXAME	ESTA...	SERVICO	SERVICO_DESC	PROCESSO	TITULO	DATA	HORA	IDENTIFICACAO	DESCRICAO	
1	RXU.511.2009.7654	3	RXU	RXU - Radiologia Urgência	600367	Tc Cranioencefálica	04-11-2009	14:32:47	_quot;Doente vítima de queda com perda de consciencia (seg...	Foram observados cortes axiais
2	RXU.511.2009.7637	3	RXU	RXU - Radiologia Urgência	1204096	Tc Cranioencefálica	03-11-2009	17:18:58	Antecedentes de macroadenoma de hipófise- operado e recid...	Cortes axiais de 2.5 e 5mm de e
3	RXU.511.2009.7115	3	RXU	RXU - Radiologia Urgência	1043231	Tc Cranioencefálica	15-10-2009	19:45:50	Dor sugestiva de nevralgia do trigêmeo	Técnica: Realizaram-se cortes i
4	RXU.514.2009.142	3	RXU	RXU - Radiologia Urgência	753736	Tc Cranioencefálica	26-10-2009	13:52:57	NULL	_#xD_#xA;Obtiveram-se image
5	RXU.511.2009.7449	3	RXU	RXU - Radiologia Urgência	993353	Tc Cranioencefálica	27-10-2009	21:34:10	MARIA CARMO RAMOS MENDES_#xD_#xA;DN 01-11-1968_...	RELATÓRIO:_#xD_#xA;#xD_
6	RXU.511.2009.7076	3	RXU	RXU - Radiologia Urgência	581580	Tc Cranioencefálica	14-10-2009	11:23:08	TCE há 1 semana com HSD interhemisférico e sobre a tenda...	Técnica: Cortes tomográficos a
7	RXU.511.2009.7404	3	RXU	RXU - Radiologia Urgência	714455	Tc Cranioencefálica	26-10-2009	14:23:32	Doente de 76 anos encontrado inconsciente- com respiração ...	TÉCNICA_#xD_#xA;Obtiveram-
8	RXU.511.2009.7424	3	RXU	RXU - Radiologia Urgência	614359	Tc Cranioencefálica	27-10-2009	4:47:49	TCE_#xD_#xA;	TÉCNICA_#xD_#xA;Cortes axia
9	RXU.511.2009.7061	3	RXU	RXU - Radiologia Urgência	714877	Tc Cranioencefálica	13-10-2009	15:08:15	queda há um mês. bateu com cabeça - prostrado e com desiq...	Cortes axiais de 2.5 e 5mm de e
10	RXU.511.2009.7029	3	RXU	RXU - Radiologia Urgência	804860	Tc Cranioencefálica	12-10-2009	15:53:18	NULL	Cortes axiais antes e após conti
11	RXU.511.2009.7082	3	RXU	RXU - Radiologia Urgência	1055580	Tc Cranioencefálica	14-10-2009	13:41:59	TCE evoluindo com cefaléia e vômitos_#xD_#xA;	Técnica: Cortes tomográficos a
12	RXU.511.2009.7687	3	RXU	RXU - Radiologia Urgência	1031155	Tc Cranioencefálica	05-11-2009	13:15:22	TCE com hemorragia perimesencefálica- angioTAC normal- c...	Actualmente não se define sanç
13	RXU.511.2009.7296	3	RXU	RXU - Radiologia Urgência	597609	Tc Cranioencefálica	22-10-2009	11:04:18	_quot; TCE em hipocogulada - Controle de 24h_quot;	Técnica: Realizados cortes axia
14	RXU.511.2009.7293	3	RXU	RXU - Radiologia Urgência	1047568	Tc Cranioencefálica	22-10-2009	0:34:01	Mulher de 32 anos- com antecedentes de mielite transversa. ...	Não se detectam alterações rel

Query executed successfully. DESKTOP-ODHHME9\SQLEXPRESS ... DESKTOP-ODHHME9\Joao (53) master 00:00:33 30904 rows

Após a criação das *views*, os dados foram extraídos do *SQL Server 2014* para folhas de cálculo(.xls) para serem analisados posteriormente pelas ferramentas escolhidas, com o mesmo nome das *views*:

- **ReportsObitosInicial.xls** – resultante da extração da *view* **ReportsObitosInicial**;
- **ReportsObitos.xls** – resultante da extração da *view* **ReportsObitos**;
- **ReportsVivos.xls** – resultante da extração da *view* **ReportsVivo**.

## 5.5. Criação do dicionário

Para fazer uma análise aos dados de uma maneira mais objetiva, foi decidido criar um dicionário com termos médicos. Irão ser criados dois dicionários, de maneira a este poder ser lido pelas duas ferramentas de análise de texto. O dicionário para o *KH Coder* é constituído por temas, e pelas palavras que correspondem a esses temas. Para o KNIME o dicionário tem as mesmas

palavras que o dicionário utilizado para o *KH Coder*, mas não está agrupado por temas porque o KNIME não faz a identificação de temas e dos grupos de palavras, apenas lê cada palavra do Dicionário como um tema.

#### *5.5.1. Processo de Criação do dicionário*

O Processo de criação do dicionário realizou-se de maneira manual com base na análise à folha de cálculo **ReportsObitosInicial.xls** resultante da *view* criada no *SQL Server*. O ficheiro contém apenas raio-x de doentes que faleceram, o que poderá permitir a criação de um dicionário que contém termos integrados nos raio-x de doentes que faleceram. A maneira manual de criação do dicionário foi escolhida com base no facto do ficheiro conter apenas 48 linhas, e por isso permitir uma análise aprofundada ao ficheiro e a extração de termos variados que existem nos relatórios de raio-x. Apesar do baixo número de relatórios de raio-x, foram extraídos bastantes termos para os dicionários.

#### *5.5.2. Dicionários Criados*

Foram criados dois dicionários que irão ser descritos de seguida. Um dicionário orientado para aplicações na ferramenta *KH Coder*, e um dicionário orientado para aplicações na ferramenta KNIME.

##### *A) Dicionário para o KH Coder em Português e Inglês*

O *KH Coder*, no que toca aos dicionários tem a vantagem de agrupar bastantes palavras num tema específico, o que facilita as análises de texto. A Figura 7 mostra como o dicionário está estruturado para o *KH Coder*. Como se pode verificar na figura seguinte já referida, as palavras que significam o mesmo que o tema, são agrupadas numa palavra apenas, o que torna mais fácil a visualização das análises feitas ao texto.

```

*morfologia
morfologia | morfologicas | morfologias

*calcificações
calcificações

*Perdas
Perda

*Volume
Volume|

*moderada
medio | moderada | medios | moderadas

*Perinasais
Perinasais

*Permeaveis
Permeaveis | Permeavel

*parenquima
parenquima

*espaço
espaço | espaços

```

Figura 7 – Exemplo do Dicionário Estruturado para o KH Coder

### B) Dicionário para o KNIME em Português

Para o KNIME, o dicionário teve que ser estruturado de maneira diferente. Como o KNIME no “*Dictionary Tagger*” não faz reconhecimento de grupos de palavras como um só tema, por isso o dicionário não continha temas, mas sim termos, o que torna o dicionário com mais palavras para analisar, e que poderá dificultar a análise do texto em certos casos.

Na Figura 8 fica um extrato do dicionário estruturado para o KNIME:

```

morfologia
morfologicas
morfologias

calcificações

Perda

Volume

medio
moderada
medios
moderadas

Perinasais

Permeaveis
Permeavel

parenquima

espaço
espaços

```

Figura 8 – Exemplo do Dicionário Estruturado para o KNIME



## 6. Criação de Modelos de Análise de Texto utilizando o *KH Coder*

Esta componente do projeto é focada na análise do texto não estruturado com vista à identificação de padrões de doentes que faleceram após a realização do raio-x. Este estudo foi realizado com base na ferramenta *KH Coder*.

Apesar do *KH Coder* ser uma ferramenta boa para fazer análises quantitativas de Texto, esta não permite criar modelos de previsão, como por exemplo, prever se o doente que fez o raio-x faleceu. Essa função é atribuída ao KNIME como está descrito mais à frente neste documento.

Este capítulo começa com a seleção dos dados, onde são selecionados os dados que irão ser analisados, a seleção dos tipos de análises, onde são escolhidas as análises que foram efetuadas aos dados, e por fim, a demonstração das análises e a discussão dos resultados.

### 6.1. Seleção dos Dados

Os dados selecionados para análise no *KH Coder* eram provenientes do ficheiro **ReportsObitosInicial.xls**, pois, o objetivo da utilização do *KH Coder* era fazer uma análise quantitativa do texto de modo a descobrir se existem padrões nos raio-x dos doentes que morreram.

Para fazer uma análise quantitativa dos relatórios de raio-x, algumas colunas foram eliminadas no Excel. O Resultado final foi o uma folha de calculo com as colunas **NUMEXAME, PROCESSO, DATA, HORA, DESCRICAO**. As tabelas eliminadas foram excluídas pois durante as análises preliminares com o *KH Coder*, apareciam resultados que se misturavam com o texto na coluna DESCRICAO (que é a coluna que contém as informações do raio-x), e para não afetar a análise essas colunas foram eliminadas. Uma das características do *KH Coder* é que esta ferramenta só consegue ler os dados armazenados num documento de texto(.txt), logo a folha de cálculo foi convertida para um documento de texto, para o *KH Coder* o conseguir interpretar.

## 6.2. Selecionar as Técnicas de Análise

Depois de um estudo mais aprofundado da ferramenta foi decidido fazer diferentes tipos de análises ao documento, com o objetivo de retirar informações relevantes do mesmo, e também, posteriormente, fazer comparações com a utilização do dicionário, se este é importante para a recolha de informação e torna esta mais fácil de compreender, ou pelo contrário. Assim sendo, abaixo encontram-se as análises escolhidas assim como uma breve descrição das mesmas:

- Frequência de Palavras (com dicionário e sem dicionário) – Que consiste numa lista de palavras extraídas pelo *KH Coder* com a frequência de vezes que aparece no documento;
- Análise Hierárquica de *Clusters* (com dicionário e sem dicionário); – Esta análise permitiu procurar, e analisar quais combinações ou grupos de palavras têm padrões de aparência semelhante usando análise de agrupamento hierárquico;
- Mapa Auto Organizacional (com dicionário e sem dicionário) – Este comando explorou as associações entre as palavras, criando um mapa auto organizacional;
- Coocorrência de Rede (com dicionário e sem dicionário); – Esta análise criou um diagrama de rede que mostra as palavras com padrões de aparência similar, ou seja, com alto grau de coocorrência, ligada por linhas. Ao contrário da escala multidimensional, a coocorrência de rede pode ser mais fácil de analisar dado que as palavras estão conectadas com linhas;
- Análise de Correspondência (com dicionário e sem dicionário) – Este comando realizou uma análise de correspondência de palavras extraídas e produziu um diagrama de dispersão bidimensional (X e Y) para ajudar a visualizar os resultados;
- Escala Multidimensional (com dicionário e sem dicionário) – Este comando permite a realização da escala multidimensional sobre as palavras extraídas e desenhar os resultados num diagrama que pode ter até três dimensões (X, Y e Z).

As análises foram realizadas com a utilização de *stopwords* e sem utilização das mesmas. As *stopwords* são palavras que estão presentes nos textos, mas que não contém qualquer conteúdo e informação para a análise dos textos. Quando são aplicadas as *stopwords* nos documentos, as palavras que se encontram presentes na lista de *stopwords* são removidas do documento com o objetivo de facilitar e de agilizar o processo de análise dos textos. Abaixo está exemplificado um excerto das *stopwords* utilizadas:

Tabela 15 – Exemplo de stopwords utilizadas

deveria	disto
deveriam	dito
devia	diz
deviam	dizem
disse	do
disso	dos

A seguir é demonstrado um exemplo de frase onde se demonstra a utilização dos *stopwords*:

- Frase Original: O doente não medicado para nenhum destes medicamentos.
- Frase com *stopwords*: doente medicado medicamento.

### 6.3. Resultados da Análise

Os resultados das análises efetuadas estão divididos em dois grupos. As análises com a utilização do dicionário e as análises sem a utilização do dicionário. Esta divisão acontece, pois, os resultados das análises com a utilização do dicionário restringem-se aos termos que o dicionário contém, e as análises sem dicionário utilizam todos os termos existentes nos documentos. Dentro de cada grupo de análises, as mesmas estão separadas por tipo de análise efetuada bem como os requisitos utilizados (número de palavras), se for o caso.

As análises descritas nas secções seguintes são as análises que foram realizadas com a ajuda de *stopwords*, pois, como os resultados eram bastante semelhantes às análises efetuadas sem as *stopwords*, foi dispensada uma análise das mesmas, estando estas em anexo.

#### 6.3.1. Análise sem Dicionário

Nesta secção serão apresentados os resultados das análises efetuados sem a utilização do dicionário.

##### A) Frequência de Palavras

Esta análise permitiu visualizar uma lista de palavras extraídas pelo *KH Coder*. Esta análise criou uma folha de cálculo que lista as palavras classificadas em *Part-of-Speech* juntamente com

a sua frequência no documento (Higuchi, 2016). A análise da frequência de palavras mostra os termos utilizados no documento, o seu tipo, e o número de vezes em que estes aparecem no documento. No Resultado que está presente neste documento, é só uma amostra do resultado total, pois este tem mais de 200 linhas. Na tabela também está presente o Tipo do *Tagger Part-of-Speech* aplicado aos termos onde o N refere-se a Nome, o R a rejeição, o AQ a adjetivo, e o V a verbo.

Como se pode verificar, as palavras não e lesão encefálico aparecem com maior frequência do que as restantes palavras, com acima de 50 presenças dessas palavras no conjunto do documento. Após essas palavras, as restantes palavras que aparecem nesta análise estão todas num nível muito similar, entre as 20 e as 40 presenças no documento.

*Tabela 16 – Extrato da análise de Frequência de Palavras.*

Palavras	<i>Part-of-Speech</i>	Frequência
não	R	66
lesão	N	52
encefálico	AQ	50
alteração	N	36
espessura	N	30
circulação	N	28
isquémico	AQ	28
normal	AQ	28
observar	V	28
liquor	N	26
corte	N	24
enfarte	N	24
esquerdo	AQ	24
hipodensidade	N	24
alargamento	N	22
axial	AQ	22
cerebral	AQ	22
direita	N	22

parênquima	N	22
via	N	22

### B) Análise Hierárquica de Clusters

Este comando permitiu procurar e analisar quais combinações ou grupos de palavras têm padrões de aparência semelhante usando análise de agrupamento hierárquico. Os resultados da análise são apresentados num dendrograma. Este comando produz resultados que são mais fáceis de interpretar do que os resultados criados com o comando da escala multidimensional de palavras. Esta análise também usa a matriz gerada com as palavras que aparecem no documento e junta a mesma com as variáveis que indicam posições e comprimentos de documentos removidos (Higuchi, 2016). Para a utilização desta análise foram testados três cenários:

- 1) Análise com termos que aparecem pelo menos vinte vezes no documento;
- 2) Análise com termos que aparecem pelo menos quinze vezes no documento;
- 3) Análise com termos que aparecem pelo menos dez vezes no documento.

#### 1) Análise com termos que aparecem pelo menos vinte vezes no documento:

Esta análise contém os termos que aparecem com mais frequência no documento, o que permite ter uma visão geral sobre o mesmo. As diferentes cores da análise permitem distinguir os *clusters*. Por exemplo na figura 9, o *cluster* que contém “parênquima, alteração, densidade”, significa que a expressão “alteração da densidade do parênquima” é frequente neste documento. O *cluster* que contém as expressões “posterior, espessura, corte, axial” representa a expressão “posterior espessura do corte axial”. A importância dos *clusters* estarem hierarquicamente organizados é essencial pois podem-se criar subgrupos de palavras dentro do próprio *cluster*, o que facilita a análise do mesmo. Esta hierarquização é visível através das linhas existentes *intra* e *extra clusters*.

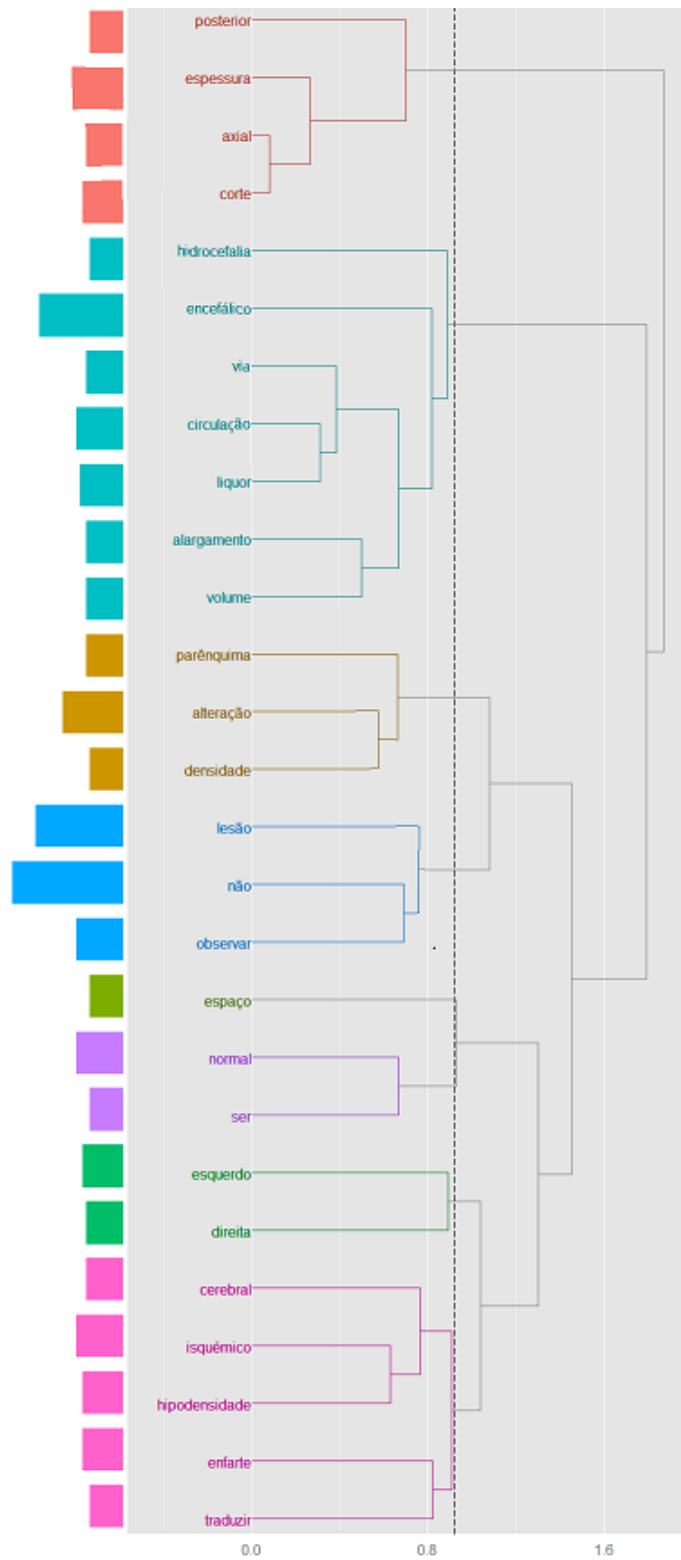


Figura 9 – Análise Hierárquica de Palavras com termos que aparecem pelo menos vinte vezes no documento

## 2) Análise com termos que aparecem pelo menos quinze vezes no documento:

A Análise com termos que aparecem pelo menos quinze vezes no documento situa-se numa posição mais intermediária em relação às análises de termos que aparecem dez e vinte vezes no documento. Com base análise da Figura 10, e em comparação com a Figura 9, pode-se concluir que houve algumas alterações na estrutura hierárquica dos *clusters*. Na Figura 9, o *cluster* que contém “posterior, espessura, axial, corte” sofreu alterações nesta análise. Esse *cluster* é agora composto por “encefálico, predomínio, alargamento, volume, hidrocefalia, via, circulação, líquor”. Esta constituição do *cluster* e a sua hierarquia, já permite fazer uma análise mais objetiva ao documento. Isto aconteceu na maioria dos *clusters*, ou seja, sofreram alterações na sua constituição que permitiram realizar uma análise mais específica ao documento. Mas por outro lado, o *cluster* “parênquima, alteração e densidade” não sofreu alterações.

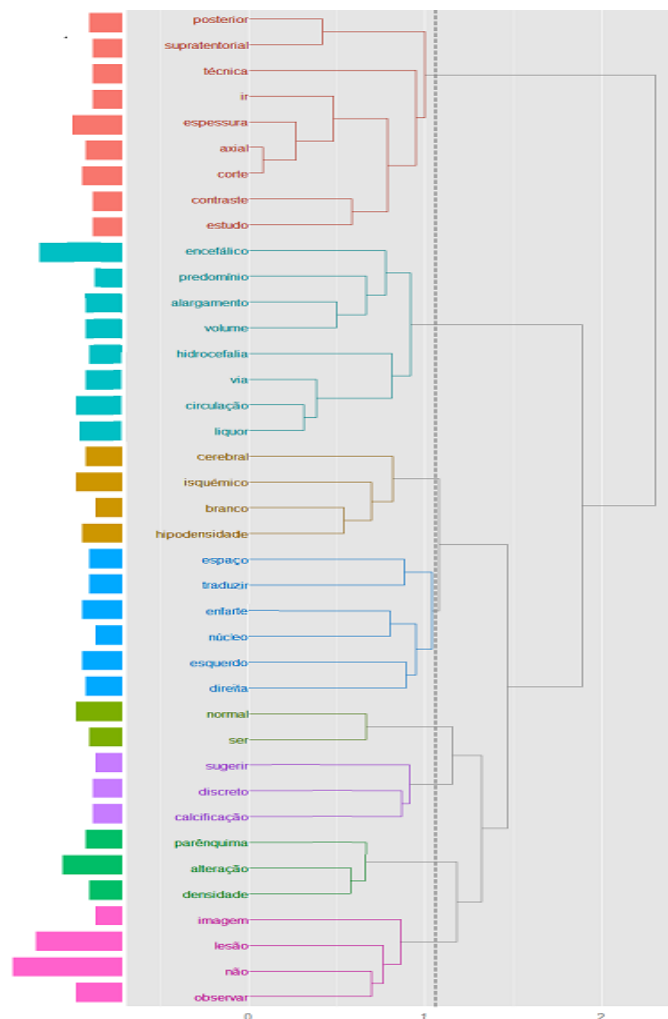


Figura 10 – Análise Hierárquica de Palavras com termos que aparecem pelo menos quinze vezes no documento

### 3) Análise com termos que aparecem pelo menos dez vezes no documento:

Nesta análise obteve-se resultados mais objetivos, pois engloba um maior número de termos que constituem o documento, o que gerou *clusters* de maiores dimensões comparados com as análises com termos que aparecem quinze ou vinte vezes no documento. Assim, consegue-se fazer análises mais específicas no documento. Com base nas análises anteriores, podemos verificar que o *cluster* “predomínio, temporal, atrofia, encefálico, via, circulação, líquido, alargamento, volume, global, hidrocefalia, redução”, contém informação mais detalhada sobre grupos de palavras que se correlacionam.

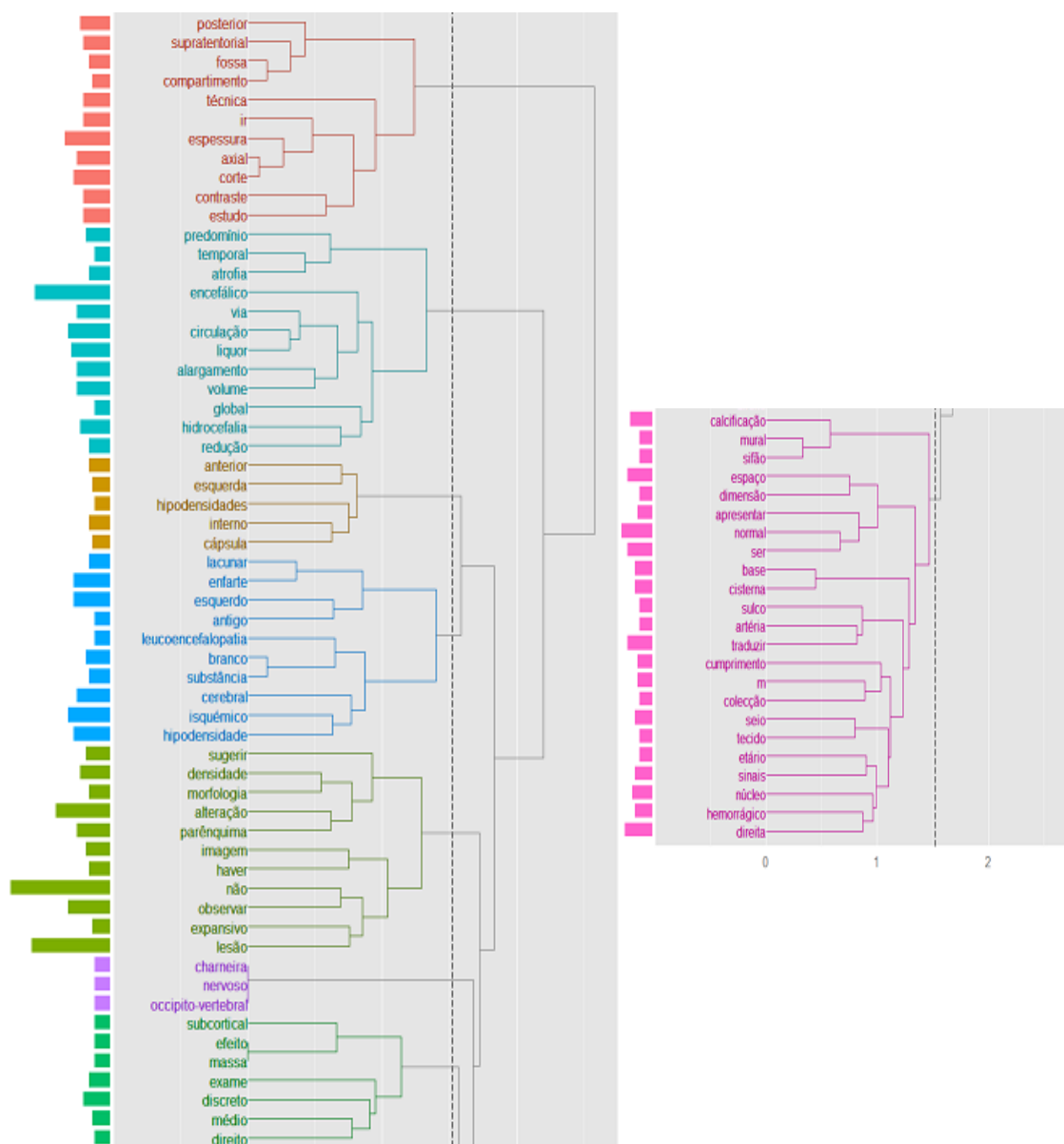


Figura 11 – Análise Hierárquica de Palavras com termos que aparecem pelo menos dez vezes no documento



### C) Mapa Auto Organizacional

Este comando explorou as associações entre as palavras, criando um mapa de auto-organização. Para criar um mapa de auto-organização no *KH Coder*, foi utilizada a matriz com as palavras utilizadas no documento juntamente com as variáveis para as posições e comprimentos de documentos removidos. No entanto, dado que a criação de um mapa de auto-organização usa distâncias euclidianas, a padronização é realizada em cada palavra, como descrito para o cálculo de distâncias euclidianas na escala multidimensional (Higuchi, 2016). Para a utilização desta análise foram testados três cenários:

- 1) Análise com termos que aparecem pelo menos vinte vezes no documento;
- 2) Análise com termos que aparecem pelo menos quinze vezes no documento;
- 3) Análise com termos que aparecem pelo menos dez vezes no documento.

#### 1) Análise com termos que aparecem pelo menos vinte vezes no documento:

Nesta análise consegue-se ter uma visão mais geral sobre o documento, onde se pode ver que os termos “isquêmico”, “hipodensidade”, “lesão” e “enfarte” constituem um *cluster*. Por outro lado, pode-se verificar que existem *clusters* que são formados apenas por um termo, como por exemplo o *cluster* que contém o termo “encefálico”. Isto acontece, pois, este mapa organizacional só contém termos que aparecem pelo menos vinte vezes no documento e agrupa os termos que se correlacionam mais vezes no mesmo *cluster*. As cores demonstram a densidade e tamanho de cada *cluster*.

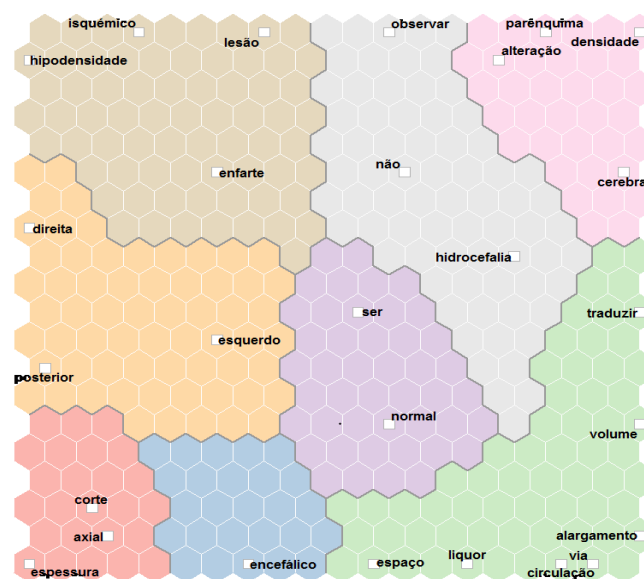


Figura 12 – Mapa Auto Organizacional com termos que aparecem pelo menos vinte vezes no documento

## 2) Análise com termos que aparecem pelo menos quinze vezes no documento:

Na análise com termos que aparecem pelo menos quinze vezes no documento pode-se verificar que os *clusters* já se encontram com outra composição, isto é, já contém mais termos dentro do mesmo, o que permite uma observação mais concreta sobre o documento, e alguns sofreram bastantes alterações. Por exemplo, na análise de *clusters* com termos que aparecem pelo menos quinze vezes no documento, verificava-se que existia um *cluster* contendo apenas um termo dentro deles, que era o *cluster* que continha o termo “encefálico”. Nesta análise esse *cluster* ficou agrupado em outro *cluster*, que contém os termos “encefálico”, “estudo”, “ir”, “corte”, “axial”, “espessura”, “contraste”.

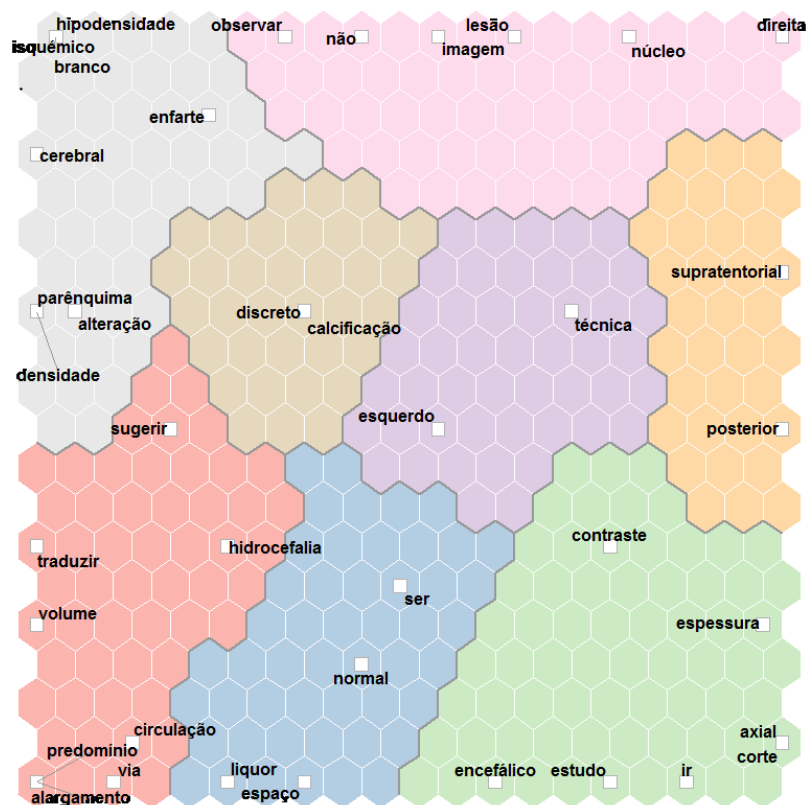


Figura 13 – Mapa Auto Organizacional com termos que aparecem pelo menos quinze vezes no documento

## 3) Análise com termos que aparecem pelo menos dez vezes no documento:

Nesta análise já se obtém uma visão mais objetiva em relação às duas análises anteriores, pois cada *cluster* já contém mais termos. Como já foi dito anteriormente esses termos dentro do *cluster* correlacionam-se entre si, como por exemplo, o *cluster* que contém os termos “calcificação”, “sifão”, “mural”, “artéria” e “esquerdo”, isto significa que na maioria das vezes

que um desses termos esteja num relatório médico, os restantes dois também estejam relacionados com esse termo.



Figura 14 – Mapa Auto Organizacional com termos que aparecem pelo menos dez vezes no documento

#### D) Coocorrência de Rede sem dicionário

Este comando criou um diagrama de rede que mostra as palavras com padrões de aparência similar, ou seja, com alto grau de coocorrência, ligada por linhas. Ao contrário da escala multidimensional, a coocorrência de rede pode ser mais fácil de analisar dado que as palavras estão conectadas com linhas. Este comando mostra a associação entre as palavras e variáveis/títulos, além das associações entre palavras. Ao contrário de resultados da escala multidimensional, as palavras próximas umas das outras nem sempre significa que eles têm uma coocorrência forte. Em vez disso, se as palavras estão ligadas com as linhas é a coocorrência é significativa (Higuchi, 2016). Para a utilização desta análise foram testados três cenários.

- 1) Análise com termos que aparecem pelo menos vinte vezes no documento;
- 2) Análise com termos que aparecem pelo menos quinze vezes no documento;
- 3) Análise com termos que aparecem pelo menos dez vezes no documento.

