Análise de Dados Textuais: Análise de Correspondências e Classificação

por

Cláudia Sofia Vieites Dias

Dissertação de Mestrado em
Modelação, Análise de Dados e Sistemas de Apoio à Decisão

Orientada por

Professora Doutora Maria Paula Brito
Doutora Conceição Nunes Rocha

Faculdade de Economia

Universidade do Porto

2015

Aos meus pais

Nota Biográfica

Cláudia Sofia Vieites Dias é natural de Vila Praia de Âncora e nasceu no dia 15 de Julho de 1991. Fascinada pelo Porto, ingressou na Faculdade de Economia da Universidade do Porto onde se licenciou em Gestão em Junho de 2013. Em Setembro do mesmo ano começou a frequentar o Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão também na Faculdade de Economia da Universidade do Porto. Durante os anos como estudante foi procurando desenvolver outras atividades como tutoria a estudantes Erasmus, voluntariado na Associação de Estudantes e foi vice-coordenadora da Comissão de Finalistas. Foi promotora bancária no Banco Santander Totta em Setembro de 2013, realizou um estágio de Verão na Caixa Geral de Depósitos em 2014 e desempenhou funções de promotora em regime de *part-time* na empresa Btrust desde Março de 2015. Neste momento integra a equipa da BIT na Sonae MC onde desenvolve um projeto na área de Supply Chain.

Agradecimentos

Começo por agradecer à Professora Doutora Paula Brito por ter aceite ser minha orientadora e por me ter sugerido este projeto. Agradeço todo o encorajamento, paciência e disponibilidade para responder a todas as minhas questões. À Doutora Conceição Rocha por toda a ajuda na compreensão dos dados e pela amabilidade e disponibilidade que sempre manifestou. Obrigado pela dedicação a este projeto, pela preocupação em todas as fases e pela motivação que me deram sempre até ao fim.

De uma forma especial, agradeço aos meus pais, a quem dedico este trabalho, pelo apoio e incentivo que me deram. Sem eles isto não seria possível.

Ao Sérgio, pelos conselhos, pela paciência e pela ajuda nos momentos mais difíceis. Obrigado por todo o interesse e por ter acreditado sempre em mim.

À Carolina por ter partilhado da minha preocupação e por me perceber melhor do que ninguém.

À Joana pela admiração e carinho que demonstrou nos nossos jantares.

À Joana, à Daniela, à Sofia e ao Diogo um obrigado por todos os momentos de distração que me proporcionaram.

À Andreia e ao Dinis por me terem ajudado a resolver alguns problemas e pela companhia neste percurso.

À Cascais e à Raquel pela constante preocupação e carinho.

A todos os meus amigos que estiveram comigo tardes na FEP, obrigado pelo apoio e pelas distrações pertinentes.

Por último, agradeço também à SAPO Labs (http://labs.sapo.pt) por disponibilizar o conjunto de notícias da agência *Lusa*.

Resumo

A extração de informação relevante a partir de dados textuais continua a colocar muitos desafios aos investigadores. Um dos métodos que pode contribuir muito para a análise de dados textuais é a Análise de Correspondências (AC), pois, é versátil e simples de implementar. O facto de impor, como única restrição, a existência de uma matriz retangular com entradas não-negativas faz desta uma técnica flexível relativamente aos requisitos dos dados. Em particular, é uma técnica adequada a dados textuais, que podem facilmente ser representados em tabelas de contingência. Por outro lado, os métodos de Análise Classificatória são complementos essenciais aos resultados obtidos pela Análise de Correspondências. De facto, quando existe um elevado número de elementos torna-se difícil perceber quais as suas posições relativas visualizando apenas o gráfico gerado pela Análise de Correspondências.

Neste trabalho, aplicam-se sucessivamente Análise de Correspondências e Análise Classificatória a três conjuntos de dados textuais. O primeiro é constituído pelas 227 notícias publicadas pela agência Lusa no dia 31 de dezembro de 2010. Este conjunto é constituído por pequenas notícias e apresenta uma grande diversidade de temas. Como a AC identifica as palavras que mais se destacam no conjunto dos dados, o uso dessas palavras como atributos das notícias para classificação das mesmas por temas é comprometido pelo aparecimento de palavras com pouco significado. Para contornar esta dificuldade, e na expectativa de melhorar os resultados, considerou-se como segundo conjunto de dados textuais a analisar a lista de entidades citadas no texto. Entidade neste contexto é todo o nome próprio ou todo o nome comum associado a determinada função ou cargo, *e.g.*, presidente ou deputado. Retiveram-se os eixos principais e seguidamente foram aplicados métodos de classificação sobre as coordenadas fatoriais. Efetuou-se uma classificação ascendente hierárquica e aplicaram-se o mapa de Kohonen e o algoritmo das K-médias, permitindo agrupar as notícias por temas. Foram assim identificados temas como Desporto e Política.

Por último, foi utilizado o livro 'Segredos da Maçonaria Portuguesa' com o intuito de descobrir as entidades que mais contribuem para os 2508 parágrafos do livro e efetuar uma comparação com um estudo realizado para o mesmo conjunto de dados usando redes sociais. O elevado número de observações e a extração das entidades a partir do livro digitalizado constituíram problemas adicionais face aos dados sobre as notícias.

Para este trabalho foi utilizado o *software* de acesso livre Dtm-Vic (*Data and Text Mining*: Visualização, Inferência, Classificação). Utilizaram-se as ferramentas Visuresp e Visutex que permitem obter um resumo do conteúdo dos dados e respetiva AC. Também se utilizou o *software* SPSS Statistics na Classificação.

Palavras-Chave: Análise de Correspondências, Classificação, Dados Textuais, Entidades

Abstract

Information retrieval from textual data still presents many challenges to researchers. Correspondence Analysis (AC) can greatly contribute to textual data analysis, since it is versatile and easy to implement. The only strict data requirement for Correspondence Analysis is a rectangular matrix with non-negative entries, which makes it a relatively flexile technique as concerns data requirements. In particular, it is a technique suitable for textual data which can easily be represented in contingency tables. Clustering methods are essential complements of the results obtained by Correspondence Analysis. In fact, when the data set under analysis has a large number of elements it becomes difficult to understand their relative positions by observing the graph generated from Correspondence Analysis.

In this work, we successively apply Correspondence Analysis and Clustering to three sets of textual data. The first consists of 227 news items published by the Lusa agency on the 31^{st} December 2010. These news are small and present a wide variety of topics. Since AC identifies the words that stand out in the data set, the use of these words as attributes to classify the news by themes is compromised by the appearance of words with low meaning. To overcome this difficulty, and hoping to improve results, a second set of textual data has been considered, consisting of the list of entities cited in the text. Entity in this context is a given name or any common name associated with a particular function or position, e.g., president or deputy. The principal axes of the AC have been retained and clustering methods have then been applied on the factorial coordinates. We conducted a hierarchical ascending classification and applied the Kohonen map and the K-means algorithm, allowing grouping the news by topics. Themes were thus identified, such as Sports and Politics.

Finally, we used the book 'Segredos da Maçonaria Portuguesa' with the objective of discovering the entities that contribute most to the 2508 paragraphs of the book, as well as make a comparison with a previous study on the same data set using social networks. The high number of observations and the extraction of entities from the digitized book constituted additional problems as compares to the news data sets.

For this study we used the free access software Dtm-Vic (Data and Text Mining: Visualization, Inference, Classification). The Visuresp and Visutex tools were used, allowing for a summary of the data and respective AC. The SPSS Statistics software has also been used for the Clustering task.

Keywords: Correspondence Analysis, Clustering, Textual Data, Entities

Índice

Nota Biográfica **ii**

Agradecimentos **iii**

Resumo **iv**

Abstract **vi**

1 Introdução **1**
1.1 Motivação 1
1.2 Objetivos 2
1.3 Contribuições 3
1.4 Organização 3

2 Estado da Arte **5**
2.1 Text Mining 5
2.1.1 Conceito e Aplicações 5
2.1.2 Métodos Utilizados 6
2.2 Análise de Correspondências 7
2.2.1 Descrição do método 7
2.2.2 Interpretação 9
2.2.3 Considerações acerca dos Dados 9
2.2.4 Aplicações 10
2.3 Análise Classificatória 10
2.3.1 Conceitos Gerais 10
2.3.2 Métodos de Classificação 11
2.3.3 Medidas de (Dis)semelhança 14
2.3.4 Métodos de Agregação 15

3 Estudo de um conjunto de notícias **16**
3.1 Descrição e análise dos dados - notícias 16
3.1.1 Análise de Correspondências 17
3.1.2 Análise Classificatória 25

3.2 Extração de entidades 30
3.3 Descrição e análise dos dados - notícias e entidades 31
3.3.1 Análise de Correspondências 32
3.3.2 Análise Classificatória 43
3.4 Discussão dos resultados 50

4 Segredos da Maçonaria Portuguesa **52**
4.1 Descrição e análise dos dados 52
4.1.1 Análise de Correspondências 52
4.1.2 Análise Classificatória 58
4.2 Discussão dos resultados 66

5 Conclusões **68**
5.1 Resultados 68
5.2 Limitações e Trabalho Futuro 69

Bibliografia **70**

Anexos **73**

A Dtm-Vic — *Data and Text Mining*: Visualização, Inferência, Classificação **73**

B Dados notícias - palavras retidas **75**

C Análise de Correspondências dos dados notícias — Histograma com os valores próprios (*output* parcial). **77**

D Classificação Hierárquica - dados notícias **78**

E Classificação Não Hierárquica - dados notícias **81**

F Mapas de Kohonen - dados notícias. **83**

G Tabela de Contingência — dados notícias e entidades **87**

H Análise de Correspondências dos dados notícias e entidades — Histograma com os valores próprios (*output* parcial). **89**

I Análise de Correspondências - Notícias e entidades **90**

J Classificação Hierárquica - dados entidades e notícias **103**

K Classificação Não Hierárquica - dados entidades e notícias **108**

L Mapas de Kohonen - dados notícias e entidades **110**

M Dados Livro - Valores próprios e inércia para os 38 primeiros eixos. **113**

N Análise de Correspondências - Livro **114**

O Classificação Hierárquica - livro **119**

P Classificação Não Hierárquica - livro **122**

Q Mapas de Kohonen - livro **126**

Lista de Tabelas

3.1 Número de palavras retidas para alguns níveis de frequência. 18
3.2 Inércia explicada para as partições com 2 a 30 classes. 27
3.3 Inércia explicada para as partições em 2 a 30 classes. 29
3.4 Frequência das entidades de acordo com o número de caracteres . . . 32
3.5 Número de entidades retidas para alguns níveis de frequência. 33
3.6 Entidades retidas e respetivas frequências 33
3.7 Inércia explicada para as partições com 2 a 25 classes. 43
3.8 Inércia explicada para as partições com 2 a 25 classes. 47

4.1 Entidades retidas e respetivas frequências 53
4.2 Valores próprios, inércia e inércia acumulada para os 38 primeiros eixos. 54
4.3 Inércia explicada para as partições de 2 até 30 classes. 60
4.4 Inércia explicada para as partições 2 até 30 classes. 63

B.1 Palavras retidas e frequências 75

D.1 Classes formadas através da aplicação da Classificação Hierárquica às 30 coordenadas fatoriais das 227 notícias — partição em 3 classes. . . 78
D.2 Classes formadas através da aplicação da Classificação Hierárquica às 30 coordenadas fatoriais das 227 notícias — partição em 23 classes. . 79

E.1 Classes formadas através da aplicação do algoritmo K-médias às 30 coordenadas fatoriais das 227 notícias — partição em 27 classes. . . . 81

I.1 Coordenadas, contribuições absolutas e relativas das 50 entidades retidas para o eixo 1. 90
I.2 Coordenadas, contribuições absolutas e relativas das 227 notícias para o eixo 1. 91
I.3 Coordenadas, contribuições absolutas e relativas das 50 entidades retidas para o eixo 2. 94
I.4 Coordenadas, contribuições absolutas e relativas das 227 notícias para o eixo 2. 95
I.5 Coordenadas, contribuições absolutas e relativas das 50 entidades retidas para o eixo 3. 98

I.6 Coordenadas, contribuições absolutas e relativas das 227 notícias para o eixo 3. 100

J.1 Classes formadas através da aplicação da Classificação Hierárquica às 24 coordenadas fatoriais das 227 notícias — partição em 2 classes. . . 103
J.2 Classes formadas através da aplicação da Classificação Hierárquica às 24 coordenadas fatoriais das 227 notícias — partição em 12 classes. . 104
J.3 Classes formadas através da aplicação da Classificação Hierárquica às 24 coordenadas fatoriais das 227 notícias — partição em 25 classes. . 106

K.1 Classes formadas através da aplicação do algoritmo K-médias às 24 coordenadas fatoriais das 227 notícias — partição em 19 classes. . . . 108

M.1 Valores próprios, inércia e inércia acumulada para os 38 primeiros eixos. 113

N.1 Coordenadas, contribuições absolutas e relativas das 56 entidades retidas para o eixo 1. 114
N.2 Coordenadas, contribuições absolutas e relativas das 56 entidades retidas para o eixo 2. 115
N.3 Coordenadas, contribuições absolutas e relativas das 56 entidades retidas para o eixo 3. 117

O.1 Classes formadas através da aplicação da Classificação Hierárquica às 30 coordenadas fatorias das 56 entidades retidas — partição em 15 classes. 119
O.2 Classes formadas através da aplicação da Classificação Hierárquica às 30 coordenadas fatorias das 56 entidades retidas — partição em 4 classes. 120

P.1 Classes formadas a partir da aplicação do algoritmo K-médias às 30 coordenadas fatoriais das 56 entidades retidas — partição em 19 classes. 122
P.2 Classes formadas a partir da aplicação do algoritmo K-médias às 30 coordenadas fatoriais das 56 entidades retidas — partição em 21 classes. 123
P.3 Classes formadas a partir da aplicação do algoritmo K-médias às 30 coordenadas fatoriais das 56 entidades retidas — partição em 4 classes. 124

Lista de Figuras

2.1 Dendrograma representando uma Análise Classificatória num conjunto de oito elementos (Lebart et al., 1998). 13
2.2 Gráfico que relaciona as 8 partições (entre 1 e 8 classes) com a inércia intra-classes de cada uma. 13

3.1 Distribuição do número de palavras por notícia. 17
3.2 Quadro resumo - Eixo 1. 20
3.3 Notícias e palavras representadas de acordo com o *ranking* no plano [1,2] 21
3.4 Quadro resumo - Eixo 2. 22
3.5 Quadro resumo - Eixo 3. 23
3.6 Notícias e palavras representadas de acordo com o *ranking* no plano [1,3]. 24
3.7 Representação através de um dendrograma da classificação hierárquica ascendente aplicada às 227 notícias descritas pelas 30 coordenadas fatoriais. 26
3.8 Inércia intra-classes para as partições em 2,..., 30 classes. 27
3.9 Quadro resumo - temas obtidos através da Classificação Hierárquica. 28
3.10 Número de entidades (5028) por notícia (227). 31
3.11 Entidades retidas representadas no plano [1,2]. 36
3.12 Entidades retidas representadas de acordo com o *ranking* no plano [1,2]. 37
3.13 As 227 notícias representadas de acordo com o *ranking* no plano [1,2]. 38
3.14 Notícias e entidades no plano [1,2] 39
3.15 Quadro resumo - Eixo 1. 40
3.16 Quadro resumo - Eixo 2. 41
3.17 Quadro resumo - Eixo 3. 41
3.18 Notícias e entidades no plano [1,3]. 42
3.19 Representação através de um dendrograma da classificação hierárquica ascendente aplicada às 227 notícias e às 24 coordenadas fatoriais. 44
3.20 Inércia intra-classes para as partições 2 a 25. 45
3.21 Quadro resumo - temas obtidos através da Classificação Hierárquica. 46
3.22 Boxplot obtido para um número de classes igual a 2. 47
3.23 Número de elementos em cada classe para K=19. 48

3.24 Quadro resumo dos temas identificados — mapa de Kohonen. 50
3.25 Quadro resumo dos temas identificados — conjunto de dados notícias. 50
3.26 Quadro resumo dos temas identificados — conjunto de dados notícias e entidades. 50

4.1 Entidades retidas do livro representadas no plano [1,2]. 56
4.2 Entidades retidas do livro, após a exclusão da entidade 'Vice', representadas no plano [1,2]. 57
4.3 Quadro resumo - Eixos 1, 2 e 3. 58
4.4 Entidades retidas do livro representadas no plano [1,3]. 59
4.5 Inércia intra-classes para as partições de 2 até 30 classes. 60
4.6 Representação através de um dendrograma da classificação hierárquica ascendente aplicada às 57 entidades retidas descritas pelas 30 coordenadas fatoriais. 61
4.7 Mapa de Kohonen (3 x 3) representando as 56 entidades do livro. . . 64
4.8 Classes relevantes obtidos a partir dos mapas de Kohonen 5x5 e 6x6 com as 56 entidades do livro. 66
4.9 As seis comunidades de maior dimensão obtidas através da aplicação de redes socais por Rocha et al. (2014). 67

A.1 Menu principal do *software* Dtm-Vic. 73
A.2 Comandos do *software* Dtm-Vic. 74

C.1 Histograma com os primeiros 43 valores próprios da AC do conjunto de dados notícias. 77

F.1 Mapa de Kohonen (4 x 4) representando as 227 notícias e as 87 entidades. 84
F.2 Mapa de Kohonen (5 x 5) representando as 227 notícias e as 87 entidades. 85
F.3 Mapa de Kohonen (6 x 6) representando as 227 notícias e as 87 entidades. 86

G.1 Notícia e frequência das 50 entidades retidas. 88

H.1 Histograma com os primeiros 25 valores próprios. 89

L.1 Mapa de Kohonen (3 x 3) representando as 227 notícias e as 50 entidades. 111
L.2 Mapa de Kohonen (4 x 4) representando as 227 notícias e as 50 entidades. 112

Q.1 Mapa de Kohonen (4 x 4) representando as 56 entidades do livro. . . 127
Q.2 Mapa de Kohonen (5 x 5) representando as 56 entidades do livro. . . 128

Q.3 Mapa de Kohonen (6 x 6) representando as 56 entidade do livro. . . . 129

Capítulo 1

Introdução

Neste capítulo pretende-se fornecer uma visão global daquilo que irá ser desenvolvido na dissertação, nomeadamente a descrição do tema, motivação, objetivos que se pretendem alcançar e a forma como esta dissertação está organizada.

1.1 Motivação

Em diversas áreas de estudo, os investigadores lidam com um grande conjunto de dados textuais que é necessário gerir e analisar cuidadosamente. Com o objetivo de processar este tipo de dados foram desenvolvidos e propostos vários métodos. Estas contribuições dividem-se em métodos desenvolvidos com origem na inteligência artificial, métodos estatísticos e técnicas de análise exploratória de dados. Apesar destas últimas apresentarem excelentes propriedades, Morin (2004a) constatou que os métodos desenvolvidos com origem na inteligência artificial são os mais utilizados.

Dentro das técnicas de análise exploratória de dados destaca-se a Análise de Correspondências. Esta técnica pode contribuir muito para a deteção e explicação de dados textuais, pois é versátil e simples de implementar. Para além disso, esta técnica apresenta características úteis a diversas investigações. Exemplo disso é a sua natureza multivariada que permite revelar interligações entre as variáveis. A representação gráfica bidimensional, gerada por este método, facilita a deteção e análise das relações entre as variáveis, entre os indivíduos e entre as variáveis e os indivíduos. Esta característica da Análise de Correspondências é uma vantagem, pois, esta dualidade não está presente noutras abordagens multivariadas de representação gráfica de dados (Hoffman e Franke, 1986).

O tipo de requisitos impostos aos dados para aplicação da Análise de Correspondências é outra das suas vantagens. O facto de impor, como única restrição, a existência de uma matriz retangular com entradas não-negativas faz desta uma técnica flexível relativamente aos requisitos dos dados. Em particular, é uma técnica adequada a dados textuais, que podem facilmente ser representados em tabelas de

contingência. A Análise de Correspondências utiliza estas tabelas de forma a cruzar termos e documentos permitindo aos autores questionarem-se sobre se há alguma proximidade entre termos, entre documentos e entre termos e documentos. Neste contexto, um termo pode ser formado por uma palavra, uma sigla ou um conjunto de palavras.

Lebart et al. (1998) observaram que os métodos de Análise Classificatória têm-se revelado complementos essenciais aos resultados obtidos pela Análise de Correspondências. Estas técnicas são uma segunda família das técnicas de análise de dados em adição aos métodos dos eixos principais (nos quais a Análise de Correspondências está incluída) e são usados para representar proximidades entre os elementos de uma tabela através do agrupamento em classes. De facto, quando existe um elevado número de elementos torna-se difícil perceber quais as suas posições visualizando apenas o gráfico gerado pela Análise de Correspondências. O mesmo se verifica quando o texto é longo. Assim, os métodos de classificação enriquecem as representações de um ponto de vista multidimensional.

Ao aplicar sucessivamente as duas técnicas (Análise de Correspondências e Análise Classificatória) ao mesmo conjunto de dados, é possível obter mais informações sobre as relações existentes entre as variáveis, possibilitando ao analista ter uma visão sistematizada dos dados.

1.2 Objetivos

Com o aparecimento das tecnologias de informação, o acesso a dados deixou de ser um problema. O grande desafio com que hoje nos confrontamos é a extração de conhecimento desses dados. Uma parcela significativa das informações disponíveis encontra-se sob a forma de textos (ou documentos) não estruturados ou semi-estruturados, tais como livros, artigos, manuais, *e-mails* e a Web. A extração de informação deste tipo de dados (dados qualitativos) tornou-se possível devido à expansão do campo relativo a *Text Mining*. Esta área dedica-se à descoberta, extração e interpretação da informação contida em documentos de texto (Petrović et al., 2009). Podem ser analisadas palavras isoladas, conjunto de palavras e documentos através das suas similaridades ou das suas relações com variáveis de interesse. Morin (2006) constatou que com um elevado volume disponível de dados textuais, é necessário descobrir formas de analisar os dados e obter informação relevante. A escolha de uma estratégia para analisar este tipo de dados só pode ser feita em função dos objetivos definidos. Que tipo de texto estamos a analisar? Que questões pretendemos responder? É o nosso objetivo classificar documentos de forma a encontrá-los mais facilmente? Estas são algumas questões que devem ser colocadas de forma a definir o tipo de abordagem a implementar.

Nesta dissertação pretende-se analisar informação textual do livro 'Segredos da Maçonaria Portuguesa' e de notícias da Web. Os dados extraídos terão que assumir

a forma de tabelas de contingência. Estas tabelas, formadas por documentos e termos, indicam que termos é que aparecem em cada documento e com que frequência. Nas linhas irão estar os documentos, onde cada notícia será um documento e no caso do livro cada parágrafo será considerado um documento. Nas colunas irão estar os termos relevantes denominados neste trabalho por entidades. Entidade neste contexto é todo o nome próprio ou todo o nome comum associado a determinada função ou cargo, *e.g.*, presidente ou deputado, citado no texto. Para analisar e interpretar os dados extraídos, serão aplicadas duas técnicas: a Análise de Correspondências, de forma a reconhecer quais os termos que mais se destacam e como se relacionam com os documentos, e a Análise Classificatória, agrupando as notícias do mesmo tema e as entidades fortemente relacionados no caso do livro.

1.3 Contribuições

Neste trabalho propõe-se duas abordagens que podem ser utilizadas para estudar e analisar dados textuais. Para tal aplicam-se dois métodos inseridos nas técnicas de análise exploratória de dados: Análise de Correspondências e Análise Classificatória. A contribuição deste trabalho consiste na aplicação sucessiva destes dois métodos a três conjuntos de dados. O primeiro conjunto é constituído por notícias numeradas de 1 a 227, o segundo é constituído pelas 5028 entidades extraídas dessas notícias e o terceiro é constituído pelas entidades dos 2508 parágrafos do livro 'Segredos da Maçonaria Portuguesa'. A utilização conjunta dos dois métodos é uma mais-valia para analisar dados textuais, proporcionando uma representação gráfica mais completa e apelativa, o que permite compreender melhor as relações existentes entre elementos de natureza textual. Com estas aplicações, pretende-se realçar o uso de entidades para extrair informações de textos. Para o efeito, será realizada uma análise ao texto completo das notícias com o objetivo de ser comparada com a análise às entidades dessas notícias. Relativamente aos dados do livro, sendo que estes foram estudados *a priori* utilizando ferramentas de *Text Mining* (Rocha et al., 2014), nesta dissertação propõe-se uma abordagem alternativa através da aplicação dos métodos de Análise de Correspondências e Análise Classificatória com o objetivo de comparar os resultados dos dois estudos realizados.

1.4 Organização

Esta dissertação está dividida em cinco capítulos. No primeiro capítulo é feita uma introdução ao tema, expondo os objetivos desta dissertação, o problema a estudar e o conjunto de dados a utilizar. No Capítulo 2 é feito um levantamento do estado da arte. Este capítulo é dividido em três secções nomeadamente *Text Mining*, Análise de Correspondências e Análise Classificatória. Os Capítulos 3 e 4 são dedicados ao estudo dos conjuntos de dados. O Capítulo 3 refere-se à aplicação dos métodos aos

dois conjuntos de dados sobre as notícias da agência Lusa e respetiva comparação. No Capítulo 4 apresenta-se a análise realizada aos dados do livro e compara-se com os resultados obtidos no estudo efetuado com recurso a redes sociais. Por fim, no Capítulo 5 apresentam-se as considerações finais onde se faz um breve resumo sobre os resultados obtidos e referindo algumas limitações.

Capítulo 2

Estado da Arte

Neste capítulo apresentam-se os métodos a utilizar: Análise de Correspondências e Análise Classificatória. Começa-se por descrever as especificidades de cada um deles e apresentam-se algumas aplicações. Antes de explorar as técnicas referidas, este capítulo conta ainda com a apresentação de diversos métodos desenvolvidos na área do *Text Mining*. Nesta exposição consideram-se apenas os aspetos pertinentes no âmbito do tema em estudo.

2.1 Text Mining

2.1.1 Conceito e Aplicações

A área de *Text Mining* (TM) tem-se desenvolvido muito nos últimos tempos devido ao grande volume de informação textual disponível. TM refere-se geralmente ao processo de extrair conhecimento bem como padrões não triviais de documentos de texto não estruturados (Tan, 1999). É também conhecido como *Intelligent Text Analysis*, *Text Data Mining* ou *Knowledge-Discovery in Text* (KDT) (Gupta e Lehal, 2009). TM é uma área do *Data Mining*, com especificidade nos dados. Enquanto as ferramentas de *Data Mining* são concebidas para lidar com dados estruturados extraídos de bases de dados, o *Text Mining* analisa dados semi-estruturados e não estruturados, tais como *e-mails*, documentos de texto, ficheiros *html*, entre outros (Gupta e Lehal, 2009).

É possível encontrar na literatura diversas aplicações de técnicas de TM em diferentes áreas. Uma das aplicações encontradas na literatura é a análise de patentes. Técnicas de classificação de texto são frequentemente aplicadas para apoiar a análise de patentes em grandes empresas através da estruturação e visualização do *corpus* estudado. Hotho et al. (2005) mencionam ainda outros campos onde é aplicado o TM, tais como classificação de texto para agências de notícias, bioinformática e filtragem de *e-mails anti-spam*. Cohen e Hersh (2005) referem na sua investigação aplicações de TM na área das Ciências Biomédicas. Gupta e Lehal

(2009) descrevem algumas aplicações nas áreas de Gestão de Recursos Humanos, *Customer Relationship Management* e Análise de Mercado, Tecnologia e aspetos do Multilinguismo.

2.1.2 Métodos Utilizados

Sumarização de documentos de texto, técnicas de redução da dimensionalidade dos dados e de extração de informação são alguns dos métodos utilizados na área do *Text Mining*. Estas técnicas focam-se essencialmente em algoritmos que permitam retirar informações a partir de diferentes tipos de dados textuais.

No âmbito do TM, um aspeto importante é a visualização inteligente dos dados. Esta visualização permite expor a estrutura latente dos dados bem como providenciar novos conhecimentos. Além disto, pode ser usada como uma etapa de pré-processamento para outras técnicas, *e.g.*, pode ser usada para determinar o número de classes na Análise Classificatória. Diversas técnicas que permitem a redução da dimensão dos dados tornaram-se recentemente muito populares, como a Análise Fatorial, a Análise Semântica Latente, a Análise em Componentes Principais e a Análise de Correspondências (Petrović et al., 2009). Estas técnicas consideram uma representação baseada em matrizes retangulares. A principal diferença entre elas assenta nos *inputs* que as matrizes usam. A Análise de Correspondências é um dos temas fulcrais do presente trabalho, o seu estudo será detalhado na Secção 2.2.

Aggarwal e Zhai (2012) apresentam alguns métodos de aprendizagem supervisionada e não supervisionada. Os métodos de aprendizagem não supervisionada têm como objetivo a observação e descoberta e por isso não requerem dados de treino. Nestes métodos não se conhecem classes *a priori* nem se sabe como é que as observações se agrupam em classes. Os dois principais métodos de aprendizagem não supervisionada utilizados no contexto de dados textuais são Análise Classificatória (em inglês, *Clustering*) e *Topic Modeling*. A Análise Classificatória é uma metodologia usada para agrupar documentos similares. Este agrupamento é feito através de uma medida de semelhança e de um método de agregação. Será aprofundada na Secção 2.3.

Os métodos de aprendizagem supervisionada utilizam dados de treino para fazer a aprendizagem de um classificador. Esta aprendizagem é usada para prever a classe de um elemento que não tenha sido considerado. Esta família inclui métodos como classificadores baseados em regras, Árvores de Decisão, método do Vizinho Mais Próximo e classificadores probabilísticos.

2.2 Análise de Correspondências

2.2.1 Descrição do método

A origem da Análise de Correspondências (AC) pode ser atribuída simultaneamente a H. O. Hartley pela Matemática e a Richardson e Kuder (1933), e Horst (1935) pelas aplicações em Psicometria. Mais tarde, desenvolvimentos matemáticos foram realizados por Fisher, Guttman e Hayashi (*c.f.* Greenacre (1984)). A álgebra subjacente ao método de AC já existe há alguns anos, mas apenas em 1973 é que a AC assumiu a forma descrita nesta dissertação. Esta surgiu em França e foi desenvolvida por J. P. Benzécri num contexto linguístico (Benzécri, 1973).

A AC é um método de análise exploratória de dados utilizado para descrever tabelas de contingência. Estas tabelas, cruzando duas variáveis qualitativas, permitem estudar as correspondências, isto é, as relações que podem eventualmente existir entre as variáveis. O objetivo principal deste método é reduzir a dimensão de um determinado problema, tornando a análise do mesmo mais simples. No entanto, esta redução da dimensionalidade não pode ser obtida sem uma certa perda de informação, por isso pretende-se restringir esta perda ao mínimo possível para que a máxima quantidade de informação seja retida (Hoffman e Franke, 1986; Greenacre, 2007). Este método oferece a possibilidade de visualizar, simultaneamente, a distância entre documentos e a distância entre termos através de planos de eixos principais. Nestas representações gráficas, dois documentos estão perto se contêm termos que são próximos uns dos outros; dois termos são próximos se são usados com frequência nos mesmos documentos (Bécue-Bertaut et al., 2005).

Consideremos uma tabela de contingência com i linhas e j colunas formada por duas variáveis qualitativas A e B que assumem as categorias $A_1, ..., A_i, ...A_m$ e $B_1, ...B_j, ...B_p$:

	B_1	...	B_j	...	B_p	Total
A_1	n_{11}	...	n_{1j}	...	n_{1p}	$n_{1.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_i	n_{i1}	...	n_{ij}	...	n_{ip}	$n_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_m	n_{m1}	...	n_{mj}	...	n_{mp}	$n_{m.}$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{.p}$	n

onde n_{ij} representa a frequência absoluta para as categorias A_i e B_j, $n_{i.}$ e $n_{.j}$ indicam a frequência absoluta das linhas e colunas (respetivamente) e n representa o número total de indivíduos. As frequências relativas $f_{ij}=\frac{n_{ij}}{n}$ representam a proporção de indivíduos que apresentam as categorias A_i e B_j simultaneamente, obtendo-se o quadro:

	B_1	…	B_j	…	B_p	Total
A_1	f_{11}	…	f_{1j}	…	f_{1p}	$f_{1.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_i	f_{i1}	…	f_{ij}	…	f_{ip}	$f_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_m	f_{m1}	…	f_{mj}	…	f_{mp}	$f_{m.}$
Total	$f_{.1}$	…	f_{j}	…	$f_{.p}$	1

onde $f_{i.}$ e $f_{j.}$ são as frequências relativas marginais. Para analisar a tabela de contingência não se utilizam as frequências absolutas, mas sim os perfis-linha e perfis-coluna representados por $f(j|i) = \frac{n_{ij}}{n_{i.}}$ (perfis-linha) e $f(i|j) = \frac{n_{ij}}{n_{.j}}$ (perfis-coluna). Os perfis-linha e os perfis-coluna representam estimativas das probabilidades condicionadas de uma categoria observada de uma variável sabendo a categoria observada da outra variável. Assim, um perfil-linha $(\frac{f_{ij}}{f_{i.}})$ indica a proporção de indivíduos que verificam a categoria B_j sabendo que se verifica a categoria A_i. Analogamente, um perfil-coluna $(\frac{f_{ij}}{f_{.j}})$ representa a proporção de indivíduos que verificam a categoria A_i sabendo que se verifica a categoria B_j.

Perfis-linha

	B_1	…	B_j	…	B_p	Total
A_1	$\frac{f_{11}}{f_{1.}}$	…	$\frac{f_{1i}}{f_{1.}}$	…	$\frac{f_{1p}}{f_{1.}}$	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_i	$\frac{f_{i1}}{f_{i.}}$	…	$\frac{f_{ij}}{f_{i.}}$	…	$\frac{f_{ip}}{f_{i.}}$	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_m	$\frac{f_{m1}}{f_{m.}}$	…	$\frac{f_{mj}}{f_{m.}}$	…	$\frac{f_{mp}}{f_{m.}}$	1

Perfis-coluna

	B_1	…	B_j	…	B_p
A_1	$\frac{f_{11}}{f_{.1}}$	…	$\frac{f_{1i}}{f_{j}}$	…	$\frac{f_{1p}}{f_{.p}}$
⋮	⋮	⋮	⋮	⋮	⋮
A_i	$\frac{f_{i1}}{f_{.1}}$	…	$\frac{f_{ij}}{f_{j}}$	…	$\frac{f_{ip}}{f_{.p}}$
⋮	⋮	⋮	⋮	⋮	⋮
A_m	$\frac{f_{m1}}{f_{.1}}$	…	$\frac{f_{mi}}{f_{j}}$	…	$\frac{f_{mp}}{f_{.m}}$
Total	1	…	1	…	1

A distância entre dois pontos-linha ou pontos-coluna é dada pela distância do qui-quadrado entre os perfis. Esta distância permite evitar que categorias mais frequentes tenham maior peso. A distância do qui-quadrado entre dois pontos perfis-linha i e i' é dada por:

$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} (\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}})^2$, sendo que $(\frac{f_{ij}}{f_{i.}})$ corresponde ao perfil da linha i. Analogamente, a distância entre dois pontos perfis-coluna j e j' é dada por:

$d^2(j,j') = \sum_{i=1}^{m} \frac{1}{f_{i.}} (\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}})^2$, sendo que $(\frac{f_{ij}}{f_{.j}})$ corresponde ao perfil da coluna j.

A AC pode ser vista como uma ACP (Análise em Componentes Principais) sobre a matriz dos perfis-linha (ou perfis-coluna). A ACP aplica-se a variáveis quantitativas e tem como objetivo identificar novas variáveis, não correlacionadas, que melhor explicam a dispersão de um conjunto de dados. Cada nova componente principal é uma combinação linear das variáveis originais, de variância máxima, e não correlacionada com as componentes principais obtidas anteriormente.

2.2.2 Interpretação

Para realizar a AC é necessário centrar os dados e analisar a matriz de variâncias e covariâncias. Os valores próprios (λ_α) desta matriz medem a variância ao longo de cada eixo principal. Um valor próprio perto de 1 assegura uma boa representação ao longo do eixo (Lebart et al., 1998). A variância ou inércia representa uma percentagem explicativa da informação 'recuperada' por cada eixo. Mede a importância relativa de cada valor próprio relativamente à soma de todos os valores próprios e calcula-se por: $\frac{\lambda_\alpha}{\sum_\alpha \lambda_\alpha}$.

Consideram-se ainda dois outros coeficientes que contribuem para a interpretação dos resultados — as contribuições absolutas e as contribuições relativas. As contribuições absolutas (CTA) medem a contribuição de cada indivíduo (linha ou coluna) para a formação de cada um dos eixos. São considerados relevantes os indivíduos que apresentam um CTA acima da média. A contribuição de um ponto linha i cujas coordenadas no eixo α são $\Psi_{\alpha i}$ é dado por: $CTA_\alpha(i) = \frac{f_{i.}\psi^2_{\alpha i}}{\lambda_\alpha}$. A soma das contribuições absolutas para um eixo α dos pontos das linhas i é igual a 1: $\sum_{i=1}^{m} CTA_\alpha(i) = 1$. Analogamente, a contribuição da coluna j para a variância do eixo α é dada por: $CTA_\alpha(j) = \frac{f_{.j}\varphi^2_{\alpha j}}{\lambda_\alpha}$, sendo que $\varphi^{\alpha j}$ é a coordenada da coluna j no eixo α e $\sum_{j=1}^{p} CTA_\alpha(j) = 1$.

As contribuições relativas (CTR) medem a qualidade de representação de cada elemento em cada eixo, representadas por $cos^2_\alpha(i)$ no caso dos elementos das linhas e por $cos^2_\alpha(j)$ no caso dos elementos das colunas. Uma variável ou um indivíduo consideram-se bem representados no eixo ou no plano se a sua CTR for superior a 0.5. A CTR para o ponto i é dada por: $cos^2_\alpha(i) = \frac{d^2_\alpha(i,G)}{d^2(i,G)} = \frac{\Psi^2_{\alpha i}}{d^2(i,G)}$, sendo $d^2_\alpha(i,G)$ o quadrado da distância do ponto i ao centro de gravidade, no eixo α: $d^2(i,G) = \sum_{j=1}^{p} \frac{1}{f_{.j}}(\frac{f_{ij}}{f_{.i}} - f_{.j})^2$. Note-se que a soma das contribuições de todos os eixos para um elemento é igual a 1: $\sum_\alpha cos^2_\alpha(i) = 1$. Analogamente, $cos^2_\alpha(j) = \frac{d^2_\alpha(j,G)}{d^2(j,G)} = \frac{\varphi^2_{\alpha j}}{d^2(j,G)}$ e $\sum_\alpha cos^2_\alpha(j) = 1$.

2.2.3 Considerações acerca dos Dados

O método da AC pode ser aplicado a qualquer matriz de dados, desde que as entradas sejam não-negativas. A matriz de dados implícita para a análise é uma tabela de contingência. Hoffman e Franke (1986) constatam que a AC é apropriada para diversos tipos de dados, como dados nominais e ordinais, questões abertas e variáveis quantitativas discretizadas.

As possíveis aplicações da técnica de AC são ilimitadas mas Lebart et al. (1984) sugeriram que três condições deveriam ser satisfeitas para a AC ser mais eficiente. Uma delas consiste na dimensão da matriz de dados. Esta deve ser suficientemente grande para que a análise visual ou a simples análise estatística não permitam revelar a sua estrutura. Outra condição diz respeito à homogeneidade das variáveis, *i.e.*,

as variáveis deverão ser do mesmo tipo. Desta forma, torna-se possível determinar a distância entre linhas e colunas e obter informações significativas a partir das distâncias calculadas. Por fim, o método deve ser aplicado a matrizes de dados cuja estrutura é desconhecida ou difícil de compreender.

2.2.4 Aplicações

A Análise de Correspondências tem sido aplicada em vários domínios. Greenacre (1984) apresenta aplicações em Genética, Psicologia Social, Educação, Criminologia, Ciência Alimentar, Linguística, Ecologia, Paleontologia e Meteorologia. Diversas aplicações na área do Marketing são mencionadas por Hoffman e Franke (1986). Também são encontradas aplicações na Biologia, como a exploração das características do genoma de um organismo por Tekaia et al. (2002) e a avaliação das componentes de proteínas por Krah et al. (2004). Koutsoupias (2002) demonstrou as capacidades oferecidas pela AC através da aplicação a um estudo acerca do comportamento dos utilizadores e as suas preferências no acesso a um *web site*.

Em contexto de tarefas de visualização de texto, Morin (2004b) demonstra como usar a AC na recuperação de informação de resumos de relatórios internos de um centro de investigação em França. Em Morin (2006) esta técnica é utilizada para analisar dados textuais de publicações na área da educação. O método de AC também foi utilizado num estudo relativo ao desenvolvimento do Inglês como língua internacional realizado por Hassall e Ganesh (2005).

2.3 Análise Classificatória

Existem diversos métodos de classificação desenvolvidos em diferentes áreas. Nesta secção o tema Análise Classificatória (em inglês, *Clustering*) apenas será abordado no contexto de análise de dados textuais.

2.3.1 Conceitos Gerais

A Análise Classificatória é uma metodologia multivariada que permite classificar elementos (objetos ou variáveis), sendo o agrupamento geralmente alcançado a partir do cálculo das similaridades entre eles. De forma geral, a Análise Classificatória permite identificar grupos, ou classes (em inglês, *clusters*), de objetos similares (El-Hamdouchi e Willett, 1989). Assume diferentes nomes em áreas diferentes (Lee, 1981). Na Biologia é referida como Taxonomia enquanto que na área de reconhecimento de padrões é chamada de aprendizagem não supervisionada. Os métodos de classificação foram inicialmente desenvolvidos para o uso em ciências sociais (El-Hamdouchi e Willett, 1989). A partir daí começaram a ser utilizados noutras áreas

de aplicação, tais como Computação, Gestão de Operações, Reconhecimento de Padrões e extração de texto.

Dois tipos de Classificação têm vindo a ser estudados no contexto de sistemas de análise de textos: Classificação de documentos, com base nos termos que estes têm em comum e a Classificação de palavras, com base nos documentos em que estas aparecem (Willett, 1988).

A Classificação é muitas vezes implementada após a aplicação da Análise de Correspondências. Neste caso, retêm-se os primeiros q fatores, que expliquem uma parte importante da inércia. Para determinar o número de fatores a reter utiliza-se, frequentemente, o critério de *Pearson*, que consiste em reter as componentes que apresentem uma percentagem de inércia explicada de, pelo menos, 80%. Posteriormente, determinam-se as coordenadas dos documentos (e dos termos) nestes fatores, e efetua-se a Classificação dos documentos (e dos termos) com base nestas coordenadas.

2.3.2 Métodos de Classificação

Tal como na Análise de Correspondências, os métodos de Classificação podem ser aplicados a tabelas de contingência. É possível agrupar em classes o conjunto de colunas (usualmente constituídas por palavras e partes de textos) e o conjunto de linhas (geralmente constituídas por diferentes partes do texto) (Lebart et al., 1998). Os métodos de classificação dividem-se em dois grandes grupos: métodos Não-Hierárquicos ou de Partição e métodos Hierárquicos (Lebart et al., 1998; Greenacre, 1984; Willett, 1988; Jain et al., 1999; Steinbach et al., 2000). Os métodos Não-Hierárquicos ou de Partição determinam uma partição dos elementos em k classes, para k fixo, que otimize um critério de homogeneidade e/ou separação das classes. Assim, necessitam como *input* o número k de classes que se vai formar. Estes métodos adaptam-se melhor em aplicações que envolvam grandes conjuntos de dados para as quais a construção de um dendrograma é computacionalmente complexa. Existem diversas técnicas deste tipo, no entanto o algoritmo K-médias (em inglês, K-*means*) é o mais utilizado em classificação de documentos (Steinbach et al., 2000). Este algoritmo baseia-se na ideia de que um ponto central pode representar uma classe e utiliza a noção de centróide, que corresponde ao ponto médio de um conjunto de pontos. O método de K-médias atribui os elementos à classe com o centróide mais próximo. Para iniciar este algoritmo, um conjunto de k pontos é selecionado (representantes das classes ou centróides). De seguida, o método calcula a distância dos indivíduos aos centróides (representantes de cada classe) e afeta cada indivíduo ao centróide 'mais semelhante'. Os centróides das classes formadas são recalculados após cada iteração. O processo continua até não existirem alterações no conjunto de classes em duas iterações sucessivas.

Similar a este algoritmo surge o algoritmo 'Self-organized map' (SOM) ou 'Kohonen map' proposto por Kohonen em 1989 (*c.f.*, Lebart et al. (1984)) sendo consi-

derado relativamente simples e com a capacidade de organizar dados complexos em *clusters*, permitindo reduzir a dimensão do conjunto de dados. Os dados originais são representados em grelhas ou redes (*e.g.*, uma grelha com 5 linhas e 5 colunas representa 25 *clusters*.) O tamanho da grelha e o número de *clusters* são definidos *a priori*. Consideremos n pontos num espaço p-dimensional. Inicialmente a cada *cluster* k é atribuído um centro provisório C_k com p componentes (*e.g.*, escolhidos aleatoriamente ou entre os primeiros elementos). Para cada etapa h, o elemento $i(h)$ é atribuído ao centro mais próximo $C_k(h)$. A abordagem mais simples usa a distância Euclideana (Friedman et al., 2001). A diferença relativamente ao algoritmo de K-médias é a atualização dos centros. Na etapa *h+1*, este centro e os centros das classes vizinhas são modificados de acordo com a expressão $C_k(h+1) = C_k(h) + \epsilon(h)[i(h) - C_k(h)]$, onde, $\epsilon(h)$ é um parâmetro de adaptação que varia entre 0 e 1 e que é uma função decrescente de h. Tal como no algoritmo K-médias, a partição obtida depende dos centros escolhidos inicialmente. As técnicas de SOM apresentam grandes vantagens pois fornecem uma visualização bidimensional dos *clusters* a serem analisados, exigem menos esforço computacional, são bastante robustas à presença de dados com ruído e/ou *outliers* e não requerem que os grupos sejam previamente identificados. Assim, estas técnicas representam um compromisso entre as representações produzidas pelas técnicas dos eixos principais (ACP, AC, ACM)[1] e as técnicas de Classificação, pois apresentam algumas das vantagens de cada um dos métodos.

Em contraste com estes métodos, encontram-se os métodos hierárquicos, que permitem obter uma série de partições encaixadas. O resultado de um algoritmo de Classificação Hierárquica pode ser graficamente representado como uma árvore, denominada de dendrograma. Esta representação gráfica põe em evidência o processo de agrupamento e as classes intermédias. No topo encontra-se uma única classe que engloba todos os elementos conforme ilustrado na Figura 2.1. Se o número de elementos a agrupar é elevado torna-se difícil examinar a representação completa do dendrograma. Uma solução para ultrapassar este obstáculo consiste em *cortar* a árvore de acordo com o número de classes pretendidas. Um *corte* no dendrograma a qualquer nível produz uma classificação em k subgrupos ($1 \leq k \leq n$). Para selecionar o corte do dendrograma a efetuar é possível recorrer ao gráfico que relaciona o número de classes com a inércia intra-classes de cada partição — a partição ideal será dada pelo ponto onde ocorre um 'cotovelo'. A título de exemplo, na Figura 2.2 é possível observar que o 'cotovelo' da curva ocorre para K=2, *i.e.*, o dendrograma deveria ser 'cortado' em duas classes. Para complementar esta análise pode ainda recorrer-se ao conceito de inércia explicada que é dada por: *inércia explicada = inércia inter-classes / inércia total* ou $\Re^2 = SQC / SQT$, onde SQC é a soma de quadrados de desvios entre classes (*Sum of Squares Between Groups*) e SQT é

[1] Análise em Componentes Principais, Análise de Correspondências, Análise de Correspondências Múltiplas.

a soma dos quadrados totais (*Total Sum of Squares*). Esta medida, calculável para cada partição do dendrograma, indica-nos qual a percentagem da variabilidade total explicada pela partição (cada solução de número de classes). Esta medida está compreendida entre 0 e 1, em que valores próximos de 1 representam uma boa solução em classes homogéneas e bem separadas.

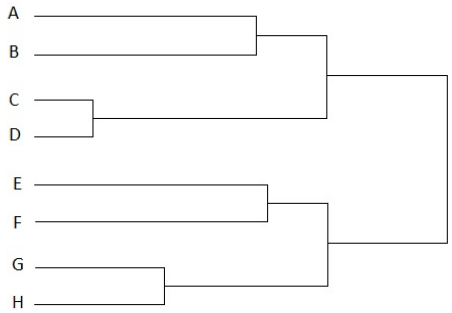

Figura 2.1: Dendrograma representando uma Análise Classificatória num conjunto de oito elementos (Lebart et al., 1998).

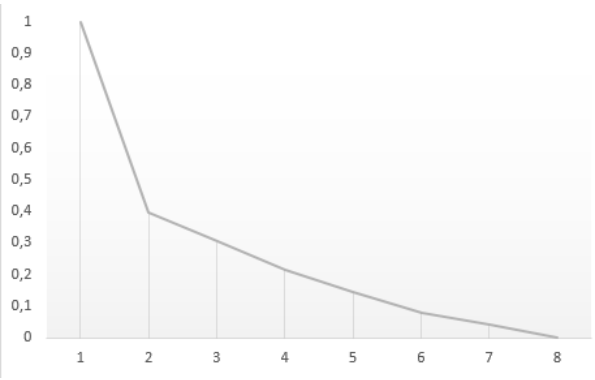

Figura 2.2: Gráfico que relaciona as 8 partições (entre 1 e 8 classes) com a inércia intra-classes de cada uma.

Na literatura encontram-se duas abordagens básicas relativamente à Classificação Hierárquica: Aglomerativa ou Ascendente e Divisiva ou Descendente. A abordagem Descendente parte de uma única classe que inclui os n elementos. As classes são sucessivamente divididas em classes 'mais pequenas' até se obterem n classes, cada uma com um elemento. A abordagem Ascendente parte de n elementos agrupados em n classes, cada classe com 1 elemento. Reúnem-se as classes sucessivamente, identificando o par de classes mais semelhante a partir da matriz de similaridade, até se obter uma única classe. Enquanto que nos métodos aglomerativos o agrupamento

de dois elementos permanece até ao final do processo, nos métodos divisivos acontece justamente o contrário: uma vez que dois elementos são separados eles jamais voltarão a fazer parte da mesma classe. Lebart et al. (1998) aplica a abordagem ascendente a um conjunto de dados textuais e El-Hamdouchi e Willett (1989) referem que esta abordagem é a mais comum visto que os cálculos para uma Classificação Hierárquica Descendente são mais complexos. De uma forma geral, o algoritmo Ascendente de Classificação Hierárquica segue os seguintes passos (Steinbach et al., 2000):

1. Calcular a semelhança entre todos os pares de elementos, isto é, calcular a matriz de similaridade;
2. Agrupar as classes mais similares;
3. Atualizar a matriz de similaridade para determinar a semelhança entre a nova classe e as classes precedentes;
4. Repetir os passos *2* e *3* até restar apenas uma classe.