### 1) Análise com termos que aparecem pelo menos vinte vezes no documento:

Nesta análise, pode-se verificar que estes termos se relacionam entre si principalmente por causa dos termos “traduzir” e “encefálico” que tem a maioria das relações. Num efeito prático, significa que os termos que se relacionam com os termos “corte”, “volume” e “alargamento”, por exemplo, tem uma relação direta com o termo “encefálico”

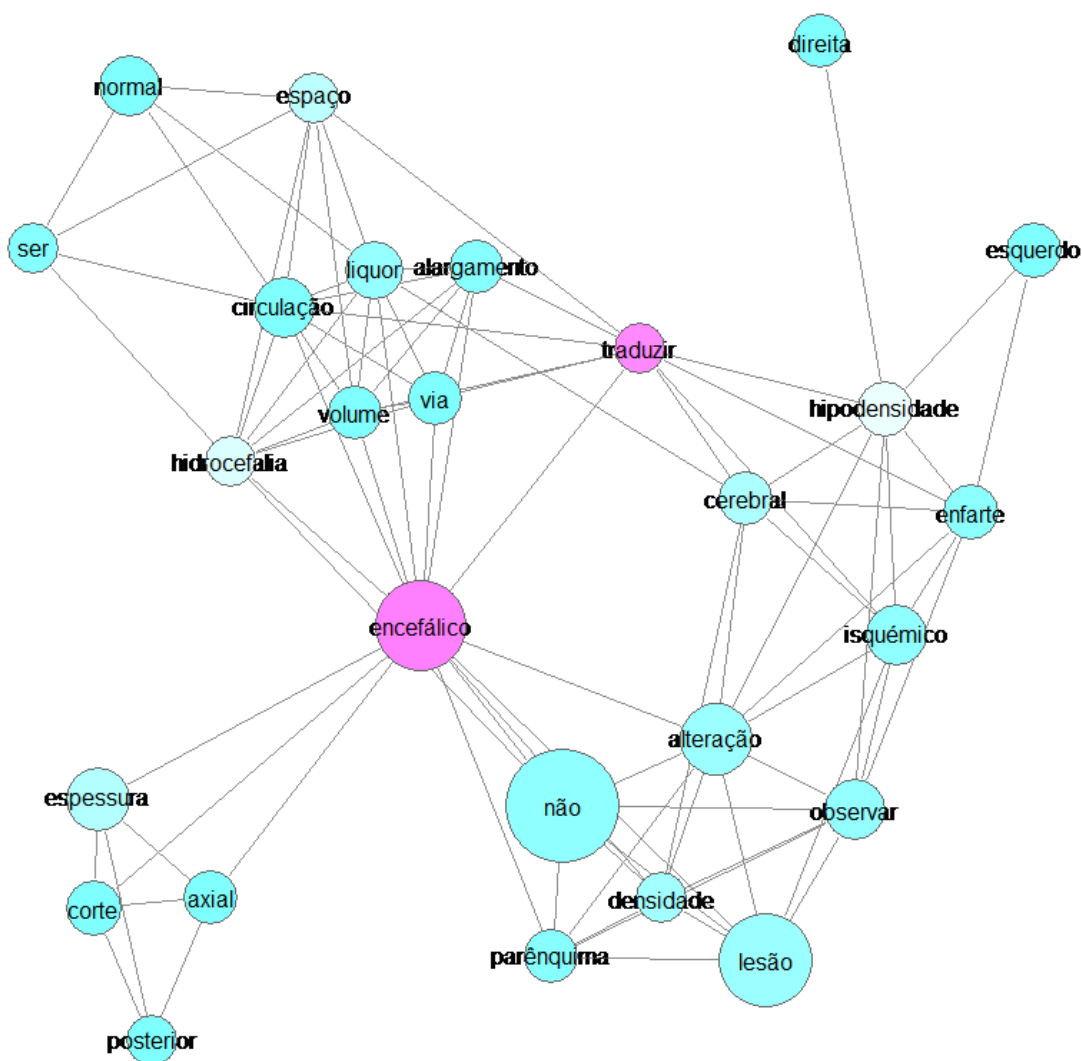


Figura 15 – Coocorrência de Rede de Palavras com termos que aparecem pelo menos vinte vezes no documento

## 2) Análise com termos que aparecem pelo menos quinze vezes no documento:

Nesta figura, como já existe um maior número de termos já se consegue fazer uma análise mais detalhada sobre a relação dos termos existentes no documento. Por exemplo, o termo “hipodensidade” relaciona-se com os termos “observar”, “isquêmico”, “cerebral”, “branco” e “enfarte”. Como já existe uma visão mais detalhada sobre os documentos, também se verifica que os termos já não se encontram todos relacionados entre si.

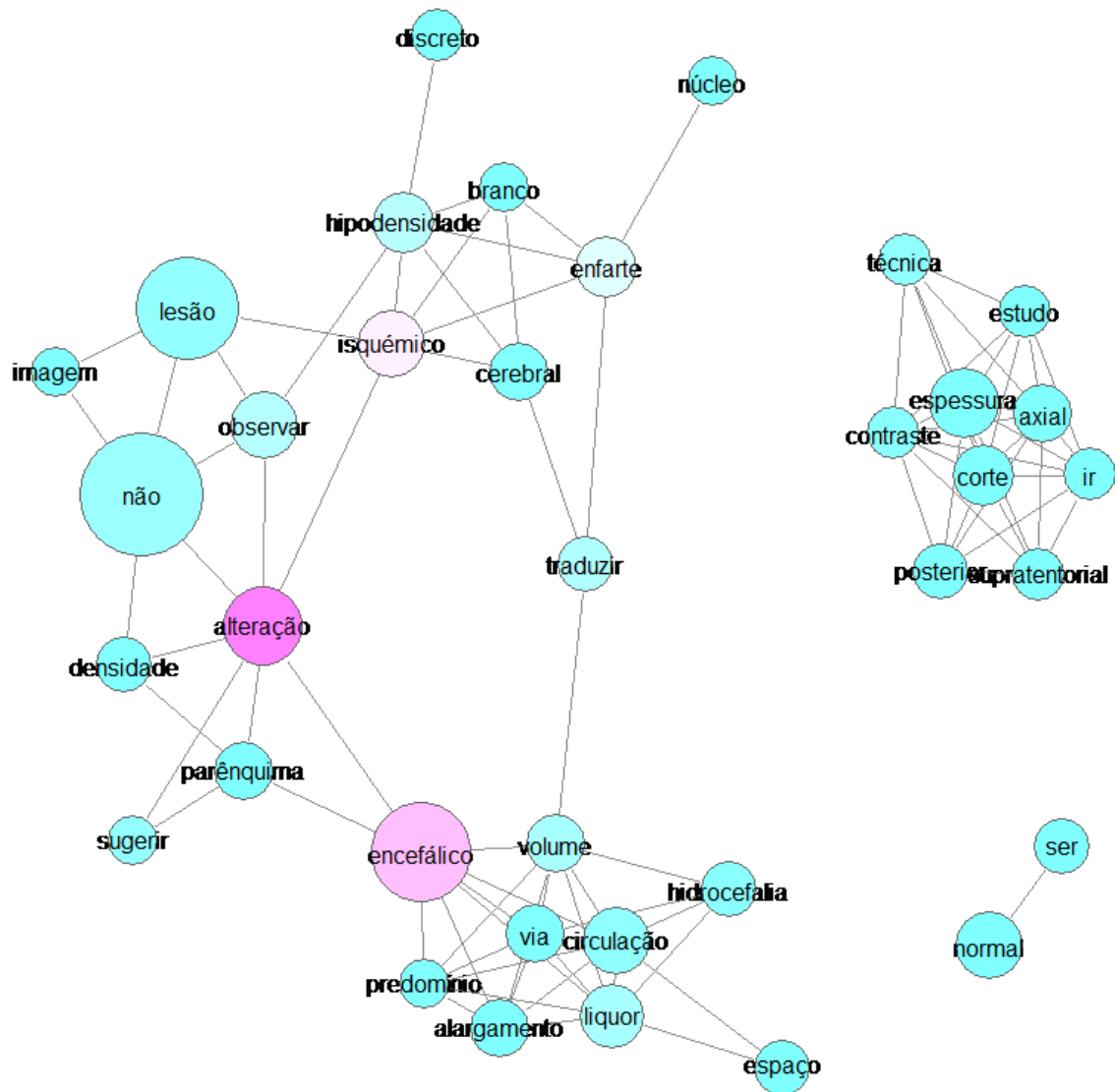


Figura 16 – Coocorrência de Rede de Palavras com termos que aparecem pelo menos vinte vezes no documento

### 3) Análise com termos que aparecem pelo menos dez vezes no documento:

Por fim, na análise com termos que aparecem pelo menos dez vezes no documento pode-se verificar que o nível de detalhe da análise é ainda maior e que existem cada vez mais conjuntos de termos sem relação a outros conjuntos de termos. Por exemplo, pode-se verificar que o termo “densidade” se relaciona com os termos “parênquima”, “alteração”, e “morfologia”, o que no documento pode significar a seguinte frase “alteração da densidade parênquima e morfologia”.

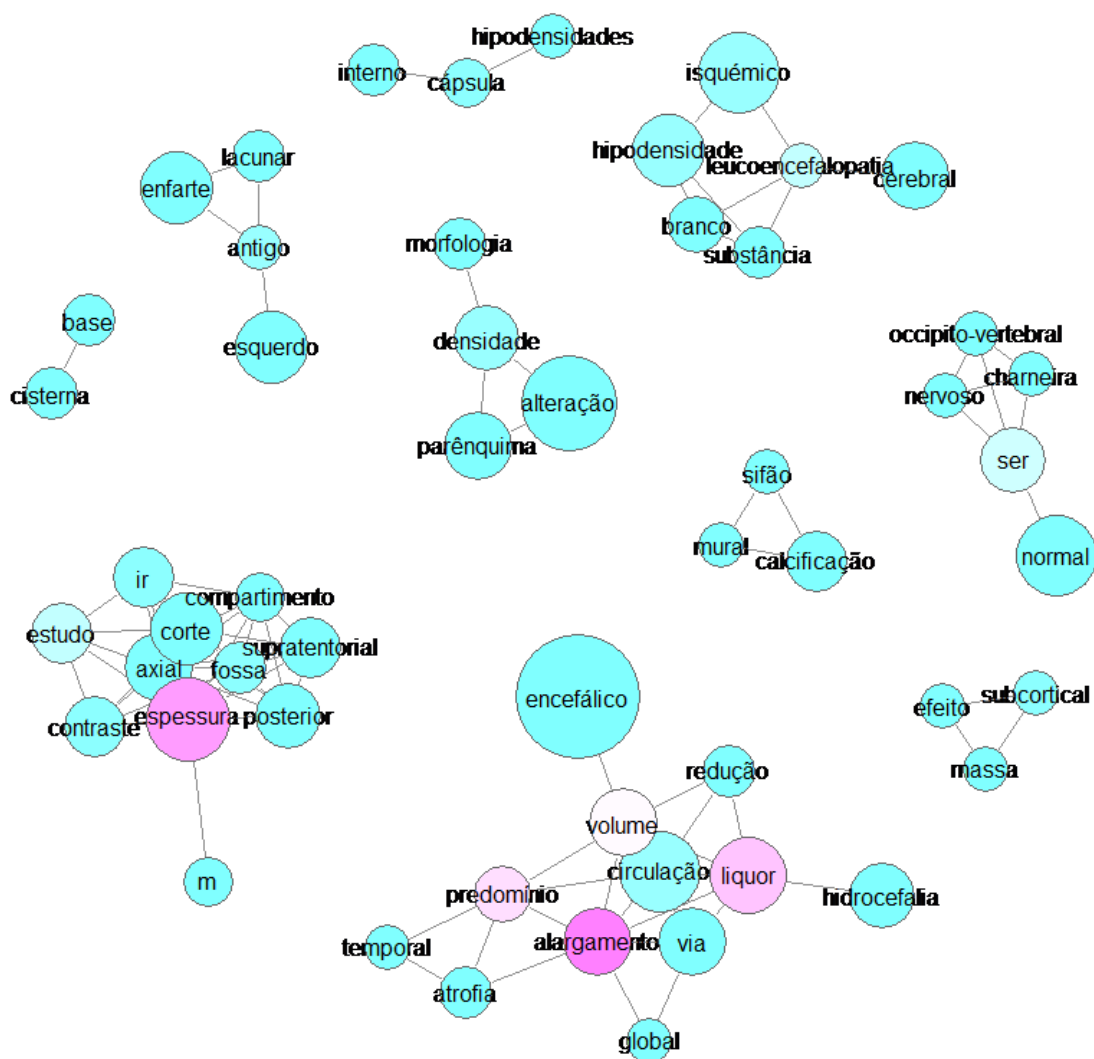


Figura 17 – Coocorrência de Rede de Palavras com termos que aparecem pelo menos dez vezes no documento

### E) Análise de Correspondência

Este comando realizou uma análise de correspondência de palavras extraídas e produziu um diagrama de dispersão bidimensional (X e Y) para ajudar a visualizar os resultados. As dimensões podem tomar valores negativos pois o valor 0 é o valor mais comum no documento, isto é, a parte central da estrutura do documento, quando os valores são negativos significa que as palavras aparecem antes do meio do relatório, e quando aparecem após o zero quer dizer que estas aparecem na parte final do documento. Esta análise é usada para explorar o que tipos de palavras tem um padrão de aparência similar (Higuchi, 2016). Para a utilização desta análise foram testados três cenários.

- 1) Análise com termos que aparecem pelo menos vinte vezes no documento;
- 2) Análise com termos que aparecem pelo menos quinze vezes no documento;
- 3) Análise com termos que aparecem pelo menos dez vezes no documento.

#### 1) Análise com termos que aparecem pelo menos vinte vezes no documento:

Na análise de correspondência com termos que aparecem pelo menos vinte vezes no documento pode-se verificar que muitas palavras têm o mesmo padrão de aparência (com cerca de 88% dos termos a situarem-se no valor -0,5 da *Dimension 1*), onde apenas os termos “posterior”, “espessura”, “corte”, “axial”, “pm” e “relatório” padrões bastante diferentes dos restantes, estando em posições diferentes na figura 18.

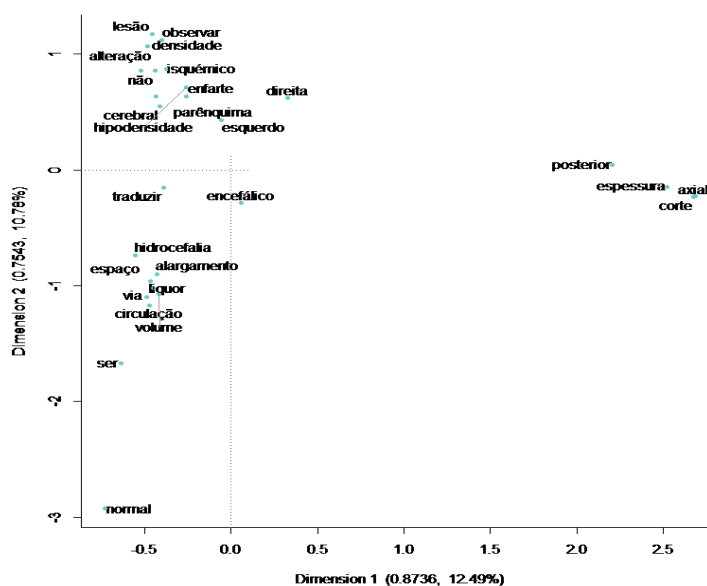


Figura 18 – Análise de Correspondência com Termos que aparecem pelo menos vinte vezes no documento

## 2) Análise com termos que aparecem pelo menos quinze vezes no documento:

Aqui, em semelhança à análise anterior, os termos situam-se nas mesmas posições, com a diferença de existirem mais termos nas posições, visto que esta análise engloba mais termos do que a análise anterior.

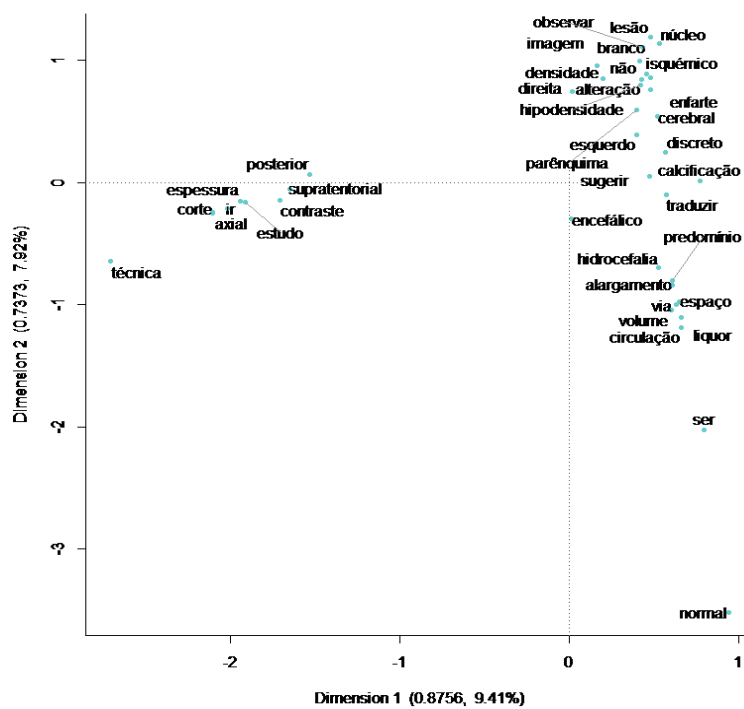


Figura 19 – Análise de Correspondência com Termos que aparecem pelo menos quinze vezes no documento



### 3) Análise com termos que aparecem pelo menos dez vezes no documento:

Na análise com termos que aparecem pelo menos dez vezes no documento, em semelhança as outras duas análises, os termos estão praticamente na mesma posição na *Dimension 1*, rondando o valor 0, mas na *Dimension 2*, 95% termos ocupam valores entre o valor -1 e o valor 1.

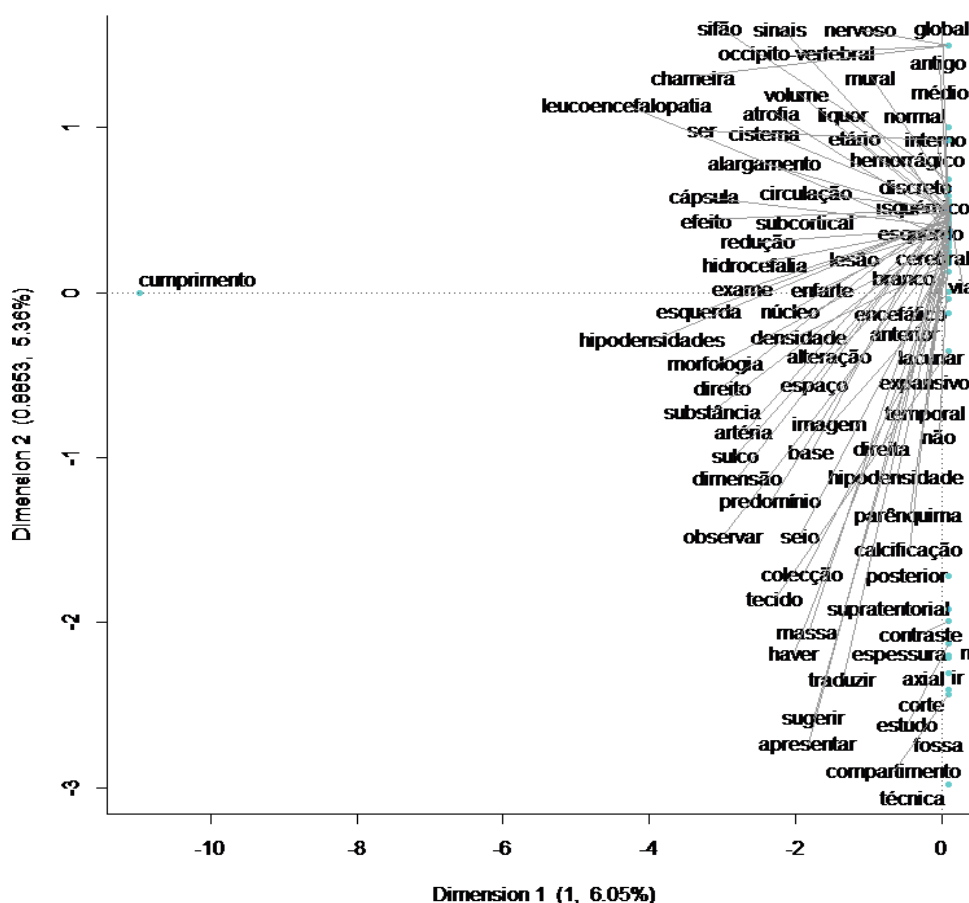


Figura 20 – Análise de Correspondência com Termos que aparecem pelo menos dez vezes no documento

### F) Escala Multidimensional de Termos

Este comando permite a realização da escala multidimensional sobre as palavras extraídas e desenhar os resultados num diagrama que pode ter até três dimensões (X, Y e Z). Esta análise pode usar esta função para encontrar combinações ou grupos de palavras que têm padrões de aparência semelhantes. Esta análise usa uma matriz gerada com as palavras que aparecem no documento e junta a mesma com as variáveis que indicam posições e comprimentos de documentos removidos. As dimensões podem tomar valores negativos pois o valor zero é o valor mais comum no documento, isto é, a parte central da estrutura do documento, quando os valores são negativos significa que os as palavras aparecem antes do meio do relatório, e quando

aparecem após o zero quer dizer que estas aparecem na parte final do documento. No entanto, se existirem pares de palavras com uma distância igual a zero, um destes pares é automaticamente retirados da análise com uma mensagem a informar esse acontecimento (Higuchi, 2016). Para a utilização desta análise foram testados três cenários:

- 1) Análise com termos que aparecem pelo menos vinte vezes no documento;
- 2) Análise com termos que aparecem pelo menos quinze vezes no documento;
- 3) Análise com termos que aparecem pelo menos dez vezes no documento.

### 1) Análise com termos que aparecem pelo menos vinte vezes no documento:

Esta análise é relativamente mais simples de fazer o estudo dos resultados pois estes apresentam uma maior facilidade de visualização dos mesmos. Os termos com maior relação entre si têm a mesma cor. A relação entre os termos verifica-se também pela posição dos mesmos consoante as coordenadas da *Dimension 1* e *Dimension 2*. Por exemplo os termos “alargamento”, “volume”, “liquor”, “via” e “circulação”, são termos que tem a mesma cor logo, pertencem ao mesmo grupo relacional.

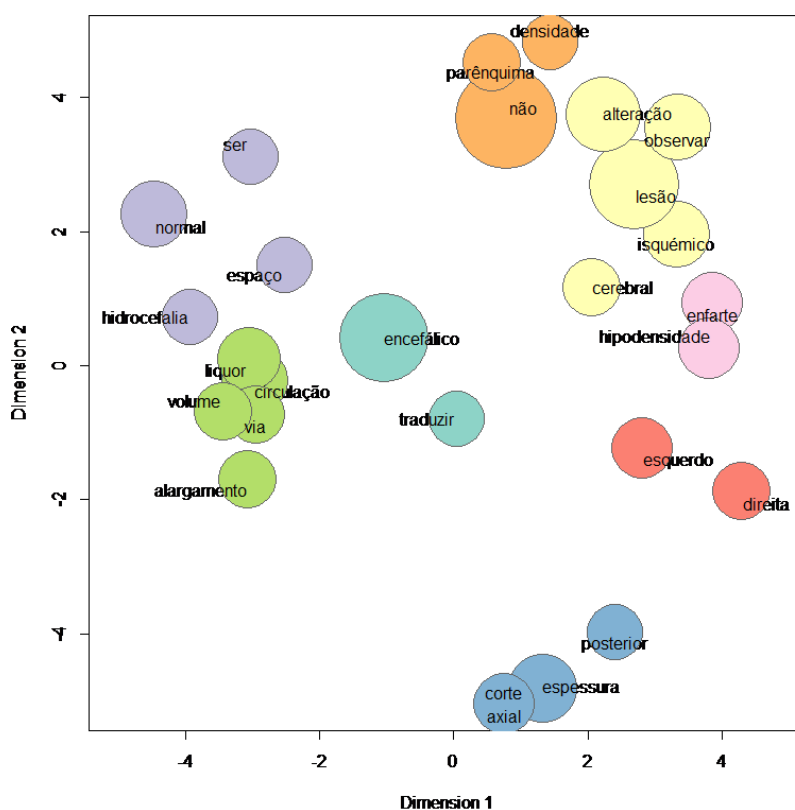


Figura 21 – Escala multidimensional de termos que aparecem pelo menos vinte vezes no documento

## 2) Análise com termos que aparecem pelo menos quinze vezes no documento:

Nesta análise, já se apresentam mais termos na análise porque existe um maior número de termos que aparecem pelo menos quinze vezes no documento. Em semelhança à análise anterior, os termos assinalados com a mesma cor fazem parte do mesmo grupo de termos que se relacionam mais vezes entre si. Por exemplo, os termos “posterior”, “supratentorial”, “axial”, “contraste”, “espessura”, “corte”, “ir”, e “estudo”, e “técnica” pertencem ao mesmo grupo de palavras.

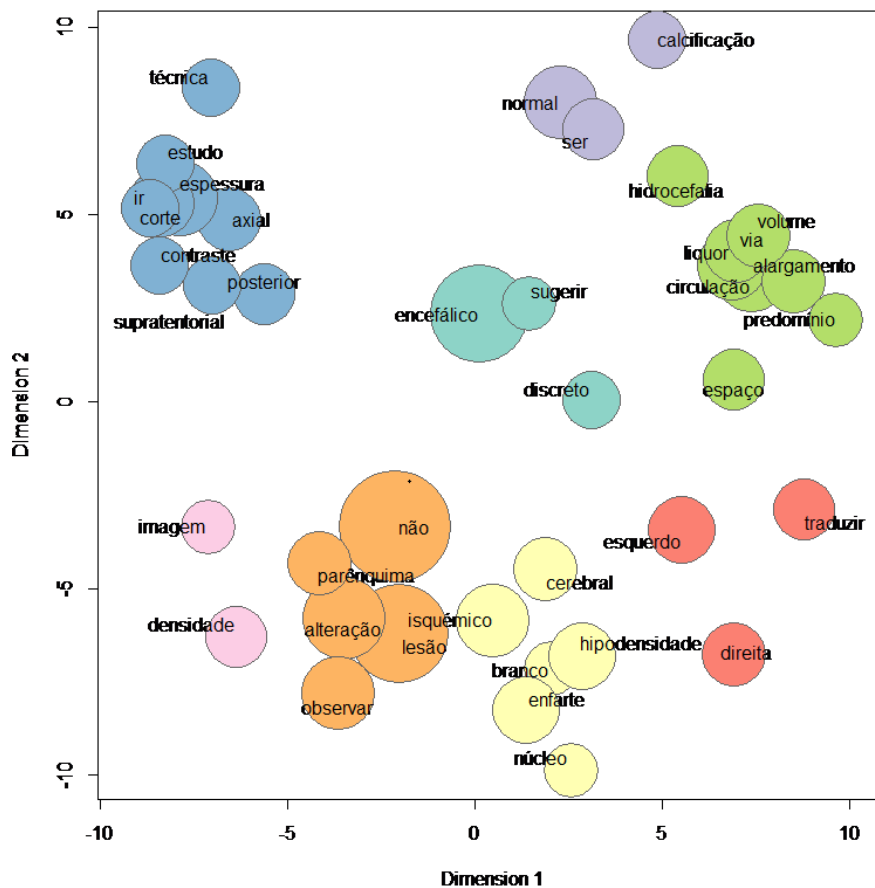


Figura 22 – Escala multidimensional de termos que aparecem pelo menos quinze vezes no documento

### 3) Análise com termos que aparecem pelo menos dez vezes no documento:

Por fim, nesta análise, consegue-se ver com maior exatidão os grupos de palavras existentes no documento, pois engloba termos que aparecem pelo menos dez vezes no documento. Os termos “cumprimento”, “coleção”, “imagem”, “observar”, “expansivo”, “exame”, “médio” e “direito” pertencem ao mesmo grupo de palavras.

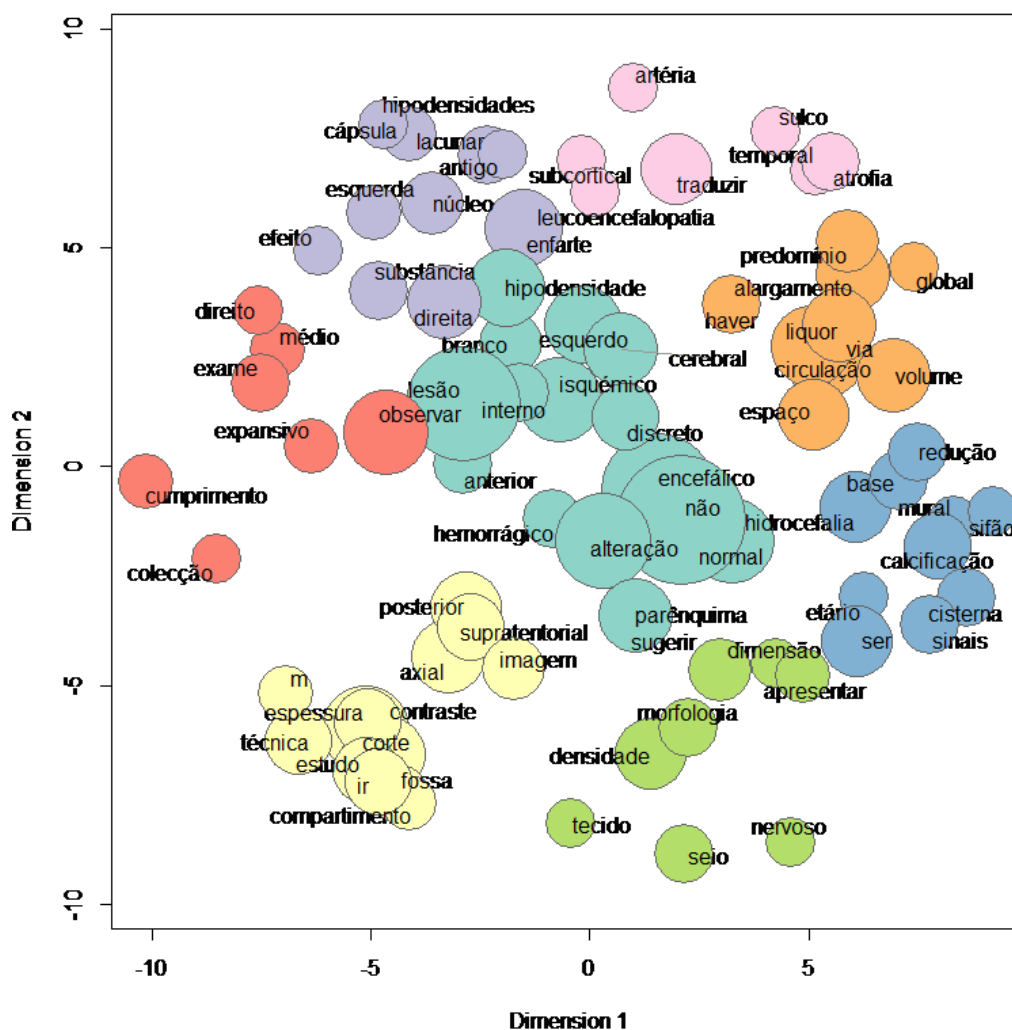


Figura 23 – Escala multidimensional de termos que aparecem pelo menos dez vezes no documento

### 6.3.2. Análise com Dicionário

De seguida serão apresentadas as análises efetuadas com a utilização do dicionário.

#### A) Frequência de Palavras com a Utilização do Dicionário

Neste comando foi executado a codificação de acordo com o conteúdo do arquivo de regras de codificação, e fornece a seguinte tabulação: o número de documentos de cada tema e aplica-se a sua percentagem do total, e o número de documentos em que nenhum tema é aplicado e a sua percentagem total (Higuchi, 2016). Os resultados apresentados representam os termos do dicionário que aparecem mais vezes no documento. Por exemplo, as palavras que englobam o tema “Negativo” são as que aparecem mais vezes no documento, com 66 aparições, com 52 e 50 presenças no documento estão os temas “Encefálico” e “lesão”.

Tabela 17 – Extrato do resultado de Frequência de Palavras do Dicionário

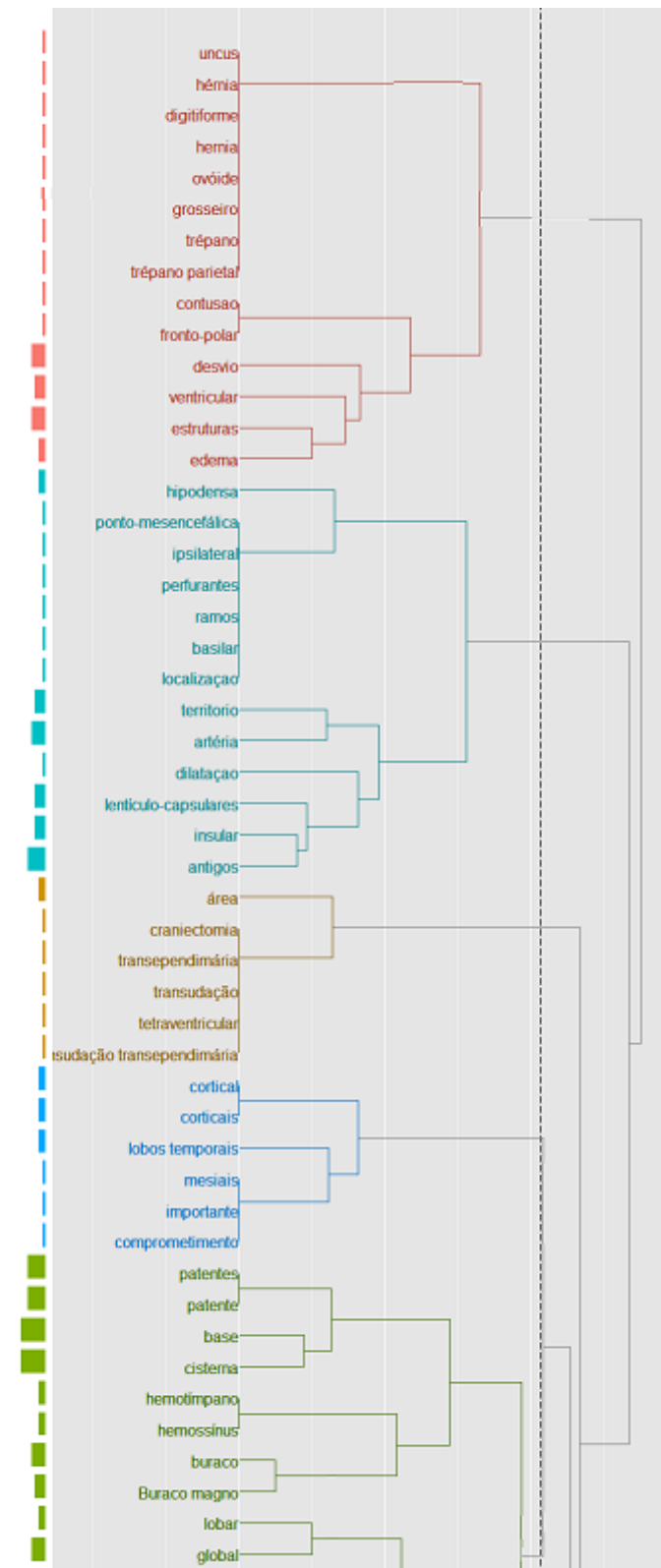
Termo	Frequência	Percentagem
*Negativo	66	15.24%
*Encefálico	52	12.01%
*lesao	50	11.55%
*Alterações	36	8.31%
*Sem Alterações	36	8.31%
*LCR	34	7.85%
*Hiposensibilidade	32	7.39%
*espessura	30	6.93%
*Normal	30	6.93%
*esquerda	28	6.47%
*Aumento	24	5.54%
*direita	22	5.08%
*Enfarte	22	5.08%
*Volume	22	5.08%
*parenquima	22	5.08%
*cerebral	22	5.08%
*Hidrocefalia	20	4.62%

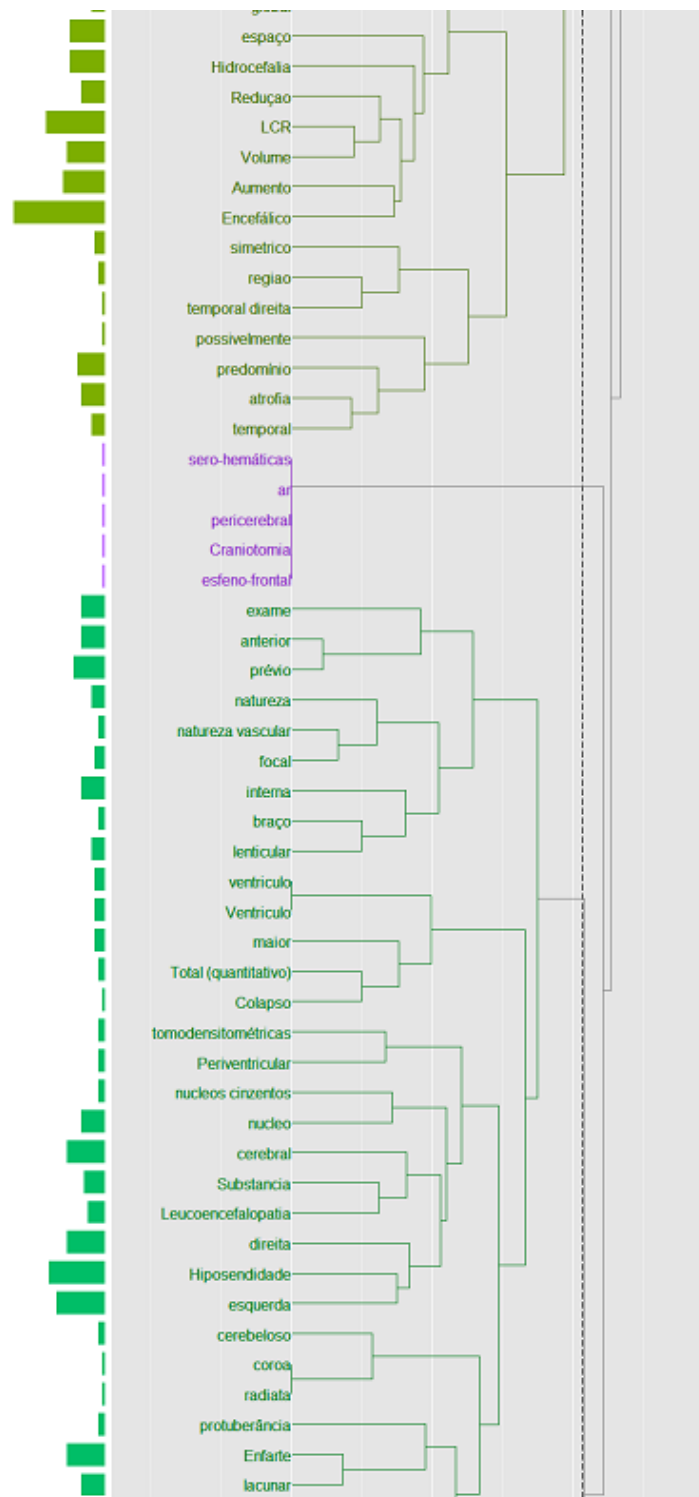
*espaço	20	4.62%
*prévio	18	4.16%
*subcorticais	18	4.16%

### B) *Análise Hierárquica de Clusters*

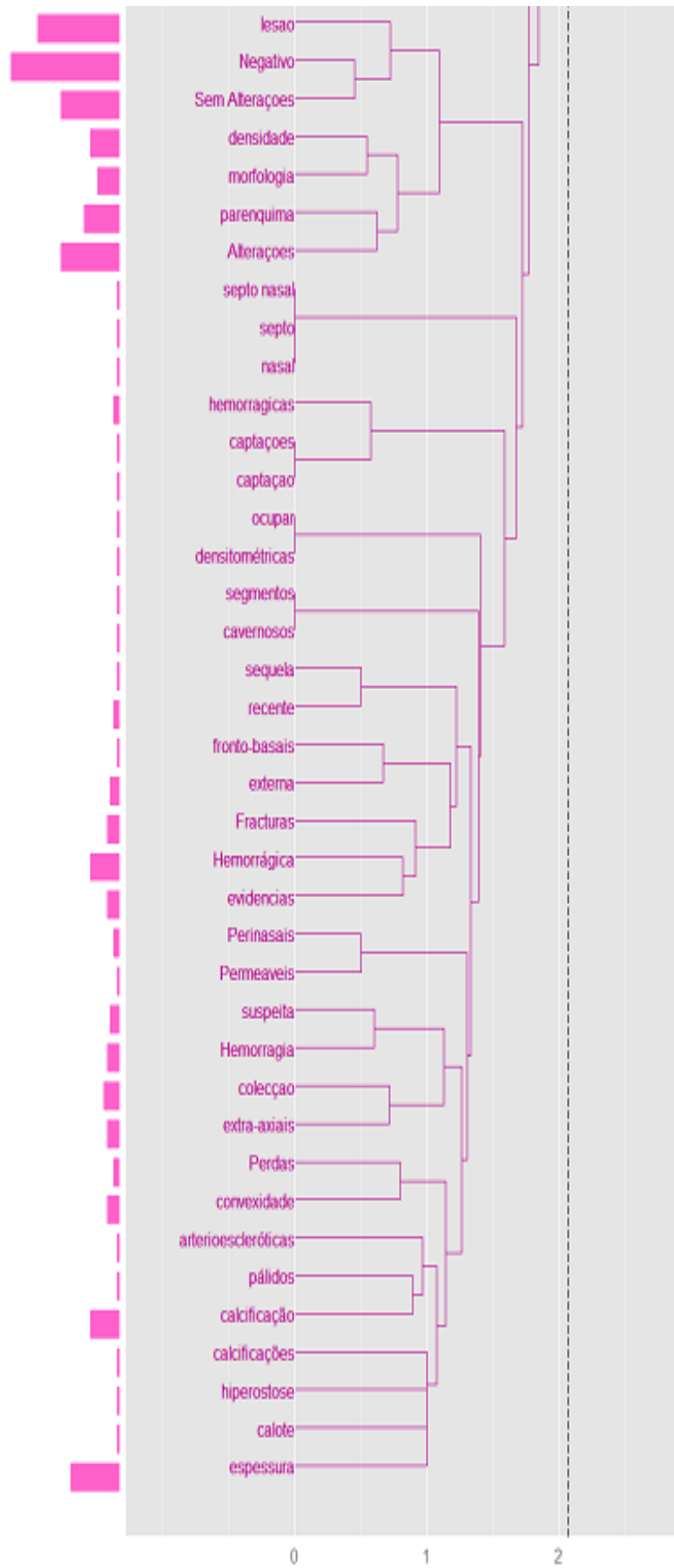
Este comando criou um dendrograma para explorar associações de código. Em alguns casos, um dendrograma pode ser mais fácil de interpretar do que outros tipos de análise. A operação e opções para este comando são semelhantes aos do Análise Hierárquica de palavras. A única diferença é que este utiliza temas em vez de palavras para a realização da análise (Higuchi, 2016).

Os resultados apresentados representam todos os termos existentes no dicionário que foram identificados no documento e foram agrupados em *clusters* com valores semelhantes, isto é, palavras que normalmente aparecem numa mesma frase.









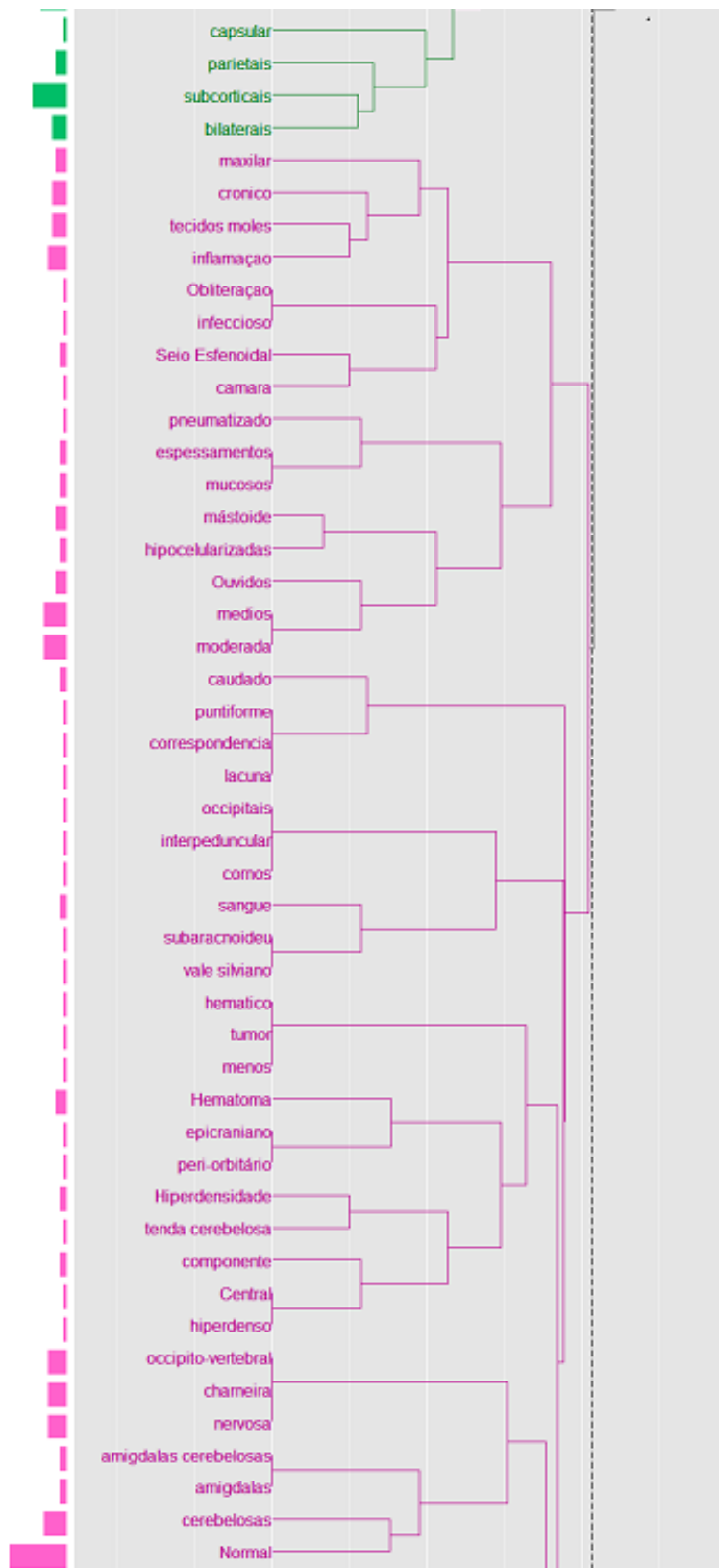


Figura 24 – Análise Hierárquica de Clusters dos Temas do Dicionário

### C) Mapa Auto Organizacional com Dicionário

Este comando possibilitou a criação de um mapa auto organizacional para explorar associações de código. A operação e opções para este comando são semelhantes àquelas para o Mapa Auto Organizacional de Palavras. A única diferença é que este utiliza temas em vez de palavras para a realização da análise (Higuchi, 2016).

No Mapa Auto Organizacional, ao contrário da análise sem Dicionário, não dá para selecionar um número mínimo de vezes em que os termos aparecem no documento. Isto dá a possibilidade de ver mais em detalhes os oito tipos de doentes que o *KH Coder* definiu. O tamanho das cores significa o tamanho do *cluster*, que neste caso é quantos termos estão presentes no *cluster*. Num dos *clusters* encontra-se definido pelas palavras camara, maxilar, espessamentos, mucosos, ouvidos, hipocelularizadas, mastoide, cronico, total (quantitativo), Obliteração, seio esfenoidal, tecidos moles e inflamação e infeccioso. No documento também se pode fazer a mesma verificação e normalmente estes termos costumam aparecer no mesmo relatório.

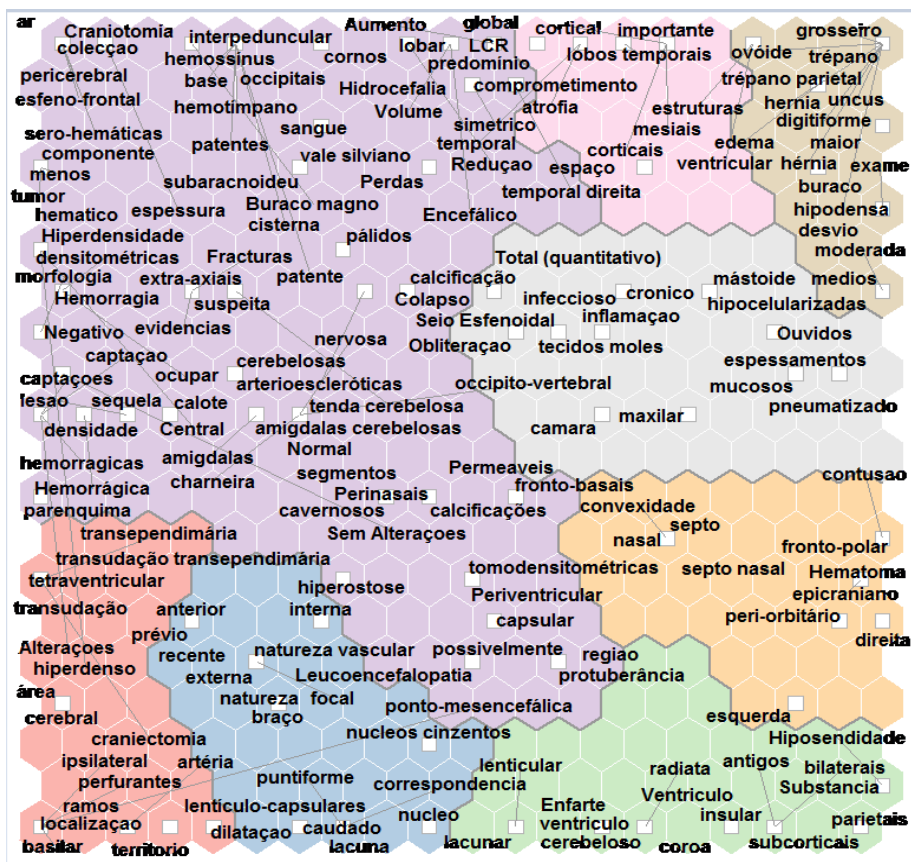


Figura 25 – Mapa Auto Organizacional dos Temas presentes no dicionário

#### D) Coocorrência de rede com Dicionário

Esse comando permitiu criar um diagrama de rede para explorar associação código. Dado que os códigos com padrões de aparência semelhante são diretamente conectados com linhas em uma rede, pode ser mais fácil de interpretar estes resultados visuais, em comparação com outras análises. A operação e as opções nesta etapa são semelhantes aos da Coocorrência de Palavras. A única diferença é que este utiliza temas em vez de palavras para a realização da análise (Higuchi, 2016).

Na Coocorrência de Rede, os temas do dicionário aparecem relativamente organizados. Os termos aparecem junto do tema ou temas que aparecem mais frequentemente no mesmo parágrafo. Por exemplo, as palavras área e hipodensa tem muitas relações, o que quer dizer que são utilizadas muitas vezes neste documento e relacionam-se com várias outras palavras neste documento. A palavra braço aparece ao lado de lenticular, e no documento verifica-se bem isso pois o conjunto “braço lenticular” aparece algumas vezes no documento. O mesmo acontece com segmentos e cavernosos.



Figura 26 – Coocorrência de Rede de Palavras do Dicionário

### E) Análise de Correspondência

Este comando executou uma análise de correspondência em temas e apresenta os resultados num diagrama de dispersão bidimensional. Esta análise é usada para explorar quais os tipos de temas que têm um padrão de aparência semelhante; quais os temas são considerados característicos para cada valor de uma variável, ou para cada capítulo ou secção de texto; e que capítulos ou valores de variáveis são semelhantes, com base nos temas que são definidas para estes dados de texto (Higuchi, 2016). Para esta análise ter uma representação legível estão representados apenas os 50 códigos com mais frequência no documento.

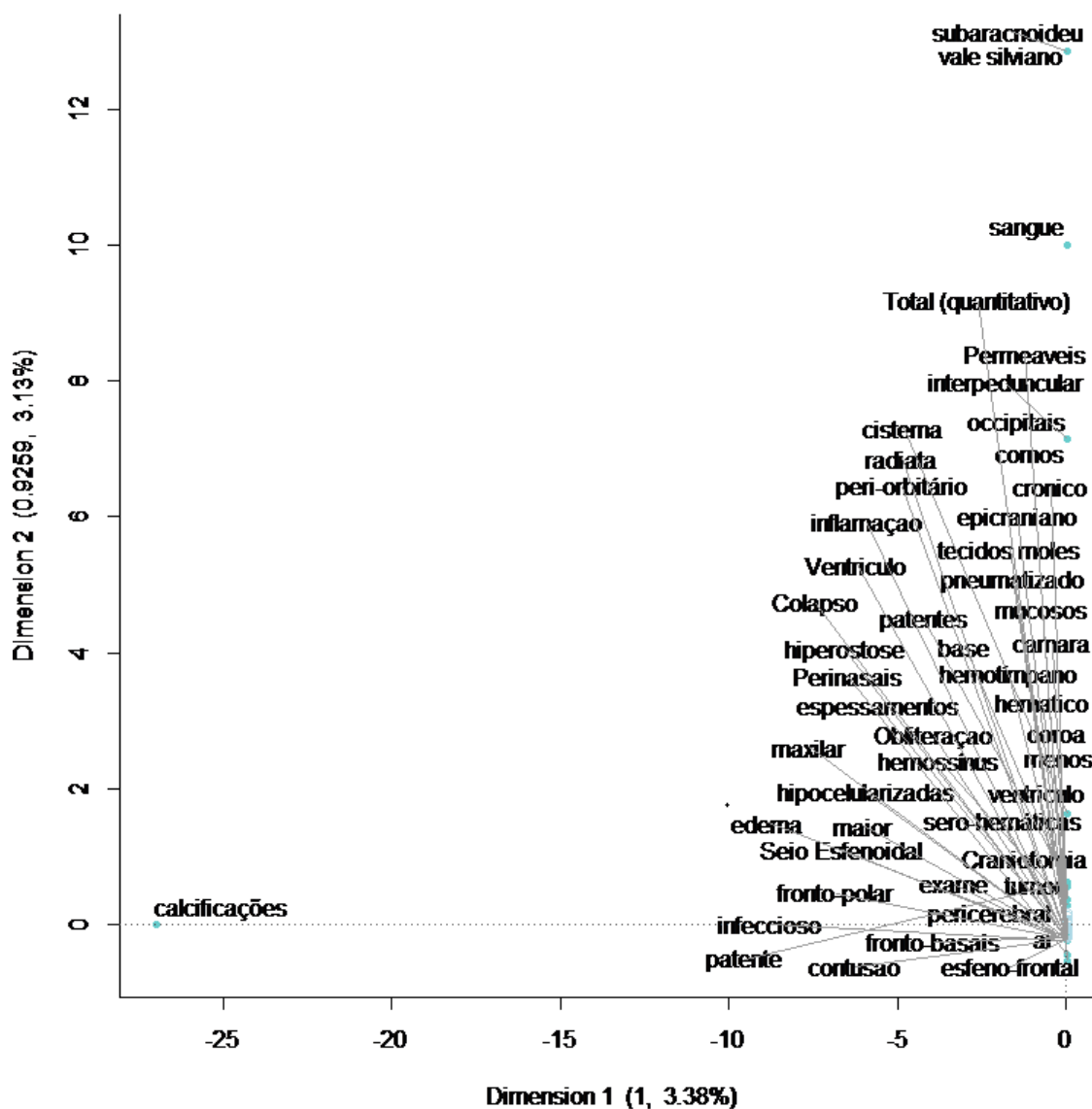


Figura 27 – Análise de Correspondência de Palavras presentes no Dicionário

### F) Escala multidimensional de Códigos

Este comando permitiu a criação de diagramas de dispersão tridimensionais para explorar as associações entre os códigos. A representação gráfica é muitas vezes mais fácil de interpretar do que os dados da matriz de similaridade. A operação e as opções neste ponto são semelhantes aos da escala multidimensional de palavras. A única diferença é que este utiliza temas em vez de palavras para a realização da análise (Higuchi, 2016).

Os grupos de palavras que se relacionam entre si apresentam a mesma cor. Por exemplo, os termos “cortical”, “base”, “LCR”, “espaço”, “buraco”, “Encefálico”, “Redução”, “patentes”, “cisterna”, “Buraco magno”, e “Normal” são temas pertencentes ao mesmo grupo de palavras, o que significa que é normal que a maioria destes termos estejam presentes na mesma frase.

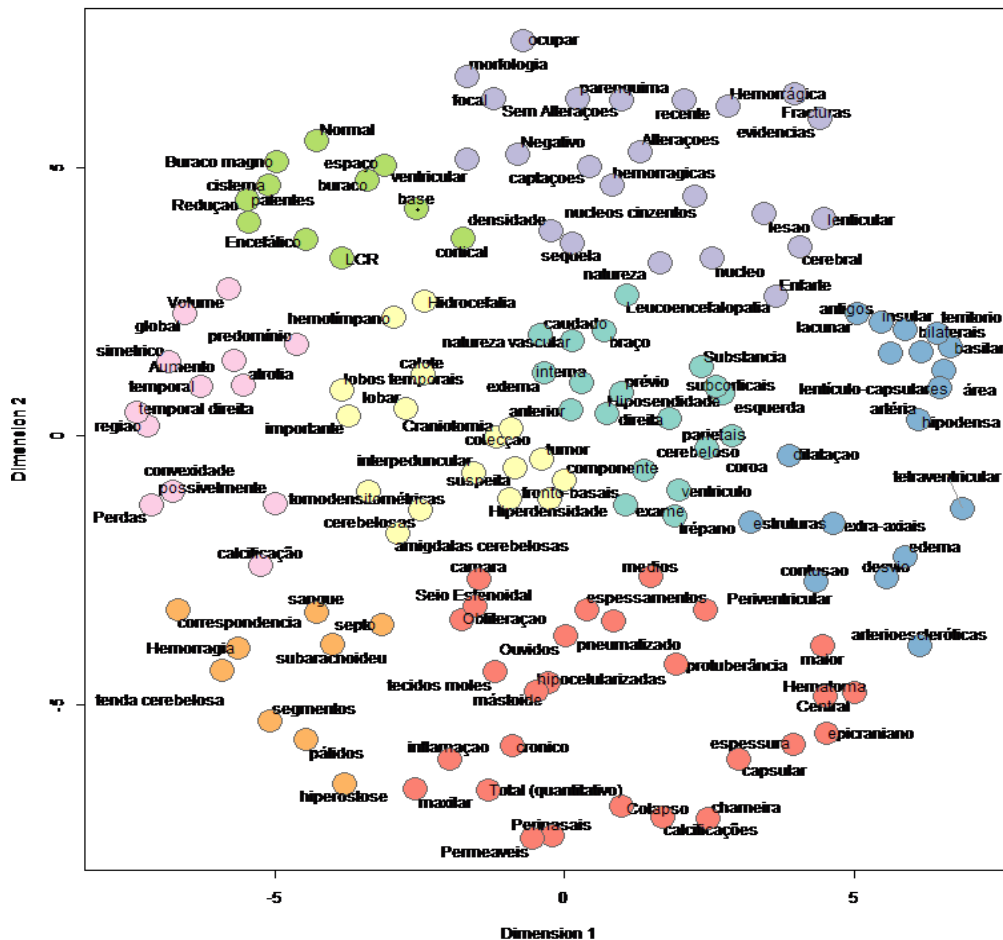


Figura 28 – Escala multidimensional de temas presentes no dicionário

# 7. Criação de uma Ontologia

A morte cerebral é um dos diagnósticos mais graves que se pode verificar num paciente. A possibilidade de detetá-los antes do seu acontecimento é um dos passos possíveis para a prevenção desse acontecimento. Os raio-x, Tomografias Computorizadas (TC), são exames bastante importantes para a deteção de morte cerebral. Neste capítulo é proposta a utilização de uma ontologia com base no registo dos raio-x efetuados aos doentes. Este trabalho foi possível através da obtenção dos dados fornecidos pelo Centro Hospitalar do Porto - Hospital de Santo António. A ontologia foi utilizada com base numa análise efetuada aos dados e com a utilização de um dicionário desenvolvido nessa mesma análise. Por fim adicionou-se os tipos de pacientes com Morte Cerebral que foram descobertos no capítulo anterior que utilizava o dicionário que esta ontologia contém.

## 7.1. Contexto do Problema

Os raio-x TC estão alojados no sistema informático da entidade prestadora de cuidados de saúde, e estes são escritas sob forma de texto livre, o que dificulta a análise em grupo de várias notas clínicas. Uma ontologia ajuda a organizar toda a informação que existe nas notas clínicas, o que pode facilitar uma análise automática com o objetivo de obter resultados de forma mais rápida e mais objetiva. Os doentes podem beneficiar destas análises, pois podem prevenir acontecimentos futuros que poderão ser prejudiciais.

Por exemplo, os raio-x TC são essenciais para a descoberta de morte cerebral nos doentes, e uma análise rápida sobre vários raio-x podem trazer identificações de tipos de pacientes que tiveram esse diagnóstico. No capítulo anterior, realizou-se um estudo onde se obteve oito tipos de padrões de doentes que tiveram morte cerebral e além disso foi desenvolvido um dicionário com termos médicos sobre esses mesmos pacientes.

## 7.2. Desenvolvimento da Ontologia

Uma ontologia desenvolvida contém todo o conhecimento obtido no capítulo anterior, nomeadamente o dicionário e os tipos de doentes que tiveram morte cerebral.

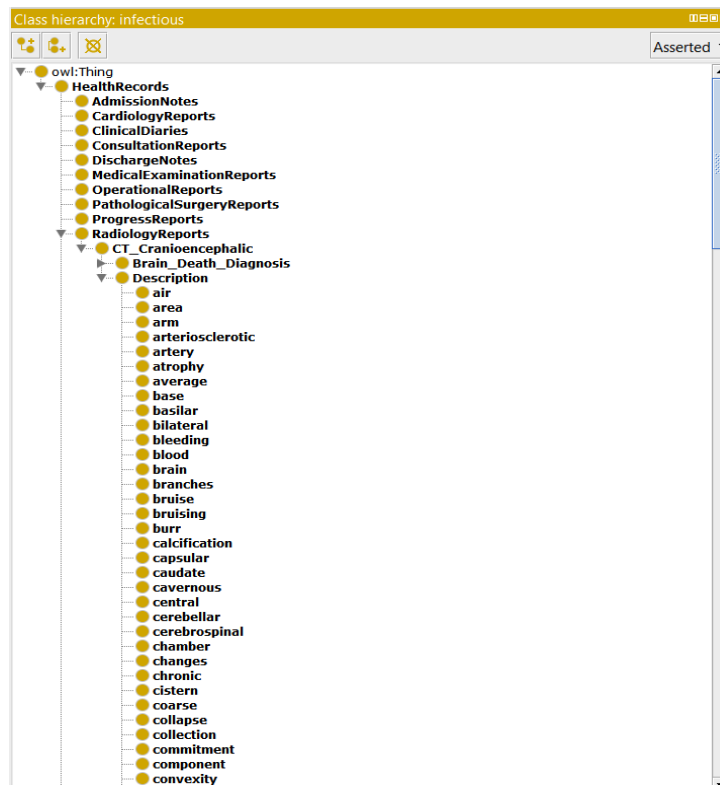


Figura 29 - Classes da Ontologia

Esta ontologia, como se pode verificar na figura 29, teve uma abordagem para a sua hierarquia de “*top-down*”, o que permite futuramente a adição de diferentes tipos de notas clínicas. A classe HealthRecords é a classe principal das notas clínicas, isto é, as classes que se encontram dentro deste são diferentes tipos de notas clínicas. Uma dessas classes é a RadiologyReports, que armazena os relatórios de raio-x. Essa classe divide-se em vários tipos de subclasses sendo uma dessas a CT\_Cranioencephalic. Esta classe tem duas subclasses, a Description e a Brain\_Death\_Diagnosis. A primeira contém todos os dados do dicionário, isto é, os dados relevantes sobre o diagnóstico dos raio-x aos doentes, e a outra classe tem oito tipos de doentes que tiveram morte cerebral. Esses tipos foram descobertos utilizando os termos na classe Description, ou seja, as duas classes estão diretamente relacionadas.

A figura 30 mostra a forma hierárquica desta ontologia, e a correlação existente entre as subclasses Description e Brain\_Death\_Diagnosis.



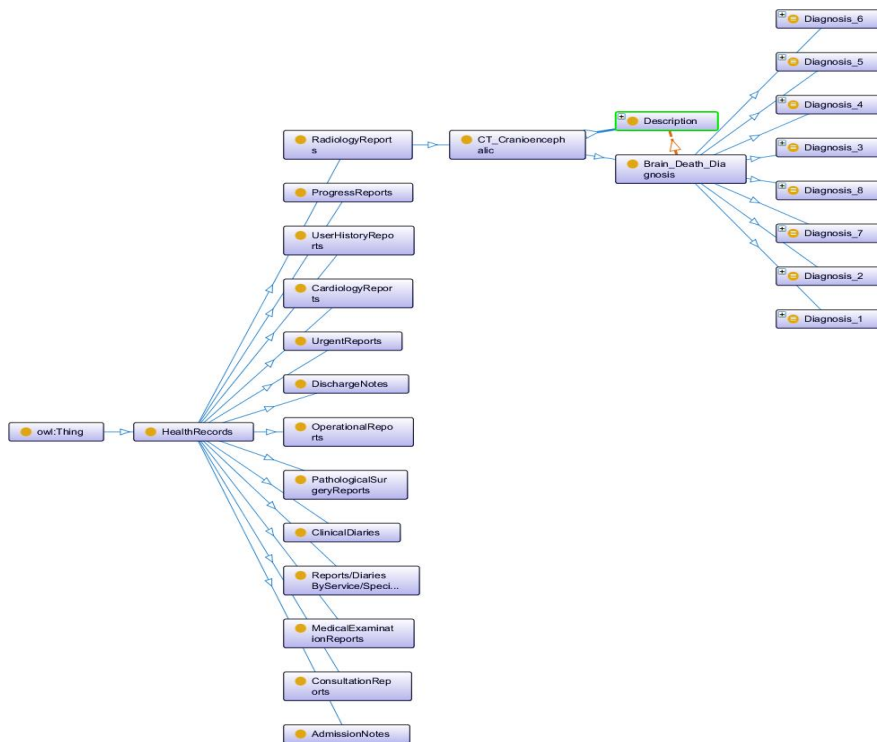


Figura 30 - Diagrama das Classes da Ontologia

### 7.3. Discussão da Ontologia

Esta ontologia foi desenvolvida sobre acontecimentos que acontecem num ambiente real, nomeadamente num hospital. Esta ontologia seguiu a hierarquia dos tipos de notas clínicas e subdividiu os mesmos nas suas vertentes mais específicas.

Com o desenvolvimento desta ontologia conseguiu-se uma melhor organização dos tipos de notas clínicas existentes nas instituições de saúde. Além disso, os dados que pertencem a cada tipo de nota clínica podem, posteriormente, fazer ligações a tipos de diagnósticos desta se for realizado um estudo de padronização de informação dessas mesmas notas clínicas.

Por fim, esta ontologia tem como maior objetivo a implementação num ambiente real. Essa implementação poderá ser realizada, pois utilizou dados reais de ambientes reais e foi hierarquizada de modo a ser adotada pelas instituições de saúde.

# 8. Criação de Modelos de Previsão com o KNIME

Neste capítulo é abordado o problema que irá ser explorado pelo processo de descoberta de conhecimento através de *Text Mining* (TM), que é a previsão de Morte cerebral depois da realização de um raio-x. Este trabalho é realizado seguindo a metodologia CRISP-DM.

## 8.1. Compreensão do Negócio

O Objetivo do negocio é a previsão de morte cerebral após a realização de um raio-x, para os técnicos de saúde conseguirem intervir antes de acontecer a Morte Cerebral com o Objetivo de a evitar.

Para isso irão ser utilizadas técnicas de *Data Mining* (DM), mas neste caso adaptadas para dados em texto. O Objetivo é que estes modelos alcancem resultados positivos na Sensibilidade, Especificidade para serem viáveis na ajuda à tomada de decisão dos técnicos de saúde.

## 8.2. Compreensão dos Dados

Os dados escolhidos para esta análise foram os ficheiros **ReportsObitos.xls** e o **ReportsVivos.xls**, pois estes dados contém a informação de raio-x de doentes que faleceram e de doentes que não faleceram, e o objetivo desta análise é criar modelos de previsão que consigam prever se o doente vai morrer – e aí os médicos poderão fazer um tratamento de modo a evitar com que possam falecer – ou se o doente vai sobrevier.

A coluna que contém a informação principal sobre os raio-x, é a coluna **DESCRICA0**, pois ela contém a descrição sobre o raio-x. Como os dados já foram tratados anteriormente, estes já se encontram praticamente prontos para o KNIME visto que os erros e incoerências dos dados já foram eliminados.