2.3.3 Medidas de (Dis)semelhança

Numa Análise Classificatória deve ser definida uma medida de (dis)semelhança para poder identificar elementos que sejam semelhantes e/ou elementos que sejam dissemelhantes. O grau de semelhança (dissemelhança) entre elementos vai depender da medida que é escolhida para avaliar essa semelhança (dissemelhança). Nas medidas de semelhança, grandes valores do índice representam elevada semelhança entre os elementos. Nas medidas de dissemelhança, grandes valores do índice representam afastamento entre os elementos. Uma medida de dissemelhança satisfaz algumas condições:

1. A dissemelhança entre dois elementos x e y tem de ser não negativa, isto é, $d(x, y) \geq 0$;
2. A dissemelhança entre dois elementos deve ser nula se e apenas se os dois elementos são idênticos, ou seja, $d(x, y) = 0$ se e apenas se $x = y$;
3. A dissemelhança tem que ser simétrica, ou seja, a dissemelhança entre x e y é a mesma que a dissemelhança entre y e x $(d(x, y) = d(y, x), \forall(x, y))$;

Se além das propriedades anteriores se verificar uma quarta condição, então a dissemelhança satisfaz as propriedades de uma medida de distância:

4. A medida tem que satisfazer a desigualdade triangular[2]: $d(x, z) \leq d(x, y) + d(y, z)$.

Huang (2008) propõe algumas medidas de similaridade para determinar o grau de semelhança entre elementos no contexto de classificação de documentos de texto: a distância Euclideana, similaridade do cosseno, o coeficiente de *Jaccard*, o coeficiente de correlação de *Pearson* e a divergência de *Kullback-Leibler*.

[2]A desigualdade triangular refere-se ao teorema que afirma que, num triângulo, o comprimento de um dos lados é sempre inferior ou igual à soma dos comprimentos dos outros dois lados

2.3.4 Métodos de Agregação

Nos métodos de Classificação Hierárquica é necessário escolher o método para determinar os pares de classes a serem agrupados (no caso da Classificação Hierárquica Aglomerativa) e as classes a serem divididas (no caso da Classificação Hierárquica Divisiva).

São encontrados na literatura diversos métodos de agregação, sendo que os mais simples e populares são o índice do mínimo (*single linkage*) e o índice do máximo (*complete linkage*) (Jain et al., 1999; Zhao et al., 2005; Willett, 1988). Estes dois algoritmos diferem na determinação da semelhança entre duas classes. No método de *single linkage*, a distância entre dois grupos é determinada pelos dois elementos mais próximos em classes diferentes. Este método também é denominado de método do Vizinho Mais Próximo (*nearest neighbor*, em inglês). Ao contrário, o método de *complete linkage*, ou método do Vizinho Mais Distante (*furthest neighbor*) usa a maior distância entre um par de elementos para definir a distância entre grupos.

O índice das Distâncias Médias entre Grupos (em inglês, *group average* ou *average link*), o índice da Mediana (*median linkage*), o índice do Centróide (*centroid linkage*) e o índice de *Ward* (*minimum variance method*) também são utilizados como métodos de agregação. Tal como o nome indica, o índice das Distâncias Médias entre grupos, consiste em considerar que a distância entre dois grupos é a média de todas as distâncias entre pares de elementos (um em cada grupo). O índice do Centróide toma a distância entre duas classes como sendo a distância entre os centros de gravidade, ou outros pontos considerados representativos (centróides). O índice de *Ward* define a dissemelhança entre duas classes A e B como o aumento de inércia quando passamos de A e B para A $\cup$ B. Este método tem provado ser altamente eficiente na formação de grupos (Greenacre, 2007). O objetivo deste índice é maximizar a inércia inter-classes, que mede a separação das classes, e minimizar a inércia intra-classes, que mede a homogeneidade das mesmas.

Capítulo 3

Estudo de um conjunto de notícias

Neste capítulo são aplicados os métodos atrás descritos a um conjunto de 227 notícias. Inicialmente será elaborada uma análise ao texto completo das notícias e posteriormente às entidades extraídas dessas mesmas notícias de modo a ser possível comparar resultados. Os programas utilizados para o efeito são o Dtm-Vic (*Data and Text Mining*: Visualização, Inferência, Classificação) (Anexo A) e o SPSS *Statistics*. Também será explicado sucintamente o processo de extração de entidades.

3.1 Descrição e análise dos dados - notícias

Os dados estudados são referentes a todas as notícias publicadas pela agência Lusa no dia 31 de Dezembro de 2010 e disponibilizadas pela SAPO Labs [1]. A seleção deste dia deve-se à expectativa de que haja uma maior diversidade de notícias e um menor número de repetições das mesmas, uma vez que é usual fazer-se um balanço dos acontecimentos mais relevantes (que foram notícias durante o ano) no último dia do ano. Cada uma das 227 notícias publicadas nesse dia é representada pela ordem de publicação, *i.e.*, um número de 1 a 227. Este conjunto de dados não só é constituído por pequenas notícias como também apresenta uma grande diversidade de temas. O conjunto compreende 34595 palavras sendo que 9660 delas são diferentes umas das outras[2].

A distribuição do número de palavras por notícia apresenta-se na Figura 3.1. Através deste gráfico é possível ver que cinco das notícias do conjunto de dados têm entre 0 a 25 palavras, outras cinco têm entre 25 a 50 e assim sucessivamente.

[1] http://labs.sapo.pt

[2] Numa fase de pré-processamento removeram-se números, símbolos (. , ; : ! () [ ] ' / \ + ? * @ º &) e as *stopwords* tais como 'a', 'ao', 'de', 'o', entre outras, através do programa `R` e de ferramentas do *Microsoft Word*, pois não iriam trazer informação relevante para a análise. Para uma análise mais cuidada, mantiveram-se palavras separadas por um hífen como 'primeiro-ministro', 'secretário-geral', etc.

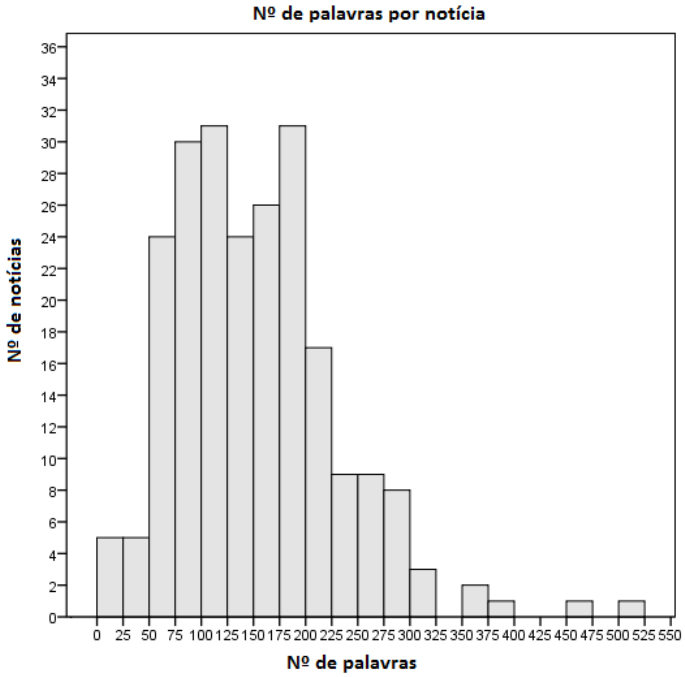

Figura 3.1: Distribuição do número de palavras por notícia.

3.1.1 Análise de Correspondências

Para aplicar o método da AC ao conjunto das 227 notícias foi utilizada a ferramenta Visutex do *software* Dtm-Vic. Esta ferramenta permite obter um resumo do conteúdo dos dados, como o número de notícias e de palavras existentes, o número de caracteres de cada palavra e a frequência de cada uma. Também possibilita a construção de tabelas de contingência e apresenta os resultados obtidos através da aplicação da AC, tais como os valores próprios, a inércia, as coordenadas, as contribuições e os planos de eixos principais. Para a aplicação dos métodos, o programa apenas mantém as palavras mais relevantes para o conjunto de dados. Para reter tais palavras usa o critério da frequência mínima. Esta frequência é decidida de modo a que o número de palavras distintas seja drasticamente reduzido. O objetivo é excluir as categorias raras, pois se são raras, não são, em princípio, pertinentes. Por defeito, o programa utilizou uma frequência mínima de 9. Com esta frequência, o número de palavras distintas retidas seria muito elevado (785) e, assim, a tabela de contingência seria muito esparsa. Por isso, decidiu-se optar por uma frequência mais elevada de forma a reter um menor número de palavras. Analisou-se o número de palavras retidas para alguns níveis de frequência, tal como se pode ver na Tabela 3.1, e decidiu-se optar por uma frequência igual a 35. É de notar que a escolha de outra frequência poderia gerar resultados diferentes. Com esta opção, eliminaram-se algumas palavras que não iriam ter grande relevância para a interpretação do con-

junto de dados — como a frequência máxima é de 241, palavras com uma frequência baixa não iriam ter grande impacto. Ao utilizar esta frequência retiveram-se 5458 palavras sendo que 87 delas são distintas. As palavras retidas são indicadas no Anexo B com as respetivas frequências. É possível observar que a palavra que surge mais vezes é a palavra 'hoje' com frequência igual a 241. Também se pode ver que o *software* diferencia entre maiúsculas e minúsculas uma vez que reteve as palavras 'Presidente' e 'presidente'.

Tabela 3.1: Número de palavras retidas para alguns níveis de frequência.

Frequência	Nº de palavras retidas	Nº de palavras retidas (distintas)
1	34.595	9.600
...	...	...
7	18.459	1.117
8	17.129	927
9	15.993	785
10	15.093	685
11	14.133	589
...	...	...
34	5.764	96
35	5.458	87
36	5.178	79
37	4.890	71
38	4.742	67

A tabela de contingência cruza 227 notícias com 87 palavras, indicando a respetiva frequência. Verifica-se que esta tabela é muito esparsa, *i.e.*, muitas das palavras não aparecem nenhuma vez numa determinada notícia[3]. Depois disto, é relevante identificar quais os eixos a reter para aplicar a Análise Classificatória. Obteve-se um histograma com os 226 eixos principais. Como é possível observar através do *output* parcial do histograma representado na Figura C.1, os valores próprios bem como as percentagens de inércia apresentam valores muito baixos. É possível constatar que à primeira componente está associado um valor próprio de 0,3893 o que corresponde a 6,01% da variância total, à segunda componente está associado um valor próprio de 0,3465 o que corresponde a 5,35% da variância total, à terceira componente está associado um valor próprio de 0,3255 o que corresponde a 5,02% da variância total, e assim sucessivamente, até explicarmos 100% da variância total. Estes valores baixos devem-se ao facto do conjunto de dados aqui analisado ser muito disperso, ou seja, é constituído por muitas palavras diferentes. Para reter os fatores importantes para a análise considerou-se utilizar o critério de *Pearson* que consiste em reter os eixos que apresentam conjuntamente uma percentagem de inércia explicada de, pelo menos,

[3]Decidiu-se não apresentar a tabela de contingência nesta dissertação por ser muito longa.

80%, o que corresponde a 40 eixos. No entanto, o *software* só consegue 'guardar' 30 coordenadas, que são essenciais para aplicar a Análise Classificatória no programa SPSS. Assim, decidiu-se reter os primeiros 30 eixos que explicam 70,11% da inércia total, que, de acordo com Naito (2007), já é uma proporção de inércia aceitável visto que se deve manter um número suficiente de eixos de modo a explicar uma proporção de inércia superior a 50%.

O próximo passo será estudar as palavras e notícias que mais influenciam a formação de cada um dos eixos. Devido ao elevado número de eixos e aos valores reduzidos da inércia, o estudo apenas será realizado entre os eixos 1 e 2 e entre 1 e 3 como exemplo de algumas correspondências.

Primeiro eixo principal

A percentagem de inércia explicada pelo primeiro eixo é de 6,01%. Nesta fase pretende-se saber quais as palavras e notícias que mais contribuem para a formação do eixo principal (neste caso, o primeiro eixo) e se têm coordenadas positivas ou negativas. A contribuição absoluta de um indivíduo para a formação de um eixo principal, isto é, para a variância explicada pelo eixo, permite evidenciar os indivíduos que apresentam caraterísticas relacionadas com o fenómeno traduzido pelo eixo principal que lhe corresponde. Costuma usar-se o critério que consiste em escolher os elementos de forma a que a soma das contribuições absolutas (CTA) seja aproximadamente igual a 80%. Basicamente, as palavras consideradas relevantes são aquelas que apresentam um CTA acima da média. Assim, as palavras que mais contribuem para a formação deste eixo são 'Social', 'ano', 'cento', 'euros', 'mil', 'milhões', 'pontos', com coordenadas positivas no primeiro eixo e 'Brasil', 'Gbagbo', 'Itália', 'Presidente', 'Silva', 'decisão', 'ministro', 'país' com coordenadas negativas. Também é necessário estudar a qualidade de representação dos pontos através das suas contribuições relativas (CTR). Uma variável ou um indivíduo consideram-se bem representados no eixo ou no plano se a sua contribuição relativa for superior a 0.5. Neste caso, este critério não foi utilizado devido à dispersão dos dados. Por isso, utilizou-se como critério a média. De acordo com este, todas elas estão bem representadas no primeiro eixo. Quanto às notícias é possível ver que as mais relevantes para a formação deste eixo são as 3, 10, 91, 120, 129, 133, 162, 174, 181, 201, 216, 221 e 222, com coordenadas positivas, e as 11, 31, 40, 41, 56, 57, 66, 88, 89, 94, 99, 100, 101, 108, 117, 137, 173, 182, 191, 206, 223 e 224 com coordenadas negativas pois têm uma contribuição absoluta acima da média. Não é possível identificar um tema relativamente às notícias com coordenadas positivas. Por isso, só se pode dizer que o eixo separa as notícias com as palavras 'Social', 'ano', 'cento', 'euros', 'mil', 'milhões', 'pontos', de notícias relacionadas com a Política — 'Brasil', 'Gbagbo', 'Itália', 'Presidente', 'Silva', 'decisão', 'ministro', 'país' — tal como se pode visualizar no quadro resumo da Figura 3.2 e no plano [1,2] da Figura 3.3. Para uma melhor visualização do plano foram atribuídos *ranks* a cada um dos pontos das palavras e

notícias. A função 'ranks' transforma as coordenadas das observações (notícias) e das categorias (palavras) em *ranks*, *i.e.*, em cada eixo, os N valores numéricos são organizados e substituídos pelos seus *ranks*. Ao valor mais pequeno é atribuído um *rank* igual a 1, a seguir é atribuído o número 2 e assim sucessivamente até que a observação com o valor mais alto no eixo tenha ordem N. Assim, uma escala aritmética substitui a original, fazendo com que as distribuições sejam fortemente distorcidas em distribuições uniformes. Esta alteração da escala permite manter a ordem dos elementos em cada eixo.

Eixo 1						
+			-			
			Política			
Palavras	Notícias		Palavras	Notícias		
Social	3	162	Brasil	11	89	173
ano	10	174	Gbagbo	31	94	182
cento	91	181	Itália	40	99	191
euros	120	201	Presidente	41	100	206
mil	129	216	Silva	56	101	223
milhões	133	221	decisão	57	108	224
pontos	137	222	ministro	66	117	
			país	88	137	

Figura 3.2: Quadro resumo - Eixo 1.

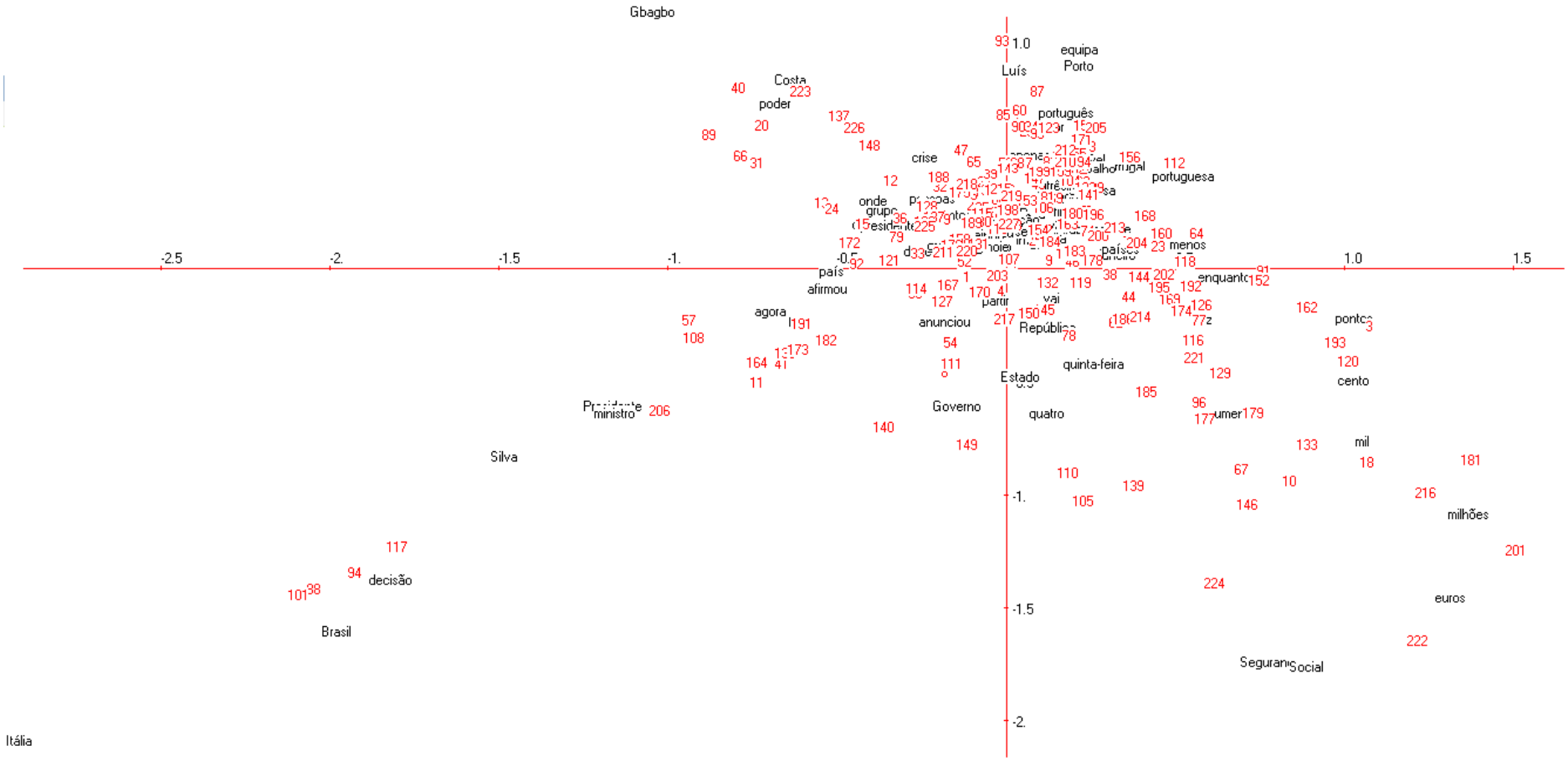

Figura 3.3: Notícias e palavras representadas de acordo com o *ranking* no plano [1,2] .

Segundo eixo principal

No segundo eixo a percentagem de inércia explicada é de 5,35%. As palavras e notícias que mais se destacam na formação deste eixo estão apresentadas na Figura 3.4. Como se pode observar, as palavras 'Costa', 'Gbagbo', 'Luís', 'Porto', 'anos' e 'equipa' têm coordenadas positivas no eixo, enquanto que as palavras 'Brasil', 'Governo', 'Itália', 'Segurança', 'Silva', 'Social', 'cento', 'decisão', 'euros', 'mil' e 'milhões' têm coordenadas negativas. Comparando com o primeiro eixo, é possível observar que a palavra 'Gbagbo' passou a ter coordenada positiva. O aparecimento da palavra 'Costa' sugere que dentro do tema Política surge um tema mais específico — Política na Costa do Marfim. Além deste, é possível identificar outros temas. Com coordenadas positivas, existem algumas notícias sobre Desporto, mais especificamente sobre Futebol, com as palavras 'Porto' e 'equipa'. Na Figura 3.4 acrescentou-se um grupo 'outros' que inclui notícias e palavras para as quais não foi possível identificar um tema específico. A palavra 'Porto' também está neste grupo pois não só aparece como clube, mas também como cidade. Com coordenadas negativas, existem algumas palavras e notícias já vistas no primeiro eixo sobre Política. Todas as notícias neste grupo são sobre a Política no Brasil e na Itália e, por isso, decidiu-se restringir um pouco o tema, ou seja, todas as notícias neste grupo são sobre Política Internacional. Para além disto, também se obtém um novo grupo sobre o Governo Português com as palavras 'Governo' e 'Segurança Social' como as mais relevantes. A palavra 'euros' está associada a este tema, mas também aparece em muitas das notícias onde não foi possível identificar um tema em comum.

Assim, o segundo eixo opõe notícias sobre a Política na Costa do Marfim (40, 66, 89, 137, 148, 223 e 226), sobre Desporto (80, 85, 87, 151 e 171), entre outras, a notícias relativas à Política Internacional (88, 94, 99, 100, 101, 117 e 140), ao Governo Português (105, 109, 110, 138, 139, 177, 222 e 224) e a outras.

Eixo 2											
+							-				
Política - Costa do Marfim		Desporto - Futebol		outros			Política Internacional		Governo Português		outros	
Palavras	Notícias	Palavras	Notícias	Palavras	Notícias		Palavras	Notícias	Palavras	Notícias	Palavras	Notícias
Costa	40	Porto	80	Luís	16	210	Brasil	88	Governo	105	euros	10
Gbagbo	66	equipa	85	Porto	17	212	Itália	94	Segurança	109	mil	133
	89		87	anos	25		Silva	99	Social	110	milhões	146
	137		93		45		decisão	100	euros	138	cento	149
	148		151		47			101		139		181
	223		171		58			117		177		185
	226				60			140		222		201
					209					224		216

Figura 3.4: Quadro resumo - Eixo 2.

Terceiro eixo principal

A percentagem de inércia explicada pelo terceiro eixo é de 5,02%. As palavras que mais contribuem para a formação do terceiro eixo são 'Brasil', 'cento', 'mil', 'milhões', 'países' e 'pontos', com coordenadas positivas, e 'Governo', 'Lusa', 'Segurança', 'Social' e 'quatro' com coordenadas negativas. A palavra 'Gbagbo' também tem uma contribuição absoluta acima da média mas está mal representada no eixo 3 pois tem uma contribuição relativa inferior à média de 0,02 (CTR médio= 0,04). O mesmo acontece com a notícia 40 com uma contribuição relativa de 0,02 (CTR médio = 0,03). Como é possível observar através da Figura 3.5, o tema sobre o Governo Português e sobre a Política Internacional mantém-se neste terceiro eixo. No entanto, as palavras relevantes para os dois temas alteraram-se — as do Governo Português aumentaram, não trazendo grande informação adicional relevante; as da Política Internacional diminuíram comparando com as do segundo eixo, mantendo-se apenas a palavra 'Brasil'. Além destes, não é possível identificar mais nenhum tema em comum entre as notícias, uma vez que estas estão representadas por palavras com pouca informação. O plano [1,3] da Figura 3.6 permite visualizar quais as notícias e palavras que o eixo opõe.

Eixo 3						
+					-	
Política Internacional					Governo Português	
Palavras	Notícias	Palavras	Notícias		Palavras	Notícias
Brasil	99	cento	3	162	Governo	105
	100	mil	10	181	Segurança	109
	101	milhões	77	192	Social	110
	117	países	91	201	Lusa	138
		pontos	120	214	quatro	139
			129	216		222
			149	221		224
			150			

Figura 3.5: Quadro resumo - Eixo 3.