### 8.3. Preparação dos Dados

Na preparação dos dados verificou-se que como os dados contém bastantes linhas, a máquina de testes não tem capacidade para conseguir ler os dados todos, por isso, foi feita uma filtragem no *Microsoft Excel 2014*, e retirou-se uma amostra dos dados que continham todos os raio-x de pessoas que faleceram e que não faleceram relativas aos anos de 2009 e 2010. O resultado desta filtragem foi a criação de dois ficheiros com o nome de **ReportsObitos20092010.xls** e o ficheiro **ReportsVivos20092010.xls**.

Os dados continham uma disparidade assinalável entre o ficheiro **ReportsObitos20092010.xls** e o ficheiro **ReportsVivos20092010.xls**, pois o ficheiro **ReportsObitos20092010.xls** continha 1094 linhas, e o ficheiro **ReportsVivos20092010.xls** continha 4434 linhas. A solução foi copiar e colar as linhas existentes no ficheiro **ReportObitos20092010.xls** de modo a atingir um número próximo das 4434 linhas do ficheiro **ReportsVivos20092010.xls**. A este método chama-se *Oversampling*, e o resultado foi a criação de um ficheiro novo, o **ReportsObitos20092010Oversampling.xls**, que contém 4373 linhas, o que já é um número de linhas semelhante ao ficheiro **ReportsVivos20092010.xls**, o que permitira fazer uma análise mais justa aos dados.

A Figura 31 e a Tabela 17 demonstram a diferença em percentagens entre doentes vivos e mortos com e sem *Oversampling*.

Tabela 18 – Tabela com informação sobre doentes que morreram com e sem *oversampling*

<i>Sem Oversampling</i>		<i>Com Oversampling</i>	
Mortos	Vivos	Mortos	Vivos
1904	4434	4373	4434

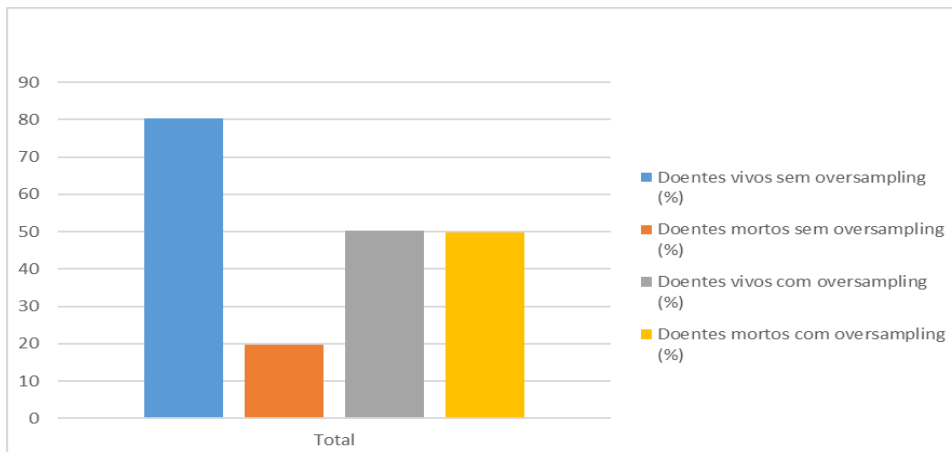


Figura 31 – Representação Gráfica da diferença de doentes que morreram com e sem oversampling

#### 8.4. Modelação

Para esta primeira análise foram utilizados os classificadores mais comuns do KNIME, e também os algoritmos que normalmente obtém melhores resultados, como se pode comprovar no teste executado por Weiss, S.M. Damerau, F. Apte (1998), que fizeram um teste com diversos algoritmos usando os dados da coleção 21578 da *Reuters*, e nesse teste os quatro melhores algoritmos foram o *Decision Tree* com 78,9%, *K-Nearest Neighbour* e o *Optimized Rule Induction* o com 82% de acuidade e o *Support Vector* com 86.3% de acuidade.

O KNIME não continha o *Optimized Rule Induction* para fazer a análise dos dados, logo, esse algoritmo foi excluído, e o *Support Vector (Support Vector Machine no KNIME)* inicialmente estava incluído nos testes, mas quando chegava a altura de fazer a análise e a criação dos modelos o algoritmo dava erro por falta de capacidade (memória RAM) da Máquina utilizada nos testes.

Com a exclusão do *Optimized Rule Induction* e do *Support Vector(SVM)*, restaram apenas o *Decision Tree*, e o *K-Nearest Neighbour* para a análise e criação dos modelos de previsão. Além disso, irá ser aplicado o *Cross Validation* nos algoritmos escolhidos, logo cada *Workflow* irá ter quatro resultados, que são os seguintes:

- Resultado com o *Decision Tree*;
- Resultado com o *K-Nearest Neighbour*;
- Resultado com *DecisionTree* com *Cross Validation*;
- Resultado com *K-Nearest Neighbour* com *Cross Validation*.

A Tabela 19 descreve as definições dos algoritmos utilizados neste teste.

Tabela 19 – Algoritmos utilizados no teste e os seus detalhes

	<i>Parâmetro</i>	<i>Valor</i>
<b>Decision Tree (DT)</b>	Amostra	30%
	<i>Number of Seeds</i>	<i>Random</i>
	<i>Quality Measure</i>	<i>Gini Index</i>
	<i>Pruning Method</i>	<i>No pruning</i>
	<i>Min Number records per node</i>	2
	<i>Number threads</i>	4
<hr/>		
<b>K-Nearest Neighbour (KNN)</b>	Amostra	30%
	<i>Number of Seeds</i>	<i>Random</i>
	<i>Number of Neighbours to consider (k)</i>	5
	<i>Weight neighbours by distance</i>	<i>True</i>
<hr/>		
<b>Decision Tree com Cross Validation (DTCV)</b>	<i>Amostra</i>	30%
	<i>Number of Seeds</i>	1441778237149
	<i>Quality Measure</i>	<i>Gini Index</i>
	<i>Pruning Method</i>	<i>No pruning</i>
	<i>Min Number records per node</i>	2
	<i>Number Threads</i>	8
	<i>Number of Validations</i>	10
<hr/>		
<b>K-Nearest Neighbour com Cross Validation (KNNCV)</b>	Amostra	30%
	<i>Number of Seeds</i>	1441778237149
	<i>Number of Neighbours to consider (k)</i>	3
	<i>Weight neighbours by distance</i>	<i>True</i>
	<i>Number of validations</i>	10

Nas Figuras 32, 33 e 34 encontra-se um exemplo de um *workflow* dividido pelas três figuras criado no *KNIME* para fazer a criação dos modelos de previsão. A Figura 32 representa o conjunto os dados que são extraídos das folhas de cálculos e são tratados e concatenados de modo a poderem ser interpretados pelo *KNIME*. O *File Reader* lê o ficheiro de texto que contém o dicionário e aplica os dados aos dados que irão ser analisados, o *POS Tagger* completa as tarefas de enriquecimento dos dados.

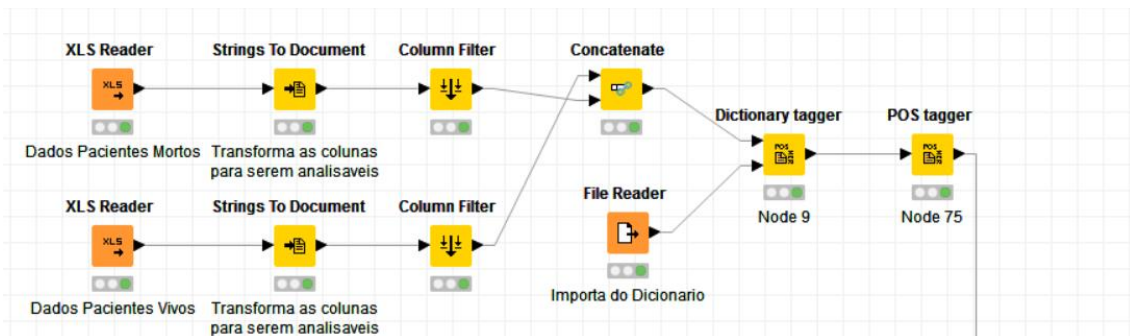


Figura 32 – Primeiro conjunto do workflow aplicado no *KNIME*

A Figura 33 contém o processo de tratamento dos dados, que consiste na uniformização dos dados para que estes possam ser analisados com uma maior eficiência e eficácia. O *Tag Filter* filtra apenas os tipos de *Tag* desejados para a análise e descarta os termos que não contém o *Tag* selecionado. Após isso, os termos filtrados irão ser agrupados no *Bag of Words Creator*, e o *TF* irá calcular a frequência de vezes que esses termos aparecem no conjunto de dados.

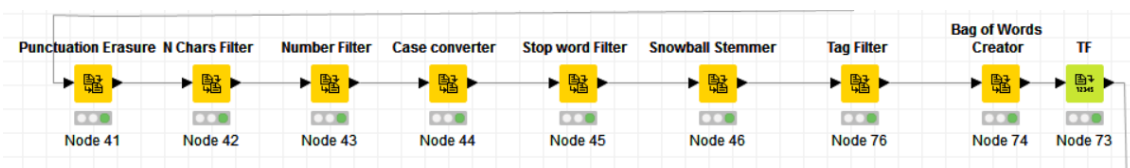


Figura 33 – Segundo conjunto do workflow aplicado no *KNIME*

Por fim, na Figura 34 está representada a última parte do *workflow*, onde as frequências de termos calculadas irão ser convertidas para valores binários pelo *Document Vector*, depois disso, o *Document Data Extractor* extrai a categoria de cada documento, neste caso Morto ou Vivo, depois é dada uma cor a assinalar cada categoria. No *Partitioning* o documento com os relatórios de raio-x é dividido em duas partes, a primeira parte contém 70% dos dados que se destina para o

treino nos algoritmos e a segunda parte com 30% dos dados, cujo objetivo é testar a previsão os dados nos algoritmos. Por fim, os dados irão ser aplicados aos quatro algoritmos de TM.

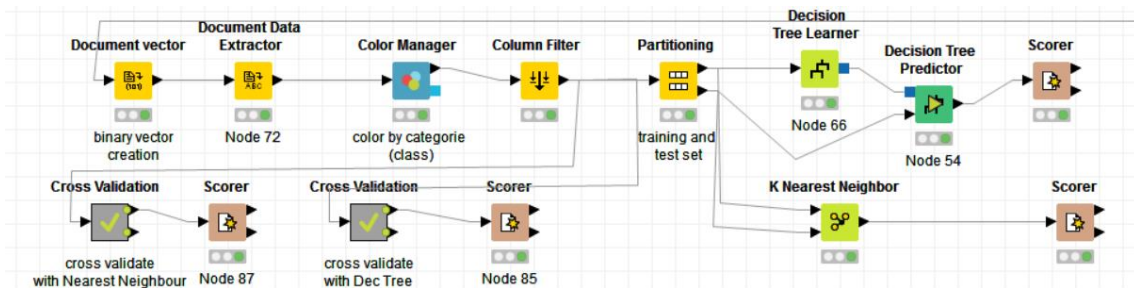


Figura 34 – Terceiro conjunto do workflow utilizado no KNIME

Abaixo segue-se uma descrição detalhada de todos os métodos que estão identificados no *Workflow*.

Tabela 20 – Métodos do Workflow

Nome	Descrição
<i>XLS Reader</i>	Lê os documentos em formato *.xls e importa-os para o programa
<i>Strings To Document</i>	Transforma as <i>strings</i> dos documentos no formato que o programa consegue interpretar
<i>Column Filter</i>	Faz a filtragem das colunas escolhidas
<i>Concatenate</i>	Junta os ficheiros escolhidos
<i>File Reader</i>	Lê os ficheiros, normalmente, documentos de texto, neste caso lê o dicionário
<i>Dictionary Tagger</i>	Marca as palavras no documento que estejam presentes no dicionário
<i>POS Tagger</i>	Identifica todas as <i>Part-of-Speech</i> existentes no documento
<i>Punctuation Erasure</i>	Elimina as pontuações dos documentos.
<i>N Chars Filter</i>	Elimina as palavras que estejam abaixo de um número de caracteres definido pelo utilizador. Neste caso eram palavras com menos de três caracteres
<i>Number Filter</i>	Elimina os números do Documento

<i>Case Converter</i>	Poe todos os caracteres do documento ao mesmo nível. Neste caso foi sem Maiúsculas
<i>Stop word Filter</i>	Utiliza um dicionário <i>stop word</i> integrado no programa e elimina as palavras que estão contidas nesse dicionário.
<i>Snowball Stemmer</i>	Agrupa todas as palavras que se assemelham numa palavra apenas
<i>Tag Filter</i>	Marca os termos escolhidos pelo utilizador. Neste caso, foram os verbos, os advérbios, os nomes os adjetivos e os termos do dicionário.
<i>Bag of Words Creator</i>	Cria o grupo de palavras filtrado no Tag filter
<i>TF</i>	Calcula a frequência de vezes em que aparece a palavra no documento
<i>Document Vector</i>	Transforma o documento num vetor binário
<i>Document Data Extractor</i>	Extrai informação específica do documento, neste caso extrai a categoria do documento (Morto ou Vivo)
<i>Color Manager</i>	Atribui uma cor a cada categoria do documento
<i>Partitioning</i>	Divide os dados em dois conjuntos. Um conjunto com 70% para ser usado no treino, e os 30% para serem usados no teste
<i>Decision Tree Learner</i>	Faz o treino dos dados
<i>Decision Tree Predictor</i>	Faz a previsão dos dados com base no treino feito anteriormente
<i>K-Nearest Neighbour</i>	Faz a previsão dos dados com o algoritmo <i>K-Nearest Neighbour</i>
<i>Cross Validation</i>	Faz a previsão dos dados, mas com <i>Cross Validation</i> . Neste caso foi definido um número de dez repetições
<i>Scorer</i>	Recolhe os resultados e permite a visualização dos mesmos

Este *Workflow* representado na Tabela 20 é o mais completo de todos os *Workflows* criados no KNIME para esta análise, pois ele contém a integração de um dicionário, a utilização de um *POS Tagger*, e a utilização de um *Tag Filter*. Estes três métodos são o que vão diferenciar os *Workflows* no KNIME. Cada *Workflow* representa um cenário, e no total foram criados sete Cenários diferentes que são os seguintes:

- S1 – {Dicionário, *POS Tagger*, *Tag Filter*};
- S2 – {Dicionário com *POS Tagger* e sem *Tag Filter*};
- S3 – {com a utilização de Dicionário sem *POS Tagger* e com *Tag Filter*};



- S4 – {com a utilização de Dicionário sem POS *Tagger* e sem *Tag Filter*};
- S5 – {Sem Dicionário, mas com a utilização de POS *Tagger* e *Tag Filter*};
- S6 – {Sem Dicionário, com a POS *Tagger* e sem *Tag Filter*};
- S7 – {Sem Dicionário, sem POS *Tagger* e sem *Tag Filter*}.

Além dos cenários criados foram utilizados dois métodos de *sampling* na folha de cálculo **ReportObitos20092010.xls**, que consistia na utilização ou não do *oversampling*. Cada cenário continha as quatro técnicas de *mining* referidas anteriormente (*Decision Tree*, *Decision Tree* com *Cross Validation*, *K-Nearest Neighbor* e *K-Nearest Neighbor* com *Cross Validation*). E o objetivo definido era o falecimento ou não dos doentes.

A equação abaixo define cada modelo de TM:

$$TMM_m = TMT_y \times A \times T_s \times S_i$$

O  $TMT_y$  representa a técnica de TM utilizada, o  $A$  representa o alvo definido, o  $T_s$  refere-se ao método de *sampling* aplicado, e o  $S_i$  representa o cenário utilizado. Abaixo está presente um exemplo de uma equação a definir um dos modelos criados neste estudo:

$$TMM_m = DT_y \times Morto \times SemOversampling_s \times 7_i$$

Foram criados um total de cinquenta e seis modelos de TM (quatro técnicas de TM x um alvo x dois métodos de *sampling* x sete cenários). Estes modelos querem-se sempre com a maior acuidade possível, mas foi definido um critério de valores mínimos (recorrendo ao conhecimento clínico) para os modelos serem aceites como fiáveis para a utilização em casos reais. Os critérios definidos são:

- A **sensibilidade** terá de ter valores superiores a 85%;
- A **especificidade** terá de ter valores superiores a 75%;
- A **acuidade** tem de ter valores superiores a 85%;
- E o **erro** nunca poderá passar os 15%.

## 8.5. Avaliação

Os resultados irão ser apresentados numa tabela com os dois melhores resultados por algoritmo utilizado em combinação com a presença ou não de *Oversampling*. Nessa tabela será possível ver o cenário utilizado, o algoritmo utilizado, bem como se os dados tinham *Oversampling*

ou não. Além destes indicadores, a tabela também irá conter os valores de Especificidade(E), Sensibilidade(S), Acuidade(A) e o Erro(Er). A tabela com os resultados dos cinquenta e seis modelos está na secção Anexos.

Os resultados presentes na tabela com os melhores resultados, os valores que atingiram os valores mínimos definidos na modelação estão assinalados a negrito na Tabela 21. Posto isto, só irão ser considerados os modelos que tenham os valores todos a negrito, isto é, atingem os resultados definidos anteriormente.

*Tabela 21 – Tabela com os melhores resultados por Algoritmo*

Algoritmo	<i>Oversampling</i>	Cenário	S	E	A	Er
DT	Não	3	27,30%	<b>85,51%</b>	74,01%	25,98%
DT	Não	5	28,05%	<b>85,49%</b>	74,13%	25,87%
DT	Sim	1	<b>91,01%</b>	<b>76,24%</b>	83,57%	16,43%
DT	Sim	2	<b>90,78%</b>	<b>76,77%</b>	83,72%	16,28%
KNN	Não	1	4,27%	<b>98,05%</b>	79,49%	20,51%
KNN	Não	2	4,27%	<b>97,52%</b>	79,07%	20,93%
KNN	Sim	4	<b>97,03%</b>	71,35%	84,10%	15,90%
KNN	Sim	5	<b>97,33%</b>	<b>75,34%</b>	<b>86,26%</b>	<b>13,74%</b>
DTCV	Não	3	27,11%	<b>85,48%</b>	73,94%	26,06%
DTCV	Não	4	26,72%	<b>84,19%</b>	72,82%	27,18%
DTCV	Sim	3	<b>98,39%</b>	<b>79,18%</b>	<b>88,72%</b>	<b>11,28%</b>
DTCV	Sim	4	<b>98,03%</b>	<b>77,96%</b>	<b>87,93%</b>	<b>12,07%</b>
KNNCV	Não	2	8,78%	<b>96,62%</b>	79,24%	20,76%
KNNCV	Não	3	6,16%	<b>97,28%</b>	79,27%	20,73%
KNNCV	Sim	1	<b>93,46%</b>	70,74%	82,02%	17,98%
KNNCV	Sim	2	<b>93,78%</b>	69,28%	81,44%	18,56%

## 8.6. Discussão dos Resultados

Com base na análise dos resultados pode-se constatar que os resultados no Algoritmo *K-Nearest Neighbour* sem *Oversampling* são bastante positivos na Especificidade com os melhores resultados a chegar aos 98% (sem *cross validation*), mas os resultados não são satisfatórios na Acuidade e Sensibilidade, pois nunca alcançaram os valores mínimos definidos anteriormente.

Nos dados com *Oversampling* já se verifica uma melhoria significativa na Acuidade e Sensibilidade, embora a especificidade tenha descido ligeiramente. Neste cenário, os modelos com o *cross validation* conseguem alcançar os valores mínimos definidos.

O algoritmo que teve melhores resultados foi o *Decision Tree* com *cross validation* e com *Oversampling*, onde na maior parte dos cenários definidos, os valores cumprem os requisitos estabelecidos.

Estes resultados são bastante positivos, pois cumprem os valores mínimos definidos para cada atributo, e não deixa de ser interessante o facto de existirem modelos de previsão com excelentes resultados mesmo não sem a utilização do dicionário. Apesar disso o dicionário verificou-se bastante útil para a análise dos dados de uma maneira mais específica.

De um ponto de vista clínico, os resultados alcançados pelos melhores modelos podem permitir a previsão de mortes cerebrais após a realização do raio-x, permitindo uma melhor prestação de tratamentos com doentes que tem problemas cerebrais. Num exemplo real, os técnicos de saúde podem ser avisados se o doente pode vir a falecer com base no diagnóstico efetuado no raio-x, para intervirem e realizarem um tratamento preventivo antes da morte cerebral acontecer.

A análise mais detalhada dos resultados atingidos é apresentada no capítulo seguinte (8) – Discussão de Resultados.

## 9. Discussão de Resultados

Neste capítulo o maior foco é a discussão dos resultados obtidos nos dois estudos desenvolvidos nesta dissertação. A discussão será dividida em duas partes, uma por cada estudo realizado neste projeto. No final de cada discussão é realizada uma análise crítica dos resultados obtidos.

### 9.1. Resultados Obtidos nos Modelos de Análise criados pelo *KH Coder*

#### *A) Frequência de Palavras*

Esta análise nunca pode incidir num assunto específico dos relatórios de raio-x pois só contabiliza a frequência de palavras, ou seja, só dá as indicações das palavras que aparecem mais frequentemente no documento. Assim, esta análise fará mais sentido se for levada apenas com fins estatísticos, e não de procura e descoberta de padronização de doentes com morte cerebral.

Com base na análise realizada à frequência de palavras presentes nos relatórios de raio-x analisados, verificou-se que a palavra mais frequente era a palavra “não”. A palavra em si não é muito objetiva, pois o termo não pode tomar muitos valores como “não se observam lesões”, ou então “não foram registadas alterações”. As palavras “lesão” e “encefálico” surgem logo a seguir com 52 e 50 aparições no documento. Neste caso aplica-se a mesma situação da palavra “não” pois a palavra “lesão” e “encefálico” podem tanto estar a demonstrar que o doente teve alguma lesão como não teve lesão.

#### *B) Análise Hierárquica de Clusters*

Esta análise permitiu ter uma visão mais abrangente sobre o documento, e tornou mais fácil a descoberta de padrões de doentes que tiveram morte cerebral, agrupando em *cluster* os termos que têm valores semelhantes de aparência no documento, tal como a Escala Multidimensional. No entanto, a Análise Hierárquica de *Clusters* permitiu uma visão mais específica da análise pois divide os resultados em diferentes ramificações do dendrograma como se irá verificar nas análises abaixo, que contêm os dados de cada *cluster*. Cada *cluster* contém um tipo de frases utilizadas nos relatórios de raio-x. Para uma melhor discussão dos resultados,

estão também abaixo descritos os dados que constituíam cada *cluster*, bem como uma discussão de resultados por cada tipo de análise realizada.

**1) Discussão da análise com termos que aparecem pelo menos vinte vezes no documento:**

*Clusters* criados por esta análise:

- *Cluster 1* – {posterior, espessura, axial, corte};
- *Cluster 2* – {hidrocefalia, encefálico, via, circulação, *líquor*, alargamento, volume};
- *Cluster 3* – {parênquima, alteração, densidade};
- *Cluster 4* – {lesão, não, observar};
- *Cluster 5* – {espaço};
- *Cluster 6* – {normal, ser};
- *Cluster 7* – {esquerdo, direita};
- *Cluster 8* – {cerebral, isquêmico, hipodensidade, enfarte, traduzir}.

Nesta fase verificou-se que a análise, por conter as palavras que aparecem mais vezes no documento, é uma análise geral, e que não se conseguem retirar resultados sólidos em termos de padronização dos doentes. Por exemplo, um *cluster* contém o método utilizado para a realização do raio-x, como é o caso do *Cluster 1*, outros *clusters* não contêm informação relevante como os *Clusters 5* e *7*. Mas também há casos de *clusters* que já contêm alguma informação relevante em relação aos relatórios, como o *Cluster 8*, em que pode conter a seguinte frase “Hipodensidade isquêmica cerebral, o que se traduziu num enfarte”, ou também como o *Cluster 3* “Alteração da densidade do Parênquima”, e o *cluster 2* “Alargamento das vias de circulação de *líquor* encefálico, hidrocefalia”. Estas frases representadas podem significar diagnósticos que os doentes tiveram antes de terem morte cerebral.

**2) Discussão da análise com termos que aparecem pelo menos quinze vezes no documento:**

*Clusters* criados por esta análise:

- *Cluster 1* – {posterior, supratentorial, técnica, ir, espessura, corte, axial, contraste estudo};
- *Cluster 2* – {encefálico, predomínio, alargamento, volume, hidrocefalia, via, circulação, *líquor*};

- *Cluster 3* – {cerebral, isquêmico, branco, hipodensidade};
- *Cluster 4* – {espaço, traduzir, enfarte, núcleo, esquerdo, direita};
- *Cluster 5* – {normal, ser};
- *Cluster 6* – {sugerir, discreto, calcificação};
- *Cluster 7* – {parênquima, alteração, densidade};
- *Cluster 8* – {imagem, lesão, não, observar}.

Nesta análise já se verifica uma maior objetividade, ou seja, já existem *clusters* que contêm mais conteúdo relevante no que toca aos diagnósticos de morte cerebral, porém ainda existem *clusters* que não contêm informação relevante, como os *clusters* 1 e 5. Os *Clusters* 2, 3, 4, 6, 7 e 8 já contêm bastante informação sobre os relatórios de raio-x, onde por exemplo o *cluster 2* tem a seguinte frase “Hidrocefalia nas vias de circulação de *líquor*, e alargamento de volume de predomínio encefálico”, e o *cluster 8* contém a frase “não se observa lesão na imagem”. Mas, para um uso mais específico, apenas os *clusters* 2, 3, 4, 7 e 8 é que contém informação mais relevante sobre os diagnósticos efetuados pelos médicos.

### 3) Discussão da análise com termos que aparecem pelo menos dez vezes no documento:

*Clusters* criados por esta análise:

- *Cluster 1* – {posterior, supratentorial, fossa, compartimento, técnica, ir, espessura, axial, corte, contraste, estudo};
- *Cluster 2* – {predomínio, temporal, atrofia, encefálico, via, circulação, *líquor*, alargamento, volume, global, hidrocefalia, redução};
- *Cluster 3* – {anterior, esquerda, hipodensidades, interno, capsula};
- *Cluster 4* – {lacunar, enfarte, esquerdo, antigo, leucoencefalopatia, branco, substancia, cerebral, isquêmico, hipodensidade};
- *Cluster 5* – {sugerir, densidade, morfologia, alteração, parênquima, imagem, haver, nao, observar, expansivo, lesão};
- *Cluster 6* – {charneira, nervoso, occipitovertebral};
- *Cluster 7* – {subcortical, efeito, massa, exame, discreto, medio, direito};
- *Cluster 8* – {calcificação, mural, sifão, espaço, dimensão, apresentar, normal, ser, base, cisterna, sulco, artéria, traduzir, cumprimento, m, coleção, seio, tecido, etário, sinais, núcleo, hemorrágico, direita}.

Por fim, esta análise contém uma grande parte dos termos utilizados nos relatórios de raio-x, e quase todos os *clusters* desta análise podem ser usados para representar frases de diagnósticos efetuados pelos médicos. Além dos diagnósticos efetuados, alguns destes *clusters* contém informações sobre as técnicas utilizadas, o que pode facilitar ou não a descrição destes diagnósticos.

Em jeito de conclusão à Análise Hierárquica de *Clusters*, podemos verificar que a análise efetuada com os termos que aparecem pelo menos vinte vezes no documento permite descobrir o tipo de diagnósticos mais comuns efetuados aos doentes, assim como as técnicas mais utilizadas para a realização desses diagnósticos. As análises com termos que aparecem pelo menos quinze e dez vezes já dão outra visão sobre os relatórios, dando a possibilidade de descobrir vários tipos de diagnósticos realizados aos doentes, tendo estes um nível de maior detalhe na análise com termos que aparecem pelo menos dez vezes no documento. Por fim, estas análises podem ser muito importantes na descoberta de diagnósticos de doentes que tem raio-x, bem como verificar quais são os diagnósticos que aparecem mais vezes nesses doentes, bem com a obtenção detalhada dos diferentes tipos de diagnósticos.

### *C) Mapa Auto Organizacional*

No Mapa Auto Organizacional obteve-se a representação da informação dos relatórios divididos em vários tipos de relatórios de raio-x. Este processo é feito através das associações de palavras, e é o tipo de análise que demorou mais a executar, cada modelo demorou em média trinta minutos a ser executado. Cada modelo vai ser discutido separadamente, e no final das discussões é realizada uma conclusão geral sobre a execução desta análise.

#### **1) Discussão da análise com termos que aparecem pelo menos vinte vezes no documento:**

*Clusters* criados por esta análise:

- *Cluster 1* – {lesão, isquémico, hipodensidade, enfarte};
- *Cluster 2* – {observar, nao, hidrocefalia};
- *Cluster 3* – {parênquima, densidade, alteração, cerebral};
- *Cluster 4* – {direita, esquerdo, posterior};
- *Cluster 5* – {ser, normal};
- *Cluster 6* – {corte, axial, espessura};

- *Cluster 7* – {encefálico};
- *Cluster 8* – {traduzir, volume, alargamento, via, circulação, *líquor*, espaço}.

Nesta análise, em semelhança com a Análise Hierárquica de *Clusters*, contém *clusters* que não contêm qualquer tipo de informação relevante. Nesta análise, os *clusters* que contêm maior informação são os *clusters* 1, 2, 3 e 8. O *cluster* 6 contém apenas uma técnica utilizada.

## 2) Discussão da análise com termos que aparecem pelo menos quinze vezes no documento:

*Clusters* criados por esta análise:

- *Cluster 1* – {hipodensidade, isquémico, branco, enfarte, cerebral, parênquima, densidade, alteração};
- *Cluster 2* – {observar, não, imagem, lesão, núcleo, direita};
- *Cluster 3* – {calcificação, discreto};
- *Cluster 4* – {técnica, esquerdo};
- *Cluster 5* – {supratentorial, posterior};
- *Cluster 6* – {sugerir, traduzir, hidrocefalia, volume, circulação, predomínio, via, alargamento};
- *Cluster 7* – {normal, ser, *líquor*, espaço};
- *Cluster 8* – {contraste, espessura, encefálico, estudo, ir, axial, corte}.

Nesta análise, já se verifica uma maior e melhor identificação de tipos de doentes que tiveram morte cerebral após a realização do raio-x. O nível de detalhe de cada tipo de doente não é grande, mas já se consegue identificar alguns tipos de doentes, como por exemplo, os *clusters* 1, e 6 já contêm informação mais detalhada. Os restantes *clusters* não contêm informação onde seja possível definir claramente um tipo de doente com morte cerebral. Os *clusters* 3, 4 e 5 são o exemplo disso mesmo, onde o nível de detalhe dos mesmos é reduzido, pois só contêm duas palavras cada um.

## 3) Discussão da análise com termos que aparecem pelo menos dez vezes no documento:

*Clusters* criados por esta análise:

- *Cluster 1* – {axial, corte, m, espessura, ir, estudo, encefálico, contraste, atrofia};



- *Cluster 2* – {sulco, sugerir, hidrocefalia, temporal, predomínio, alargamento, global, via, circulação, volume, redução, traduzir, *líquor*, discreto, etário, dimensão, espaço, apresentar, normal, subcortical};
- *Cluster 3* – {cumprimento, exame, seio, sinais, direito, cisterna, base, medio, efeito, massa};
- *Cluster 4* – {calcificação, mural, sifão, esquerdo, artéria};
- *Cluster 5* – {densidade, morfologia, parênquima, alteração, cerebral, substancia, núcleo, leucoencefalopatia, hipodensidade, branco, isquêmico, lacunar, enfarte, antigo, hipodensidades};
- *Cluster 6* – {fossa, posterior, compartimento, supratentorial};
- *Cluster 7* – {técnica, direita, tecido, hemorrágico, lesão, haver, coleção, não, imagem, observar, expansivo, esquerda, interno, anterior};
- *Cluster 8* – {occipitovertebral, nervoso, ser, charneira, capsula}.

Na análise com termos que aparecem mais de dez vezes no documento já se pode verificar que todos os *clusters* incluem diagnósticos efetuados aos doentes, alguns deles até com grande detalhe como por exemplo, o *cluster 5*.

A análise do Mapa Auto Dimensional permitiu que fossem realizadas análises imediatas aos dados das mesmas devido à facilidade em que esta dispôs dos dados. Na parte da análise em si, verificou-se tanto as análises com termos que apareciam vinte e quinze vezes no documento não forneciam um número razoável de tipos de doentes que tiveram morte cerebral. Apenas a análise com termos que aparecem pelo menos dez vezes permitiu uma identificação de oito tipos de doentes, um por cada *cluster* definido. Estes resultados podem servir para criar os padrões de doentes que tiveram morte cerebral.

#### *D) Coocorrência de Rede*

Esta análise permitiu visualizar as relações existentes entre os termos do documento. Isto é, quando existem ligações entre os termos, isto significa que essa relação está presente nos relatórios de raio-x. Esta discussão irá ser dividida em três grupos com base nas análises efetuadas, sendo feita no fim uma síntese das análises efetuadas.

**1) Discussão da análise com termos que aparecem pelo menos vinte vezes no documento:**

Esta análise resultou num grupo de palavras onde todos os termos se relacionavam entre si, ou seja, não haviam grupos de palavras separados ou sem qualquer relação. As relações existentes estão presentes em frases dos relatórios de raio-x, sendo estas frases o resultado das relações presentes nesta análise:

- Não se observou alteração na densidade do parênquima encefálico;
- Alargamento das vias do volume de circulação de *líquor*;
- Observou-se alteração da densidade cerebral isquêmica, hipodensidade;
- Posterior espessura do corte axial encefálico.

**2) Discussão da análise com termos que aparecem pelo menos quinze vezes no documento:**

Esta análise resultou em três grandes grupos de palavras onde os termos se relacionavam entre si. As relações existentes estão presentes em frases dos relatórios de raio-x, sendo estas frases o resultado das relações presentes nesta análise:

- Alteração da densidade do parênquima encefálico;
- Não se observou lesão na imagem;
- Alteração isquêmica;
- Estudo da espessura posterior supratentorial.

**3) Discussão da análise com termos que aparecem pelo menos dez vezes no documento:**

Esta análise resultou em onze grandes grupos de palavras onde os termos se relacionavam entre si. As relações existentes estão presentes em frases dos relatórios de raio-x, sendo estas frases o resultado das relações presentes nesta análise:

- Enfarte lacunar antigo no lado esquerdo.
- Hipodensidades na cápsula interna.
- Charneira nervosa, occipitovertebral está normal.
- Alteração da densidade morfológica do parênquima.
- Hipodensidade da substância branca, leucoencefalopatia cerebral.
- Calcificação do sifão mural.

- Efeito da massa subcortical.
- Redução de volume encefálico de predomínio temporal, atrofia.
- Alargamento global das vias de circulação de *líquor*, hidrocefalia.
- Corte axial de espessura posterior.

Como acontece com as outras análises, o nível de detalhe das mesmas aumenta quando se aumenta o número de palavras presentes no documento (tendo por base o número mínimo de ocorrências). Os resultados são bastante interessantes, pois também permite ver com outra visão os tipos de diagnósticos efetuados aos doentes, e quantas vertentes diferentes pode ter cada diagnóstico.

### *E) Análise de Correspondência*

A análise de correspondência tem como objetivo a demonstração gráfica da localização das palavras no documento e a distância entre ambas no documento. Esta análise dividiu-se em três análises efetuadas sobre termos que aparecem dez ou mais vezes no documento, quinze ou mais vezes no documento e vinte ou mais vezes no documento. Esta análise foi realizada desta forma com o objetivo de ter resultados mais concretos sobre as localizações das palavras no documento. Esta discussão divide-se em três partes, sendo cada parte referente a cada tipo de análise efetuada. Por fim, os resultados são discutidos em conjunto.

#### **1) Discussão da análise com termos que aparecem pelo menos vinte vezes no documento:**

Nesta análise verificou-se principalmente que existiam três grandes grupos de palavras. Um grupo continha os termos “posterior”, “espessura”, “corte” e “axial”, em semelhança às análises anteriores, costumam aparecer quase sempre juntos, e isto indica uma técnica utilizada pelos médicos para a realização do raio-x. Outro grupo de palavras continha “hidrocefalia”, “espaço”, “alargamento”, “*líquor*”, “via”, “circulação” e “volume”. Isto significa que estas palavras aparecem sempre perto umas das outras. O mesmo se aplica aos restantes grupos de palavras. Também se pode afirmar, que o grupo de palavras “posterior”, “espessura”, “axial”, e “corte”, distanciam-se bastante dos restantes grupos de palavras, o mesmo aplica-se ao termo “normal”.

## **2) Discussão da análise com termos que aparecem pelo menos quinze vezes no documento:**

Esta análise já é mais difícil retirar conclusões mais concretas sobre o documento, pois apenas existem dois grupos de palavras. Um dos grupos contém palavras que descrevem as técnicas utilizadas para a realização do raio-x, e outro grupo de palavras representa as palavras que aparecem nos relatórios de raio-x, pois, na descrição do diagnóstico de raio-x, estas aparecem todas praticamente juntas umas das outras.

## **3) Discussão da análise com termos que aparecem pelo menos dez vezes no documento:**

Na análise com termos que aparecem pelo menos dez vezes no documento a análise ao documento torna-se ainda mais complicada, pois existem demasiados termos presentes no gráfico, por isso a limitação de termos presentes no dicionário ficou-se em 50 termos para garantir a elegibilidade do mesmo. Posto isto, pode-se verificar novamente dois grupos de palavras, que se referem às técnicas utilizadas para a realização do raio-x e os termos que identificam o diagnóstico efetuado aos doentes.

Por fim, pode-se dizer que a análise de correspondência ajudou a identificar quais eram os termos que identificavam as técnicas para a realização do raio-x e as palavras utilizadas para fazer o diagnóstico do raio-x.

### *F) Escala Multidimensional de Termos*

Na escala multidimensional de termos estão, os resultados estão demonstrados com base nos seus padrões de aparência, onde os termos que aparecem mais perto uns dos outros são os que tem padrões de aparência similar, e podem significar, por exemplo, diagnósticos efetuados aos doentes.

Os resultados obtidos nesta análise são praticamente iguais aos resultados obtidos pela Análise Hierárquica de *Clusters*, porém, esta análise não tem o nível tao detalhado como a Análise Hierárquica de *Clusters*.

Esta análise foi repartida em três análises, tendo como base a frequência de termos presentes no documento, onde o objetivo é obter visões gerais e mais específicas do documento. Assim, a discussão fica também dividida em três análises e no final é realizada uma crítica aos resultados obtidos.

**1) Discussão da análise com termos que aparecem pelo menos vinte vezes no documento:**

*Clusters* criados por esta análise:

- *Cluster 1* – {normal, ser, hidrocefalia, espaço};
- *Cluster 2* – {alargamento, via, circulação, *líquor*, volume};
- *Cluster 3* – {encefálico, traduzir};
- *Cluster 4* – {corte, axial, espessura, posterior};
- *Cluster 5* – {esquerda, direita};
- *Cluster 6* – {enfarte, hipodensidade};
- *Cluster 7* – {densidade, parênquima, não};
- *Cluster 8* – {alteração, observar, lesão, isquêmico, cerebral}.

Esta análise já contém os *clusters* que identificam os tipos de diagnósticos efetuados aos doentes, como os *Clusters 2* e *8*, embora também hajam *clusters* que identifiquem as técnicas utilizadas na realização do raio-x, como o *cluster 4*. Existem também *clusters* que não contém informação relevante no que toca aos diagnósticos, como o *cluster 3* e o *4*.

**2) Discussão da análise com termos que aparecem pelo menos quinze vezes no documento:**

*Clusters* criados por esta análise:

- *Cluster 1* – {técnica, estudo, espessura, ir, corte, axial, contraste, supratentorial, posterior};
- *Cluster 2* – {encefálico, sugerir, discreto};
- *Cluster 3* – {calcificação, normal, ser};
- *Cluster 4* – {imagem, densidade};
- *Cluster 5* – {não, parênquima, alteração, lesão, observar};
- *Cluster 6* – {cerebral, isquêmico, hipodensidade, branco, enfarte, núcleo};
- *Cluster 7* – {esquerdo, traduzir, direita};
- *Cluster 8* – {hidrocefalia, *líquor*, via, volume, circulação, alargamento, domínio, espaço}.

Neste ponto, já é possível ver com mais detalhe os diagnósticos efetuados aos doentes, bem como a existência de mais diagnósticos realizados. Os *clusters 3, 5, 6 e 8* já contém

informação relevante sobre os diagnósticos efetuados aos doentes. Já pelo contrário, os *clusters* 1, 2, 4, e 7 não contém informação relevante, pois ou contém termos bastante vagos, ou termos que descrevem as técnicas utilizadas nos raio-x.

### 3) Discussão da análise com termos que aparecem pelo menos dez vezes no documento:

*Clusters* criados por esta análise:

- *Cluster* 1 – {hipodensidades, capsula, lacunar, antigo, esquerda, núcleo, enfarte, substancia, direita};
- *Cluster* 2 – {direito, medio, exame, expansivo, observar, cumprimento, coleção};
- *Cluster* 3 – {hipodensidade, branco, lesão, anterior, interno, esquerdo, cerebral. isquémico, discreto, encefálico, não, normal, parênquima, hemorrágico};
- *Cluster* 4 – {m, espessura, técnica, estudo, ir, compartimento, fossa, corte, contraste, axial, posterior, supratentorial, imagem};
- *Cluster* 5 – {tecido, seio, nervoso, densidade, morfologia, dimensão, sugerir, apresentar};
- *Cluster* 6 – {redução, base, mural, hidrocefalia, sifão, calcificação, cisterna, sinais, etário, ser};
- *Cluster* 7 – {haver, alargamento, predomínio, global, líquido, via, circulação, espaço, volume};
- *Cluster* 8 – {subcortical, leucoencefalopatia, artéria, traduzir, temporal, sulco, atrofia}.

Por fim, nesta análise já se conseguem identificar vários tipos de diagnósticos efetuados aos doentes. Os *clusters* 1, 2, 3, 5, 6, 7 e 8 já contém vários tipos de diagnósticos em doentes que tiveram morte cerebral. Esta análise, em relação as análises com vinte e quinze termos, a melhor para identificar diagnósticos de doentes com morte cerebral, onde existiu apenas 1 *cluster* que não continha informação relevante para a análise.

Para concluir, esta análise foi bastante útil para identificar tipos de diagnósticos de morte cerebral nos doentes, principalmente na análise com termos que aparecem pelo menos dez vezes no documento, onde se identificaram sete *clusters* com informação relevante para o estudo. Porém, em comparação com a Análise Hierárquica de *Clusters*, esta última permite uma análise mais eficaz dos *clusters* pois ordena os mesmos de maneira hierárquica, sendo mais fácil estruturar uma frase que exista dentro de um *cluster*.

### *G) Frequência de Palavras com Dicionário*

Na frequência de palavras usando o dicionário, o conjunto de palavras que apareceu mais vezes nos relatórios de raio-x foram as palavras que se agrupavam no tema “Negativo”. Após isso, os temas “Encefálico” e “Lesão” aparecem na casa das 50 presenças no documento. Curiosamente, os temas “Alterações” e “Sem Alterações” apresentam-se com 36 presenças no documento cada um.

Por fim, referindo o que já foi dito nas frequências de palavras sem o dicionário, esta análise tem um fim meramente estatístico pois só contabiliza a frequência dos termos no documento, não conseguindo retirar qualquer tipo de diagnóstico efetuado aos doentes ou qualquer tipo de padrão de doentes.

### *H) Análise Hierárquica de Clusters com Dicionário*

*Clusters* criados por esta análise:

- *Cluster 1* – {úncus, hérnia, digitiforme, hernia, ovoide, grosseiro, trepano, trepano parietal, contusão, fronto-polar, desvio, ventricular, estruturas, edema};
- *Cluster 2* – {hipodensa, ponto-mesocefálica, ipsilateral, perfurantes, ramos, basilar, localização, território, artéria, dilatação, lenticulo-capsulares, insular, antigos};
- *Cluster 3* – {área, craniectomia, transpendimária, transudação. tetraventricuçar, transudação primaria};
- *Cluster 4* – {cortical, corticais, lobos temporais, mesiais, importanste, comprometimento};
- *Cluster 5* – {patentes, patente, base, cisterna, hemotímpano, hemossinus, buraco, buraco magno, lobar, global, espaço, hidrocefalia, redução, LCR, volume, aumento, encefálico, simétrico, região, temporal direita, possivelmente, predomínio, atrofia, temporal};
- *Cluster 6* – {sero-hemáticas, ar, pericerebral, craniotomia, esfeno-frontal};
- *Cluster 7* – {exame, anterior, prévio, natureza, natureza vascular, focal, interna, braço, lenticular, ventrículo, ventrículo, maior, total (quantitativo), colapso, tomodensitometricas, periventricular, núcleos cinzentos, núcleo, cerebral, substância, leucoencefalopatia, direita, hipodensidade, esquerda, cerebeloso, coroa, radiata, protuberância, enfarte, lacunar, capsular, parietais, subcorticais, bilaterais};

- *Cluster 8* – {maxilar, crônico, tecidos moles, inflamação, obliteração, infeccioso, seio esfenoidal, câmara, pneumatizado, espessamentos, mucosos, mastoide, hipocelularizadas, ouvidos, médios, moderada, caudado, puntiforme, correspondência, lacuna, occipitais, interpenduncular, cornos, sangue, subaracnoideu, vale silviano, hemático, tumor, menos, hematoma, epicraniano, peri-orbitário, hiperdensidade, tenda cerebelosa, componente, central, hiperdenso, occipitovertebral, charneira, nervosa, amígdalas cerebelosas, amígdalas, cerebelosas, normal, lesão, negativo, sem alterações, densidade, morfologia, parênquima, alterações, septo nasal, septo, nasal, hemorrágicas, captações, captação, ocupar, densitométricas, segmentos, cavernosos, sequela, recente, fronto-basais, externa, fraturas, hemorrágica, evidências, perinasais, permeáveis, suspeita, hemorragia, coleção, extra-axiais, perdas, convexidade, arterioscleróticas, pálidos, calcificação, calcificações, hiperosteose, calote, espessura}.

Os *clusters* criados com base no dicionário referem-se a tipos de diagnósticos efetuados aos doentes. Ao contrário da análise sem dicionário, esta análise contém apenas os dados do dicionário, o que automaticamente filtra as técnicas utilizadas pelo médico para a realização do raio-x, ou seja, os dados existentes são praticamente todos dados referentes ao diagnóstico efetuado no raio-x. Além disto, não foi utilizada a frequência de palavras, por isso todos os termos existentes no relatório foram considerados para a análise, o que permitiu a criação de oito *clusters* com informações detalhadas sobre os diagnósticos efetuados pelos médicos. Logo os oito *clusters* definidos, podem ser considerados como tipos de diagnósticos efetuados aos doentes que tiveram morte cerebral.

#### *1) Mapa Auto Organizacional com Dicionário*

*Clusters* criados por esta análise:

- *Cluster 1* – {craniotomia, coleção, interpenducular, aumento, global, LCR, domínio, lobar, cornos, hemossinus, pericerebral, base, occipitais, hidrocefalia, comprometimento, simétrico, volume, hemotímpano, sero-hemáticas, sangue, componente, patentes, vale silviano, temporal, redução, encefálico, temporal direita, perdas, pálidos, buraco magno, subaracnoideu, menos, tumor, hemático, espessura, cisterna, colapso, fraturas, patente, extra-axiais, densitométricas, morfologia, hemorragia, suspeita, negativo, evidências, captação, captações,



ocupar, lesão, sequela, calote, densidade, central, nervosa, cerebelosas, arterioscleróticas, tenta cerebela, amígdalas cerebelosas, normal segmentos perinasais, amígdalas, hemorrágicas, hemorrágica, parênquima, cavernosos, sem alterações, permeáveis, fronto-basais, calcificações, tomografias, periventricular, capsular, possivelmente, região};

- *Cluster 2* – {cortical, importante, lobos temporais, atrofia, estruturas mesiais, corticais, espaço};
- *Cluster 3* – {ovoide, grosseiro, trepano, trepano parietal, hérnia, úncus, digitiforme, maior, hérnia, exame, buraco, hipodensa, desvio, moderada, médios};
- *Cluster 4* – {total(quantitativo), infeccioso, crônico, mastoide, hipocelularizadas, ouvidos, espessamentos, mucosos, pneumatizado, câmara, maxilar, tecidos moles, obliteração, seio esfenoidal, inflamação};
- *Cluster 5* – {contusão, convexidade, septo, nasal, septo nasal, fronto-polar, hematoma, epicraniano, peri-orbitário, direita, esquerda};
- *Cluster 6* – {transependimária, transudação transependimária, tetraventricular, transudação, alterações, hiperdenso, área, cerebral, craniectomia, ipsilateral, perfurantes, ramos, localização, basilar, território, dilatação};
- *Cluster 7* – {anterior, interna, recente, prévio, recente, natureza vascular, externa, natureza, leucoencefalopatia, focal, braço, núcleos cinzentos, puntiforme, correspondência, caudado, lacuna, núcleo};
- *Cluster 8* – {lenticular, lacunar, enfarte, ventrículo, cerebeloso, coroa, radiata, insular, antigos, hipodensidade, bilaterais, substância, parietais, subcorticais}.

O Mapa Auto Organizacional com dicionário faz uma demonstração gráfica dos dados existentes no relatório, dividindo-os por cada tipo de dados, neste caso, a divisão é feita por tipo de doente que teve morte cerebral.

Os *clusters* criados com base no dicionário referem-se a tipos de diagnósticos efetuados aos doentes. Ao contrário da análise sem dicionário, esta análise contém apenas os dados do dicionário, o que automaticamente filtra as técnicas utilizadas pelo médico para a realização do raio-x, ou seja, os dados existentes são praticamente todos dados referentes ao diagnóstico efetuado no raio-x. Além disto, não foi utilizada a frequência de palavras, por isso todos os termos existentes no relatório foram considerados para a análise, o que permitiu a criação de oito *clusters*

com informações detalhadas sobre os diagnósticos efetuados pelos médicos. Logo os oito *clusters* definidos, podem ser considerados como tipos de doentes que tiveram morte cerebral.

#### *J) Coocorrência de Rede com Dicionário*

A Coocorrência de redes com dicionário teve variados grupos de palavras que com bastantes termos que se relacionam entre si, assim como grupos de palavras sem qualquer relação com outros termos. As frases abaixo servem de exemplos aos resultados obtidos por esta análise:

- Suspeita de hemorragia prévia.
- Sem alterações na densidade do parênquima.
- Aumento do volume do LCR.
- Captações Hemorrágicas.
- Espessamentos mucosos externos fronto-basais.

Esta análise é bastante útil para ver que termos se relacionam entre si, o que pode servir para identificar diagnósticos detalhados, assim como ver quais são os diferentes tipos de diagnósticos efetuados pelos médicos. Ao contrário da análise sem dicionário, esta análise contém apenas os dados do dicionário, o que automaticamente filtra as técnicas utilizadas pelo médico para a realização do raio-x, ou seja, os dados existentes são praticamente todos dados referentes ao diagnóstico efetuado no raio-x. Além disto, não foi utilizada a frequência de palavras, por isso todos os termos existentes no relatório foram considerados para a análise, o que permitiu a criação de oito *clusters* com informações detalhadas sobre os diagnósticos efetuados pelos médicos.

#### *K) Análise de Correspondência com Dicionário*

Esta análise não teve resultados assinaláveis, pois como o dicionário apenas contém termos que existem no relatório de raio-x, estes termos apareceram todos praticamente na mesma posição do gráfico, que indica isso mesmo, que estão todos no relatório de raio-x. Posto isto, esta análise com a utilização do dicionário não teve os resultados esperados para este estudo.

## L) Escala Multidimensional de Códigos

*Clusters* criados por esta análise:

- *Cluster 1* – {Buraco Magno, normal, cisterna, redução, patentes, encefálico, LCR, cortical, base, buraco, espaço, normal};
- *Cluster 2* – {ocupar, morfologia, focal, sem alterações, parênquima, recente, hemorrágica, evidencias, fraturas, alterações, negativo, ventricular, captações, núcleos cinzentos, densidade, sequela, natureza, núcleo, lesão, lenticular, cerebral, enfarte};
- *Cluster 3* – {antigos, lacunar, insular, território, bilaterais, basilar, lenticulo-capsulares, área, artéria, hipodensa, dilatação, tetraventricular, estruturas, extra-axiais, contusão, desvio, edema, arterioscleróticas};
- *Cluster 4* – {hidrocefalia, hemotímpano, calote, lobos temporais, lobar, importante, craniotomia, coleção, suspeita, tumor, interpeduncular, tomografiométricas, cerebelosas, hiperdensidades, frontobasais, tumor};
- *Cluster 5* – {volume, global, predomínio, simétrico, aumento, atrofia, temporal, temporal direita, região, convexidade, possivelmente, perdas, calcificação};
- *Cluster 6* – {natureza vascular, caudado, braço, interna, externa, anterior, prévio, hipodensidade, substância, subcorticais, esquerda, parietais, cerebeloso, ventrículo, trepano, exame, componente, direita};
- *Cluster 7* – {correspondência, sangue, septo, subaracnoideu, hemorragia, tenda cerebelosa, segmentos, pálidos, hiperosteose};
- *Cluster 8* – {camara, seio esfenoidal, obliteração, ouvidos, pneumatizado, espessamentos, médios, periventricular, maior, protuberância, hematoma, central, epicraniano, espessura, capsular, charneira, calcificações, colapso, perinsanais, permeáveis, total (quantitativo), maxila, inflamação, crônico, mastoide, hipocelularizadas, tecidos moles}.

Na escala multidimensional os resultados apresentados estão limitados para garantir a elegibilidade do gráfico, mas estão praticamente todos os temas utilizados no dicionário.

Aqui, em semelhança à Análise Hierárquica de *Clusters* com dicionário, e apesar de não ter o mesmo nível de detalhe como esta, os resultados apresentados identificam tipos de diagnósticos efetuados pelos médicos. Todos os *clusters* contêm informação bastante relevante e

podem ser considerados para a criação de padrões de diagnóstico em doentes que tiveram morte cerebral.

Ao contrário da análise sem dicionário, esta análise contém apenas os dados do dicionário, o que automaticamente filtra as técnicas utilizadas pelo médico para a realização do raio-x, ou seja, os dados existentes são praticamente todos dados referentes ao diagnóstico efetuado no raio-x.

Os resultados obtidos foram bastante satisfatórios, pois os resultados podem ser considerados para a criação de padrões de diagnóstico em doentes que tiveram morte cerebral, assim como a criação de tipos de doentes que tiveram morte cerebral. Estes resultados só são obtidos na sua maioria nas análises mais detalhadas, onde os termos aparecem pelo menos dez vezes no documento, ou com a utilização do dicionário. Também podem ser identificadas as técnicas utilizadas para a realização de raio-x, como pode ser verificado na Análise de correspondência.

Para concluir este capítulo, resta fazer uma comparação entre as técnicas utilizadas com a utilização do dicionário e sem a utilização deste. Podemos afirmar que se podem fazer análises mais objetivas com o dicionário, pois este método dá a oportunidade de utilizar um dicionário que seja mais orientado aos objetivos que se quer, neste caso um tipo específico de descoberta de informação ou padrões. A análise sem dicionário apenas permite a realização de análises sobre todo o documento, mudando apenas o tipo de frequência das suas palavras, onde se conseguem realizar análises mais gerais ou mais específicas sobre o documento, isto é, não permite a obtenção de informação de um tipo de informação do documento, mas sim de toda a informação do mesmo.

## 9.2. Resultados Obtidos pelos Modelos de Previsão do KNIME

Os modelos de previsão criados pelo KNIME tiveram resultados bastante satisfatórios. A discussão dos resultados foi dividida em quatro partes, uma por cada algoritmo utilizado, e em cada uma dessas partes foram identificados os modelos com melhores valores na Especificidade, Sensibilidade, Acuidade e no erro. No final das análises foi feita uma crítica geral sobre os resultados obtidos.

### *A) Resultados obtidos com a utilização do algoritmo Decision Tree*

Com o algoritmo *Decision Tree* (DT) os resultados foram satisfatórios, mas não existiu nenhum modelo que conseguisse cumprir os quatro requisitos mínimos definidos para garantir a viabilidade dos modelos. O melhor resultado na sensibilidade verificou-se no modelo com o cenário 1 e com a utilização de *oversampling*, com cerca de 91% de sensibilidade. Na Especificidade o máximo alcançado foi de 76,8% utilizando o cenário 2 com a utilização de *oversampling*. Os resultados na acuidade e no erro nunca chegaram aos 85% e aos 15% respectivamente, estando os melhores valores a rondar os 83,7% e os 16,3% com a utilização do cenário 2 com *oversampling*.

Por fim, sabendo que nenhum dos modelos com o algoritmo DT alcançou os resultados mínimos estipulados, é importante referir que os modelos que utilizaram os dados provenientes do *oversampling* tiveram resultados francamente superiores, principalmente no campo da sensibilidade, do que os modelos que não utilizaram *oversampling*.

#### *B) Resultados obtidos com a utilização do algoritmo K-Nearest Neighbor*

Com o algoritmo *K-Nearest Neighbor* (KNN) os resultados têm valores que variam nos 98,3% na especificidade e os 97,3% na Sensibilidade. Este algoritmo teve um modelo que alcançou os resultados mínimos definidos, para garantir a viabilidade dos mesmos, esse modelo utiliza o cenário 5 e o *oversampling*. Neste modelo os valores da Sensibilidade rondam os 97,3%, na especificidade os 75,3%, na Acuidade os 86,3% e no erro 13,7%.

Para concluir, os resultados com o KNN são bastante bons no campo da Especificidade sem a utilização de *oversampling*, mas os resultados são bastante negativos na sensibilidade com valores a chegar aos 3% e nunca passando os 4%. Com *oversampling* os resultados já ficaram mais equilibrados rondando sempre a casa dos 97% na sensibilidade, e sempre acima dos 70% na especificidade. Apesar dos valores altos, apenas um desses modelos cumpriu os requisitos estabelecidos.

#### *C) Resultados obtidos com a utilização do algoritmo Decision Tree com cross validation*

Os resultados com o algoritmo *Decision Tree* com *cross validation* (DTCV) são bastante semelhantes aos resultados obtidos pelo algoritmo DT, apenas melhorando alguns pontos os resultados obtidos, e com isso com a utilização do DTCV, sete modelos atingiram os resultados obtidos e são considerados para a utilização em ambiente real.

Esses modelos alcançaram o resultado com a utilização do *oversampling* em todos os cenários criados. Isto quer dizer que os algoritmos DTCV alcançou sempre os resultados mínimos com a utilização de *oversampling*. Os melhores resultados destes modelos tiveram percentagens a rondar os 98,7% na sensibilidade, os 79% na especificidade, os 88% na acuidade e os 11% no erro. Sem a utilização de *oversampling*, os valores não ultrapassavam os 27% na sensibilidade, e os 73% na acuidade.

*D) Resultados obtidos com a utilização do algoritmo K-Nearest Neighbor com cross validation*

Os resultados obtidos pelo algoritmo *K-Nearest Neighbor* com *cross validation* (KNNCV) foram bastante bons na especificidade quando os modelos não utilizavam o *oversampling*, com valores a rondar os 97%, mas quando se utilizava o *oversampling*, estes valores desciam para a casa dos 70%. No campo da Sensibilidade os resultados foram excelentes quando a utilização dos modelos com o *oversampling*, com os resultados a rondar os 94%, mas os modelos sem a utilização de *Oversampling* obtiveram valores a rondar os 8% na sensibilidade. Posto isto, nenhum dos modelos do KNNCV alcançaram os valores mínimos estabelecidos e por isso não irão ser considerados para a utilização dos mesmos num ambiente real.

# 10. Conclusão

Neste capítulo é realizada a conclusão desta dissertação. Este capítulo contém as considerações finais sobre a dissertação, as dificuldades encontradas para a realização do mesmo, e trabalho que poderá ser feito a partir do que foi desenvolvido neste projeto.

## 10.1. Considerações finais

Os dados fornecidos pelo Centro Hospitalar do Porto (CHP) – Hospital de Santo António continham dados reais e com informações sobre os doentes que realizaram raio-x e faleceram, ou não, no Hospital de Santo António, mas apesar disso a confidencialidade dos dados foi garantida. As tabelas com os dados pessoais dos doentes foram removidas para não haver problemas de confidencialidade. Apesar dessa remoção das tabelas, os dados não sofreram de tratamento dos mesmos pois o objetivo era analisar os dados em bruto, e só foram removidos os dados que continham valores nulos. Após esse tratamento, os dados foram estudados com as ferramentas selecionadas, e como se pode verificar nos objetivos definidos inicialmente e nos resultados atingidos neste projeto, os resultados finais são bastante positivos, pois, estes resultados confirmam que se pode utilizar o *Text Mining* (TM) e o Processamento de Linguagem Natural (PLN) para obtenção de informação e previsão de acontecimentos utilizando as Notas Clínicas.

Estes resultados só foram atingidos com a utilização de um dicionário. Dicionário esse que foi criado durante o desenvolvimento deste projeto, visto não existirem ferramentas de interpretação e de categorização de termos médicos na língua portuguesa. O dicionário foi criado com base na análise de um *dataset* inicial que continha apenas doentes que morreram após a realização de um raio-x de modo a criar padrões de informação sobre esses doentes.

Por fim, e respondendo à questão de investigação que foi estruturada no início desta dissertação que era a seguinte “*De que forma o processo de decisão clínico poderá beneficiar da introdução do TM e Linguagem natural para a análise de texto livre?*”, o TM e o PLN podem ajudar no processo de decisão clínico tanto na previsão como na padronização de tipos de doentes. Os estudos realizados permitiram a criação de padrões de tipos de doentes que

tiveram morte cerebral, bem como a identificação de diferentes diagnósticos aplicados a esses mesmos doentes, também foi possível identificar quais foram as técnicas utilizadas para a realização do exame aos doentes. Identificaram-se também quais os termos mais utilizados nos relatórios de raio-x, apesar deste último ser uma informação mais virada para a estatística. Por fim, foi possível a criação de modelos de previsão com valores que tornam possível a aplicação dos mesmos num ambiente real.

Para cada um dos objetivos definidos aquando do começo desta dissertação, os resultados foram os seguintes:

- Criação de um sistema que apoie a decisão na Saúde – Foram criados modelos de TM e PLN que ajudaram no apoio na decisão na saúde, tanto na criação de padrões como na previsão de acontecimentos com base na análise ao texto das suas notas clínicas;
- Ajuda na padronização dos tratamentos aos doentes – Com o *KH Coder* foi possível a criação de vários tipos de análise que ajudam a criar diversos padrões, mas principalmente, foram descobertos padrões nos diagnósticos feitos aos doentes que tiveram morte cerebral. Isto pode ajudar a padronização do tratamento feito aos mesmos com o objetivo de evitar a morte cerebral;
- Introdução do TM e do PLN na área da Saúde – Nesta dissertação foi possível verificar que se pode introduzir o TM e o PLN na área da Saúde nomeadamente na padronização de diagnósticos feitos aos mesmos, bem como na previsão de acontecimentos utilizando as notas clínicas para esse efeito;
- Criação de um sistema que utilize TM – Tanto os modelos criados para a obtenção e padronização de informação bem como os modelos de previsão desenvolvidos utilizaram todos o TM, sendo os modelos de previsão aqueles que tiveram uma maior utilização de TM. Esses modelos de previsão tiveram resultados bastante satisfatórios, atingindo os 88% na acuidade, 98% na sensibilidade, e 79% na especificidade, o que torna os modelos válidos para a aplicação dos mesmos num ambiente real;
- Criação de um sistema que faça a análise de texto livre – os modelos de análises criados no *KH Coder* permitem fazer uma análise de texto livre, neste caso as notas clínicas, mas pode ser estendida a outros ramos e a outros tipos de texto sendo a utilização de um dicionário específico, essencial para o sucesso na análise do texto.



Estes resultados só foram atingidos com a utilização das duas metodologias aplicadas nesta dissertação. Com o *Design Science Research* (DSR) e com o apoio do *Cross Industry Standard Process for Data Mining* (CRISP-DM), e seguindo os passos de forma minuciosa, foi possível a criação de um estado da arte que continha toda a informação de todos os conceitos e técnicas abordados neste projeto, que foi bastante útil para a parte prática desta dissertação através do desenvolvimento de vários artefactos que permitem responder ao problema inicialmente identificado. Após o levantamento do estado da arte, foram desenvolvidos os artefactos que contribuíram para a resolução deste projeto. Os artefactos resolvidos foram a criação de modelos de análise de texto com o *KH Coder*, o que permitiu a análise e a descoberta de padrões nas notas clínicas, e por fim, a criação de modelos de previsão de acontecimentos com o KNIME, usando a informação textual das notas clínicas. Estes modelos de previsão permitiram a previsão da morte cerebral nos doentes. Estes foram os artefactos desenvolvidos com a utilização das metodologias escolhidas para a realização do projeto, e estes são a resolução do problema que motivou o nascimento desta dissertação.

Com a aplicação destes resultados, tanto dos modelos de padronização como os modelos de previsão, é possível uma melhoria significativa no processo de tomada de decisão clínica, o que viabiliza a integração do TM e do PLN na área clínica.

## 10.2. Dificuldades Encontradas

Durante a realização deste projeto foram encontradas algumas dificuldades. A primeira das dificuldades existentes foi a pouca informação disponível sobre TM e PLN aplicado na área da medicina. O TM e o PLN ainda são dois métodos relativamente recentes e a sua disseminação pelas várias áreas ainda não começou verdadeiramente. Posto isto, a pesquisa teve de ser realizada sobre aplicação do TM e PLN em várias áreas sem ser a medicina.

Outra das dificuldades foi a dificuldade em encontrar um *software* que permita fazer recolha de informação e padronização da mesma e também criar modelos de previsão sobre essa informação. Alguns *softwares* existentes no mercado eram versões pagas e o preço a pagar pelas mesmas era bastante elevado, por isso foi optado a via por ferramentas *open source* apesar de não existir um *software* que permita realizar as duas tarefas. Apesar da procura dos *softwares* ter sido bem-sucedida estes tinham outro problema, a impossibilidade de interpretar a língua portuguesa de forma nativa e a categorização automática de termos

específicos. Essa categorização teve de ser feita com a utilização de um dicionário criado especificamente para o caso. Após isso existiu um processo de aprendizagem às mesmas, que demorou algumas semanas, mas que apesar disso o projeto não ficou atrasado.

### 10.3. Trabalho Futuro

Para o trabalho futuro, este projeto dá perspectivas bastante positivas sobre o TM e o PLN aplicados à área da medicina e em outras áreas, pois os resultados foram extremamente positivos, e podem dar ideias para outros tipos de investigação na medicina preventiva por exemplo.

A categorização de informação pode ser bastante útil para criar padrões de informação sobre um certo problema e conseguir uma melhor resolução do mesmo problema. Sobre a previsão de acontecimentos, esta pode ser ainda mais útil, pois, no caso desta dissertação, incidiu sobre a morte de doentes ou não, e por isso mesmo pode ser essencial para evitar a morte ou acontecimentos negativos para os doentes. E nos tempos que correm, a previsão de acontecimentos começa a ser bastante utilizada hoje em dia, principalmente pela necessidade de evitar males maiores, mas a conjugação da previsão de eventos com a categorização da informação desses eventos pode levar a uma melhor previsão e gestão de possíveis ocorrências. Neste caso, a previsão de morte cerebral sobre o doente, pode ser evitada se o doente pertencer a um conjunto de informação presente num dos padrões criados.

Por fim aconselha-se a continuação da investigação sobre o TM e PLN tanto na área da medicina, usando outro tipo de Notas Clínicas, como em outras áreas de investigação.