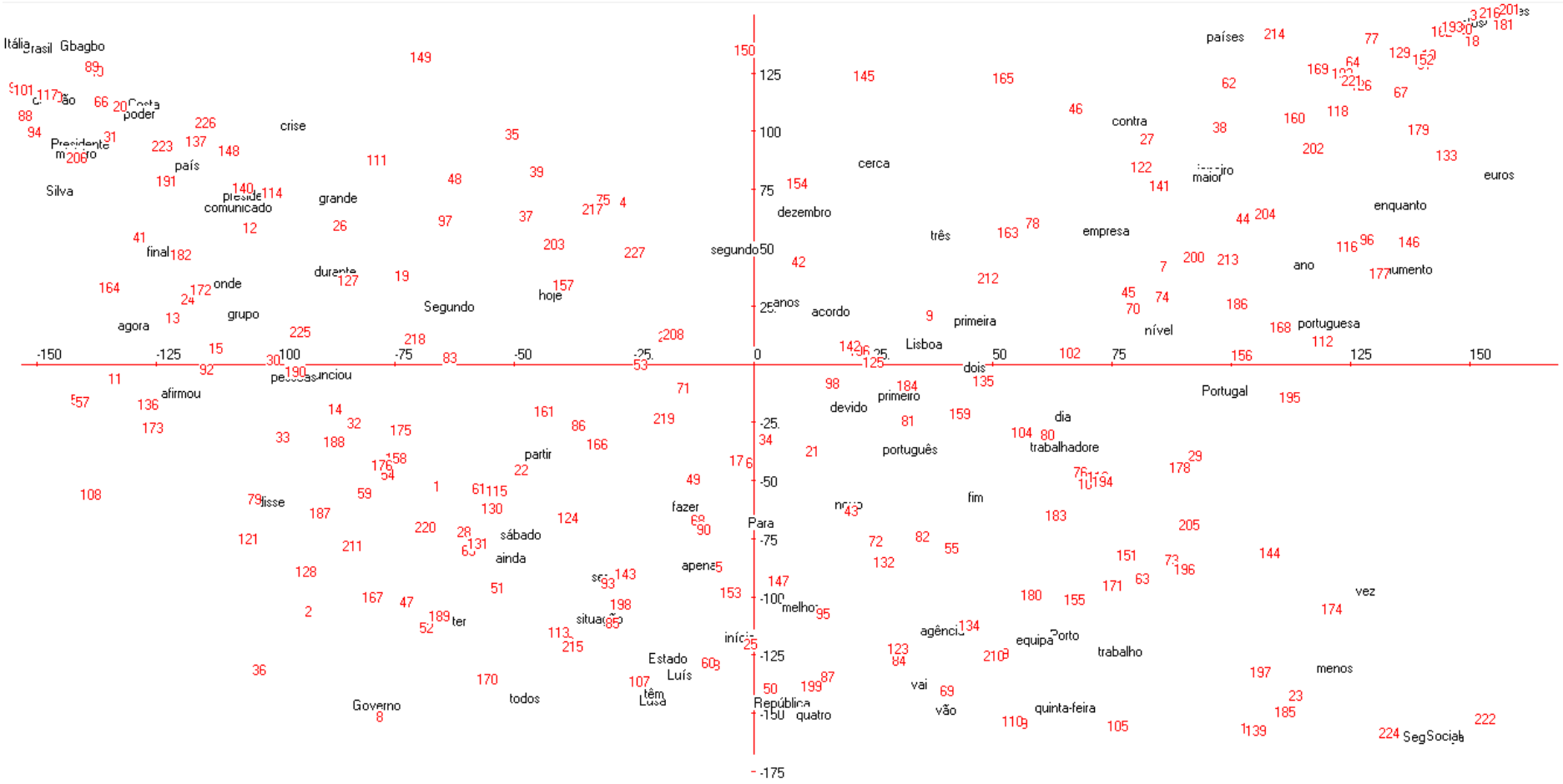

Figura 3.6: Notícias e palavras representadas de acordo com o *ranking* no plano [1,3].

3.1.2 Análise Classificatória

Nesta fase pretende-se agrupar o conjunto das notícias em classes recorrendo à classificação hierárquica e não hierárquica. Cada classe considerada deve ter notícias o mais semelhantes possível entre si, de forma a que seja homogénea. Para notícias de classes diferentes supõe-se que estas sejam o mais distintas possível. Visto que já se determinaram os fatores principais, as variáveis a utilizar serão as coordenadas nos 30 eixos retidos. O programa a utilizar para aplicar a classificação hierárquica e a classificação não hierárquica (*K-means*) será o SPSS *Statistics*. A análise através do mapa de *Kohonen* será realizada recorrendo ao *software* Dtm-Vic.

Classificação Hierárquica

Tal como foi visto na Secção 2.2 a classificação ascendente hierárquica parte em geral da matriz de proximidades entre indivíduos e agrega sucessivamente as classes em grupos homogéneos até à existência de apenas uma classe. Para a determinação desta matriz existem várias medidas de semelhança e dissemelhança à disposição. Decidiu-se utilizar como medida de dissemelhança o quadrado da distância Euclideana. A escolha desta medida para a construção da matriz deve-se ao facto de aumentar as distâncias elevadas, ressaltando a diferença entre classes. Como método escolheu-se o índice de Ward para evitar efeitos de cadeia obtendo classes compactas. Com base nestes parâmetros obteve-se o dendrograma da Figura 3.7. É visível a existência de três classes distintas. As notícias em cada uma das classes pode ser vista no Anexo D. A classe 1 é constituída por 180 notícias, a classe 2 por 38 notícias e a classe 3 por 9 notícias. Claramente a classe 3 agrupa as notícias sobre Política na Costa do Marfim. Devido ao elevado número de notícias nas outras duas classes não é possível identificar uma caraterística em comum entre elas. Desta forma, decidiu-se analisar outras partições para determinar qual o corte que define a partição apropriada para definição do número de classes. Assim, efetuou-se o cálculo da inércia intra-classes[4] de modo a construir o gráfico que permite visualizar a curva e definir assim a partição mais adequada — Figura 3.8.

No entanto, através da visualização do gráfico também não é possível identificar um ponto de destaque. Para ultrapassar o problema, calculou-se a inércia explicada[5] para várias partições como se pode ver na Tabela 3.2. Os valores da inércia explicada têm uma tendência crescente. Recordando que a inércia explicada é o rácio entre a inércia inter-classes e a inércia total e que a inércia inter-classes mede a separação das classes, quanto maior for o número de *clusters*, mais elevada será a inércia explicada. Para 227 classes a inércia explicada é igual a 1 pois cada notícia é considerada um *cluster*.

[4] A inércia intra-classes foi calculada através das tabelas ANOVA geradas a partir do SPSS.
[5] A inércia explicada foi calculada através das tabelas ANOVA geradas a partir do SPSS.

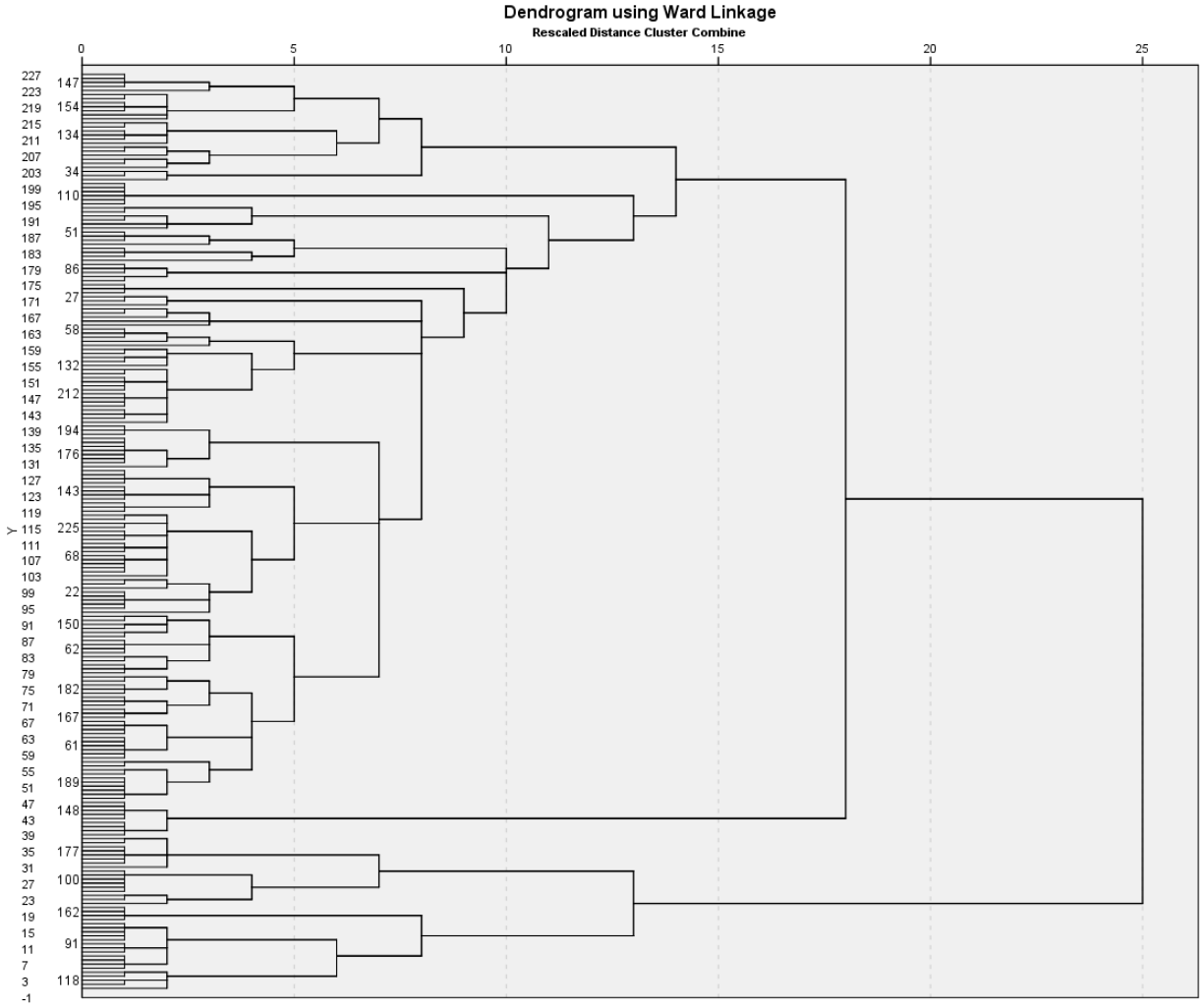

Figura 3.7: Representação através de um dendrograma da classificação hierárquica ascendente aplicada às 227 notícias descritas pelas 30 coordenadas fatoriais.

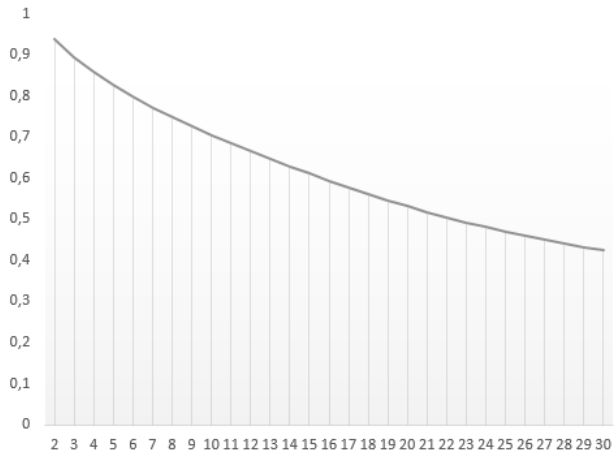

Figura 3.8: Inércia intra-classes para as partições em 2,..., 30 classes.

Tabela 3.2: Inércia explicada para as partições com 2 a 30 classes.

Nº de classes	Inércia explicada	Nº de classes	Inércia explicada
2	0,062062193	17	0,422675505
3	0,106251514	18	0,438498078
4	0,140310329	19	0,454215235
5	0,171054545	20	0,468505083
6	0,200916945	21	0,482710815
7	0,227895095	22	0,495078729
8	0,252158612	23	0,507148791
9	0,274992046	24	0,518500775
10	0,296003318	25	0,529631937
11	0,315136995	26	0,540289796
12	0,334160777	27	0,550081603
13	0,352442772	28	0,559236979
14	0,370608556	29	0,567907719
15	0,38843881	30	0,576418081
16	0,405772551		

A inércia explicada para as partições com 23, ... , 30 classes indicam uma melhor homogeneidade e uma maior separação entre as classes relativamente à classificação em 2, 3, ... , 22 classes pois a inércia explicada é mais baixa nestas partições. Ao fazer uma análise mais pormenorizada à partição 23 (ver Anexo D) é possível identificar algumas classes que sugerem temas diferentes. Na classe 6 surge um novo tema — Casinos — constituído pelas notícias 6, 50, 51 e 130. A notícia 50 é um *outlier*. Talvez tenha sido incluída na classe por conter a palavra 'Lusa'. Na classe 11 observam-se dois temas, Política Internacional (88, 94, 99, 100, 101 e 117), já

identificado na AC, e Mercado Chinês (201 e 216). Estas são as únicas duas notícias no conjunto das 227 que falam sobre este tema. Também nesta classes se observa um *outlier* (18). Identificou-se uma classe (classe 12) sobre a Política na Costa do Marfim, tema também encontrado na AC, com as notícias 20, 31, 40, 66, 89, 137, 148, 223 e 226. Além destes, também surgem outros temas já vistos — Estado Português (classe 20 com as notícias 56, 57 e 108), Desporto (classe 22 com as notícias 80, 85, 95, 123, 134 e 156) e Governo Português (classe 23 com as notícias 105, 109, 110, 138, 139 e 224). Assim, foram identificados alguns temas associados a sete destas 23 classes. A Figura 3.9 refere esses temas.

Temas
n= 23
Casinos
Política Internacional
Mercado Chinês
Política na Costa do Marfim
Estado Português
Desporto
Governo Português

Figura 3.9: Quadro resumo - temas obtidos através da Classificação Hierárquica.

A consideração de mais classes (pelo menos até 30 classes) não traz alterações significativas aos temas identificados nas classes por isso decidiu-se passar ao estudo da Classificação Não Hierárquica com o objetivo de identificar mais temas ou apenas destacar aqueles já identificados.

Classificação Não Hierárquica

A classificação não hierárquica é um processo iterativo que permite a atribuição final de cada notícia a uma classe eventualmente diferente da que poderá ter sido considerada anteriormente na análise hierárquica.

- K-médias

Efetuou-se uma classificação não hierárquica por recurso ao método das K-médias para K=2 até K=30. Para duas classes, obtém-se 226 notícias na primeira classe e 1 notícia na segunda classe — 37. Esta notícia apresenta coordenadas fatoriais muito diferentes relativamente às restantes. Refere-se a uma ação judicial retirada pela Ensitel e das discussões que gerou nas redes sociais. Como não apresenta um tema em comum com as restantes, esta notícia vai continuar a aparecer isolada, formando uma classe. Para determinar o número ideal de classes recorreu-se ao cálculo da inércia explicada para K=2 até K=30 representada na Tabela 3.3.

Tabela 3.3: Inércia explicada para as partições em 2 a 30 classes.

Nº de classes	Inércia explicada	Nº de classes	Inércia explicada
2	0,01528925	17	0,359478468
3	0,038208737	18	0,361565078
4	0,075372391	19	0,386387081
5	0,09877052	20	0,419148304
6	0,119863325	21	0,414893362
7	0,178223763	22	0,454825572
8	0,185062289	23	0,442507788
9	0,207376223	24	0,464406998
10	0,253856081	25	0,479492948
11	0,26114362	26	0,480323245
12	0,304410847	27	0,499954123
13	0,271667191	28	0,489282416
14	0,315306212	29	0,525072525
15	0,351388775	30	0,533707267
16	0,337208944		

É possível identificar uma redução da inércia explicada para K igual a 28, o que sugere que esta partição não separa tão bem os elementos como a divisão em 27 classes (ver Anexo E). Nesta partição foram identificados alguns temas já vistos na Classificação Hierárquica — Governo Português (classe 2) com 6 elementos e Política na Costa do Marfim (classe 20) com 9 elementos. Na classe 4 encontram-se dois temas — Política Internacional (94, 99, 100, 101 e 117) e Mercado Chinês (201 e 216). Para além destas existem mais três notícias que não se enquadram nestes temas. Observa-se ainda uma classe com 3 notícias (56, 57 e 108). Estas são as únicas notícias onde as palavras 'Cavaco' e 'Silva' aparecem em conjunto diversas vezes. A notícia 71 também inclui este nome, mas apenas uma vez. Nesta partição observam-se oito classes com apenas uma notícia e uma classe com cem notícias. Com o aumento do número de classes, esta classe seria dividida em mais classes, mas iria continuar a existir uma classe com muitas notícias e a informação dos restantes seria perdida.

- Mapas de Kohonen

Além do método de K-médias também se utilizou o mapa de Kohonen. Este mapa permite visualizar notícias e palavras agrupadas em classes. É uma boa ferramenta de visualização pois permite ver, além das notícias, as palavras associadas às classes. Começou-se por construir um mapa 3x3. No entanto, devido à existência de muitas notícias e entidades não se identificou nenhum tema nas 9 classes formadas. Assim,

analisaram-se os mapas 4x4 (16 classes), 5x5 (25 classes) e 6x6 (36 classes)[6] que estão apresentados no Anexo F.

A classe 1 do mapa 4x4, constituído por 18 notícias e pelas palavras 'primeira', 'equipa', 'Porto' e 'Luís', diz respeito ao tema Desporto, desde futebol português (FC Porto, Benfica, Sporting e Vitória de Guimarães) até à divisão inglesa de futebol e basquetebol americano (NBA). No entanto, apresenta alguns *outliers* — 5, 25, 55, 58, 60 — pois as palavras 'Porto' e 'Luís' nem sempre estão associados ao Desporto. A classe 21 do mapa 5x5 é constituído por mais uma notícia relativamente ao anterior, notícia 90 sobre futebol. No mapa 6x6 este tema também aparece na classe 6, com menos duas notícias relativamente ao mapa 4x4 — a 55, que é um *outlier* e a 156, a única notícia nesta classe sobre NBA. Assim, esta classe continua a ser sobre Desporto, mais ligado ao Futebol. Assim, verificou-se uma melhoria relativamente a esta classe com o aumento do número de classes.

No mapa 4x4 identifica-se uma classe com 21 notícias e com as palavras 'presidente', 'poder', 'país', 'ministro', 'decisão', 'crise', 'afirmou', 'Presidente', 'Itália', 'Gbagbo', 'Costa' e 'Brasil' sobre Política Internacional. Esta classe também evolui ao longo dos mapas. No mapa 5x5 esta classe foi separada em dois, originando um outro tema — Política na Costa do Marfim — com as palavras 'poder', 'crise', 'Gbagbo' e 'Costa' e com as notícias 89, 66, 40, 31, 226, 223, 20, 148 e 137. A outra classe continua a ser sobre Política Internacional, mas com algumas notícias sobre o Estado Português — 56, 57 e 108. No mapa 6x6 a classe sobre a Política na Costa do Marfim surge novamente, e a última classe foi separada, excluindo as notícias sobre o Estado Português. Inclui as palavras 'decisão', 'Itália' e 'Brasil' e as notícias 99, 94, 88, 206, 117, 101 e 100.

Por último, foi identificado o tema relativo ao Governo Português. No primeiro mapa, esta classe (classe 13) ainda contém algumas notícias não relacionadas com o tema devido às palavras 'milhões', 'mil', 'euros' e 'cento'. A separação é feita no mapa 5x5, formando uma classe com 8 notícias e com as palavras 'quinta-feira', 'quatro', 'Social', 'Segurança', 'República' e 'Estado'. Como a palavra 'Governo' deixa de aparecer nesta classe, pode-se dizer que esta classe é sobre o Estado Português. Ainda inclui um *outlier*, a notícia 185, que deixa de aparecer no mapa 6x6 (grupo 31).

3.2 Extração de entidades

Na análise anterior, palavras como 'cento', 'mil', 'milhões', 'euros' surgiram no conjunto das palavras mais relevantes. Com o objetivo de contornar essas referências e focar a análise na informação relevante presente nos textos decidiu-se substituir o texto de todas as notícias pela lista das entidades citadas nas mesmas. As entidades, neste contexto, são palavras ou conjuntos de palavras e dizem respeito a

[6]Os mapas foram numerados para facilitar a identificação das classes na análise.

todo o nome próprio — seja de pessoas, cidades, países, clubes, etc. — ou todo o nome comum associado a determinada função ou cargo como, *e.g.*, deputado e presidente. A tarefa de extrair as entidades não foi alvo do trabalho desenvolvido nesta dissertação. Para a obtenção das listas de entidades associadas a cada notícia utilizou-se um programa desenvolvido, por Rocha et al. (2014), especificamente para extrair entidades citadas em textos escritos em português e que se encontra implementado em `R` (R Core Team, 2014). O processo de extração das entidades baseia-se na correspondência de padrões, na marcação da categoria morfo-sintática de cada palavra, em regras lexicais e na distância entre os nomes das entidades.

3.3 Descrição e análise dos dados - notícias e entidades

No conjunto das notícias foram extraídas 5028 referências a 2121 entidades distintas. A distribuição do número de entidades por notícia apresenta-se na Figura 3.10, *i.e.*, no conjunto de dados existe uma notícia com duas entidades, quatro notícias com três entidades cada e assim sucessivamente.

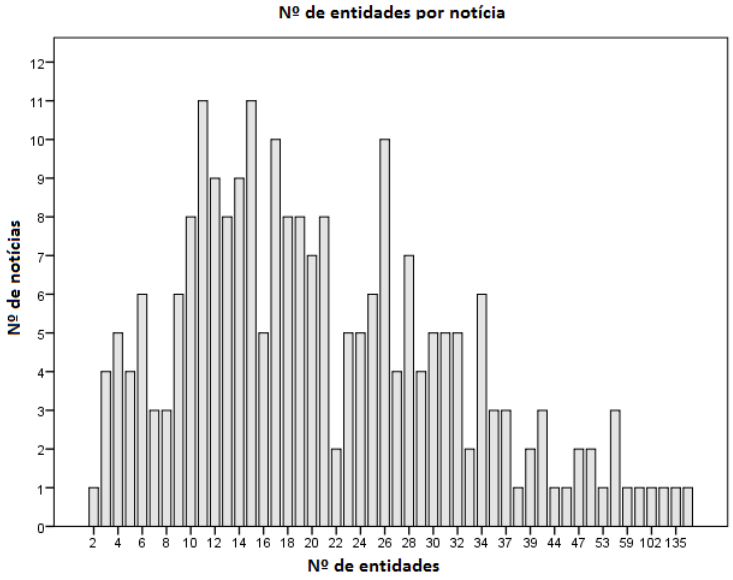

Figura 3.10: Número de entidades (5028) por notícia (227).

Como é possível observar na Tabela 3.4, o *software* apenas considera 2089 entidades distintas, quando na realidade existem 2121. Nesta tabela estão representadas as frequências observadas das entidades com um número de caracteres de 1 até '20 ou mais'. Como se pode constatar a frequência observada real e a frequência obtida através do programa Dtm-Vic apresentam valores iguais exceto para entidades com

20 ou mais letras. Esta diferença deve-se ao facto do programa não considerar palavras com mais de 20 caracteres. O programa retira as letras acima de 20, sendo que em alguns casos duas entidades diferentes são consideradas iguais. Um exemplo onde isto acontece pode ser facilmente observado na redução de caracteres das entidades 'Bombeiros*Voluntários'[7] e 'Bombeiros*Voluntários*de*Porto*de*Mós' onde são transformadas em 'Bombeiros*Voluntário'. No entanto, esta limitação não é impedimento para continuar a análise pois após a implementação dos métodos é possível reconhecer quais das entidades retidas sofreram redução de caracteres.

Tabela 3.4: Frequência das entidades de acordo com o número de caracteres

Nº de caracteres	Freq obs real	Freq obs Dtm-Vic
1	0	0
2	39	39
3	138	138
4	73	73
5	126	126
6	130	130
7	113	113
8	110	110
9	86	86
10	110	110
11	87	87
12	97	97
13	117	117
14	103	103
15	94	94
16	62	62
17	81	81
18	45	45
19	65	65
≥ 20	445	413
Total	2121	2089

3.3.1 Análise de Correspondências

Utilizou-se a ferramenta Visutex e foram retidas 1133 entidades, sendo que 50 delas são distintas. O programa manteve automaticamente as entidades que apresentam

[7]Sendo que o *software* faz a contagem das palavras existentes e o objetivo é saber o número de entidades, o símbolo '*' foi acrescentado a todas as entidades por forma a transformá-las em apenas uma palavra.

uma frequência igual ou superior a 12. Tal como no capítulo anterior, analisou-se o número de entidades retidas para alguns níveis de frequência mínima como se pode ver na Tabela 3.5 e optou-se por continuar a análise com esta frequência, pois considerou-se que 50 entidades retidas distintas já é um valor relevante para a análise.

Tabela 3.5: Número de entidades retidas para alguns níveis de frequência.

Frequência	Nº de entidades retidas	Nº de entidades retidas (distintas)
1	5028	2089
2	3790	851
3	2992	452
4	2521	295
5	2113	193
6	1888	148
7	1702	117
8	1541	94
9	1421	79
10	1331	69
11	1221	58
12	1133	50
13	1037	42
14	920	33
15	906	32

Na Tabela 3.6 é possível observar quais as entidades retidas e a frequência com que aparecem no conjunto das notícias. As frequências das entidades variam entre 12 e 95.

Tabela 3.6: Entidades retidas e respetivas frequências

Entidades	Frequência	Entidades	Frequência
África*do*Sul	20	Itália	31
Agência*Brasil	19	Laurent*Gbagbo	37
Agência*Lusa	13	Lisboa	36
Alassane*Ouattar	12	Lousã	12
Ano*Novo	20	Lusa	67
BPN	28	Moçambique	12
Benfica	21	ONU	12
Brasil	36	PS	13
Brasília	16	PSD	13
Caixa*Geral*de*Aposentações	12	PSI	14
Caixa*Geral*de*Depósitos	13	Porto	15

Cavaco*Silva	25	Portugal	68
Cesare*Battisti	34	Presidente	27
China	25	Presidente*Lula*da*Silva	13
Coimbra	16	Presidente*da*República	24
Costa*do*Marfim	24	primeiro-ministro	13
Diário*da*República	21	RN	15
Espanha	20	Reino*Unido	17
Estado	35	Rússia	13
Europa	16	SNGB	13
ex-ativista	16	Sara*Moreira	13
FC*Porto	17	Segurança*Social	21
França	12	Supremo*Tribunal*Federal	18
Governo	95	União*Europeia	26
Guarda	12	Varzim*Sol	12

Nesta lista de entidades é possível observar que várias delas têm uma frequência igual a 12, sendo elas: 'Alassane Ouattar', 'Caixa Geral de Aposentações', 'França', 'Guarda', 'Lousã', 'Moçambique', 'ONU' e 'Varzim Sol'.

A tabela de contingência é formada por 50 linhas e 227 colunas. Como é uma tabela formada por muitos zeros, optou-se por apresentar, para cada umas das 50 entidades retidas, o número da notícia e respetiva frequência (diferente de zero) para cada uma delas (no Anexo G). Torna-se agora fundamental identificar quais os eixos a reter para posteriormente aplicar a Análise Classificatória. Obteve-se um histograma com os 226 valores próprios. Como é possível observar através do *output* parcial do histograma representado no Anexo H, os valores próprios bem como as percentagens de inércia apresentam valores muito baixos. É possível constatar que à primeira componente está associado um valor próprio de 0,8706 o que corresponde a 5,17% da variância total, à segunda componente está associado um valor próprio de 0,8504 o que corresponde a 5,05% da variância total, à terceira componente está associado um valor próprio de 0,8121 o que corresponde a 4,83% da variância total, e assim sucessivamente, até explicarmos 100% da variância total. Tal como já foi explicado, estes valores baixos devem-se à existência de muitas entidades diferentes.

De acordo com o critério de *Pearson* retiveram-se os primeiros 24 eixos que explicam 81,72% da inércia total. O próximo passo será estudar as entidades e notícias que mais influenciam a formação de cada um dos eixos. Devido ao elevado número de eixos e aos valores reduzidos da inércia, o estudo apenas será efetuado entre os eixos 1 e 2 e entre os eixos 1 e 3 como forma de demonstração de algumas correspondências.

Primeiro eixo principal

A percentagem de inércia explicada pelo primeiro eixo é de 5,17%. Como é possível constatar através da Tabela I.1 no Anexo I, as entidades 'Benfica', 'Europa', 'FC Porto', 'PSI' e 'Sara Moreira' têm coordenadas negativas no primeiro eixo e as entidades 'Agência Brasil', 'Brasil', 'Cesare Battisti', 'Itália', 'Presidente Lula da Silva', 'Supremo Tribunal Federal' e 'ex-ativista' têm coordenadas positivas.

Além disto, procura-se também interpretar a contribuição de cada entidade para a formação de cada eixo. As entidades a negrito na Tabela I.1 são as que têm mais destaque na formação do primeiro eixo pois apresentam um CTA acima da média. Quanto à qualidade de representação, as entidades que apresentam uma CTR superior à média são consideradas bem representadas. À exceção de 'Benfica', 'FC Porto' e 'PSI' todas as entidades identificadas acima estão bem representadas no eixo 1.

Como podemos ver através do plano [1,2] representado na Figura 3.11, o primeiro eixo separa entidades relacionadas com a Política Internacional ('Agência Brasil', 'Brasil', 'Cesare Battisti', 'Itália','Presidente Lula da Silva', 'Supremo Tribunal Federal' e 'ex-ativista') de entidades relacionadas com o Desporto ('Sara Moreira' com a maior contribuição e 'Europa').

Para uma melhor visualização foram atribuídos *ranks* aos pontos apresentados no plano como se pode observar na Figura 3.12.

Analisando agora as notícias obtêm-se os resultados apresentados na Tabela I.2. As notícias a negrito são aquelas que mais contribuem para o primeiro eixo, sendo que as notícias 88, 94, 99, 100, 101, 117 e 140 têm coordenadas positivas enquanto que as notícias 16, 17, 29, 56, 57, 87, 91, 93, 108, 120, 151, 171, 181, 193 e 205 têm coordenadas negativas no eixo. É possível observar que as notícias 51 e 87 também têm uma contribuição na formação do primeiro eixo mas não estão bem representadas pois apresentam uma contribuição relativa inferior à média e igual a 0,02 (CTR média= 0,03).

Através da visualização do plano [1,2] da Figura 3.13 obtém-se uma conclusão semelhante ao estudo para as entidades. A única diferença é que além das notícias relacionadas com Desporto, este também engloba algumas notícias relacionadas com o Estado Português (56, 57 e 108) e com o Mercado Accionista (91, 120, 181, 193). Assim, o eixo 1 opõe notícias sobre a Política Internacional (88, 94, 99, 100, 101, 117, 140) com coordenadas positivas no eixo, às notícias ligadas ao Desporto (16, 17, 29, 93, 151, 171, 205), ao Mercado Acionista e ao Estado Português.

Através do plano de eixos principais representado na Figura 3.14 é possível visualizar as proximidades entre notícias e entidades. Esta proximidade também é visível no quadro resumo da Figura 3.15. A partir deste quadro também é possível reter quais os temas que o primeiro eixo separa.

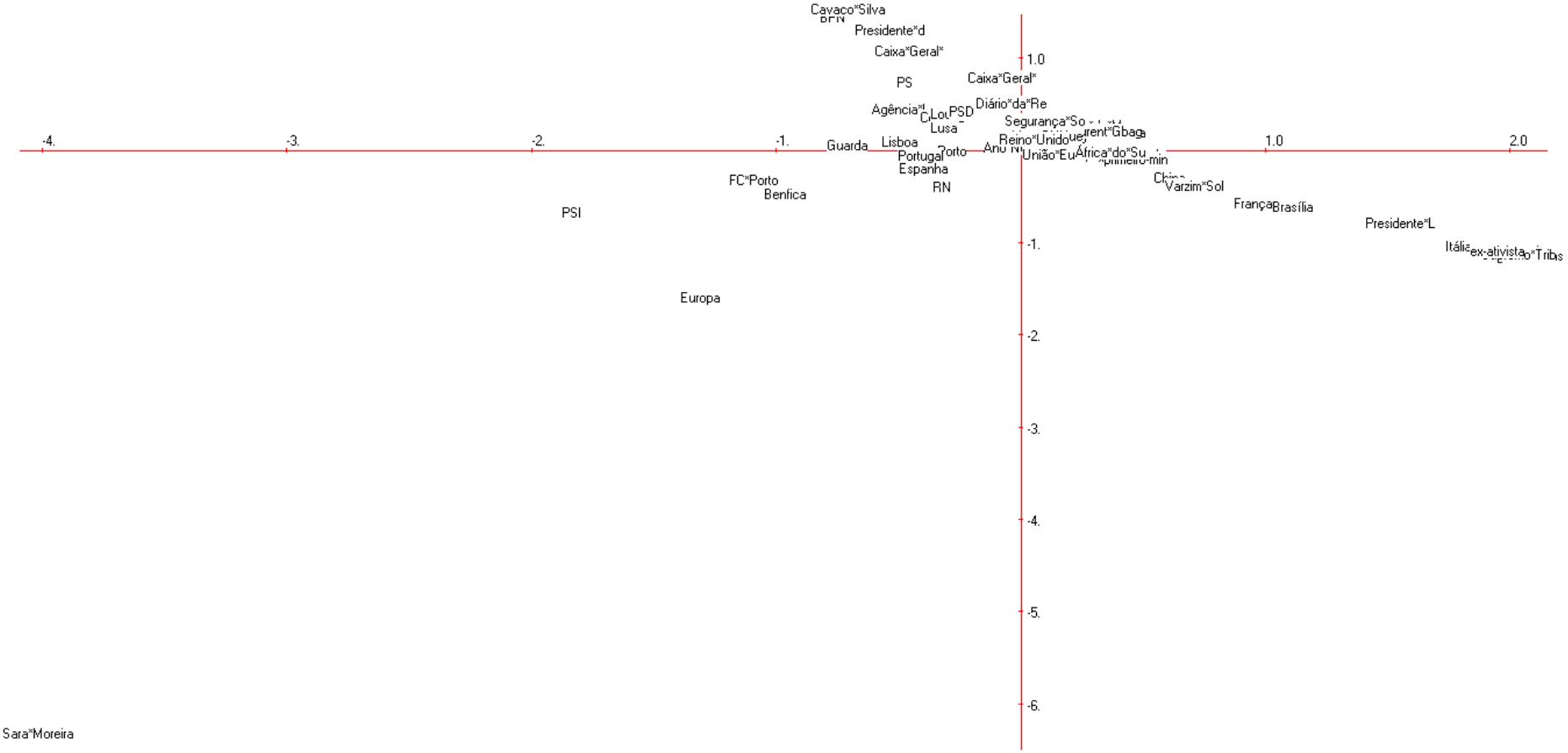

Figura 3.11: Entidades retidas representadas no plano [1,2].

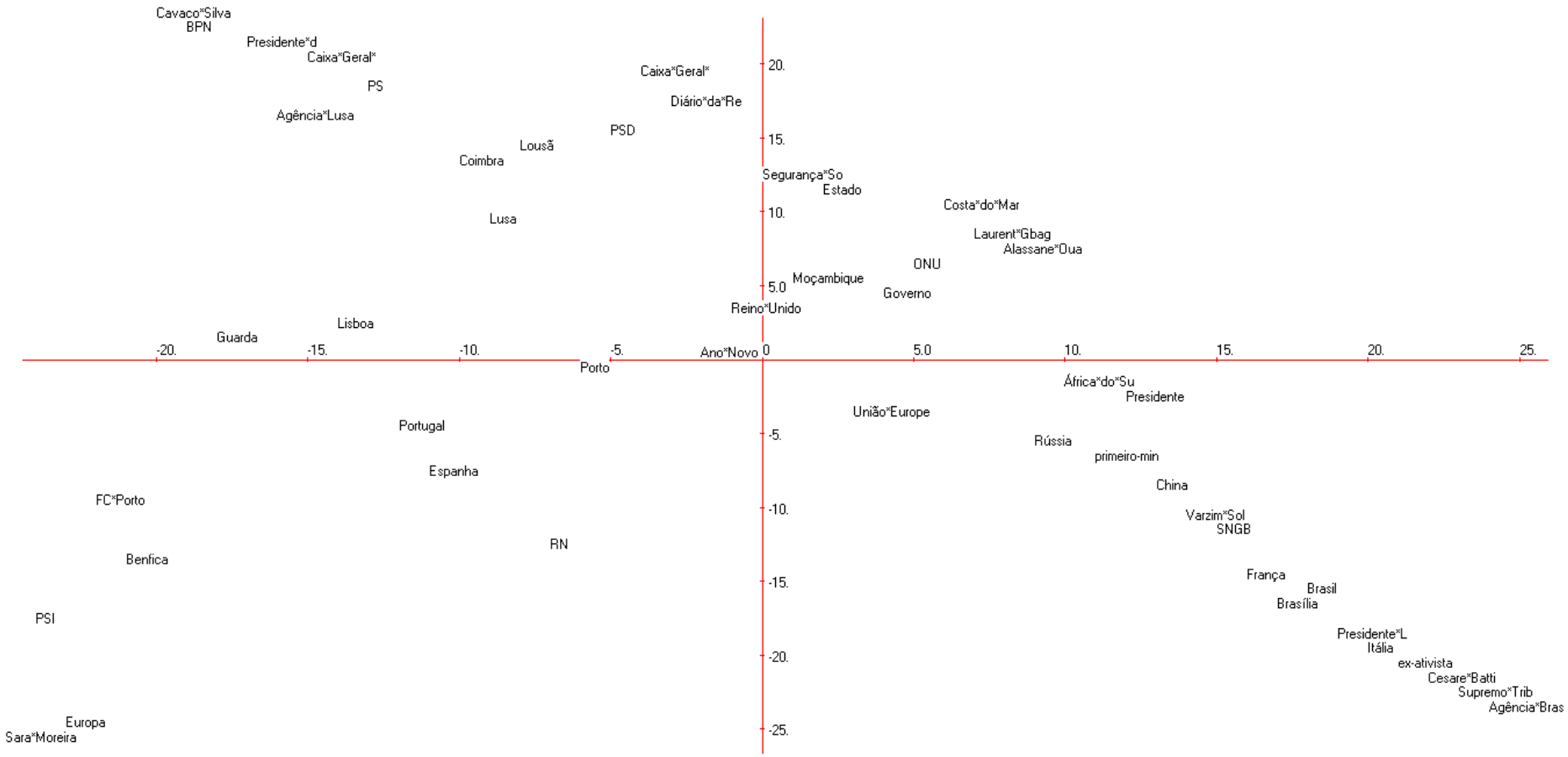

Figura 3.12: Entidades retidas representadas de acordo com o *ranking* no plano [1,2].

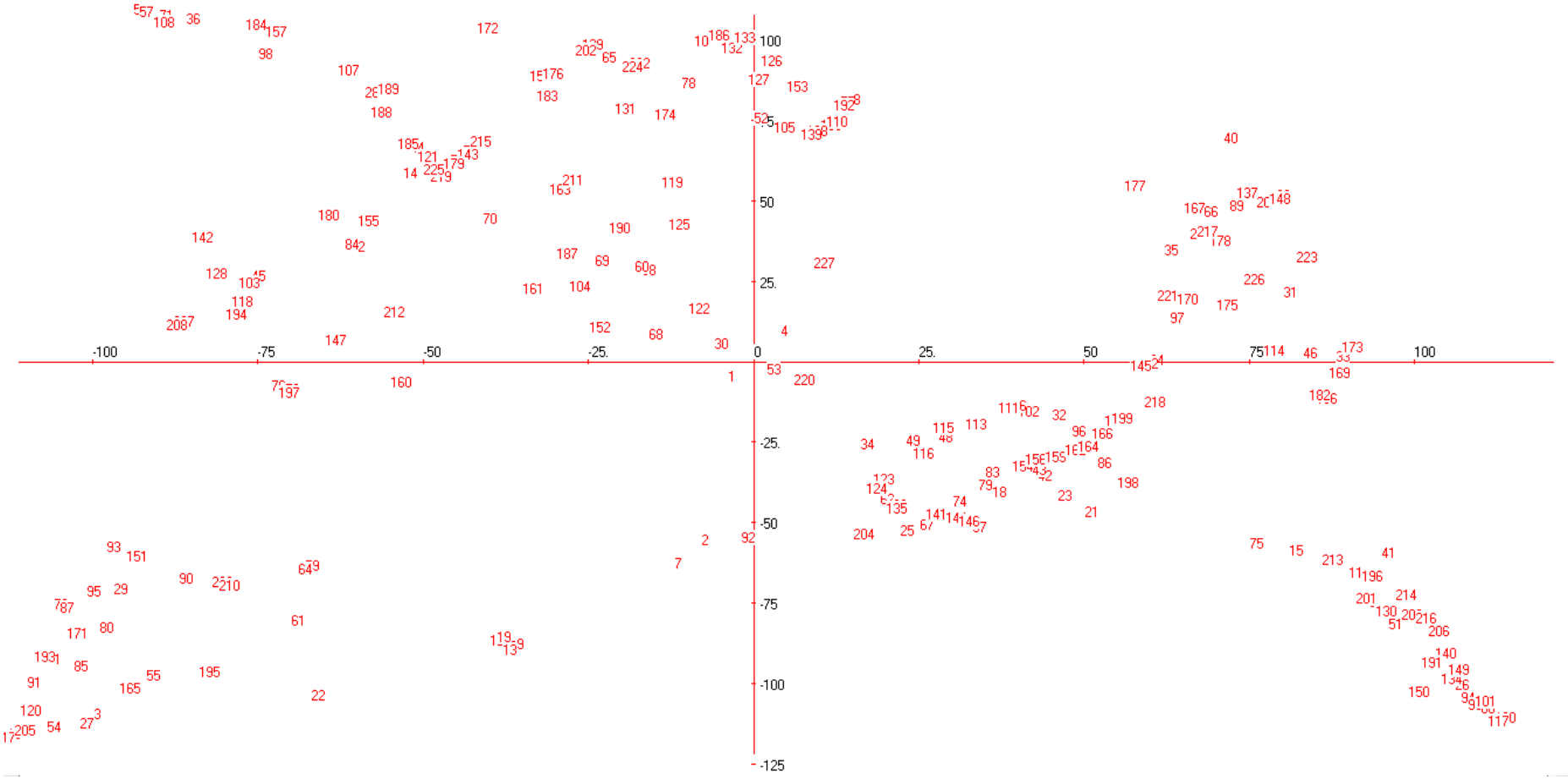

Figura 3.13: As 227 notícias representadas de acordo com o *ranking* no plano [1,2].

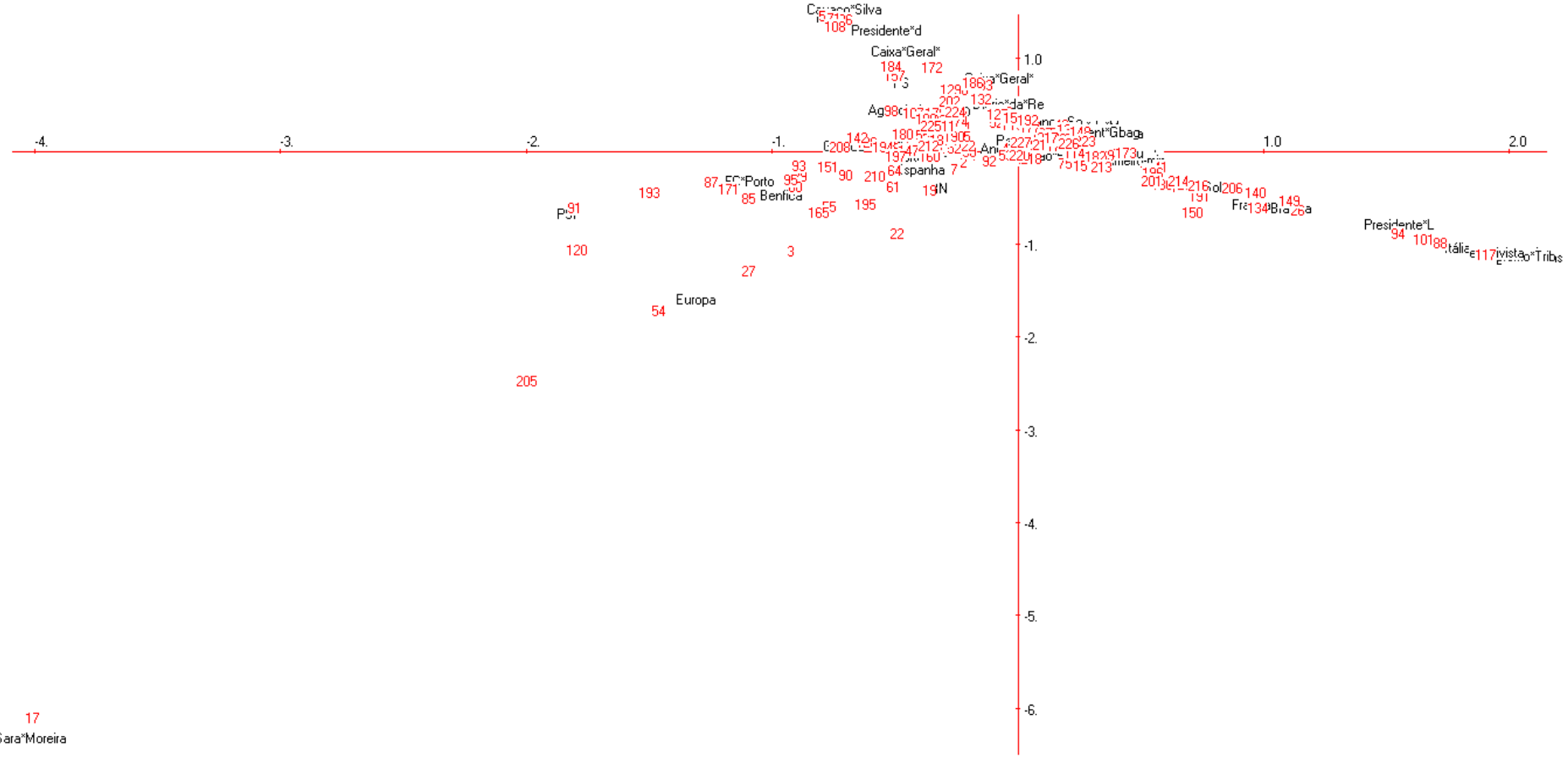

Figura 3.14: Notícias e entidades no plano [1,2]

Eixo 1					
+				-	
Política Internacional		Desporto		Mercado Acionista	Estado Português
Entidades	Notícias	Entidades	Notícias	Notícias	Notícias
Agência Brasil	88	Sara Moreira	16	91	56
Supremo Tribunal Federal	94	Europa	17	120	57
Cesare Battisti	99		29	181	108
ex-ativista	100		93	193	
Itália	101		151		
Presidente Lula da Silva	117		171		
Brasil	140		205		

Figura 3.15: Quadro resumo - Eixo 1.

Segundo eixo principal

A percentagem de inércia explicada pelo segundo eixo é de 5,05%. As entidades 'BPN', 'Cavaco Silva' e 'Presidente da República' têm coordenadas positivas no eixo 2 enquanto que as entidades 'Agência Brasil', 'Cesare Battisti', 'Europa', 'Itália' e 'Sara Moreira' têm coordenadas negativas. A Tabela I.3 mostra as coordenadas e as contribuições de cada uma das entidades. Aquelas a negrito são as que têm mais importância na formação do segundo eixo. Além disto, todas apresentam um CTR superior à média e por isso estão bem representadas no segundo eixo. Este eixo opõe entidades relacionadas com o Estado Português ('BPN', 'Cavaco Silva' e 'Presidente da República') de entidades relacionadas com o Desporto ('Sara Moreira' com a maior contribuição e 'Europa' tal como no primeiro eixo) e com a Política Internacional ('Agência Brasil', 'Cesare Battisti' e 'Itália').

Através da Tabela I.4 é possível observar quais as notícias que mais contribuem para a formação do eixo 2. O segundo eixo separa notícias sobre o Estado Português (56, 57, 71, 108, 184 e 186) com coordenadas positivas, de notícias ligadas ao Desporto (16, 17 e 205) e à Política Internacional (88, 94, 99, 100, 101 e 117) com coordenadas negativas.

A notícia 133 apresenta uma CTR abaixo da média e por isso não está bem representada no eixo em análise. As notícias 36 e 172 não estão incorporadas em nenhum destes temas. No entanto, têm destaque na formação deste eixo pois a entidade 'Presidente da República' aparece nestas notícias. Uma outra notícia que contribui para este eixo é a 27 com coordenadas negativa. Esta notícia surge pois apresenta a entidade 'Europa', não ligada com o Desporto.

Terceiro eixo principal

No terceiro eixo a percentagem de inércia explicada é de 4,83%. As entidades que mais se destacam na formação do terceiro eixo estão apresentadas na Tabela I.5 do Anexo I a negrito e são elas: 'Alassane Ouattar', 'Benfica', 'Costa do Marfim', 'FC Porto', 'Laurent Gbagbo', 'ONU', 'PSI', 'Presidente', 'SNGB', 'Sara Moreira' e 'Varzim Sol'. Todas elas estão bem representadas à exceção da entidade 'Benfica'

Eixo 2					
+				-	
Estado Português		Desporto		Política Internacional	
Entidades	Notícias	Entidades	Notícias	Entidades	Notícias
BPN	56	Sara Moreira	16	Agência Brasil	88
Cavaco Silva	57		17	Cesare Battisti	94
Presidente da República	71		205	Itália	99
	108				100
	184				101
	186				117

Figura 3.16: Quadro resumo - Eixo 2.

pois apresenta uma contribuição relativa inferior à média (média = 0,03). Através da segunda coluna da Tabela I.5 é possível constatar se uma determinada entidade tem coordenada positiva ou negativa no terceiro eixo.

Na Tabela I.6 observa-se quais as notícias que mais contribuem para a formação deste eixo. Ainda é possível constatar que as notícias 29 e 80 não estão bem representadas. A notícia 114 não se enquadra em nenhum dos temas identificados, apenas tem contribuição para a formação deste eixo pois a entidade 'Presidente' aparece diversas vezes ao longo desta notícia.

Tal como se pode ver no plano [1,3] representado na Figura 3.18 e no quadro resumo da Figura 3.17, o terceiro eixo opõe as entidades e notícias relacionadas com Casinos ('Varzim Sol' e 'SNGB'), com o Mercado Acionista ('PSI') e com o Desporto ('FC Porto') com coordenadas negativas com as entidades sobre Política na Costa do Marfim ('Alassane Ouattar', 'Costa do Marfim', 'Laurent Gbagbo' e 'ONU') e sobre Desporto, mais especificamente Atletismo ('Sara Moreira').

Eixo 3											
+						-					
Desporto - Atletismo		Política - Costa do Marfim				Casinos		Mercado Acionista		Desporto - Futebol	
Entidades	Notícias	Entidades	Notícias			Entidades	Notícias	Entidades	Notícias	Entidades	Notícias
Sara Moreira	16	Alassane Ouattar	20	137		Varzim Sol	6	PSI	91	FC Porto	87
	17	Costa do Marfim	31	148		SNGB	51		120		93
		Laurent Gbagbo	39	172			130		181		151
		ONU	40	173					193		171
		Presidente	66	223							
			89	226							
			136								

Figura 3.17: Quadro resumo - Eixo 3.