# 11. Referências

- Abney, S. (1996). Part-of-Speech Tagging and Partial Parsing. *Corpus-Based Methods in Language and Speech*, 118–136. <http://doi.org/10.1.1.12.622>
- António Silva, Filipe Portela, Manuel Filipe Santos, António Abelha and José Machado. Predicting Brain Deaths using Text Mining and X-Rays clinical notes. *Lecture Notes in Computer Science (LNCS) - Computational Science and Its Applications - MIKE 2016*. Springer. (2016).
- António Silva, Filipe Portela, Manuel Filipe Santos, José Machado and António Abelha. Towards of automatically detecting Brain Death patterns through Text Mining. *IEEE Conference on Business Informatics - ISA'HEALTH@CBI'2016 - Intelligent Systems and Applications in Healthcare Workshop*. Paris, France. IEEE. (2016).
- Akman, V., & Surav, M. (1996). Steps toward Formalizing Context. *AI Magazine*, 17(3), 55–72. <http://doi.org/10.1.1.48.3474>
- Alanazi, H. O., Jalab, H. A., Alam, G. M., Zaidan, B. B., & Zaidan, A. A. (2010). Securing electronic medical records transmissions over unsecured communications: An overview for better medical governance. *Journal of Medicinal Plants Research*, 4(19), 2059–2074. <http://doi.org/10.5897/JMPR10.325>
- ALPAC. (1966). Language and Machines. *Computers in Translation and Linguistics*. 1–138.
- Arighi, C. N., Roberts, P. M., Agarwal, S., Bhattacharya, S., Cesareni, G., Chatr-Aryamontri, A., ... Wu, C. H. (2011). BioCreative III interactive task: an overview. *BMC Bioinformatics*, 12 Suppl 8(SUPPL. 8), S4. <http://doi.org/10.1186/1471-2105-12-S8-S4>
- Bellman, R. E. (1978). *Artificial intelligence: Can Computers Think?*
- Binnie, C. D., & Prior, P. F. (1994). *Electroencephalography*, 1308–1319.
- Brank, J., Grobelnik, M., & Mladeni, D. (2007). Automatic Evaluation of Ontologies. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing (BioNLP '07)*, 1(June), 193–219. Retrieved from <http://ailab.ijs.si/dunja/papers/semantic/OntEval.pdf>

- Brock, D. (1999). The Role of the public in public policy on the definition of death. In *The Definition of death: contemporary controversies* (pp. 293–307). Baltimore. Retrieved from <https://books.google.pt/books?id=eQ6OzvpZvp8C&pg=PA293&ots=u7loUoIG02&dq=The role of the public in public policy on the definition of death brock&hl=pt-PT&pg=PA293#v=onepage&q=The role of the public in public policy on the definition of death brock&f=>
- Burton-Jones, A., Storey, V. C., Sugumaran, V., & Ahluwalia, P. (2004). A semiotic metrics suite for assessing the quality of ontologies. *Data and Knowledge Engineering*, 55(1), 84–102. <http://doi.org/10.1016/j.datak.2004.11.010>
- Centro Hospitalar do Porto. (2015). Retrieved February 21, 2016, from <http://www.chporto.pt/>
- Cerrito, P., & Cerrito, J. C. (2011). Data and Text Mining the Electronic Medical Record to Improve Care and to Lower Costs. *Data Mining and Predictive Modeling*, 1–20.
- Charniak, E., & McDermott, D. (1985). *Introduction to Artificial intelligence*. Saudi Med J.
- Chauhan Shrihari R, A. D. (2015). A Review on Knowledge Discovery using Text Classification Techniques in Text Mining. *International Journal of Computer Applications*, 111(6), 12–15.
- Chen, E. S., Hripcsak, G., Xu, H., Markatou, M., & Friedman, C. (2008). Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *Journal of the American Medical Informatics Association : JAMIA*, 15(1), 87–98. <http://doi.org/10.1197/jamia.M2401>
- Chinchor, N., Hirschman, L., & Lewis, D. D. (1993). Evaluating message understanding systems: An analysis of the third Message Understanding Conference (MUC-3). *Computational Linguistics*, 19(3), 409–449. Retrieved from <http://portal.acm.org/citation.cfm?id=972488>
- Chumsky, N. (1957). *Syntactic Structures*.
- Coden, A., Savova, G., Sominsky, I., Tanenblatt, M., Masanz, J., Schuler, K., ... de Groen, P. C. (2009). Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *Journal of Biomedical Informatics*, 42(5), 937–49. <http://doi.org/10.1016/j.jbi.2008.12.005>

- Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Information Retrieval*, 6(1), 57–72. <http://doi.org/10.1093/bib/6.1.57>
- Columbia University. (2006). Guidelines for progress notes, 2–4. Retrieved from <http://www.columbia.edu/itc/hs/medical/medicine/GuidelinesforProgressNotes.pdf>
- Cormack, J., Nath, C., Milward, D., Raja, K., & Jonnalagadda, S. R. (2015). Agile text mining for the 2014 i2b2 / UTHealth Cardiac risk factors challenge. *JOURNAL OF BIOMEDICAL INFORMATICS*. <http://doi.org/10.1016/j.jbi.2015.06.030>
- Cowie, J., Lehnert, W., & Wilks, Y. (1996). Information extraction. *Communications of the ACM*, 39(1), 80–91. <http://doi.org/10.1145/234173.234209>
- Eelco, F. M., & Wijdicks. (2007). Brain death worldwide.
- Ehrig, M., & Haase, P. (2005). Similarity for ontologies-a comprehensive framework. ... *Modelling and Ontology ...*, 1509–1518. Retrieved from [http://cms.dke.univie.ac.at/fileadmin/DKEHP/publikationen/metamodell/Proceedings\\_Workshop1\\_PAKM2004.pdf#page=13](http://cms.dke.univie.ac.at/fileadmin/DKEHP/publikationen/metamodell/Proceedings_Workshop1_PAKM2004.pdf#page=13)
- El-Sappagh, S. H., & El-Masri, S. (2014). A distributed clinical decision support system architecture. *Journal of King Saud University - Computer and Information Sciences*, 26(1), 69–78. <http://doi.org/10.1016/j.jksuci.2013.03.005>
- Etzioni, O. (1996). The World Wide Web: quagmire or gold mine? *Communications of the ACM*, 3, 1–15. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=DEC5AF8E9A6950F01C5869B0A0882D3F?doi=10.1.1.53.4031&rep=rep1&type=pdf>
- European Commission. (2012). On content in the Digital Single Market (COM/2012/0789 final). Retrieved from <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2012:0789:FIN>
- Feldman, R., & Dagan, I. (1995). Knowledge Discovery in Textual Databases (KDT). *International Conference on Knowledge Discovery and Data Mining (Kdd)*, 112–117. <http://doi.org/10.1.1.47.7462>
- Feldman, R., Dagan, I., & Hirsh, H. (1998). Mining text using keyword distributions. *Journal of Intelligent Information Systems*, 10(3), 281–300.

<http://doi.org/10.1023/A:1008623632443>

Feldman, R., & Hirsh, H. (1997). Exploiting Background Information in Knowledge Discovery from Text. *Journal of Intelligent Information Systems*, 9(1), 83–97.  
<http://doi.org/10.1145/2536536.2536556>

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook. Imagine*.  
<http://doi.org/10.1017/CBO9780511546914>

Finkel, J. R., & Manning, C. D. (2009). Hierarchical bayesian domain adaptation. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (ACL)*, (June), 602–610.  
<http://doi.org/10.3115/1620754.1620842>

Fox, M. S., Barbuceanu, M., Gruninger, M., & Lin, J. (1998). An Organization Ontology for Enterprise Modelling. *Simulating Organizations: Computational Models of Institutions and Groups*, 131–152. [http://doi.org/10.1016/0166-3615\(95\)00079-8](http://doi.org/10.1016/0166-3615(95)00079-8)

Franklin, D. (2002). Data Miners: New Software Instantly Connects Key Bits Of Data That Once Eluded Teams Of Researchers Time. *Time*. Retrieved from <http://content.time.com/time/magazine/article/0,9171,400017-1,00.html>

Frawley, W. J., Piatetsky-shapiro, G., & Matheus, C. J. (1991). Knowledge Discovery in Databases : An Overview. *AI Magazine*, 13(3), 57–70.  
<http://doi.org/10.1609/aimag.v13i3.1011>

Friedman, C. (2005). Semantic Text Parsing for Patient Records. In *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*. NY: Springer.  
<http://doi.org/10.1007/b135955>

Friedman, C., Kra, P., & Rzhetsky, A. (2002). Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4), 222–235.  
[http://doi.org/10.1016/S1532-0464\(03\)00012-1](http://doi.org/10.1016/S1532-0464(03)00012-1)

Gago, P., Santos, M. F., Silva, Á., Cortez, P., Neves, J., & Gomes, L. (2002). INTCare : A Knowledge Discovery based Intelligent Decision Support System for Intensive Care Medicine.

Gamberger, D., Prcela, M., Jovi, A., Šmuc, T., Candelieri, A., Conforti, D., & Guido, R. (2008).

- MEDICAL KNOWLEDGE REPRESENTATION WITHIN HEARTFAID PLATFORM. In *Proceedings of the First International Conference on Health Informatics* (pp. 307–314). SciTePress - Science and Technology Publications. <http://doi.org/10.5220/0001045203070314>
- Garofalakis, M., & Rastogi, R. (1999). Data mining and the Web: past, present and future. *Proceedings of the 2nd International Workshop on Web Information and Data Management*, 43–47. Retrieved from <http://dl.acm.org/citation.cfm?id=319781>
- Gómez-Pérez, A. (1994). Some ideas and examples to evaluate ontologies. *Knowledge Systems Laboratory, Stanford University*, 299. <http://doi.org/10.1109/CAIA.1995.378808>
- Gómez-Pérez, A. (1996). Towards a framework to verify knowledge sharing technology. *Expert Systems with Applications*, 11, 519–529. [http://doi.org/10.1016/S0957-4174\(96\)00067-X](http://doi.org/10.1016/S0957-4174(96)00067-X)
- Goryachev, S., Kim, H., & Zeng-Treitler, Q. (2008). Identification and extraction of family history information from clinical reports. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, 247–51. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2656021&tool=pmcentrez&rendertype=abstract>
- Greer, D. M., Varelas, P. N., Haque, S., Eelco, F. M., & Wijdicks. (2008). Variability of brain death determination guidelines in leading US neurologic institutions.
- Halkidi, M., & Vazirgiannis, M. (2005). Quality Assessment Approaches in Data Mining. In *Data Mining and Knowledge Discovery Handbook* (pp. 613–640).
- Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*.
- Hazlehurst, B., Frost, H. R., Sittig, D. F., & Stevens, V. J. (2005). MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. *Journal of the American Medical Informatics Association: JAMIA*, 12(5), 517–29. <http://doi.org/10.1197/jamia.M1771>
- Hazlehurst, B., Mullooly, J., Naleway, A., & Crane, B. (2005). Detecting possible vaccination reactions in clinical notes. *AMIA Annual Symposium Proceedings*, 306–10.
- Hearst, M. (1994). Context and Structure in Automated Full-Text Information Access. Retrieved

- from [http://people.ischool.berkeley.edu/~hearst/papers/hearst\\_dissertation\\_1994.pdf](http://people.ischool.berkeley.edu/~hearst/papers/hearst_dissertation_1994.pdf)
- Hearst, M. a. (1999). Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* - (pp. 3–10). Morristown, NJ, USA: Association for Computational Linguistics. <http://doi.org/10.3115/1034678.1034679>
- Hevner, A. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, 19(2), 87–92.
- Hevner, A., March, S., & Park, J. (2004). Design Science in Information Systems Research. *MIS Quarterly*, 28(1), 75–105. <http://doi.org/10.2307/25148625>
- Higuchi, K. (2016). KH Coder 3 Reference Manual.
- Hirschman, L., & Sager, N. (1982). Automatic information formatting of a medical sublanguage. *Sublanguage: Studies of Language in Restricted Semantic Domains*. Retrieved from <papers2://publication/uuid/F8988480-23FE-4877-AD3F-C73B92670D75>
- Hurtado, M. P., Swift, E. K., & Corrigan, M. J. (2001). Crossing the quality chasm. *Oncology (Williston Park, N.Y.)*, 21(5), 620. <http://doi.org/10.1200/JCO.2003.01.044>
- Hutchins, J. (2005). The history of machine translation in a nutshell. Retrieved December, 20(11), 2009. Retrieved from <http://hutchinsweb.me.uk/Nutshell-2005.pdf>
- Ishikiriya, C. S., Miro, D., Francisco, C., & Gomes, S. (2015). Text Mining Business Intelligence : a small sample of what words can say. *Procedia - Procedia Computer Science*, 55(1), 261–267. <http://doi.org/10.1016/j.procs.2015.07.044>
- Jonnagaddala, J., Liaw, S.-T., Ray, P., Kumar, M., Chang, N.-W., & Dai, H.-J. (2015). Coronary artery disease risk assessment from unstructured electronic health records using text mining. *Journal of Biomedical Informatics*. <http://doi.org/10.1016/j.jbi.2015.08.003>
- KH Coder Index Page. (2016). Retrieved from <http://khc.sourceforge.net/en/>
- Kolluru, B., Nakjang, S., Hirt, R. P., Wipat, A., & Ananiadou, S. (2011). Automatic extraction of microorganisms and their habitats from free text using text mining workflows. *Journal of Integrative Bioinformatics*, 8(2), 184. <http://doi.org/10.2390/biecoll-jib-2011-184>
- Kukafka, R., Bales, M. E., Burkhardt, A., & Friedman, C. (2006). Human and automated coding



- of rehabilitation discharge summaries according to the International Classification of Functioning, Disability, and Health. *Journal of the American Medical Informatics Association : JAMIA*, 13(5), 508–15. <http://doi.org/10.1197/jamia.M2107>
- Kurzweil, R. (1990). *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*.
- Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, 8(June), 282–289. <http://doi.org/10.1038/nprot.2006.61>
- Leser, U., & Hakenberg, J. (2005). What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, 6(4), 357–69. <http://doi.org/10.1093/bib/6.4.357>
- Liddy, E. D. (2001). Natural Language Processing. *Annual Review of Applied Linguistics*, 16, 70. <http://doi.org/10.1017/S0267190500001446>
- Lin, S.-H., Shih, C.-S., Chen, M. C., Ho, J.-M., Ko, M.-T., & Huang, Y.-M. (1998). Extracting classification knowledge of Internet documents with mining term associations: A semantic approach. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 241–249. <http://doi.org/10.1145/290941.291001>
- Loh, S., Wives, L. K., & de Oliveira, J. P. M. (2000). Concept-based knowledge discovery in texts extracted from the Web. *ACM SIGKDD Explorations Newsletter*, 2(1), 29–39. <http://doi.org/10.1145/360402.360414>
- Mahmooda, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular diseases: A historical perspective. *Lancet*, 383(9921), 1933–1945. [http://doi.org/10.1016/S0140-6736\(13\)61752-3](http://doi.org/10.1016/S0140-6736(13)61752-3)
- MALLET. (2013). Retrieved from <http://mallet.cs.umass.edu/>
- Matos, S., Arrais, J. P., Maia-Rodrigues, J., & Oliveira, J. L. (2010). Concept-based query expansion for retrieving gene related publications from MEDLINE. *BMC Bioinformatics*, 11, 212. <http://doi.org/10.1186/1471-2105-11-212>
- Mattox, D., Seligman, L., & Smith, K. (1999). Rapper. *Proceedings of the Second International*

- Workshop on Web Information and Data Management - WIDM '99*, (January), 6–11.  
<http://doi.org/10.1145/319759.319766>
- McDonald, D. D., & Pustejovsky, J. D. (1985). Description-Directed Natural Language Generation. *IJCAI'85 Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 2, 2*, 799–805. Retrieved from <http://ijcai.org/PastProceedings/IJCAI-85-VOL2/PDF/023.pdf>
- McDonald, R., & Pereira, F. (2005). Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*, *6 Suppl 1*(Suppl 1), S6.  
<http://doi.org/10.1186/1471-2105-6-S1-S6>
- McKeown, K. (1985). Discourse strategies for generating natural-language text. *Artificial Intelligence*, *27*(1), 1–41. [http://doi.org/10.1016/0004-3702\(85\)90082-7](http://doi.org/10.1016/0004-3702(85)90082-7)
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1993). Five Papers on Wordnet. <http://doi.org/10.1093/ijl/3.4.235>
- Miller, R. H., & Sim, I. (2004). Physicians' Use Of Electronic Medical Records: Barriers And Solutions. *Health Affairs*, *23*(2), 116–126. <http://doi.org/10.1377/hlthaff.23.2.116>
- Nilsson, N. J. (1998). Artificial Intelligence: A New Synthesis. *Artificial Intelligence*.
- Pakhomov, S., Buntrock, J., & Duffy, P. (2005). High throughput modularized NLP system for clinical text. *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions - ACL '05*, 25–28. <http://doi.org/10.3115/1225753.1225760>
- Patel, C., Supekar, K., Lee, Y., & Park, E. (2004). OntoKhoj: A Semantic Web Portal for Ontology Searching, Ranking, and Classification. *Proceedings of the 5th ACM CIKM International Workshop on Web Information and Data Management (WIDM 2004)*, 58–61.  
<http://doi.org/10.1145/956699.956712>
- Penz, J. F. E., Wilcox, A. B., & Hurdle, J. F. (2007). Automated identification of adverse events related to central venous catheters. *Journal of Biomedical Informatics*, *40*(2), 174–182.  
<http://doi.org/10.1016/j.jbi.2006.06.003>
- Perrin, P., & Petry, F. (1998). Contextual text representation for unsupervised knowledge discovery in texts. In *Research and Development in Knowledge Discovery and Data Mining* (p. 12). Springer Berlin Heidelberg.

- Pestian, J. P., Brew, C., Matykiewicz, P., Hovermale, D. J., Johnson, N., Cohen, K. B., & Duch, W. (2007). A Shared Task Involving Multi-label Classification of Clinical Free Text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing (BioNLP '07)*, 1(June), 97–104. Retrieved from <http://www.aclweb.org/anthology/W07-1013>
- Piedra, D., & Ferrer, A. (2014). Text Mining and Medicine : Usefulness in Respiratory Diseases □. *Archivos de Bronconeumología*, 50(3), 113–119. <http://doi.org/10.1016/j.arbr.2014.02.008>
- Poole, D., Mackworth, A., & Goebel, R. (1998). Computational Intelligence. *Computational Intelligence*.
- Popowich, F. (2005). Using Text Mining and Natural Language Processing for Health Care Claims Processing. *ACM SIGKDD Explorations Newsletter*, 7(1), 59–66. <http://doi.org/10.1145/1089815.1089824>
- Porzel, R., & Malaka, R. (2004). A Task-based Approach for Ontology Evaluation, 9–16. Retrieved from <http://olp.dfki.de/ecai04/final-porzel.pdf>
- Raja, U., Mitchell, T., Day, T., & Hardin, J. M. (2008). Text Mining in Healthcare, 22(3). Retrieved from <http://www.himss.org/content/files/Raja.pdf>
- Rebholz-Schuhmann, D., Jimeno Yepes, A., Li, C., Kafkas, S., Lewin, I., Kang, N., ... Hahn, U. (2011). Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. *Journal of Biomedical Semantics*, 2 Suppl 5(Suppl 5), S11. <http://doi.org/10.1186/2041-1480-2-S5-S11>
- Rich, E., & Knight, K. (1991). Artificial Intelligence (3rd Edition).
- Russel, S., & Norvig, P. (2003). *Artificial Intelligence a Modern Approach*. <http://doi.org/10.1017/S0269888900007724>
- Sager, N., Lyman, M., Bucknall, C., Nhan, N., & Tick, L. J. (1994). Natural Language Processing and the Representation of Clinical Data. *Journal of the American Medical Informatics Association*, 1(2), 142–160. <http://doi.org/10.1136/jamia.1994.95236145>
- Santos, D. (2001). Introdução ao processamento de linguagem natural através das aplicações. In E. Ranchhod (Ed.), *Tratamento das Línguas por Computador. Uma introdução à*

*linguística computacional e suas aplicações* (pp. 229–259). Lisboa: Caminho.

Santos, M. F., & Azevedo, C. S. (2005). *Data Mining : Descoberta de Conhecimento em Bases de Dados*. FCA Editores.

Sasaki, Y., Tsuruoka, Y., McNaught, J., & Ananiadou, S. (2008). How to make the most of NE dictionaries in statistical NER. *BMC Bioinformatics*, 9 Suppl 11, S5. <http://doi.org/10.1186/1471-2105-9-S11-S5>

Shu, J. (2005). Free Text Phrase Encoding and Information Extraction from Medical Notes by *Interface*. Retrieved from [http://groups.csail.mit.edu/medg/ftp/jshu/Shu\\_Jennifer\\_thesis.pdf](http://groups.csail.mit.edu/medg/ftp/jshu/Shu_Jennifer_thesis.pdf)

Soderland, S. (1997). Learning to Extract Text-Based Information from the World Wide Web. *Knowledge Discovery and Data Mining*, 251–254. Retrieved from <https://www.aaai.org/Papers/KDD/1997/KDD97-052.pdf>

Spyns, P. (1996). Natural language processing in medicine: An overview. *Methods of Information in Medicine*.

Stetson, P. D., Johnson, S. B., Scotch, M., & Hripcsak, G. (2002). The sublanguage of cross-coverage. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 742–6. <http://doi.org/D020002450> [pii]

Supekar, K. (2005). A peer-review approach for ontology evaluation. *8th Int. Protege Conf*, 77–79. <http://doi.org/citeulike-article-id:300085>

Tan, A.-H. (1999). Text Mining: The state of the art and the challenges. *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, 8, 65–70. <http://doi.org/10.1.1.38.7672>

Truyens, M., & Van Eecke, P. (2014). Legal aspects of text mining. *Computer Law & Security Review*, 30(2), 153–170. <http://doi.org/10.1016/j.clsr.2014.01.009>

Tsuruoka, Y., & Tsujii, J. (2005). Bidirectional inference with the easiest-first strategy for tagging sequence data. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language, HLT/EMNLP*, (October), 467–474. <http://doi.org/10.3115/1220575.1220634>

- Voorhees, E. M., & Hersh, W. (2012). Overview of the TREC 2012 Medical Records Track. *Trec.Nist.Gov*, (lcd). Retrieved from <http://trec.nist.gov/pubs/trec21/papers/MED12OVERVIEW.pdf>
- Weaver, W. (1949). Translation. *MIT Press*, 12. Retrieved from <http://www.mt-archive.info/Weaver-1949.pdf>
- Weikum, G. (2002). Foundations of statistical natural language processing. *ACM SIGMOD Record*, 31(3), 37. <http://doi.org/10.1145/601858.601867>
- Weiss, S.M. Damerau, F. Apte, C. (1998). Text Mining with Decision Trees and Decision Rules , June 1998. *Conference on Automated Learning and Discovery Carnegie-Mellon University*.
- Wijdicks, E. F. M. (1995). Determining brain death in adults, (May).
- Winston, H. (1992). Artificial Intelligence : A Perspective !
- Witten, I., Frank, E., & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3<sup>rd</sup> ed.). Retrieved from <http://www.amazon.com/exec/obidos/ASIN/0123748569/departmentofcompute>
- World Health Organization. (2009). World Health Statistics 2009. Retrieved from [http://encompass.library.cornell.edu/cgi-bin/checkIP.cgi?access=gateway\\_standard&url=http://find.galegroup.com/openurl/openurl?url\\_ver=Z39.88-2004&url\\_ctx\\_fmt=info:ofi/fmt:kev:mtx:ctx&res\\_id=info:sid/gale:AONE&ctx\\_enc=info:ofi:enc:UTF-8&rft\\_val\\_fmt=info](http://encompass.library.cornell.edu/cgi-bin/checkIP.cgi?access=gateway_standard&url=http://find.galegroup.com/openurl/openurl?url_ver=Z39.88-2004&url_ctx_fmt=info:ofi/fmt:kev:mtx:ctx&res_id=info:sid/gale:AONE&ctx_enc=info:ofi:enc:UTF-8&rft_val_fmt=info)
- Yue, X., Di, G., Yu, Y., Wang, W., & Shi, H. (2012). Analysis of the combination of natural language processing and search engine technology. *Procedia Engineering*, 29, 1636–1639. <http://doi.org/10.1016/j.proeng.2012.01.186>
- Zeng, Q. T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S. N., & Lazarus, R. (2006). Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making*, 6, 30. <http://doi.org/10.1186/1472-6947-6-30>
- Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., & Chan, J. (2013). Biomedical text mining

and its applications in cancer research. *Journal of Biomedical Informatics*, 46(2), 200–211. <http://doi.org/10.1016/j.jbi.2012.10.007>

## ANEXOS

### Resultados das Análises Efetuadas no *KH Coder* sem *Stopwords*

Análise sem Dicionário

#### A) Frequência de Palavras

*Tabela A1 – Frequência de Palavras*

Palavras	<i>Part-of-Speech</i>	Frequência
não	R	66
lesão	N	52
encefálico	AQ	50
relatório	N	40
alteração	N	36
espessura	N	30
circulação	N	28
isquêmico	AQ	28
normal	AQ	28
observar	V	28
liquor	N	26
corte	N	24
enfarte	N	24
esquerdo	AQ	24
hipodensidade	N	24
alargamento	N	22
axial	AQ	22
cerebral	AQ	22
direita	N	22
parênquima	N	22

B) Análise Hierárquica de *Clusters*

Análise com termos que aparecem pelo menos vinte vezes no documento:

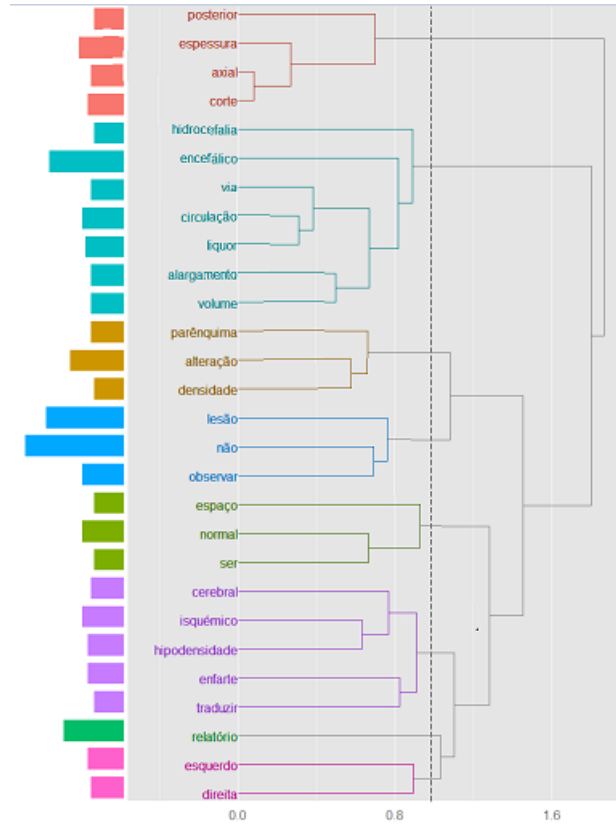


Figura A1 – Análise com termos que aparecem pelo menos vinte vezes no documento



Análise com termos que aparecem pelo menos quinze vezes no documento:

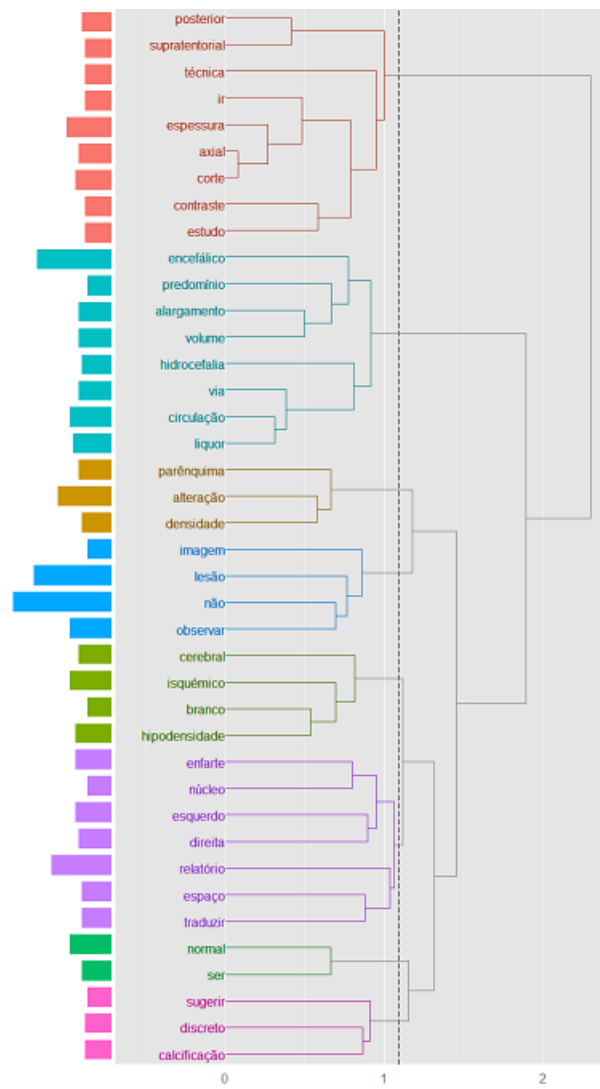


Figura A2 – Análise com termos que aparecem pelo menos quinze vezes no documento

Análise com termos que aparecem pelo menos dez vezes no documento:

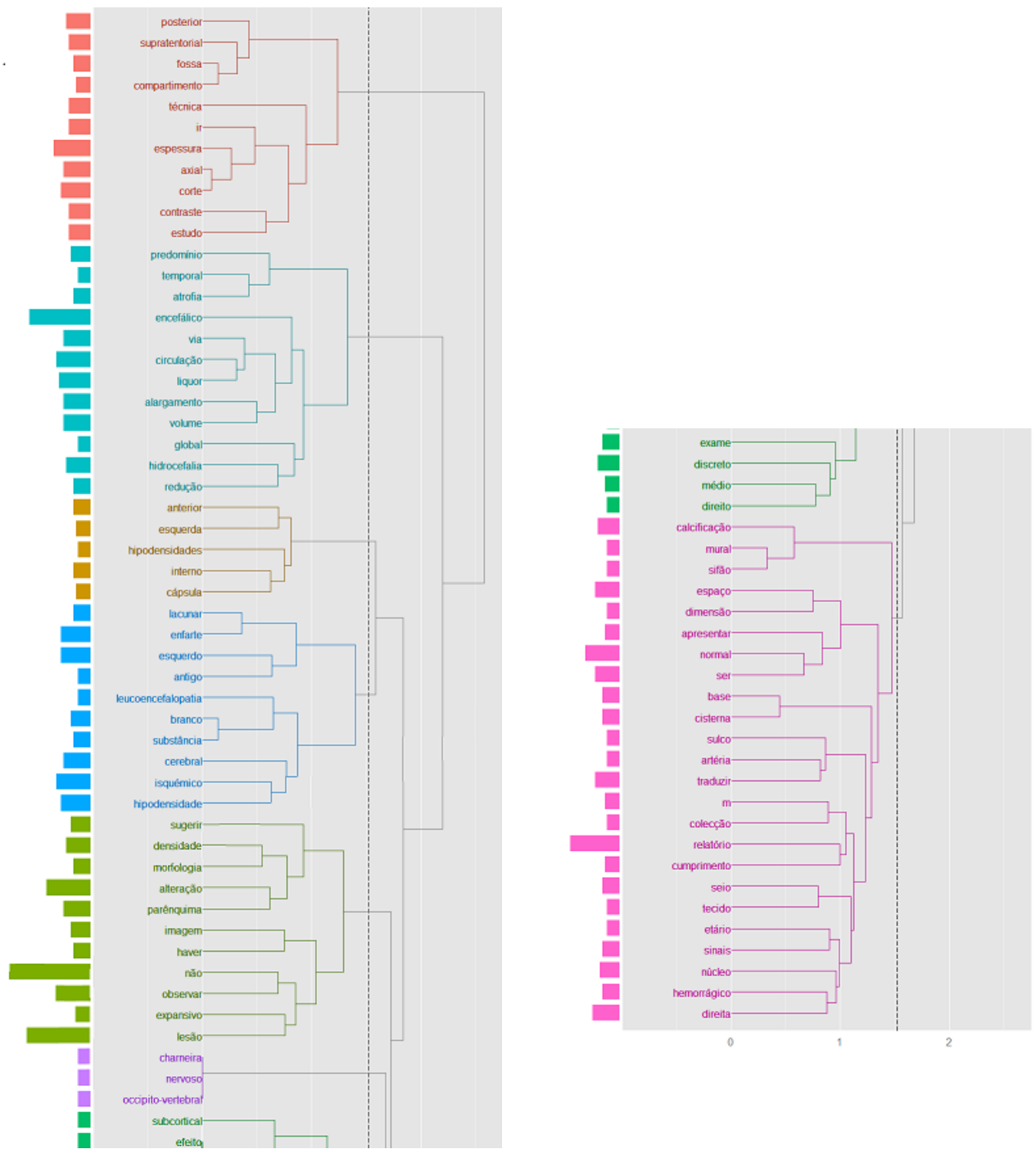


Figura A3 – Análise com termos que aparecem pelo menos dez vezes no documento

C) Mapa Auto Organizacional

Análise com termos que aparecem pelo menos vinte vezes no documento:



Figura A4 – Análise com termos que aparecem pelo menos vinte vezes no documento

Análise com termos que aparecem pelo menos quinze vezes no documento:

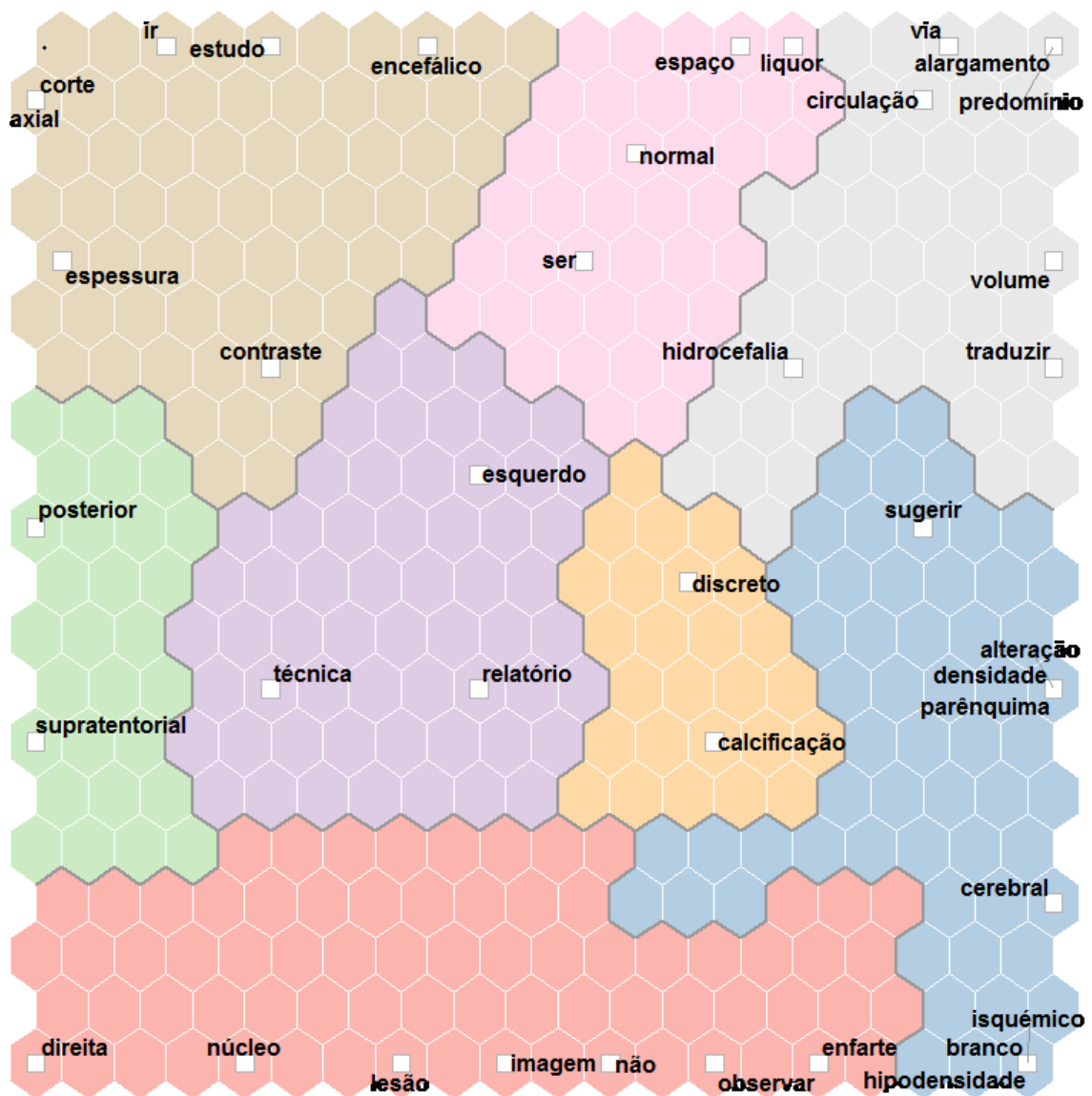


Figura A5 – Análise com termos que aparecem pelo menos quinze vezes no documento

Análise com termos que aparecem pelo menos dez vezes no documento:



Figura A6 – Análise com termos que aparecem pelo menos dez vezes no documento

D) Coocorrência de Rede sem dicionário

Análise com termos que aparecem pelo menos vinte vezes no documento:

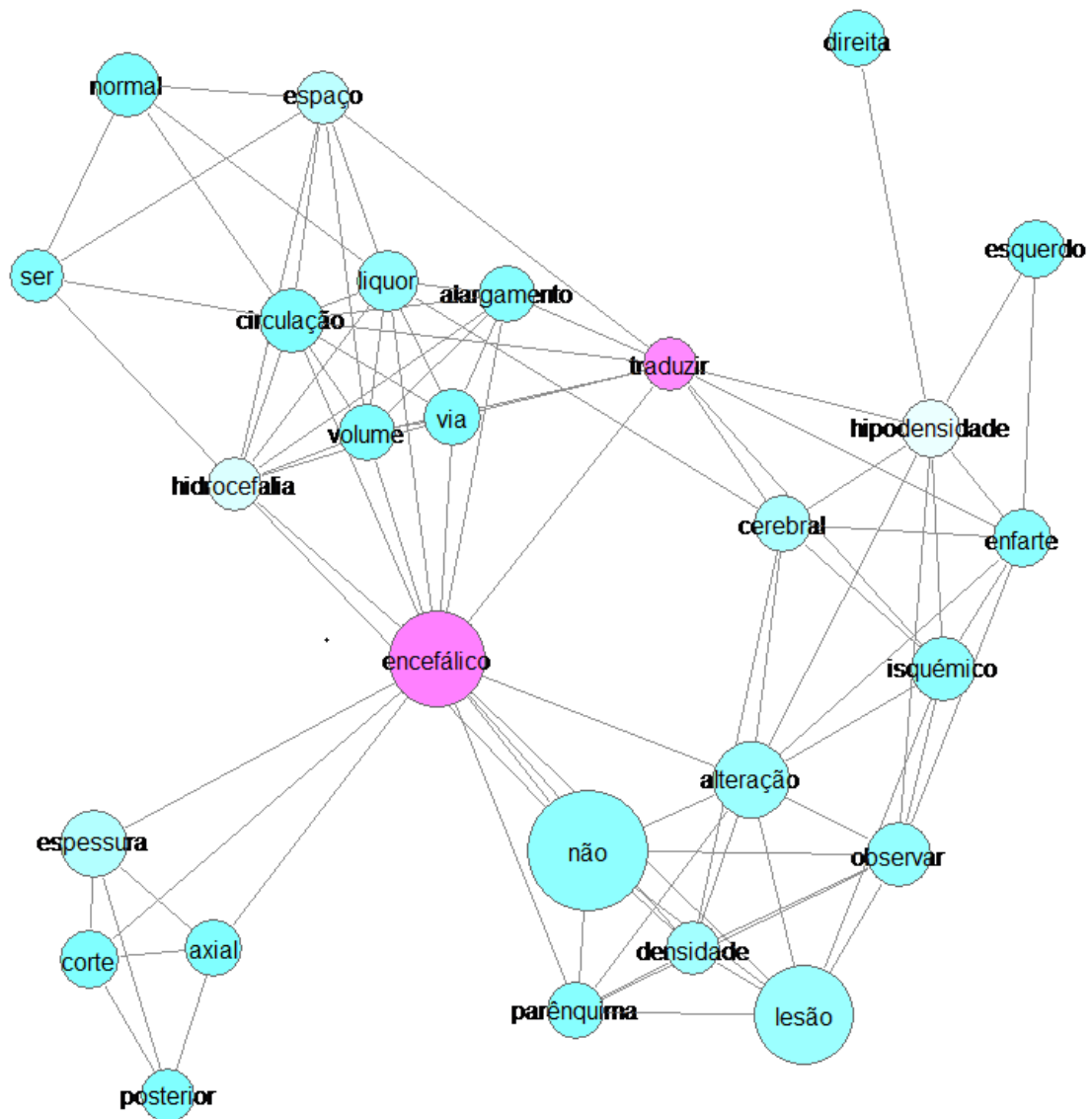


Figura A7 – Análise com termos que aparecem pelo menos vinte vezes no documento

Análise com termos que aparecem pelo menos quinze vezes no documento:

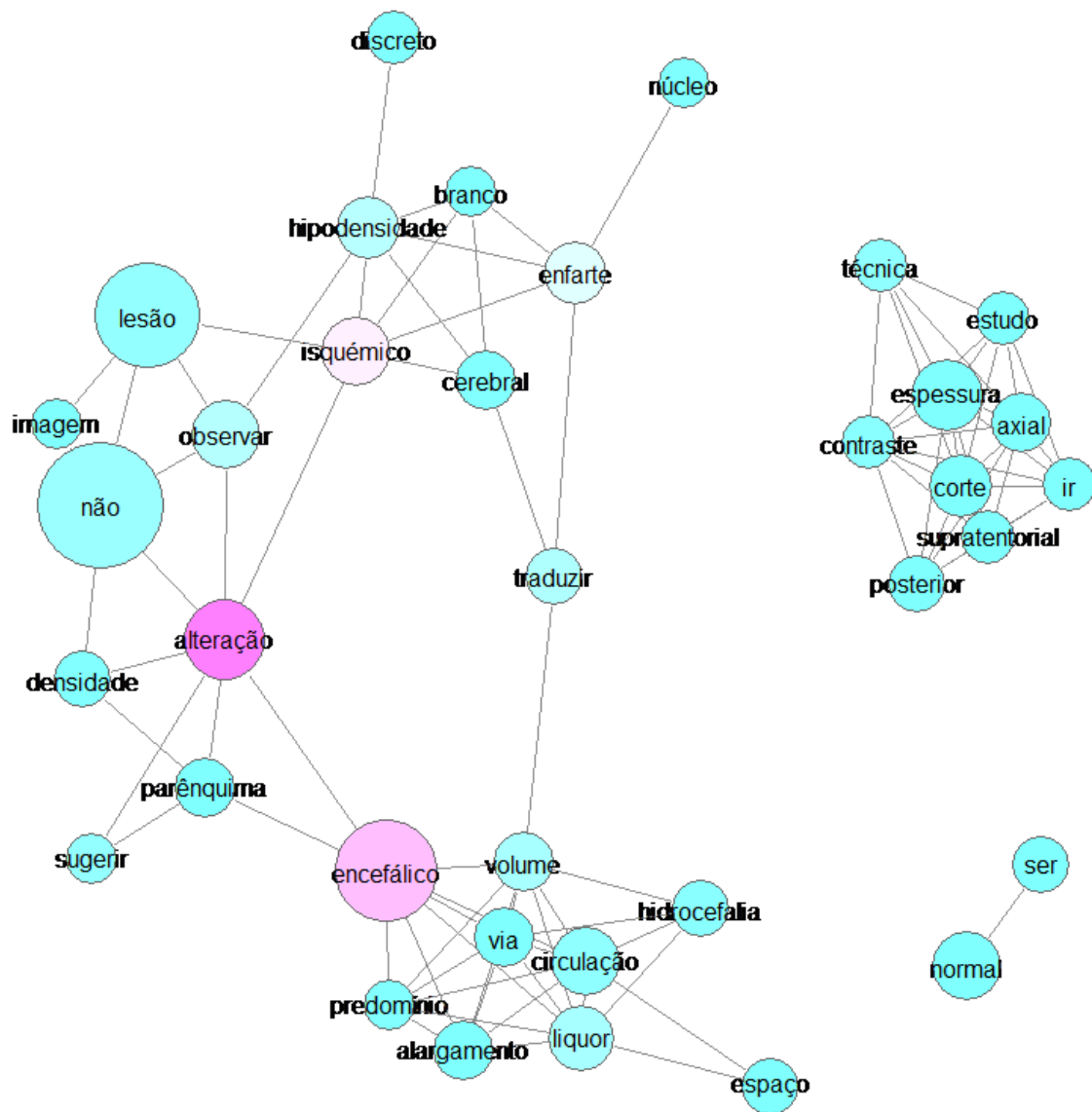


Figura A8 – Análise com termos que aparecem pelo menos quinze vezes no documento

Análise com termos que aparecem pelo menos dez vezes no documento:

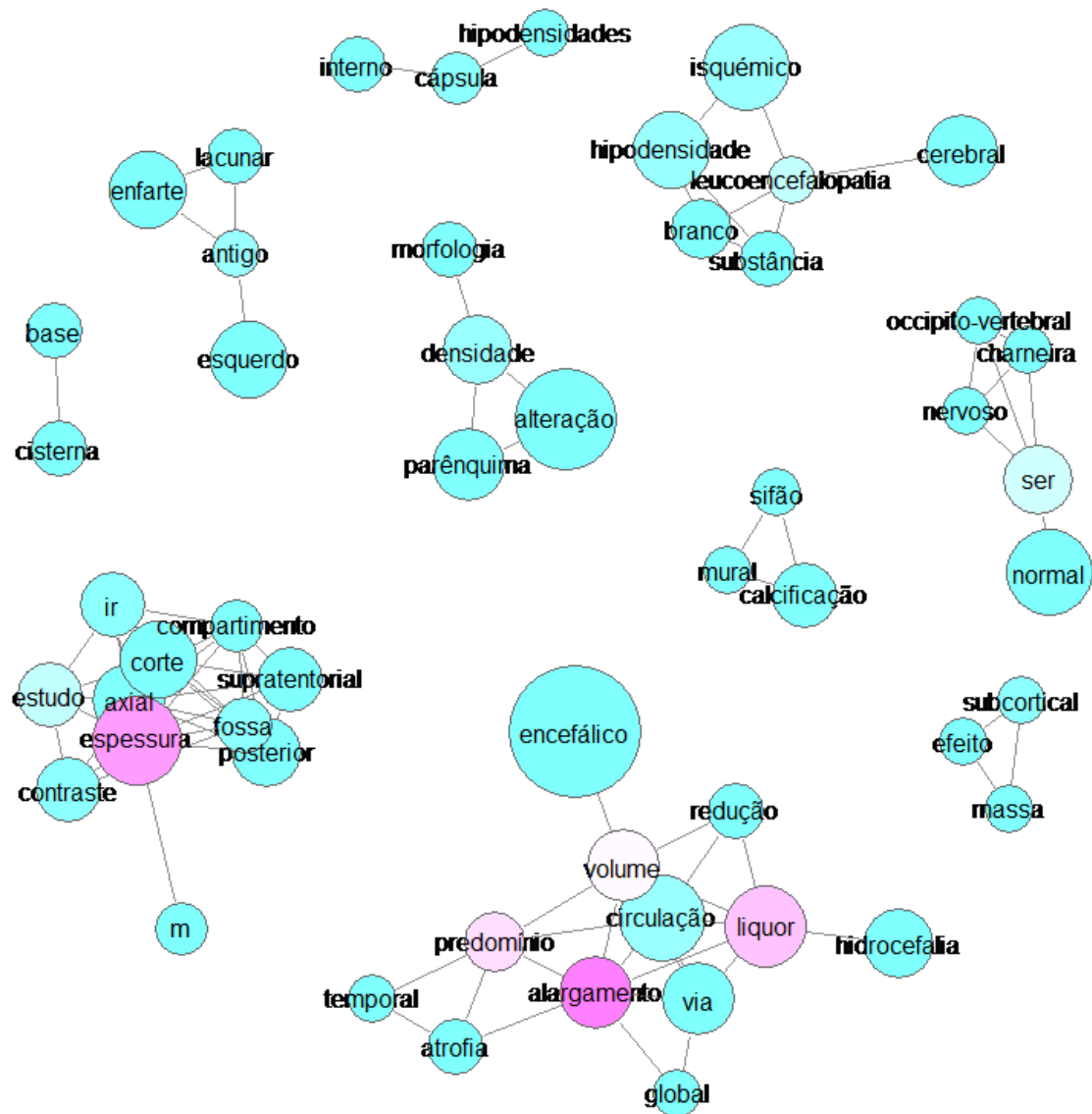


Figura A9 – Análise com termos que aparecem pelo menos dez vezes no documento



E) Análise de Correspondência

Análise com termos que aparecem pelo menos vinte vezes no documento:

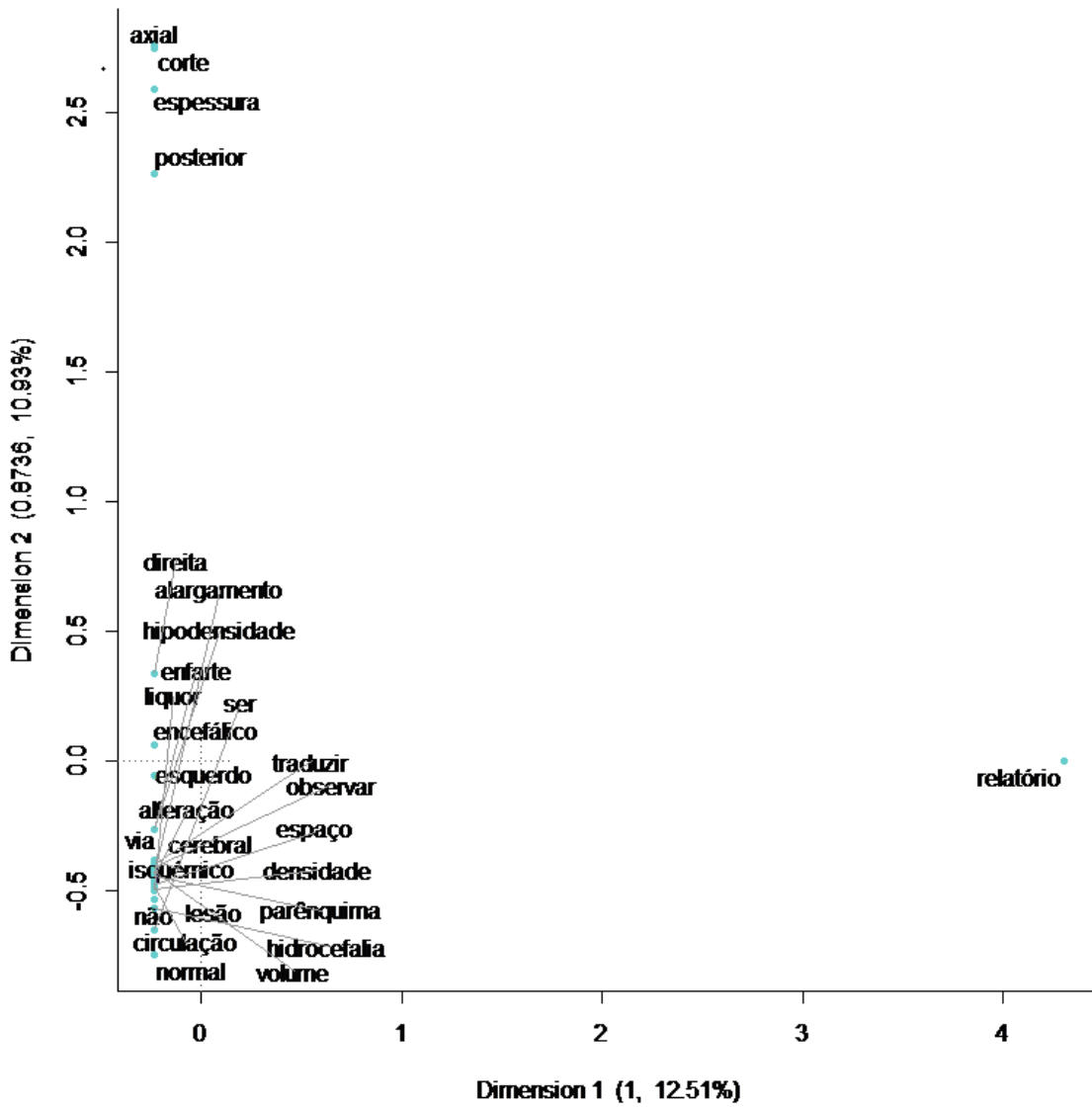


Figura A10 – Análise com termos que aparecem pelo menos dez vezes no documento

Análise com termos que aparecem pelo menos quinze vezes no documento:

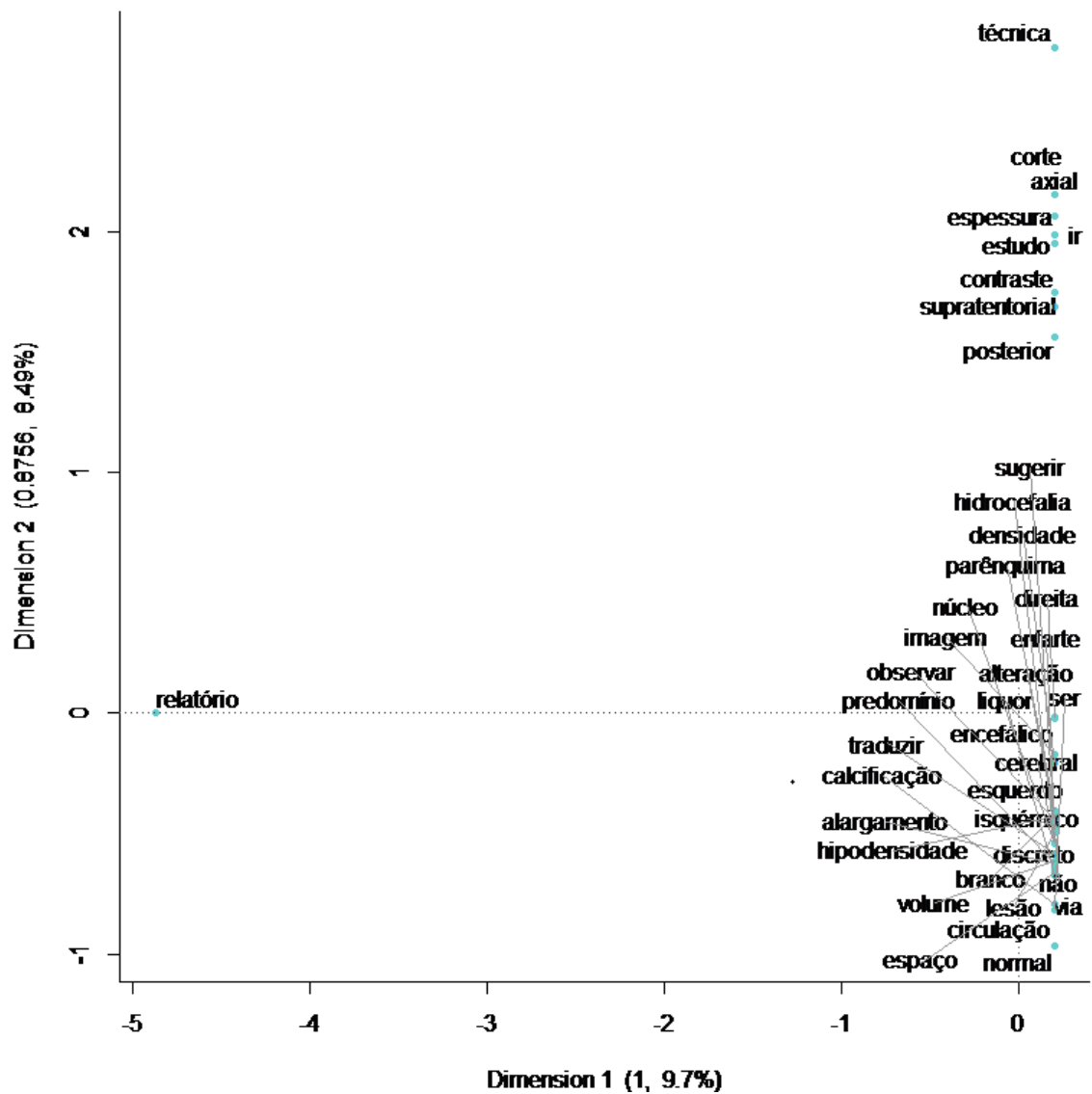


Figura A11 – Análise com termos que aparecem pelo menos quinze vezes no documento

Análise com termos que aparecem pelo menos dez vezes no documento:

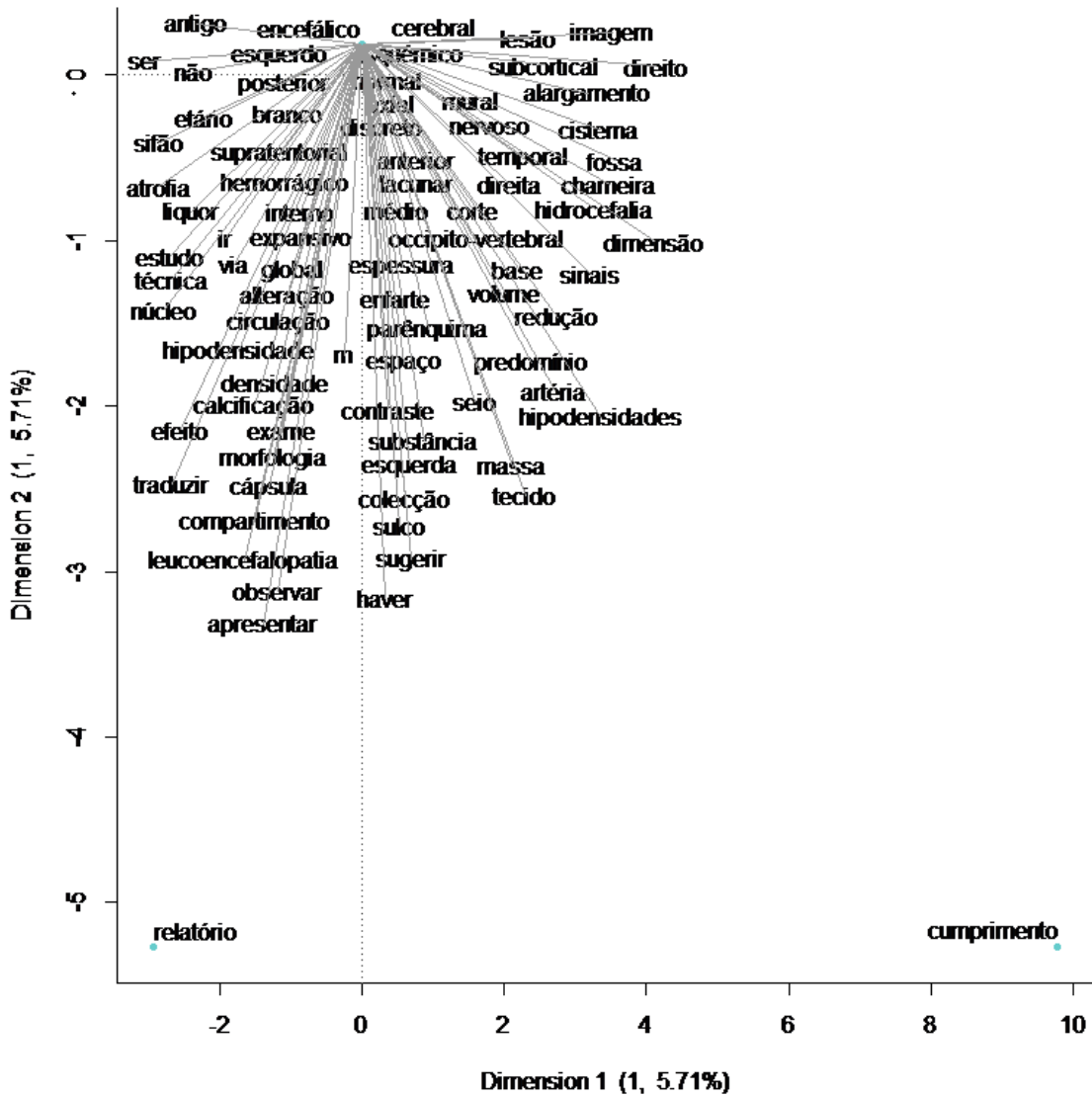


Figura A12 – Análise com termos que aparecem pelo menos dez vezes no documento

F) Escala Multidimensional de Termos

Análise com termos que aparecem pelo menos vinte vezes no documento:

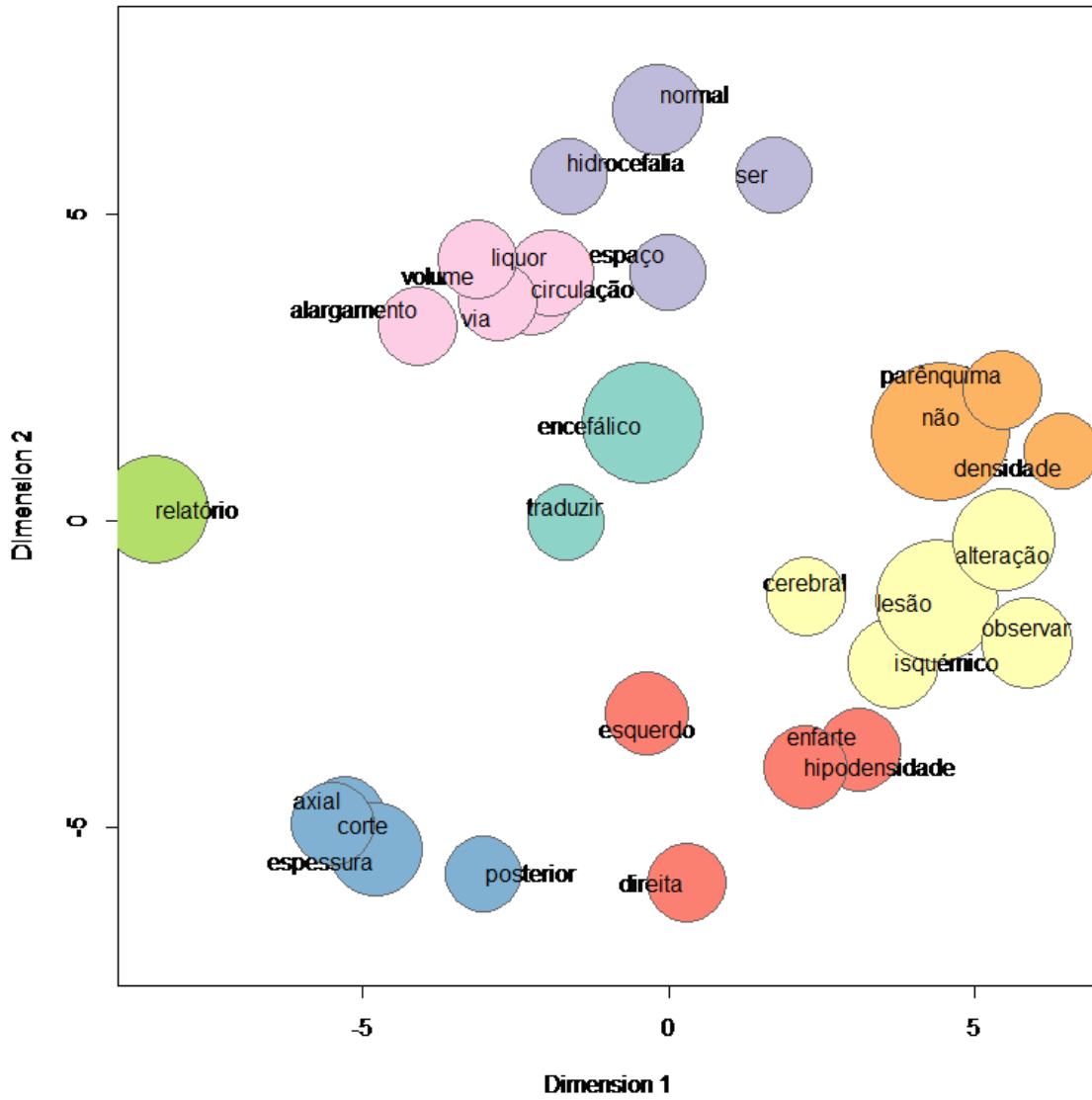


Figura A13 – Análise com termos que aparecem pelo menos vinte vezes no documento

Análise com termos que aparecem pelo menos quinze vezes no documento:

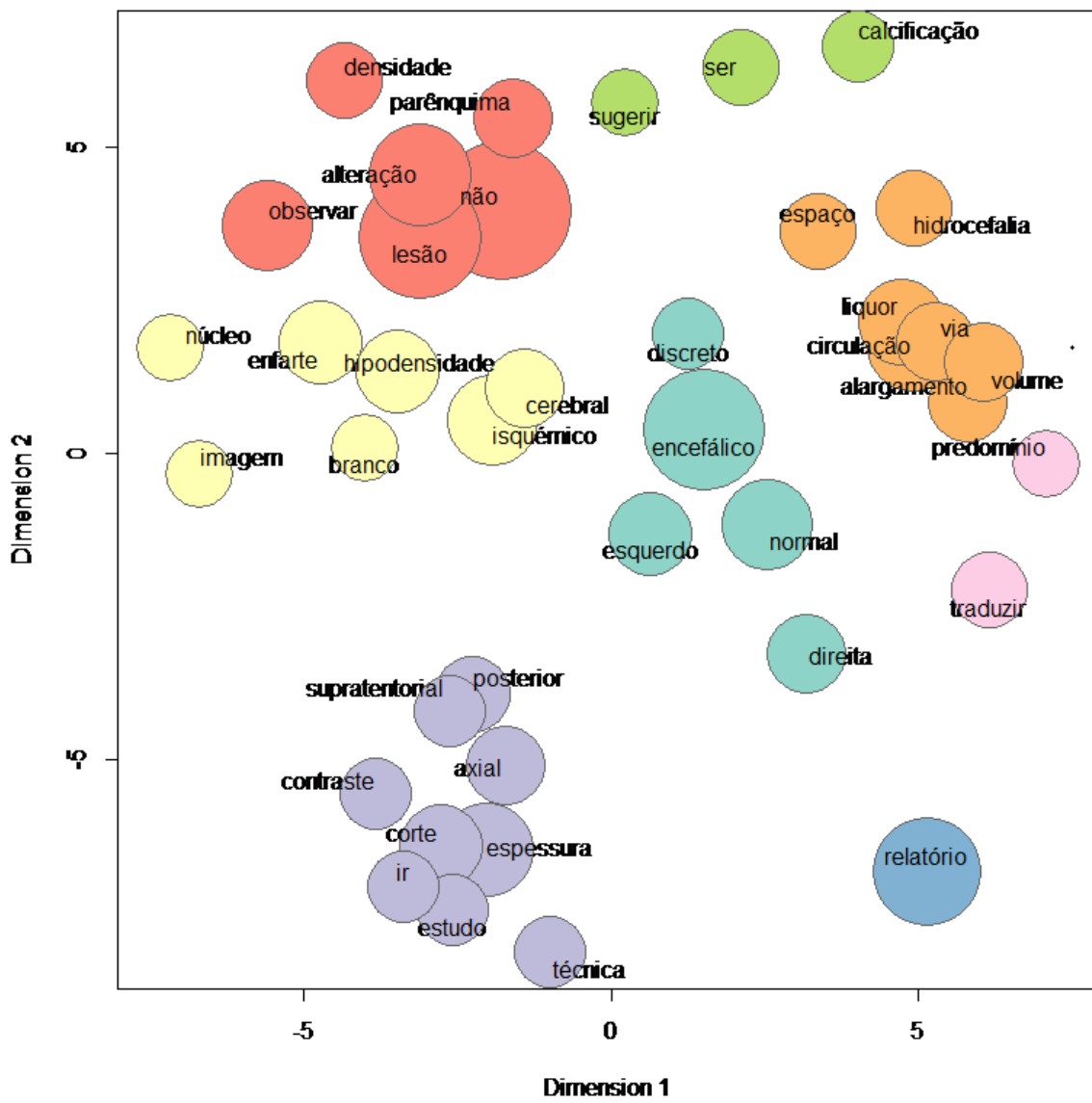


Figura A14 – Análise com termos que aparecem pelo menos quinze vezes no documento

Análise com termos que aparecem pelo menos dez vezes no documento:

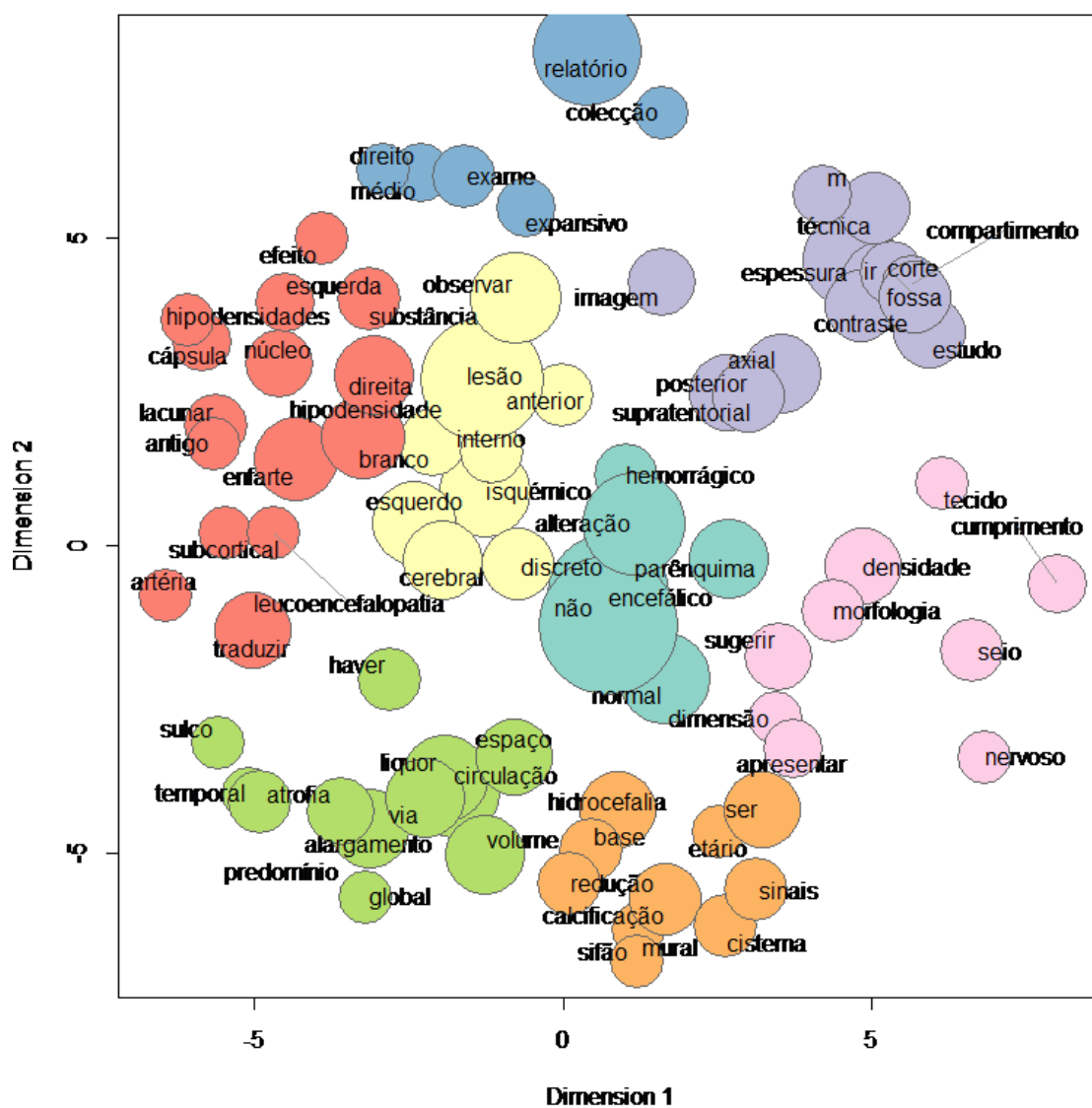


Figura A15 – Análise com termos que aparecem pelo menos dez vezes no documento

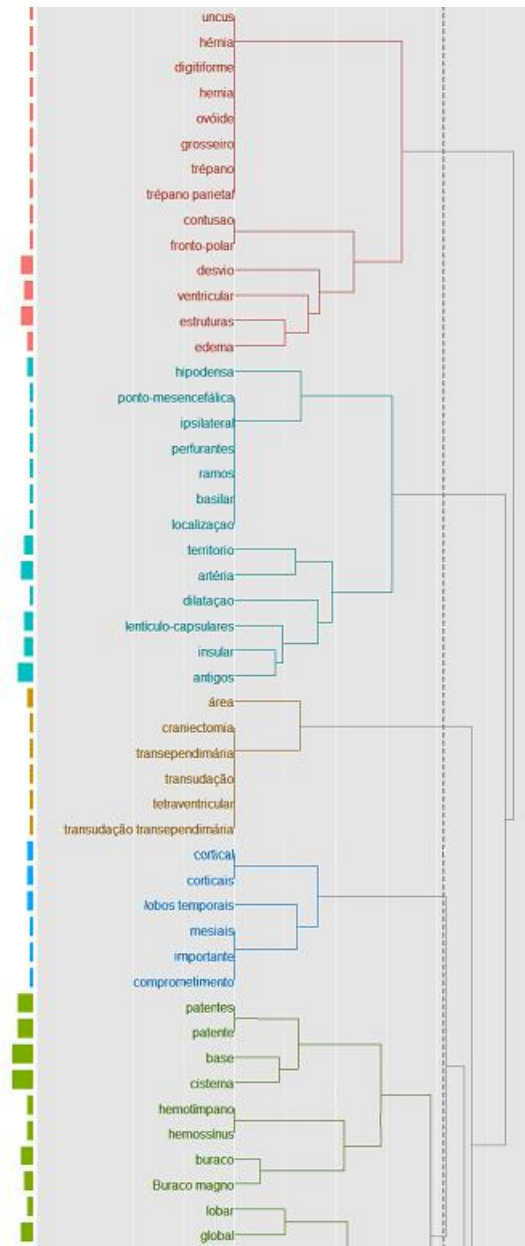
## Análise com Dicionário

### A) Frequência de Palavras com a Utilização do Dicionário

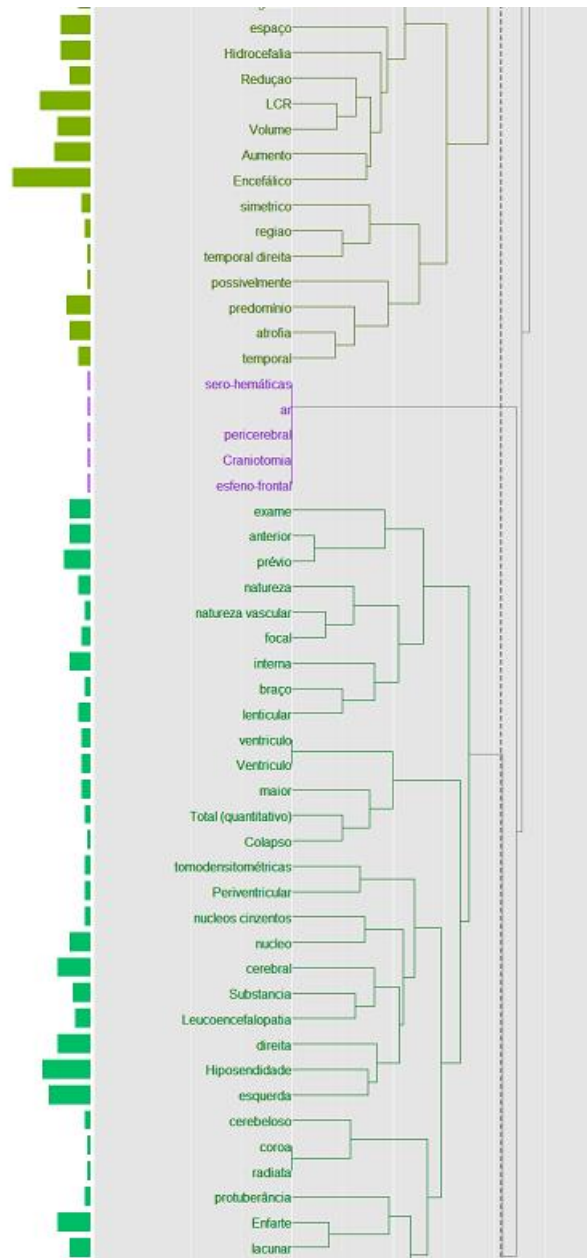
*Tabela A2 – Frequência de Palavras com a Utilização do Dicionário*

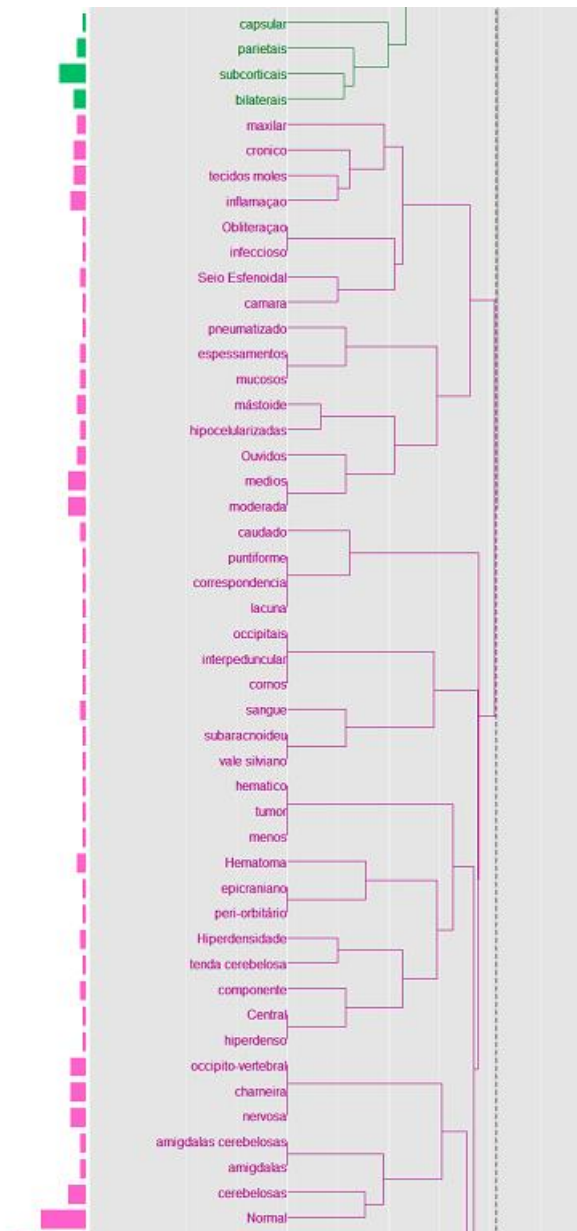
<b>Termo</b>	<b>Frequência</b>	<b>Porcentagem</b>
*Negativo	66	15.24%
*Encefálico	52	12.01%
*lesão	50	11.55%
*Alterações	36	8.31%
*Sem Alterações	36	8.31%
*LCR	34	7.85%
*Hiposensibilidade	32	7.39%
*espessura	30	6.93%
*Normal	30	6.93%
*esquerda	28	6.47%
*Aumento	24	5.54%
*direita	22	5.08%
*Enfarte	22	5.08%
*Volume	22	5.08%
*parênquima	22	5.08%
*cerebral	22	5.08%
*Hidrocefalia	20	4.62%
*espaço	20	4.62%
*prévio	18	4.16%
*subcorticais	18	4.16%

B) Análise Hierárquica de *Clusters*









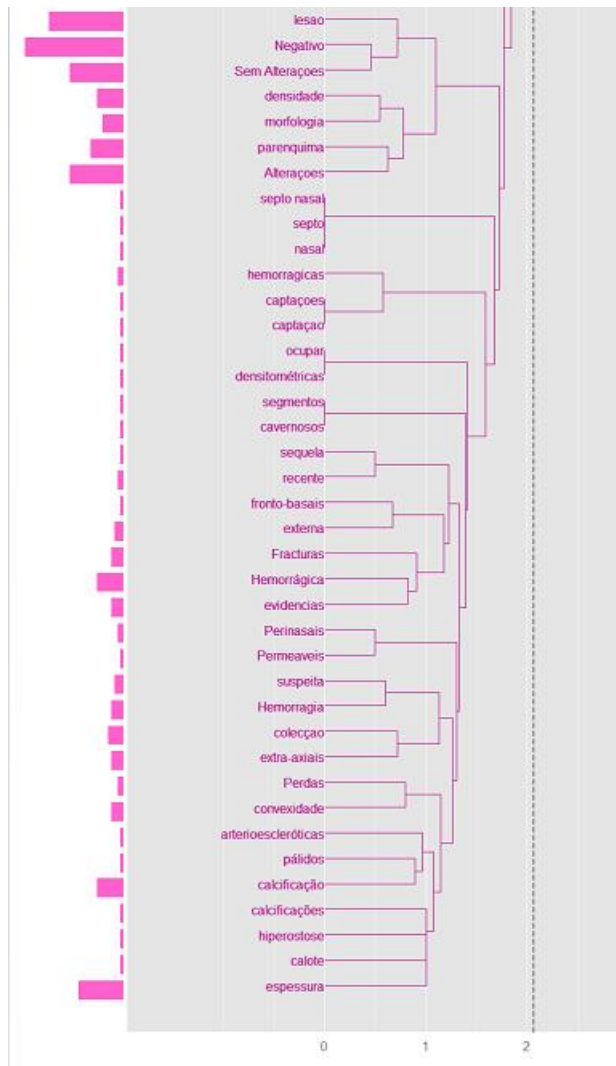


Figura A16 – Análise Hierárquica de Clusters

C) Mapa Auto Organizacional com Dicionário

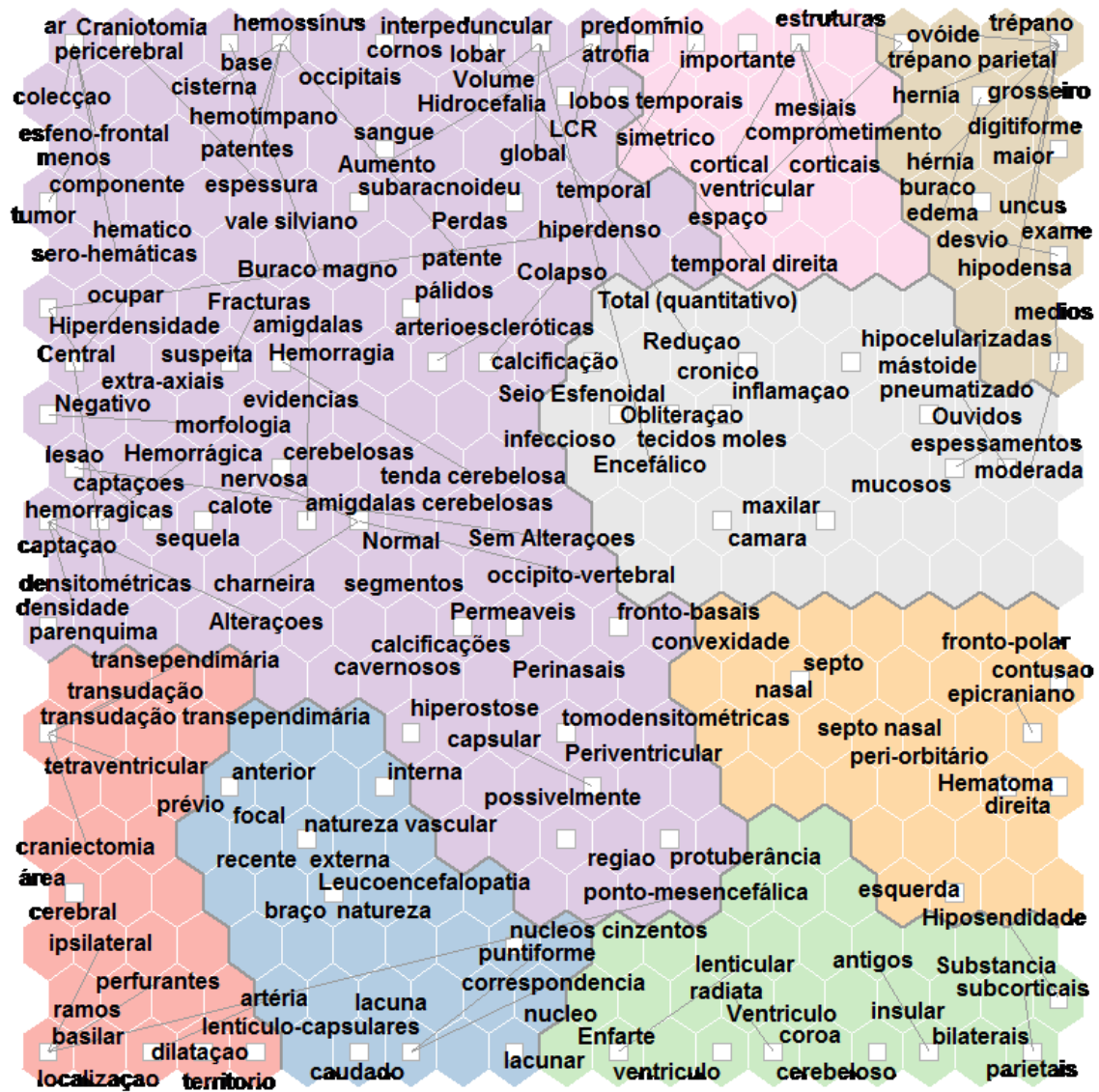


Figura A17 – Mapa Auto Organizacional com Dicionário



E) Análise de Correspondência

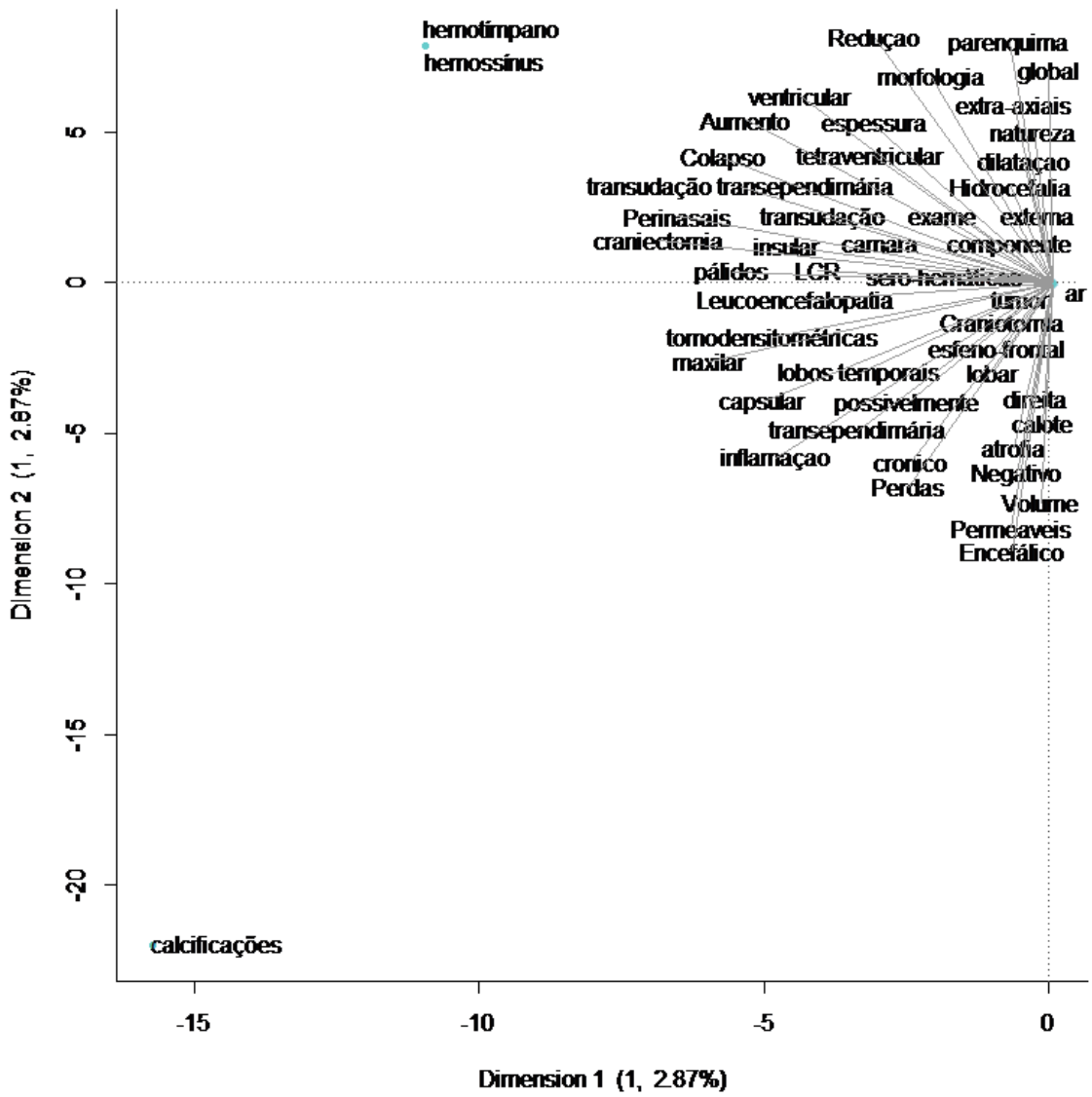


Figura A19 – Análise de Correspondência

F) Escala multidimensional de Códigos

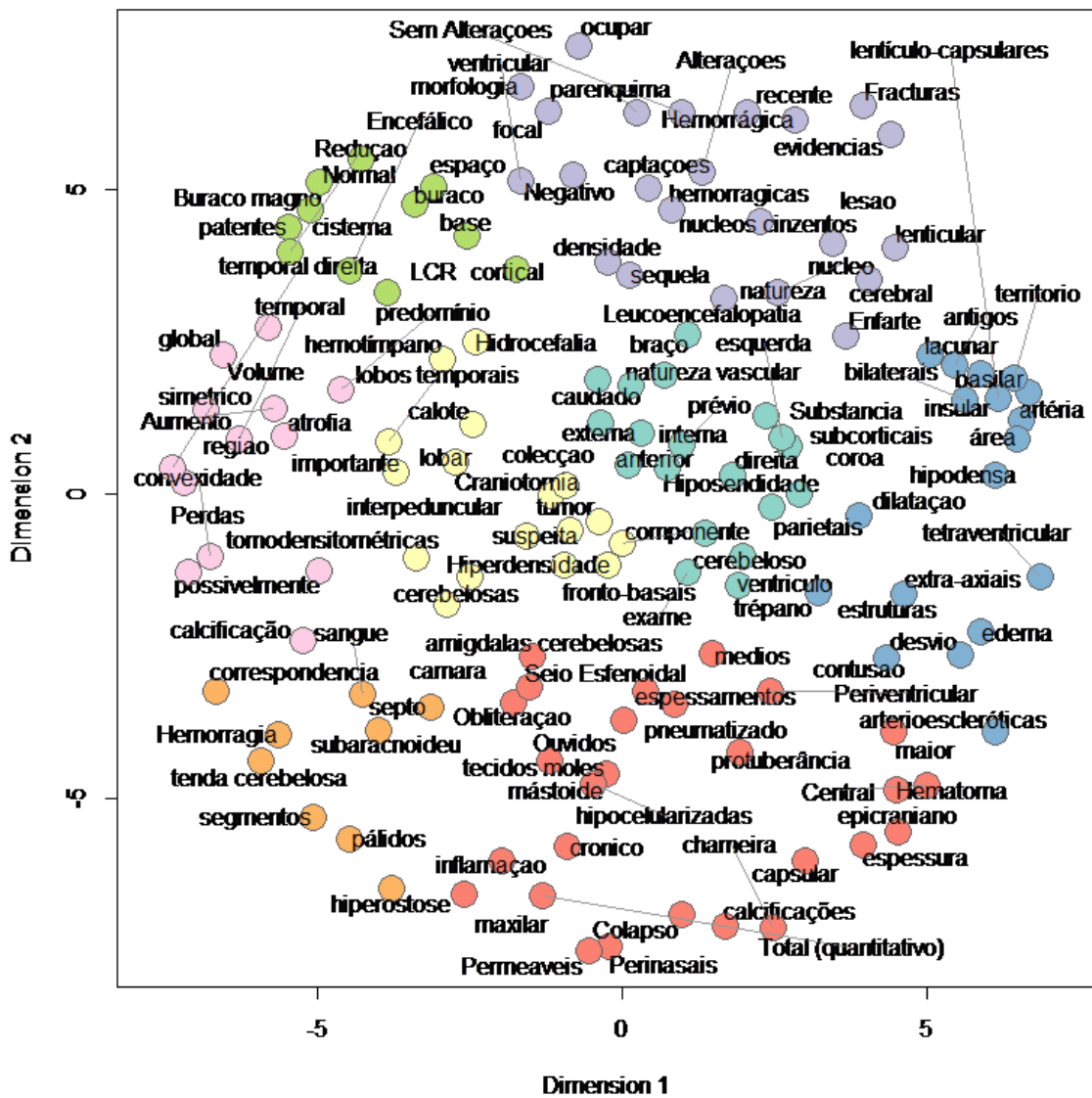


Figura A20 – Escala multidimensional de Códigos

A) Tabela com os Resultados da utilização do Algoritmo *Decision Tree*

Tabela A3 – Tabela com os Resultados da utilização do Algoritmo *Decision Tree*

Algoritmo	<i>Oversampling</i>	Cenário	S	E	A	Er
DT	Não	1	23,17%	<b>83,38%</b>	71,47%	28,53%
DT	Não	2	25%	<b>84,14%</b>	72,44%	27,56%
DT	Não	3	27,30%	<b>85,51%</b>	74,02%	25,98%
DT	Não	4	27,13%	<b>83,31%</b>	72,20%	27,80%
DT	Não	5	28,05%	<b>85,49%</b>	74,13%	25,87%
DT	Não	6	21,04%	<b>83,38%</b>	71,05%	28,95%
DT	Não	7	21,04%	<b>83,38%</b>	71,05%	28,95%
DT	Sim	1	<b>91,01%</b>	<b>76,24%</b>	83,57%	16,43%
DT	Sim	2	<b>90,78%</b>	<b>76,77%</b>	83,72%	16,28%
DT	Sim	3	<b>89,43%</b>	<b>76,38%</b>	82,86%	17,14%
DT	Sim	4	<b>89,10%</b>	<b>75,86%</b>	82,44%	17,56%
DT	Sim	5	<b>90,24%</b>	<b>76,54%</b>	83,35%	16,65%
DT	Sim	6	<b>90,02%</b>	<b>76,02%</b>	82,97%	17,03%
DT	Sim	7	90,02%	76,02%	82,97%	17,03%

B) Tabela com os Resultados da utilização do Algoritmo *K-Nearest Neighbour*

Tabela A4 – Tabela com os Resultados da utilização do Algoritmo *K-Nearest Neighbour*

Algoritmo	<i>Oversampling</i>	Cenário	S	E	A	Er
KNN	Não	1	4,27%	<b>98,05%</b>	79,49%	20,51%
KNN	Não	2	4,27%	<b>97,52%</b>	79,07%	20,93%
KNN	Não	3	4,60%	<b>98,34%</b>	79,83%	20,17%
KNN	Não	4	3,96%	<b>98,05%</b>	79,43%	20,57%
KNN	Não	5	3,05%	<b>98,12%</b>	79,31%	20,69%
KNN	Não	6	4,27%	<b>97,51%</b>	79,07%	20,93%
KNN	Não	7	4,27%	<b>97,51%</b>	79,07%	20,93%
KNN	Sim	1	<b>96,72%</b>	74,21%	<b>85,39%</b>	<b>14,61%</b>



KNN	Sim	2	<b>97,03%</b>	70,68%	83,76%	16,24%
KNN	Sim	3	<b>96,78%</b>	74,72%	<b>85,67%</b>	<b>14,33%</b>
KNN	Sim	4	<b>97,03%</b>	71,35%	84,10%	15,90%
KNN	Sim	5	<b>97,33%</b>	<b>75,34%</b>	<b>86,26%</b>	<b>13,74%</b>
KNN	Sim	6	<b>97,33%</b>	70,53%	83,84%	16,16%
KNN	Sim	7	<b>97,33%</b>	70,53%	83,84%	16,16%

C) Tabela com os Resultados da utilização do Algoritmo *Decision Tree* com *Cross Validation*

Tabela A5 – Tabela com os Resultados da utilização do Algoritmo *Decision Tree* com *Cross Validation*

Algoritmo	<i>Oversampling</i>	Cenário	S	E	A	Er
DTCV	Não	1	26,53%	<b>83,76%</b>	72,44%	27,56%
DTCV	Não	2	26,62%	<b>84,98%</b>	73,43%	26,57%
DTCV	Não	3	27,11%	<b>85,48%</b>	73,94%	26,06%
DTCV	Não	4	26,72%	<b>84,19%</b>	72,82%	27,18%
DTCV	Não	5	27,90%	<b>84,64%</b>	73,42%	26,58%
DTCV	Não	6	24,70%	<b>85,02%</b>	73,09%	26,91%
DTCV	Não	7	27,63%	<b>84,46%</b>	73,22%	26,78%
DTCV	Sim	1	<b>98,65%</b>	<b>77,24%</b>	<b>87,87%</b>	<b>12,13%</b>
DTCV	Sim	2	<b>97,60%</b>	<b>79,13%</b>	<b>88,30%</b>	<b>11,70%</b>
DTCV	Sim	3	<b>98,39%</b>	<b>79,18%</b>	<b>88,72%</b>	<b>11,28%</b>
DTCV	Sim	4	<b>98,03%</b>	<b>77,96%</b>	<b>87,93%</b>	<b>12,07%</b>
DTCV	Sim	5	<b>97,96%</b>	<b>79,36%</b>	<b>88,60%</b>	<b>11,40%</b>
DTCV	Sim	6	<b>98,92%</b>	<b>77,55%</b>	<b>88,17%</b>	<b>11,83%</b>
DTCV	Sim	7	<b>98,81%</b>	<b>78,77%</b>	<b>88,72%</b>	<b>11,28%</b>

D) Tabela com os Resultados da utilização do Algoritmo *K-Nearest Neighbour* com *Cross Validation*

Tabela A6 – Tabela com os Resultados da utilização do Algoritmo *K-Nearest Neighbour* com *Cross Validation*

Algoritmo	<i>Oversampling</i>	Cenário	S	E	A	Er
KNNCV	Não	1	5,03%	<b>96,07%</b>	78,07%	21,93%
KNNCV	Não	2	8,78%	<b>96,62%</b>	79,24%	20,76%
KNNCV	Não	3	6,16%	<b>97,28%</b>	79,27%	20,73%
KNNCV	Não	4	7,14%	<b>96,19%</b>	78,57%	21,43%
KNNCV	Não	5	5,03%	<b>96,77%</b>	78,63%	21,37%
KNNCV	Não	6	8,23%	<b>96,53%</b>	79,06%	20,94%
KNNCV	Não	7	8,42%	<b>96,37%</b>	78,99%	21,01%
KNNCV	Sim	1	<b>93,46%</b>	70,74%	82,02%	17,98%
KNNCV	Sim	2	<b>93,78%</b>	69,28%	81,44%	18,56%
KNNCV	Sim	3	<b>92,60%</b>	71,40%	81,92%	18,08%
KNNCV	Sim	4	<b>94,90%</b>	67,88%	81,29%	18,71%
KNNCV	Sim	5	<b>93,62%</b>	70,72%	82,09%	17,91%
KNNCV	Sim	6	<b>94,81%</b>	68,58%	81,60%	18,40%
KNNCV	Sim	7	<b>94,65%</b>	68,80%	81,64%	18,36%

Publicações Científicas

# *Towards of automatically detecting Brain Death patterns through Text Mining*

**Autores:** António Silva, Filipe Portela, Manuel Filipe Santos, José Machado e António Abelha

**Conferência:** *IEEE Conference on Business Informatics - ISA'HEALTH@CBI'2016 - Intelligent Systems and Applications in Healthcare Workshop*

**Livro/Editora:** *IEEE*

**Ano:** 2016

**Estado:** aceite para publicação

**Abstract:** *in the area of medicine, x-rays are very useful to check if the patient suffers from brain death. Their diagnosis is made using free text. This type of record difficult the process of making qualitative analysis in order to automatically detect possible brain problems. This project aims to make qualitatively and quantitatively analysis of Brain Computed Tomography (CT) diagnosis using text analysis tools as is Natural Language Processing and Text Mining. In this work a set of related words that can means patterns in CT reports was detected. The datasets were provided by the Centro Hospitalar do Porto- Hospital de Santo António and the contained information about patient deaths and CT done to the brain. With the analysis made, a new research and analysis perspectives of structured and unstructured texts in this field was opened.*

# *Predicting Brain Deaths using Text Mining and X-Rays clinical notes*

**Autores:** António Silva, Filipe Portela, Manuel Filipe Santos, António Abelha e José Machado

**Conferência:** *Lecture Notes in Computer Science (LNCS) - Computational Science and Its Applications - MIKE 2016*

**Livro/Editora:** *Springer*

**Ano:** 2016

**Estado:** aceite para publicação

**Abstract:** *The prediction of events is a task associated to the Data Science area. In the health, this method is extremely useful to predict critical events that may occur in people, or in a specific area. The Text Mining is a technique that consists in retrieving information from text files. In the Medical Field, the Data Mining and Text Mining solutions can help to prevent the occurrence of certain events to a patient. This project involves the use of Text Mining to predict the Brain Death by using the X-Ray clinical notes. This project is creating reliable predictive models with non-structured text. This project was developed using real data provided by Centro Hospitalar do Porto. The results achieved are very good reaching a sensitivity of 98% and a specificity of 88%.*

# *A mining analysis of Brain Death using Computed Tomography*

**Autores:** António Silva, Filipe Portela, Manuel Filipe Santos, José Machado e António Abelha

**Livro/Editora:** *Information Journal*

**Ano:** 2016

**Estado:** aceite para publicação

**Abstract:** *one of the ways to discover if the patient suffered from brain death is the use of x-rays. The x-rays are written in a free text way. The use of free text in the x-rays makes more difficult the process of making qualitative analysis in order to automatically detect possible brain problems. This project used text analysis tools as is Natural Language Processing and Text Mining to make qualitatively and quantitatively analysis of Brain Computed Tomography (CT). With this project was possible to identify a group of words that have relations with each other's and these relations can mean patterns of patients that had suffer brain death in de CT reports. The datasets that were used in this project contained information about patient deaths and CT done to the brain and were provided by the Centro Hospitalar do Porto- Hospital de Santo António. The analysis performed in this project was the Themes Frequency, the Self-Organizing Map of Codes, the Multidimensional Scaling of Codes Co-occurrence network, and the Correspondence Analysis The results of this project opens new research and analysis perspectives of structured and non-structured text.*

# *Towards an Ontology for X-Ray Reports*

**Autores:** António Silva, Filipe Portela, Manuel Filipe Santos, José Machado e António Abelha

**Conferência:** *WorldCist'17 - 5th World Conference on Information Systems and Technologies*

**Livro/Editora:** *Springer*

**Ano:** 2017

**Estado:** em avaliação.

**Abstract:** *Brain death is one of the most serious diagnoses that can be diagnosed in a patient. The possibility to detect it before its happening is one of the possible steps for the prevention of this event. The X-rays – Computed Tomography(CT) scans, are a very important test for the detection of this diagnosis. This paper proposes the use of an ontology on the registration of x-rays made to patients. This work was performed through the data provided by the Centro Hospitalar do Porto - Hospital de Santo António. The ontology was used based on an analysis made to the data and with the use of a dictionary developed in the same analysis. Finally, we added to the ontology the types of patients with brain death that were discovered in a previous work that used the dictionary that is present in this ontology.*