É possível concluir que a aplicação da Análise de Correspondências a este conjunto de dados permite visualizar através dos eixos principais as notícias e entidades que mais contribuem para a formação destes, permitindo identificar alguns temas. No entanto, como os dados são muito dispersos e, consequentemente, o número de eixos retido é elevado, a separação por temas em alguns eixos é pouco clara. Para complementar esta análise, aplica-se de seguida a Análise Classificatória de forma a obter grupos de notícias melhor definidos.

Figura 3.18: Notícias e entidades no plano [1,3].

3.3.2 Análise Classificatória

Esta secção irá seguir os mesmos procedimentos da Secção 3.1.2. As variáveis a utilizar serão as coordenadas nos 24 eixos retidos.

Classificação Hierárquica

Aplicou-se a classificação hierárquica ascendente ao conjunto de dados recorrendo ao quadrado da distância Euclideana e ao índice de Ward e obteve-se o dendrograma da Figura 3.19. São visíveis duas classes distintas — uma com 5 notícias e outra com 222 notícias (ver Tabela J.1). No entanto, não foi possível identificar qual a característica que as distingue. Tal como na análise anterior, devido ao elevado número de notícias, verificou-se particularmente difícil determinar através da visualização do dendrograma qual o corte ideal. Efetuou-se assim o cálculo da inércia intra-classes[8] de modo a obter o gráfico que permite visualizar a curva e definir a partição ideal para este conjunto de dados (ver Figura 3.20). Como se pode ver pelo gráfico, também não é possível identificar um ponto específico. Para ultrapassar esta situação, calculou-se a inércia explicada[9] tal como foi feito anteriormente (ver Tabela 3.7).

Tabela 3.7: Inércia explicada para as partições com 2 a 25 classes.

Nº de classes	Inércia explicada	Nº de classes	Inércia explicada
2	0,097694859	14	0,580238644
3	0,153653105	15	0,604652247
4	0,207484264	16	0,628646473
5	0,25767029	17	0,651518732
6	0,306692851	18	0,673250107
7	0,347906847	19	0,694649328
8	0,388485658	20	0,713848198
9	0,428673956	21	0,732658703
10	0,466350004	22	0,748981506
11	0,499697257	23	0,765045244
12	0,527910091	24	0,779912571
13	0,554584156	25	0,794135895

De acordo com a inércia explicada obtida para as partições em 2, 3,...,11 classes, os valores obtidos são relativamente baixos, o que significa que as classes não são muito homogéneas nem estão bem separadas. Os valores apresentados para as restantes partições já são aceitáveis. A inércia explicada para estas partições indicam uma melhor homogeneidade e uma maior separação entre as classes relativamente à classificação em 2,3,...,11 classes.

[8] A inércia intra-classes foi calculada através das tabelas ANOVA geradas a partir do SPSS.
[9] A inércia explicada foi calculada através das tabelas ANOVA geradas a partir do SPSS.

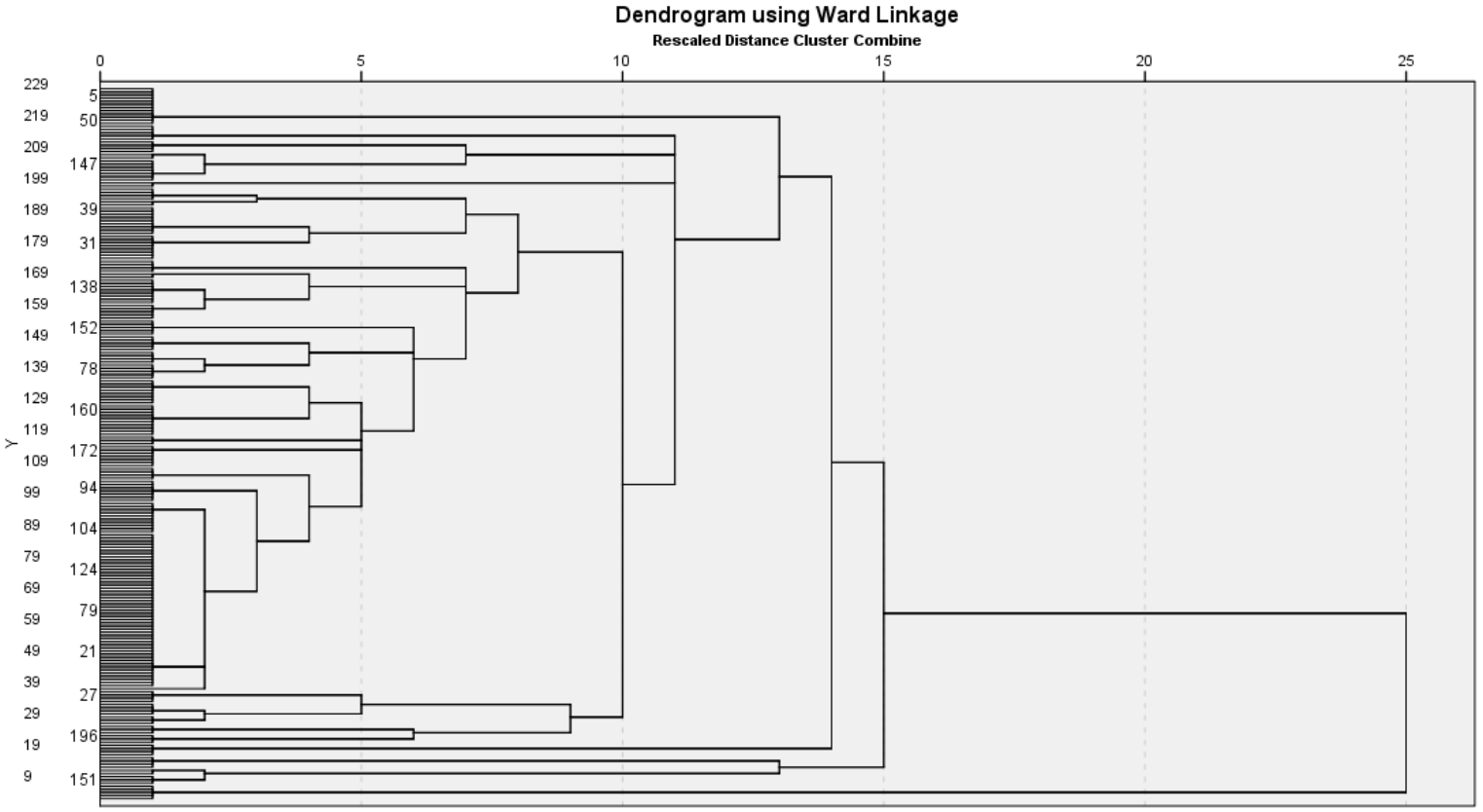

Figura 3.19: Representação através de um dendrograma da classificação hierárquica ascendente aplicada às 227 notícias e às 24 coordenadas fatoriais.

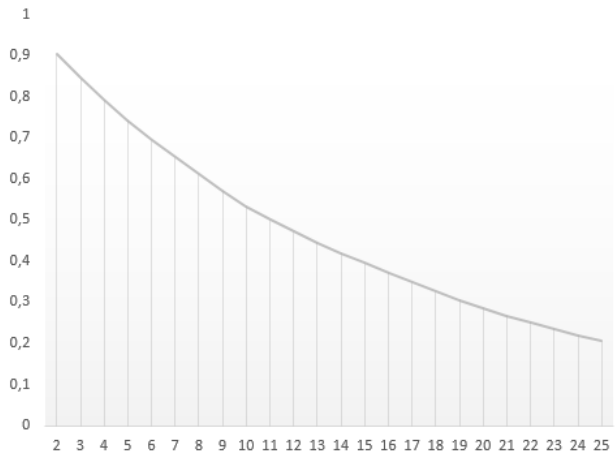

Figura 3.20: Inércia intra-classes para as partições 2 a 25.

Através da partição do dendrograma em 12 classes (ver Tabela J.2) é possível identificar alguns temas que já surgiram anteriormente na AC — grupo com 2 elementos (16 e 17) sobre a atleta Sara Moreira que apareceu na formação do eixo 3, grupo formado por 5 elementos com notícias sobre Desporto, mais especificamente sobre o FC Porto (73, 87, 93, 151 e 171) e grupo com 4 elementos (91, 120, 181 e 193) sobre o Mercado Acionista que também surgiu no terceiro eixo. Além destes, mais dois novos temas foram identificados na partição em 12 classes. Um grupo formado por 5 elementos (29, 61, 80, 85 e 95) sobre a entidade 'Benfica' e outro com 4 notícias (128, 142, 207 e 208) sobre o distrito da Guarda. Os elementos, e consequentemente os temas, das partições seguintes não iriam sofrer grandes alterações, por isso decidiu-se optar por analisar a partição 18, por ser uma partição com mais classes e, desta forma, alguns dos elementos foram alocados a grupos diferentes dos constituídos anteriormente para as outras partições. Obteve-se um grupo com 4 notícias (11, 46, 92 e 175) sobre a África do Sul. Reparou-se que esta entidade foi a que reuniu as notícias neste grupo. No entanto, através da tabela de contingência observou-se que para o grupo estar completo também teria de incluir as notícias 35 e 228. Também se encontrou um grupo, não completamente homogéneo visto que apresenta dois *outliers* (200 e 217), sobre o Governo Português com 15 notícias (105, 109, 110, 126, 132, 133, 138, 139, 153, 177, 186, 200, 217, 222 e 224). Neste grupo, tal como na situação anterior, não estão todas as notícias que falam sobre este tema. Na AC encontrou-se um tema que se denominou de 'Estado Português'. Estes dois temas estão relacionados, mas enquanto que o da AC está interligado com as entidades 'BPN', 'Cavaco Silva' e 'Presidente da República', este é mais abrangente e relaciona-se com as entidades 'Governo', 'Diário da República' e 'Segurança Social'. A classe 18 é formada por notícias acerca do mercado na China, nomeadamente sobre o comércio, cotação e PIB. Contém 4 elementos — 201, 213, 214 e 216. Decidiu-se também analisar as classes da partição 25 (ver Tabela J.3)

pela mesma razão apresentada acima. Comparativamente aos resultados já obtidos nesta análise, surgiu outro tema — Casinos — tema que já tinha sido identificado no terceiro eixo da AC. É formado por três elementos (6, 51 e 130). Outro grupo que se obteve foi o grupo 10, constituído maioritariamente por notícias sobre a Política na Costa do Marfim (20, 31, 40, 39, 66, 89, 136, 137, 148, 173, 223, 226), apesar de existirem algumas notícias que não se encaixam no assunto (24, 33, 53, 114 e 178). Este tema também já tinha sido encontrado na AC. Também se obtém um grupo de notícias referentes a Portugal com especial destaque nas cidades do Porto e Lisboa (45, 55, 68, 69, 72, 76, 82, 84, 103, 107, 118, 160, 163, 180, 194, 195 e 212). Identificou-se uma classe que agrupa as notícias sobre a Linha da Lousã (131, 158, 176 e 221). A notícia 28 também foi incluída nesta classe, possivelmente porque contém a entidade 'Coimbra', entidade comum a estas notícias. Também se obteve um grupo com duas notícias (133 e 186). No conjunto de todas as notícias são as únicas que têm a entidade 'Caixa Geral de Aposentações'.

Temas		
n= 12	n =18	n=25
Atletismo (Sara Moreira)	Atletismo (Sara Moreira)	África do Sul
Futebol (FC Porto)	Futebol (Benfica)	Futebol (Benfica)
Mercado Accionista (PSI)	Futebol (FC Porto)	Futebol (FC Porto)
Futebol (Benfica)	Mercado Accionista (PSI)	Mercado Accionista (PSI)
Distrito da Guarda	África do Sul	Governo Português
	Governo Português	Distrito da Guarda
	Mercado na China	Mercado na China
		Portugal (Porto e Lisboa)
		Casinos
		Política na Costa do Marfim
		Linha da Lousã
		Caixa Geral de Aposentações

Figura 3.21: Quadro resumo - temas obtidos através da Classificação Hierárquica.

Classificação Não Hierárquica

- K-médias

Efetuou-se uma classificação não hierárquica por recurso ao método das K-médias. Irá ser aplicado este método para K=1, K=2,...,K=25 para posteriormente ser possível comparar os resultados obtidos nas partições. Numa partição em duas classes 12 indivíduos estão na primeira classe e 215 estão na segunda classe. Os resultados foram obtidos após 4 iterações.

Analisando os indivíduos do *cluster* 1, identifica-se uma característica em comum — a existência da entidade 'RN' exceto no elemento 54. Este indivíduo é um *outlier* tal como se pode ver na *boxplot* da Figura 3.22. Retirando este indivíduo da análise, a notícia 22 deixaria de pertencer à classe 1. Supõe-se que é este elemento que faz com que o indivíduo 54 faça parte da classe 1[10]. A inércia explicada para a divisão

[10]A análise continuará com as 227 notícias.

em duas classes é de 0,114454. O valor obtido é relativamente baixo, o que significa que as classes são heterogéneas e não estão bem separadas. Os valores da inércia explicada para as restantes partições estão apresentados na Tabela 3.8.

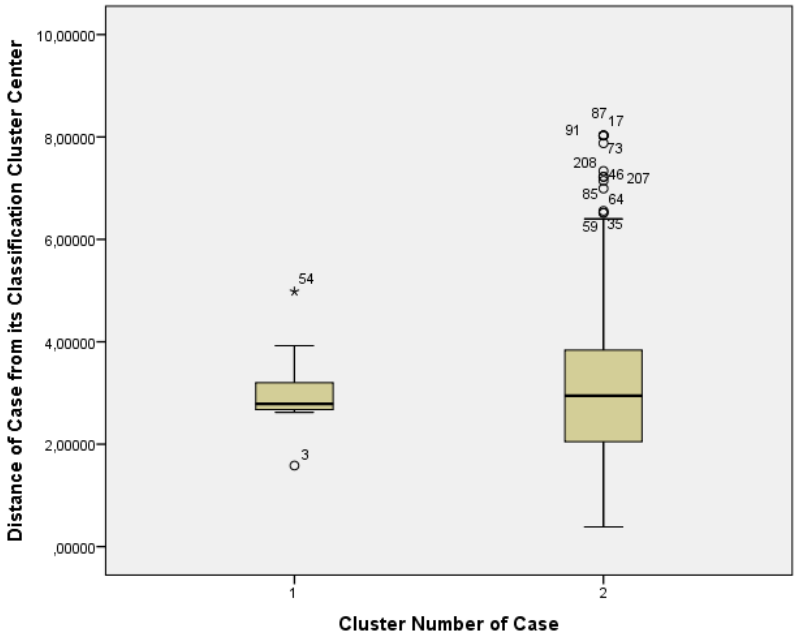

Figura 3.22: Boxplot obtido para um número de classes igual a 2.

Tabela 3.8: Inércia explicada para as partições com 2 a 25 classes.

Nº de classes	Inércia explicada	Nº de classes	Inércia explicada
2	0,114453967	14	0,585589484
3	0,092397432	15	0,607276269
4	0,205697267	16	0,623101832
5	0,26095672	17	0,641032421
6	0,314379301	18	0,658089726
7	0,35268246	19	0,683369257
8	0,371910667	20	0,674790448
9	0,415810107	21	0,688884175
10	0,463507389	22	0,703168242
11	0,483420949	23	0,751407795
12	0,515814363	24	0,773556173
13	0,536224458	25	0,787259124

Através desta tabela podemos ver que a inércia explicada sofre uma redução na partição 3 e na partição 20, sugerindo que estas partições não separam bem as

notícias. A divisão em 19 classes (ver Tabela P.1) indica uma certa homogeneidade das classes pois o valor da inércia explicada já é relativamente elevado. O número de notícias em cada classe está representado na Figura P.1. Nesta partição são identificados alguns temas iguais aos obtidos na Classificação Hierárquica — Distrito da Guarda (grupo 1), Governo Português (grupo 5), Casinos (grupo 6), África do Sul (grupo 7), FC Porto (grupo 8), Mercado Acionista (grupo 10), Atletismo (grupo 16) e Benfica (grupo 17). No grupo 15 os elementos são agrupados a partir da entidade 'União Europeia'. Pode-se dizer que esta classe diz respeito a notícias relacionadas com a Cultura e a Política na União Europeia, apesar de não incluir todas as notícias acerca deste tema no conjunto global das 227 notícias. É de notar uma classe com 109 notícias (classe 12). Engloba notícias muito diversificadas das restantes. Se aumentarmos o número de classes pretendidas esta classe iria dividir-se em classes mais pequenas. No entanto, iria continuar a existir uma classe com muitos elementos e a informação das restantes seria perdida.

Número de casos em cada cluster

Cluster	1	4,000
	2	3,000
	3	4,000
	4	2,000
	5	4,000
	6	3,000
	7	3,000
	8	7,000
	9	5,000
	10	4,000
	11	6,000
	12	109,000
	13	10,000
	14	4,000
	15	4,000
	16	2,000
	17	5,000
	18	24,000
	19	24,000
Válido		227,000
Ausente		,000

Figura 3.23: Número de elementos em cada classe para K=19.

- Mapas de Kohonen

Através do mapa de Kohonen podem visualizar-se os elementos das classes formados. O mapa obtido está representado na Figura L.1 e é constituído por 3 linhas e 3 colunas o que corresponde a 9 *clusters*. Decidiu-se incluir as entidades no mapa de modo a facilitar a sua leitura. Devido à dispersão dos dados é de esperar que existam diversos temas associados às notícias em questão. A partir deste mapa ainda se vêem

temas diferentes agrupados na mesma classe. De seguida, decidiu-se alargar a análise e construir um mapa[11] com 4 linhas e 4 colunas, como representado na Figura L.2. Neste caso não se obtiveram 16 classes como seria de esperar mas sim 15. A partir deste mapa já é possível identificar alguns temas. A classe 2 agrupa as notícias que falam da atleta Sara Moreira. Este tema também surgiu na formação dos três primeiros eixos principais na AC e na Classificação. No entanto, esta entidade tem mais notícias associadas a ela para além das notícias 16 e 17 que surgem nesta classe. Outra classe que desperta a atenção e cujo tema foi obtido na formação do terceiro eixo na AC e na análise classificatória, é a classe 3. Esta engloba as notícias sobre o 'FC Porto'. Nesta análise surge mais uma notícia relacionada com o tema comparativamente aos resultados obtidos na AC (notícia 73). A classe 4 apresenta algumas notícias sobre a política na Costa do Marfim, tema obtido também na formação do eixo 3 e na classificação. No entanto, também incorpora algumas notícias em que a entidade 'França' aparece (134, 196, 203). A classe 7 é constituída por apenas duas notícias. As entidades 'Benfica' e 'Europeu' surgem nestas duas notícias. Uma delas é sobre o Benfica e a outra é sobre a atleta Sara Moreira. Ambas são sobre o tema Desporto. No entanto, foi criada outra classe (classe 8) que agrupou as notícias onde a palavra 'Benfica' aparecia. Nesta análise surge um tema diferente daqueles identificados pela AC. A classe 11 diz respeito a notícias referentes ao Porto e a Lisboa. Também engloba algumas em que estas entidades não aparecem mas que têm em comum a entidade Portugal (76, 72, 197, 160, 152, 122, 103). Pode-se dizer que esta classe identifica notícias sobre o tema Portugal com especial enfoque nas grandes cidades do país: Porto e Lisboa. Com alguns elementos diferentes, esta classe já surgiu na Classificação Hierárquica. A classe 12 é constituída por notícias que têm em comum a entidade 'Lusa'. Apesar de todas elas serem publicadas pela agência Lusa, nem todas têm esta entidade associada. As notícias sobre partidos (PS e PSD) também estão nesta classe sendo que grande parte delas também incluem a entidade 'Lusa'. Umas das classes que tem outro tema já conhecido é a classe 15. Esta engloba as notícias sobre o mercado acionista, mais especificamente sobre o 'PSI'. Na AC também se identificou este assunto relativamente ao terceiro eixo. As restantes classes apresentam uma certa heterogeneidade. Por exemplo, a classe 5 é constituída por algumas entidades e notícias referentes ao tema da Política Internacional encontrado no primeiro eixo na AC. Apesar deste tema ser muito abrangente não é possível classificar todas as restantes entidades e notícias presentes nesta classe como sendo sobre Política Internacional. A classe 6 também é muito heterogénea pois apresenta não só notícias relacionadas com a 'África do Sul' e 'Moçambique' mas também outras notícias relacionadas com temas diferentes. Na classe 9 foram agrupados dois temas identificados na AC - Estado Português (eixo 2) e Casinos (eixo 3). Nesta classe, a entidade 'Espanha' e notícias relacionadas com a mesma também foram incluídas. Assim, considera-se uma classe heterogéneo. Nas

[11]Foram inseridos números nos grupos do mapa de forma a ser mais fácil identificar cada classe.

classes 1, 10, 13 e 14 também não é possível identificar temas concretos. De seguida apresenta-se o quadro resumo na Figura 3.24 dos temas obtidos para o mapa de Kohonen.

Mapa de Kohonen (4x4)		
classe 2	**classe 3**	**classe 4**
Atletismo (Sara Moreira)	Futebol (FC Porto)	Política na Costa do Marfim
classe 8	**classe 11**	**classe 15**
Futebol (Benfica)	Portugal (Porto e Lisboa)	Mercado Accionista (PSI)

Figura 3.24: Quadro resumo dos temas identificados — mapa de Kohonen.

3.4 Discussão dos resultados

As técnicas utilizadas — Análise de Correspondências, Classificação Hierárquica e Não Hierárquica — permitiram, tal como se pretendia, identificar alguns dos temas presentes nas 227 notícias como se pode ver nas Figuras 3.25 e 3.26.

Temas - Notícias				
Desporto	Economia	País	Mundo	Política
Futebol	Governo Português	Casinos	Mercado Chinês	Internacional
	Estado Português			Costa do Marfim
	Cavaco Silva			

Figura 3.25: Quadro resumo dos temas identificados — conjunto de dados notícias.

Temas - Notícias e entidades				
Desporto	Economia	País	Mundo	Política
Atletismo	Mercado Accionista	Guarda	África do Sul	Internacional
Futebol	Estado Português	Casinos	Mercado da China	Costa do Marfim
	Caixa Geral de Aposentações	Linha da Lousã	Cultura e Política da União Europeia	
		Portugal		

Figura 3.26: Quadro resumo dos temas identificados — conjunto de dados notícias e entidades.

Através destes quadros é visível uma maior informação extraída do segundo conjunto de dados. De facto, houve uma maior dificuldade em encontrar temas no primeiro conjunto de dados, em grande parte devido à dificuldade em retirar informação das palavras retidas. É notório que a utilização das entidades é uma mais valia. Em primeiro lugar evita que palavras como 'cento', 'mil', 'milhões', 'euros' apareçam na lista de palavras com mais relevância possibilitando uma análise mais cuidada e sem ruído. Além disto, as entidades são boas ferramentas para identificar temas com mais facilidade como se pode ver através do mapa de Kohonen. Com a visualização das entidades torna-se mais fácil saber *a priori* a que diz respeito cada grupo de notícias.

Relativamente aos métodos observou-se uma importante contribuição dos métodos de Classificação. Apesar de existirem alguns grupos não completamente homogéneos na Classificação, conclui-se que estes métodos foram um complemento importante à Análise de Correspondências. Identificaram-se muitos temas semelhantes àqueles obtidos na AC, mas também foi possível identificar novos temas, contribuindo para um conhecimento mais aprofundado acerca do conjunto das notícias.

Ainda existe alguma dificuldade em manusear o texto e interpretar os resultados. Há dificuldade na identificação de temas pois na Análise Classificatória são obtidas classes com muitos elementos e torna-se difícil saber qual o elo de ligação dessas notícias. Mais uma vez, devido à visualização das entidades, o mapa de Kohonen permite ultrapassar grande parte deste problema. Outra dificuldade diz respeito à diversidade do conjunto de dados. Por ser um conjunto de dados com muitas entidades diferentes, a AC não é tão clara e os grupos formados na Classificação não são tão homogéneos como pretendido. Para ultrapassar esta limitação seria necessário um conjunto de dados com notícias que tivessem um número razoavelmente elevado de entidades semelhantes.

Capítulo 4

Segredos da Maçonaria Portuguesa

Este capítulo é dedicado ao estudo do livro 'Segredos da Maçonaria Portuguesa' a partir das entidades extraídas de cada parágrafo. Tal como no capítulo anterior, os *softwares* a utilizar para o efeito são o Dtm-Vic e o SPSS *Statistics*. Apresenta-se também uma comparação dos resultados com aqueles obtidos por uma abordagem de redes sociais a este conjunto de dados.

4.1 Descrição e análise dos dados

O conjunto de dados a estudar é constituído por 13971 referências às 5612 entidades presentes nos 2508 parágrafos analisados. A extração de entidades foi realizada a partir do livro digitalizado, o que faz com que apareçam entidades escritas de forma irregular. Logo no primeiro parágrafo aparece a entidade 'Km Segredos da Maçonaria Portuguesa' — 'K' na realidade é um 'E'. No entanto, o livro digitalizado não impede a análise pois caso apareça alguma entidade deste género, conhece-se, geralmente, o seu significado. O processo de extração segue as mesmas regras descritas no Capítulo 3, Secção 3.2.

4.1.1 Análise de Correspondências

Como a ferramenta Visutex, utilizada no Capítulo 3, tem uma limitação de 1000 textos, utilizou-se uma ferramenta semelhante denominada de Visuresp que tem um limite de 30000 textos. Basicamente o *software* considera as entidades como respostas a cada parágrafo e por isso foi possível adaptar o conjunto de dados. Esta ferramenta também fornece algumas informações sobre o conteúdo dos dados e ainda apresenta alguns resultados da Classificação Hierárquica. No entanto, a Classificação irá continuar a ser efetuada no programa SPSS pois, para além desta ferramenta agrupar parágrafos em vez de entidades, como se pretende, o SPSS fornece mais informação permitindo uma análise mais completa.

Começou-se por analisar quais as entidades a manter, ou seja, aquelas que são mais frequentes. Após a análise de algumas frequências, optou-se por reter aquelas entidades que apresentam uma frequência igual ou superior a 25. Assim, mantiveram-se 3844 citações das 57 entidades. Como já foi referido, o programa apenas considera 20 caracteres nas entidades, cortando as restantes letras. Neste caso também foi possível decifrar quais as letras que faltavam e continuar a análise. As entidades retidas apresentam-se na Tabela 4.1 e verifica-se que as frequências das mesmas variam entre 25 e 423. Pode-se observar que a entidade que aparece mais vezes é 'GOL'. Assim, a tabela de contingência[1] cruza 2330[2] parágrafos com 57 entidades.

Tabela 4.1: Entidades retidas e respetivas frequências

Entidades	Frequência	Entidades	Frequência
Abel*Pinheiro	37	Lisboa	177
António*Arnaut	25	Loja*Mercúrio	25
António*José*Vilela	70	Loja*Universalis	25
António*Reis	78	Maçonaria	49
Bairro*Alto	29	Mercúrio	27
CO	37	Mário*Martin*Guia	27
Carbonária	38	NUIPC	26
Cf	125	Nuno*Vasconcellos	47
Coimbra	33	Ongoing	30
Conselho*da*Ordem	45	PS	63
EUA	25	PSD	50
GLLP	204	Paulo*Portas	28
GLRP	322	País	32
GOL	423	Porto	46
Governo	40	Portugal	89
Grande*Dieta	144	Presidente	39
Grande*Loja	37	presidente	32
Grande*Loja*Legal*de*Portugal	25	Público	29
Grande*Loja*Regular*de*Portugal	33	Representante	96
Grande*Oriente*Lusitano	125	Sábado	68
Grão	64	secretário	50
grão-mestre	154	secretário*de*Estado	25
grão-mestre*do*GOL	30	SIED	28
Irmão	48	SIS	31
Irmãos	207	Silva*Carvalho	33

[1]Como a tabela é formada por muitos zeros e é muito extensa decidiu-se não a apresentar nesta dissertação.

[2]Os parágrafos estão numerados até 2508 mas alguns deles não foram considerados visto que não se identificaram entidades.

Isaltino*Morais	32	TDLSB	27
Jorge*Silva*Carvalho	57	Venerável	61
José*Moreno	35	Vice	27
Justiça	35		

Após a AC, e para posteriormente aplicar a Análise Classificatória, é necessário identificar o número de eixos a reter. Ao utilizar o critério de *Pearson* retêm-se 38 eixos que explicam 81,11% da inércia total (ver Tabela 4.2). No entanto, atendendo a que o *software* não 'guarda' mais do que 30 coordenadas, essenciais para aplicar a Análise Classificatória no programa SPSS, optou-se por reter os 30 eixos que explicam 68,80% da inércia total. Esta percentagem de inércia já é aceitável pois, como já foi referido no capítulo anterior, deve manter-se um número suficiente de eixos de modo a explicar uma proporção de inércia superior a 50% (Naito, 2007), o que se verifica para 30 eixos. Pela tabela, pode constatar-se que os valores próprios são baixos. Isto deve-se à existência de muitas entidades diferentes no conjunto dos parágrafos.

Tabela 4.2: Valores próprios, inércia e inércia acumulada para os 38 primeiros eixos.

Eixo	λ	Inércia (%)	% acumulada	Eixo	λ	Inércia (%)	% acumulada
1	0,8953	3,32	3,32	20	0,5479	2,03	50,29
2	0,8597	3,19	6,51	21	0,5418	2,01	52,30
3	0,8199	3,04	9,56	22	0,5300	1,97	54,27
4	0,7719	2,86	12,42	23	0,5209	1,93	56,20
5	0,7596	2,82	15,24	24	0,5061	1,88	58,08
6	0,7400	2,75	17,99	25	0,5034	1,87	59,95
7	0,7232	2,68	20,67	26	0,4972	1,85	61,80
8	0,7043	2,61	23,29	27	0,4813	1,79	63,58
9	0,6850	2,54	25,83	28	0,4767	1,77	65,35
10	0,6660	2,47	28,30	29	0,4727	1,75	67,11
11	0,6539	2,43	30,73	30	0,4548	1,69	68,80
12	0,6230	2,31	33,04	31	0,4419	1,64	70,44
13	0,6110	2,27	35,31	32	0,4358	1,62	72,05
14	0,6082	2,26	37,57	33	0,4276	1,59	73,64
15	0,5937	2,20	39,77	34	0,4164	1,55	75,19
16	0,5843	2,17	41,94	35	0,4081	1,51	76,70
17	0,5742	2,13	44,07	36	0,4033	1,50	78,20
18	0,5700	2,12	46,19	37	0,3954	1,47	79,66
19	0,5587	2,07	48,26	38	0,3907	1,45	81,11

Recordando que o objetivo neste capítulo é estudar, principalmente, as entidades do livro, não teria grande interesse analisar quais os parágrafos que separam melhor os eixos. O que se pretende é identificar quais as entidades que o fazem e como contribuem para a formação de cada um dos eixos. Desta forma, a análise seguinte será realizada apenas tendo em consideração as entidades. Irá analisar-se a localização das suas coordenadas e as suas contribuições, tanto relativas como absolutas. Este estudo só será feito para os eixos 1, 2 e 3 a título ilustrativo.

Primeiro eixo principal

A percentagem de inércia explicada pelo primeiro eixo é de 3,32%. A entidade que mais contribui para a formação deste eixo é 'Vice', com coordenada positiva no eixo, com uma contribuição absoluta de 97,4 e uma contribuição relativa de 0,99. Como é de esperar, as outras entidades têm uma importância muito baixa neste eixo. Pelo plano [1,2] representado na Figura 4.1 vê-se realmente que a entidade 'Vice' está muito afastada das restantes, com coordenada positiva. Assim, este eixo é completamente explicado pela entidade 'Vice'. Como não se retira nenhuma informação relevante a partir desta entidade, decidiu-se retirá-la da análise pois é um *outlier*. Sem esta entidade, a percentagem de inércia e os valores próprios alteram ligeiramente (ver Tabela M.1). Assim, a percentagem explicada pelo primeiro eixo é de 3,30%. Opõe as entidades 'GLRP'[3], 'NUIPC'[4] e 'TDLSB'[5], com coordenadas negativas, às entidades 'Grande Dieta' e 'Representante' com coordenadas positivas. Todas elas estão bem representadas no plano pois apresentam uma CTR acima da média. As entidades do lado positivo do eixo contribuem mais para a sua formação — 'Representante' com uma contribuição absoluta de 58,2 e 'Grande Dieta' com 20,5 (Tabela N.1) como se pode ver pelo plano [1,2] da Figura 4.2.

Segundo eixo principal

O segundo eixo explica 3,14% da variabilidade total e separa entidades como 'NUIPC', 'Nuno Vasconcellos', 'Ongoing', 'Representante', 'Silva Carvalho' e 'TDLSB' com coordenadas positivas, de entidades como 'GLLP'[6] e 'GLRP', com coordenadas negativas. Todas elas estão bem representadas. Na Tabela N.2 podem identificar-se estas entidades e respetivas contribuições.

[3]Grande Loja Regular de Portugal
[4]Número único de identificação do processo de crime
[5]Tribunal da Relação de Lisboa
[6]Grande Loja Legal de Portugal

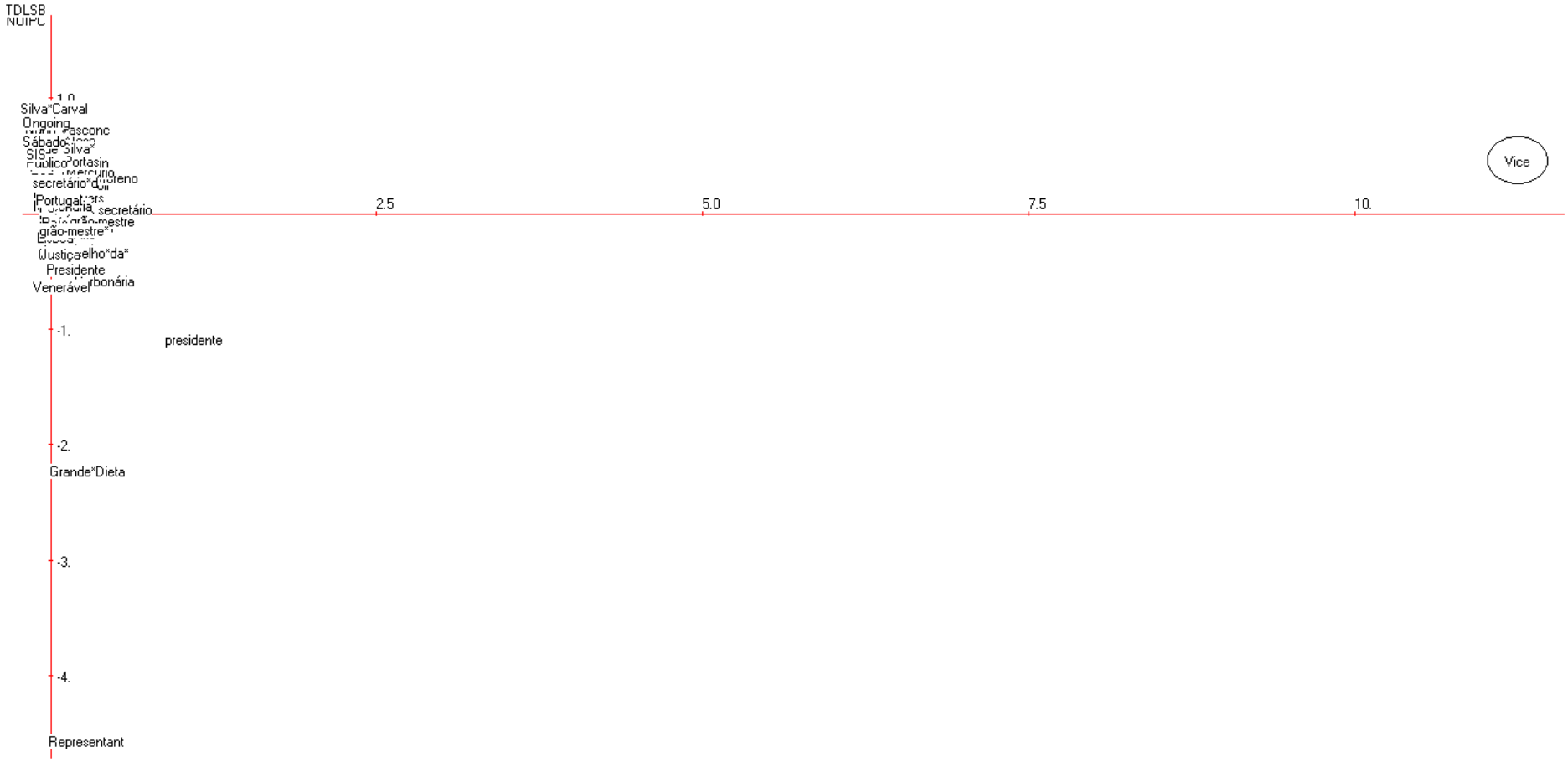

Figura 4.1: Entidades retidas do livro representadas no plano [1,2].

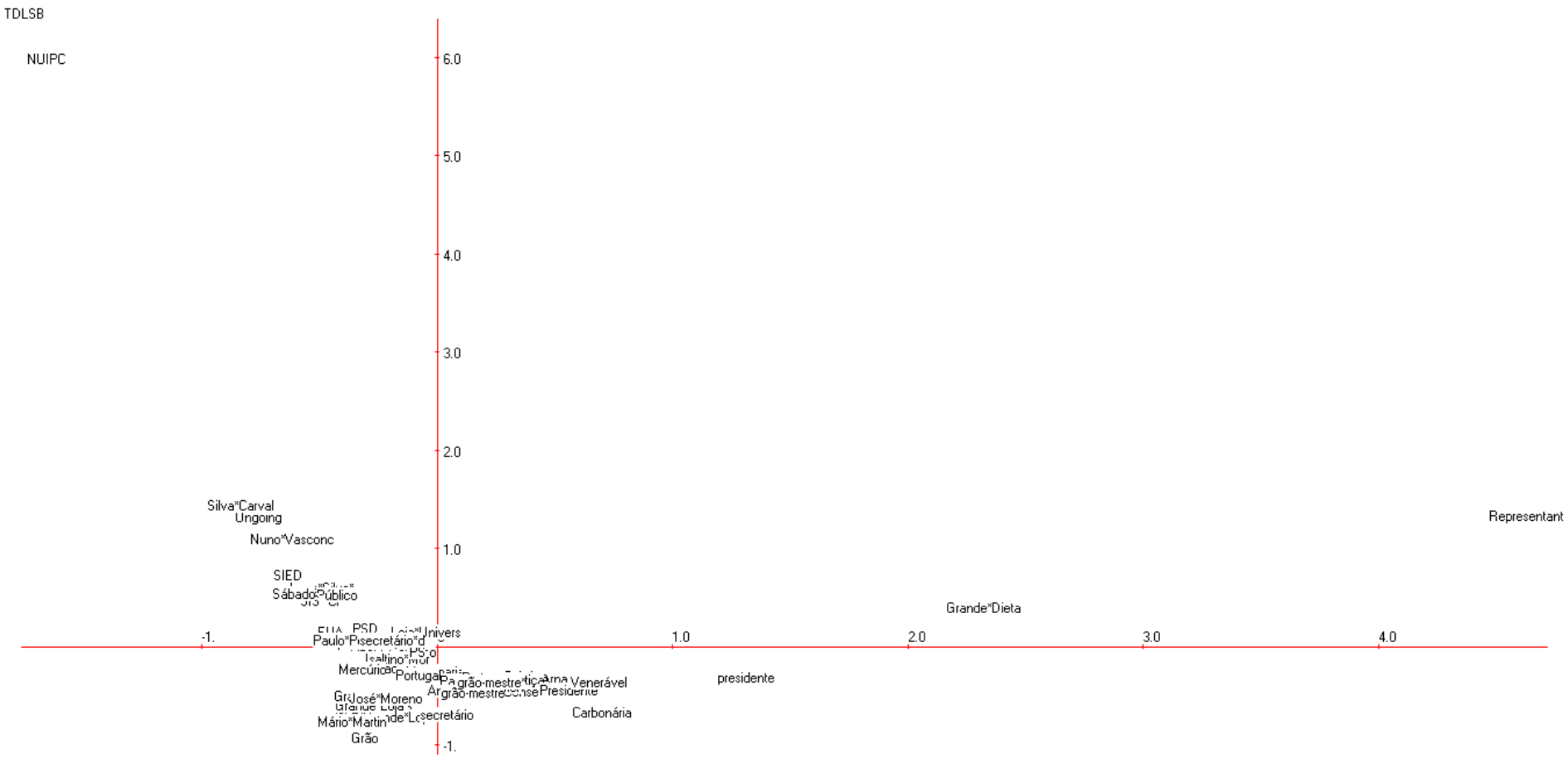

Figura 4.2: Entidades retidas do livro, após a exclusão da entidade 'Vice', representadas no plano [1,2].

Terceiro eixo principal

O terceiro eixo explica 2,96% da inércia e opõe as entidades 'Carbonária', 'Lisboa', 'Presidente', 'Venerável' e 'presidente', com coordenadas positivas, às entidades 'GLLP', 'GLRP', 'Grande Loja', 'Grão' e 'Mário Martin Guia', com coordenadas negativas. Através da Tabela N.3 e do plano [1,3] da Figura 4.4 é possível analisar a proximidade entre estas entidades e respetivas contribuições.

Eixo 1		Eixo 2		Eixo 3	
+	-	+	-	+	-
Grande Dieta	GLRP	NUIPC	GLLP	Carbonária	GLLP
Representante	NUIPC	Nuno Vasconcellos	GLRP	Lisboa	GLRP
		Ongoing		Presidente	Grande Loja
	TDLSB	Representante		Venerável	Grão
		Silva Carvalho		presidente	Mário Martin Guia
		TDLSB			

Figura 4.3: Quadro resumo - Eixos 1, 2 e 3.

4.1.2 Análise Classificatória

Nesta secção realizar-se-á o agrupamento das 57 entidades retidas em classes através da Classificação Hierárquica e Não Hierárquica. O objetivo é identificar grupos de entidades e descobrir se esses grupos são semelhantes ou não aos grupos formados a partir das redes sociais. As variáveis a utilizar serão as coordenadas nos 30 eixos fatoriais retidos.

Classificação Hierárquica

Utilizando como medidas o quadrado da distância Euclideana e o índice de Ward, aplicou-se uma classificação hierárquica ascendente e obteve-se o dendrograma da Figura 4.6. No dendrograma visualiza-se uma partição em duas classes. Uma delas contém duas entidades — 'NUIPC' e 'TDLSB' e a outra as restantes. Esta partição separa as entidades relacionadas com questões legais de todas as outras. No entanto, a primeira classe com 54 entidades pode ser dividida em mais classes e, por isso, decidiu-se calcular a inércia intra-classes e a inércia explicada[7] para determinar qual o corte adequado a fazer. Através da inércia intra-classes é possível desenhar um gráfico que permite ver a curva para as partições de 2 até 30 classes apresentado na Figura 4.5. Não é claro o ponto onde se observa o 'cotovelo' por isso optou-se por analisar a partição 15 pois apresenta um valor de inércia explicada relevante de 0,5884 como se pode ver na Tabela 4.3.

[7]As inércias foram calculadas através das tabelas ANOVA geradas pelo SPSS.

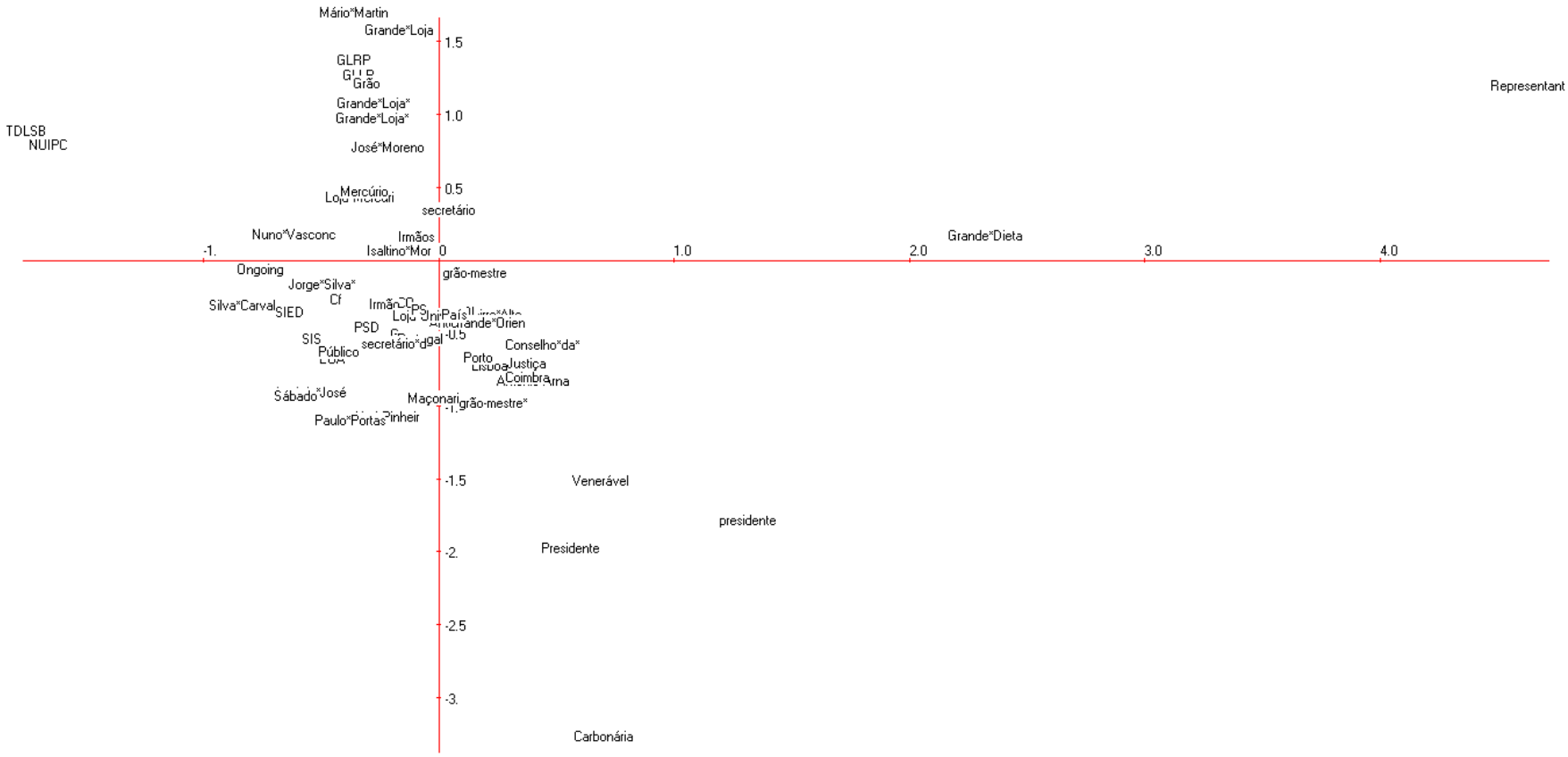

Figura 4.4: Entidades retidas do livro representadas no plano [1,3].

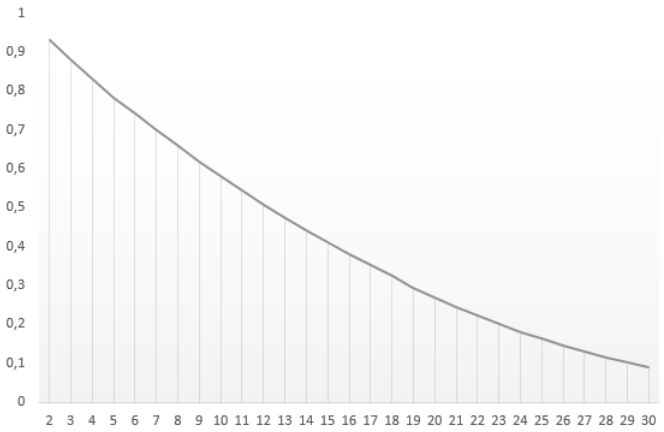

Figura 4.5: Inércia intra-classes para as partições de 2 até 30 classes.

Tabela 4.3: Inércia explicada para as partições de 2 até 30 classes.

Nº de classes	Inércia explicada	Nº de classes	Inércia explicada
2	0,070193172	17	0,647197899
3	0,120001132	18	0,675944503
4	0,168743628	19	0,704640612
5	0,216139175	20	0,730457282
6	0,258780201	21	0,754885945
7	0,301124383	22	0,778083796
8	0,342427956	23	0,798864694
9	0,380907564	24	0,819035705
10	0,418831669	25	0,836912501
11	0,456692525	26	0,853798271
12	0,492466393	27	0,870038082
13	0,526474685	28	0,885481285
14	0,557480148	29	0,897532932
15	0,588360736	30	0,90901608
16	0,618277147		

Na partição em 15 classes (ver Tabela O.1), é formado um grupo com as entidades 'Grande Loja Legal de Portugal' e 'Grande Loja Regular de Portugal'. Estas entidades dizem respeito à única organização maçónica portuguesa internacionalmente reconhecida como Regular. Um outro grupo é formado pelas entidades 'Loja Mercúrio', 'Mercúrio', 'José Moreno' e 'Isaltino Morais'. José Moreno foi fundador da Loja Mercúrio com Isaltino Morais.

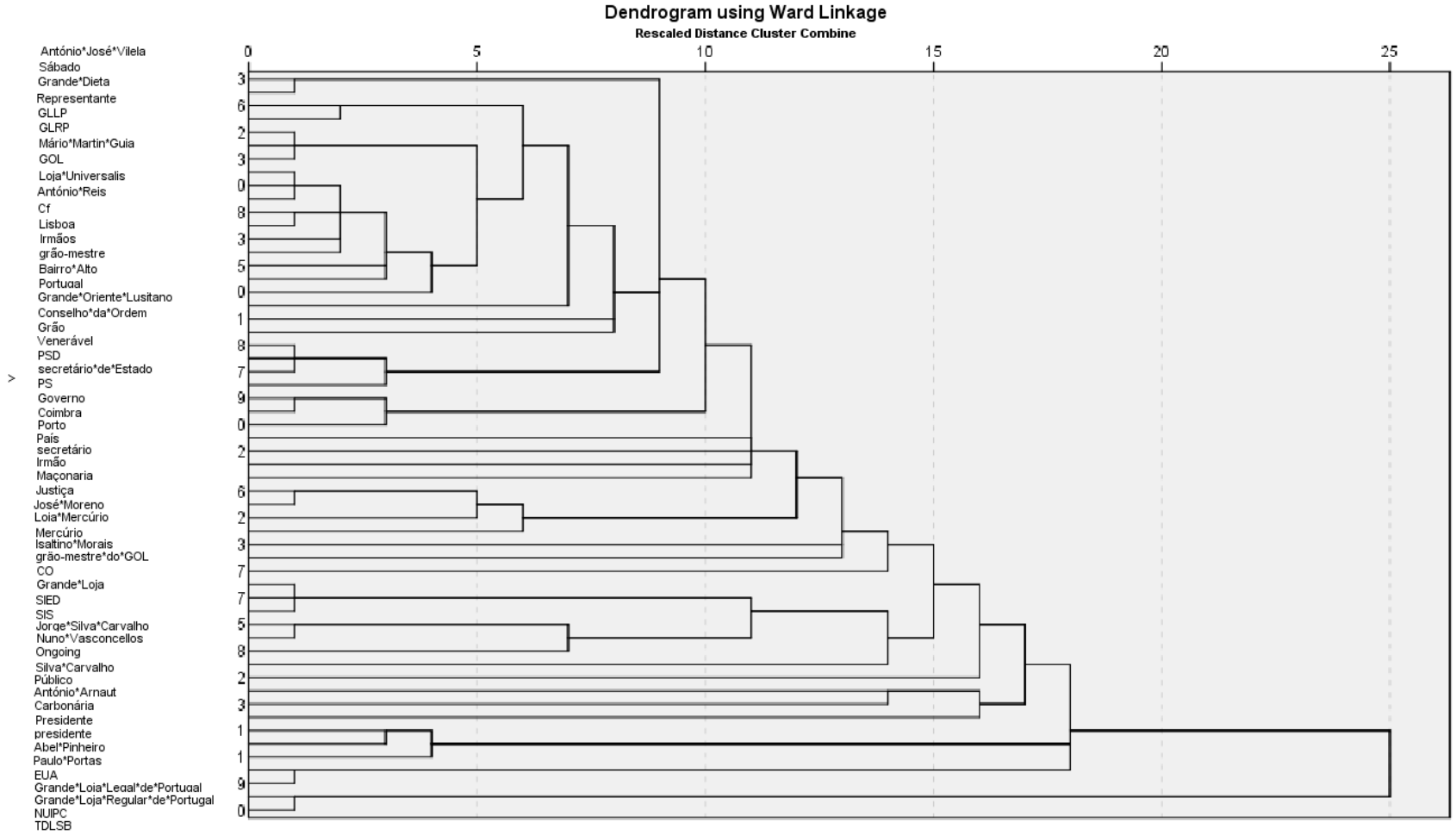

Figura 4.6: Representação através de um dendrograma da classificação hierárquica ascendente aplicada às 57 entidades retidas descritas pelas 30 coordenadas fatoriais.

Uma outra classe que se identificou inclui as entidades 'Jorge Silva Carvalho', 'SIS'[8] e 'SIED'[9]. Jorge Silva Carvalho exerceu funções dirigentes na SIS e na SIED. Outro dos grupos obtidos inclui as entidades 'TDLSB' e 'NUIPC'. O significado destas siglas sugere que este grupo está associado a questões legais. As entidades 'Nuno Vasconcellos', 'Ongoing' 'Silva Carvalho' também surgiram como uma classe. Silva Carvalho integrou os quadros da empresa *Ongoing Strategy Investments* onde Nuno Vasconcellos é o presidente. Nesta partição encontram-se grupos muito pequenos, alguns apenas com uma entidade, e um grupo muito grande com 31 entidades. Parece não existir uma estrutura classificatória forte nestes dados, pois as classes claramente definidas por temas são classes pequenas, *i.e.*, com poucas entidades. De facto, há algumas entidades que se destacam das outras, formando as tais classes pequenas e agrupando as restantes numa mesma classe. Estas classes são também são visíveis no dendrograma obtido, onde não existem grandes grupos bem destacados, e parece até ocorrer um efeito de cadeia. Assim, decidiu-se analisar uma partição mais pequena (com 4 classes) para detetar grupos maiores e identificar qual a relação entre as entidades dessas classes (ver Tabela O.2). Obtém-se 4 classes, duas com 3 entidades ('Abel Pinheiro', 'Paulo Portas', 'EUA' e 'Carbonária', 'presidente', 'Presidente'), uma com 2 entidades ('Grande Loja Legal de Portugal', 'Grande Loja Regular de Portugal') e outra com as restantes. Mais uma vez surge uma classe com muitas entidades e classes pequenas o que reforça uma vez mais que não existe uma estrutura de classes bem definida nestes dados. A classe 1 da partição em 4 classes é igual à obtida na partição em 15 classes e a classe 4 é igual à 7. A classe 3, com 3 entidades, foi dividida em 3 classes na partição em 15 classes. Assim, com o aumento do número de classes, as entidades do grupo grande vão sendo divididas em grupos mais pequenos, mantendo-se sempre uma classe com muitas entidades relativamente às restantes. Este efeito deve-se ao facto das entidades analisadas serem todas muito semelhantes, *i.e.*, são todas sobre o mesmo tema em geral — maçonaria portuguesa. Por isso, torna-se difícil obter classes com muitas entidades e bem definidas por temas.

Classificação Não Hierárquica

- K-médias

Aplicou-se a classificação não hierárquica através do algoritmo K-médias para K=2,...,K=30 tal como foi feito na classificação hierárquica. Para estudar qual a partição adequada, calculou-se a inércia explicada apresentada na Tabela 4.4.

[8]Serviço de Informações de Segurança
[9]Serviço de Informações Estratégicas de Defesa

Tabela 4.4: Inércia explicada para as partições 2 até 30 classes.

Nº de classes	Inércia explicada	Nº de classes	Inércia explicada
2	0,045695951	17	0,613567593
3	0,088073397	18	0,660967893
4	0,158435196	19	0,682608425
5	0,168425973	20	0,675334247
6	0,237055223	21	0,717774142
7	0,27947146	22	0,708317604
8	0,284888231	23	0,73342323
9	0,355040699	24	0,763844864
10	0,389692009	25	0,809353063
11	0,420443998	26	0,82207046
12	0,450953765	27	0,845353239
13	0,480058394	28	0,874782672
14	0,509375657	29	0,886442511
15	0,538213604	30	0,889801545
16	0,56436181		

Verificam-se duas quebras do valor da inércia explicada — na partição 20 e na partição 22, o que sugere que as partições 19 e 21 (ver Tabelas P.1 e P.2) separam melhor as classes. De facto, estas partições têm um valor de inércia explicada de, aproximadamente, 0,6826 e 0,71777, respetivamente, o que já são valores relevantes. Nestas partições identificam-se dois grupos semelhantes aos obtidos na classificação hierárquica com as entidades:

- 'NUIPC' e 'TDLSB';
- 'Grande Loja Legal de Portugal' e 'Grande Loja Regular de Portugal'.

Tal como na classificação hierárquica existem muitos grupos pequenos à exceção de um que contém 34 (partição em 19 classes) e 33 (partição em 21 classes) entidades. Como já se observou anteriormente, o aumento do número de classes iria fazer com que o grupo se dividisse. No entanto, iria continuar a existir um grupo com muitas entidades relativamente aos restantes formados. Por isso, decidiu-se também analisar a partição em 4 classes (Tabela P.3). As entidades 'NUIPC' e 'TDLSB' foram agrupadas numa classe. Ainda se observam duas classes com apenas uma entidade — uma com 'António Arnaut' e outra com 'presidente'. A outra classe contém as restantes 52 entidades. Tal como na Classificação Hierárquica, existe sempre uma classe com bastantes entidades e muitas classes pequenas.

- Mapas de Kohonen

A partir do mapa de Kohonen 3x3 representado na Figura 4.7 já se conseguem observar classes com mais entidades.

presidente Representant Presidente Grande*Dieta Carbonária	grão-mestre Mário*Martin Mercúrio José*Moreno Irmãos Grande*Loja* Grande*Loja* Grande*Loja GLRP GLLP	secretário Grão
secretário*d Sábado Silva*Carval SIS SIED Público PSD Ongoing Nuno*Vasconc Loja*Univers Loja*Mercúri Jorge*Silva* Irmão Governo António*José	Venerável Porto País PS Lisboa Justiça Isaltino*Mor GOL Coimbra	
TDLSB Paulo*Portas NUIPC EUA Abel*Pinheir	Cf	grão-mestre* Portugal Maçonaria Grande*Orien Conselho*da* CO Bairro*Alto António*Reis António*Arna

Figura 4.7: Mapa de Kohonen (3 x 3) representando as 56 entidades do livro.

Apesar de não ser possível identificar temas específicos, estas classes maiores permitem observar quais as entidades que se assemelham mais entre si. Neste mapa, ainda aparece uma classe com apenas uma entidade — 'Cf'. Esta entidade não está associada às restantes pois, apesar de aparecer algumas vezes no texto, não se trata realmente de uma entidade. O aparecimento de falsas entidades deve-se ao facto do algoritmo que extrai as entidades apresentar uma precisão superior a 95% mas inferior a 100%.

No mapa 4x4 (Anexo Q) já começam a surgir classes mais pequenas em contraste com uma classe com mais entidades. Apesar de se saber que o aumento do número de classes origina a formação de classes mais pequenas, decidiu-se analisar os mapas 5x5 e 6x6 [10] para perceber alguns dos temas inerentes à maçonaria portuguesa. Estes mapas estão apresentados no Anexo Q. A visualização dos dois mapas permite observar que têm duas classes em comum. Uma deles inclui as entidades 'TDLSB' e 'NUIPC', classe esta também obtida na classificação hierárquica. As entidades 'Jorge Silva Carvalho', 'SIS' e 'SIED' também pertencem à mesma classe em ambos os mapas. Jorge Silva Carvalho exerceu funções dirigentes na SIS e na SIED tal como foi referido na classificação hierárquica. A classe 2 do mapa 5x5 é constituído pelas entidades 'Silva Carvalho', 'Ongoing', 'Nuno Vasconcellos' e 'CO'. À excepção da entidade 'CO', esta classe já foi identificada na Classificação. Como já vimos, Nuno Vasconcellos é o presidente da *Ongoing Strategy Investments* e Silva Carvalho integrou os quadros desta empresa onde assumiu diversos cargos de administração. A classe 3 do mapa 5x5 parece estar associado ao Governo pois inclui as entidades 'Representante', 'PSD', 'PS' e 'Governo'. Mais uma vez a entidade 'Isaltino Morais' aparece associada à entidade 'Loja Mercúrio' na classe 6 do mapa 5x5. Esta classe também inclui a entidade 'Grão' neste mapa enquanto que no mapa 6x6 só inclui as primeiras duas entidades referidas. As entidades 'Grande Loja Legal de Portugal' e 'Grande Loja Regular de Portugal' aparecem na classe 5 do mapa 6x6. Observa-se também uma classe (classe 6) constituída pelas entidades 'Porto' e 'Coimbra' no mapa 6x6. A classe acerca do Governo parece dividir-se e gerar duas sub-classes — um constituído pelas entidades 'País', 'PS' e 'Governo' (classe 8) e outro constituído pelas entidades 'secretário de Estado' e 'PSD' (classe 9). No mapa 6x6, a classe 14 é formada pelas entidades 'grão-mestre', 'Representante', 'Mário Martin Guia', 'Mercúrio', 'Grande Dieta', 'GLRP' e 'GLLP'. Mário Martin Guia foi eleito como Grão-mestre perante uma assembleia da GLRP/GLLP. Na Figura 4.8 apresenta-se um quadro resumo com as classes identificadas nestes dois mapas.
De acordo com os resultados obtidos na Classificação, é possível observar que foi identificado um maior número de classes a partir dos mapas de Kohonen do que através dos resultados obtidos na classificação hierárquica e no algoritmo K-médias. Estes mapas permitem visualizar de forma mais rápida e simples as classes formadas.

[10] Foram inseridos números nas classes do mapa de forma a ser mais fácil identificar cada classe. Algumas entidades aparecem cortadas devido à limitação de 20 caracteres do *software*.

Temas dos grupos	Questões legais	Organização maçónica portuguesa	Loja Mercúrio	Locais
Entidades	TDLSB	Grande Loja Regular de Portugal	Loja Mercúrio	Porto
	NUIPC	Grande Loja Legal de Portugal	Isaltino Morais	Coimbra
Temas dos grupos	**Governo**	**Jorge Silva Carvalho**	**Empresa Ongoing Strategy Investments**	**Mário Martin Guia**
Entidades	secretário de Estado	SIS	Silva Carvalho	presidente
	PSD	SIED	Ongoing	Representante
	PS	Silva Carvalho	Nuno Vasconcellos	Mário Martin Guia
	Governo			Grande Dieta
				GLRP
				GLLP

Figura 4.8: Classes relevantes obtidos a partir dos mapas de Kohonen 5x5 e 6x6 com as 56 entidades do livro.

Além disto, a partir do mapa 3x3 obtiveram-se algumas classes com mais entidades agrupadas o que permitiu ver a proximidade entre elas. Como já foi referido, não existe uma estrutura classificatória bem definida nestes dados e, por isso, torna-se difícil encontrar classes com um número relevante de entidades bem definidas por temas.

4.2 Discussão dos resultados

O objetivo nesta secção é analisar se os resultados obtidos neste trabalho complementam aqueles obtidos através da aplicação de redes sociais, e/ou se apresentam alguma semelhança entre eles. É de notar que foram utilizadas abordagens diferentes e, por isso, não se pode comparar uma análise elaborada a partir de uma rede com a análise realizada através dos métodos da AC e Classificação pois as características e finalidades dos métodos são distintas. Assim, enquanto que neste trabalho se estudam as entidades mais frequentes no livro, na análise de redes sociais estudaram-se as entidades e as relações entre elas, *i.e.*, a rede em estudo é formada pelos vértices, que são as entidades, e pelas ligações entre os vértices, que no caso do livro representam a co-ocorrência das duas entidades na mesma frase. Esta é logo à partida uma das razões para o aparecimento de alguns termos neste trabalho que não são mencionados na rede social. Termos como CO, Cf e Vice, que não são entidades tal como foram atrás definidas, aparecem com alguma frequência no livro mas não associados a outras entidades o que justifica o seu aparecimento neste trabalho mas não na análise da rede formada.

Na Figura 4.9 podem observar-se seis comunidades, com algumas das entidades desses grupos, e também as relações entre elas. Observam-se várias entidades que já foram surgindo no presente estudo, bem como algumas comunidades que sugerem temas semelhantes aos obtidos na Classificação — no topo da figura temos uma comunidade sobre o GOL, e no lado esquerdo temos duas comunidades, uma sobre

locais, como se denominou acima, e outra sobre o Governo. Comparando as entidades destas comunidades com as classes obtidas na Classificação, temos no grupo sobre a GOL as entidades 'GOL' e 'Irmãos' em comum, no segundo grupo as entidades 'Porto' e 'Coimbra' e no terceiro 'PS', 'PSD' e 'Governo'. Neste trabalho, algumas das entidades das restantes três comunidades foram divididas em classes mais pequenas, *e.g.*, 'GLRP' e 'GLLP' formam, muitas vezes, apenas uma classe.

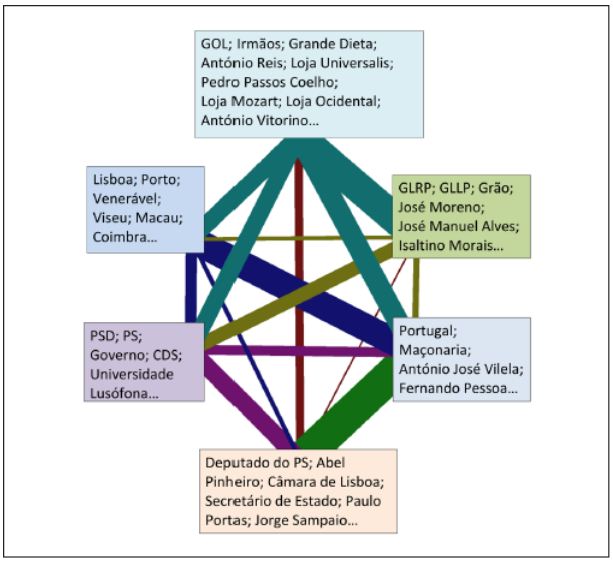

Figura 4.9: As seis comunidades de maior dimensão obtidas através da aplicação de redes socais por Rocha et al. (2014).

Pode-se concluir que existem algumas semelhanças entre os dois estudos realizados, pois identificam-se três grupos com algumas entidades em comum. No entanto, esses grupos incluem mais entidades no estudo das redes, entidades essas que não aparecem na lista das mais frequentes no presente estudo, como 'Pedro Passos Coelho', 'Viseu', 'Macau', 'CDS', 'Universidade Lusófona'. Assim, estes estudos complementam-se entre si.

Capítulo 5

Conclusões

Neste capítulo apresentam-se as principais conclusões, limitações e perspetivas de desenvolvimento do presente trabalho.

5.1 Resultados

Nesta dissertação analisaram-se três conjuntos de dados textuais. O primeiro constituído por 227 notícias, o segundo constituído pelas entidades dessas mesmas notícias e o terceiro com as entidades do livro 'Segredos da Maçonaria Portuguesa'.

No primeiro conjunto de dados foram utilizadas 87 palavras para descrever o conjunto das notícias. Retiveram-se os primeiros 30 eixos da AC que explicam 70,11% da inércia total. Através do *software* Dtm-Vic, explorou-se a AC nos primeiros três eixos principais e foram identificados os seguintes temas no conjunto das 227 notícias — Política (Política na Costa do Marfim e Política Internacional), Desporto (Futebol) e Governo Português. Exportaram-se as 30 coordenadas fatoriais para o programa SPSS e aplicou-se uma Classificação Hierárquica recorrendo ao método de Ward e ao quadrado da distância Euclideana. Através do cálculo das inércias (intra-classes, inter-classes e explicada) estudaram-se as partições e identificaram-se os temas Casinos, Política Internacional, Mercado Chinês, Política na Costa do Marfim, Estado Português, Desporto e Governo Português para a partição em 23 classes. De seguida, utilizou-se o algoritmo das K-médias e o mapa de Kohonen na Classificação Não Hierárquica. No algoritmo das K-médias os temas que surgiram foram Governo Português, Política na Costa do Marfim, Política Internacional, Mercado Chinês e ainda um grupo sobre Cavaco Silva. No mapa de Kohonen foram identificados quatro temas já referidos anteriormente — Desporto, Política Internacional, Governo e Estado Português.

No segundo conjunto de dados utilizaram-se 50 entidades (frequência mínima de 11). Na AC identificaram-se os temas Política Internacional, Desporto (Atletismo e Futebol), Mercado Accionista, Estado Português, Política na Costa do Marfim e

Casinos. Mantiveram-se os primeiros 24 eixos principais de acordo com o critério de *Pearson*. Aplicaram-se métodos de Classificação às coordenadas nos eixos e identificaram-se temas como África do Sul, Futebol (Benfica), Futebol (FC Porto), Mercado Acionista (PSI), Governo Português, Distrito da Guarda, Mercado Chinês, Portugal (Porto e Lisboa), Casinos, Política na Costa do Marfim, Linha da Lousã e Caixa Geral de Aposentações na Classificação Hierárquica. Através do algoritmo K-médias identificaram-se os temas — Distrito da Guarda, Governo Português, Casinos, África do Sul, FC Porto, Mercado Acionista, Atletismo e Benfica, já obtidos na Classificação Hierárquica. Ainda se identificou um novo tema, Cultura e Política na UE. No mapa de Kohonen, os temas que apareceram foram análogos aos já referidos — Desporto (Atletismo e Futebol), Portugal, Mercado Acionista e Política na Costa do Marfim.

Comparando os resultados do primeiro conjunto de dados com o do segundo, vemos uma diversidade de temas muito maior no segundo conjunto. Isto deveu-se à mais-valia do processo de extração de entidades que permitiu remover palavras com pouco significado como 'mil', 'milhões', 'euro', 'cento', 'quatro', entre outras. O mapa de Kohonen foi uma ferramenta muito interessante pois permitiu visualizar de forma mais rápida as classes formadas devido à possibilidade de integração das palavras e entidades no mapa juntamente com as notícias.

Para o conjunto de dados do livro utilizaram-se os mesmo métodos com o objetivo de agrupar as entidades por temas. Na AC revelou-se particularmente difícil identificar temas, mas com a Classificação identificaram-se grupos sobre Questões Legais, Organização Maçónica Portuguesa, Loja Mercúrio, Locais, Empresa *Ongoing*, Jorge Silva Carvalho, Governo e Mário Martin Guia. Comparando este estudo com o estudo sobre redes sociais, concluiu-se que para além de apresentarem algumas semelhanças, existe uma complementaridade entre eles.

Por fim podemos afirmar que os métodos de Classificação revelaram-se realmente um complemento essencial aos métodos de Análise de Correspondências, especialmente no caso do livro. Sem eles, não teria sido possível identificar temas.

5.2 Limitações e Trabalho Futuro

Por ser um trabalho muito dependente das escolhas realizadas, nomeadamente o número de palavras/entidades a manter, o número de eixos retidos e o número de classes na Classificação, seria interessante abordar este tema com diferentes parâmetros de forma a averiguar se os resultados seriam muito distintos. Devido a algumas limitações já referidas ao longo da dissertação, seria interessante desenvolver o *software* de modo a que entidades com mais de 20 caracteres fossem contabilizadas e que fosse possível 'guardar' as coordenadas fatoriais para mais de 30 eixos.

Bibliografia

Aggarwal, C. C. e Zhai, C. (2012). *Mining Text Data*. Springer.

Bécue-Bertaut, M., Rajman, M., Lebart, L., e Gaussier, E. (2005). Extraction of the Useful Words from a Decisional Corpus. Contribution of Correspondence Analysis. Em *Knowledge Mining*, pages 159–179. Springer.

Benzécri, J. (1973). *L' Analyse des Données, Tome 2: L' Analyse des Correspondances, 1973*. Dunod, Paris.

Cohen, A. M. e Hersh, W. R. (2005). A Survey of Current Work in Biomedical Text Mining. *Briefings in bioinformatics*, 6(1):57–71. Oxford Univ Press.

El-Hamdouchi, A. e Willett, P. (1989). Comparison of Hierarchic Agglomerative Clustering Methods for Document Retrieval. *The Computer Journal*, 32(3):220–227.

Friedman, J., Hastie, T., e Tibshirani, R. (2001). *The Elements of Statistical Learning*, volume 1. Springer series in Statistics Springer, Berlin.

Greenacre, M. (2007). *Correspondence Analysis in Practice*. Taylor & Francis Ltd.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press INC.

Gupta, V. e Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications. *Journal of emerging technologies in web intelligence*, 1(1):60–76.

Gupta, V. e Lehal, G. S. (2010). A Survey of Text Summarization Extractive Techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3):258–268.

Hassall, P. e Ganesh, S. (2005). Correspondence Analysis in Attitudinal Research: The Case of World Englishes and Teaching English as an International Language. *Teaching and Learning in Higher Education: Gulf Perspectives*, 2:1–23.

Hoffman, D. L. e Franke, G. R. (1986). Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research. *Journal of Marketing Research*, pages 213–227.

Hotho, A., Nürnberger, A., e Paaß, G. (2005). A Brief Survey of Text Mining. In *Ldv Forum*, volume 20, pages 19–62.

Huang, A. (2008). Similarity Measures for Text Document Clustering. Em *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56.

Jain, A. K., Murty, M. N., e Flynn, P. J. (1999). Data Clustering: A Review. *ACM computing surveys (CSUR)*, 31(3):264–323. ACM.

Koutsoupias, N. (2002). Exploring Web Access Logs with Correspondence Analysis. Em *Proc. Second Hellenic Conf. Methods and Applications of Artificial Intelligence*.

Krah, A., Wessel, R., e Pleißner, K.-P. (2004). Assessment of Protein Spot Components Applying Correspondence Analysis for Peptide Mass Fingerprint Data. *Proteomics*, 4(10):2982–2986. Wiley Online Library.

Lebart, L., Morineau, A., e Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. Wiley series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley.

Lebart, L., Salem, A., e Berry, L. (1998). *Exploring Textual Data*, volume 4. Kluwer Academic Publishers.

Lee, R. (1981). Clustering Analysis and its Applications. In *Advances in Information Systems Science*, pages 169–292. Springer.

Morin, A. (2004a). Correspondence Analysis for Data Mining with Applications in Medicine. IRISA, Université de Rennes.

Morin, A. (2004b). Intensive Use of Correspondence Analysis for Information Retrieval. Em *Information Technology Interfaces, 2004. 26th International Conference on*, pages 255–258. IEEE.

Morin, A. (2006). Intensive Use of Factorial Correspondence Analysis for Text Mining: Application with Statistical Education Publications. In *ICOTS-7 (International Conference on Teaching Statistics), Salvador, Bahia, Brazil*.

Naito, S. D. N. P. (2007). Análise de Correspondências Generalizada. *Dissertação de Mestrado da Faculdade de Ciências da Universidade de Lisboa*, Capítulo 3.

Petrović, S., Bašić, B. D., Morin, A., Zupan, B., e Chauchat, J.-H. (2009). Textual Features for Corpus Visualization using Correspondence Analysis. *Intelligent Data Analysis*, 13(5):795–813.

Rocha, C., Jorge, A. M., Oliveira, M., Brito, P., Gama, J., Pimenta, C., et al. (2014). From Entity Extraction to Network Analysis: A Method and an Application to a Portuguese Textual Source. Technical Report 32, OBEGEF-Observatório de Economia e Gestão de Fraude & OBEGEF Working Papers on Fraud and Corruption.

Steinbach, M., Karypis, G., Kumar, V., et al. (2000). A Comparison of Document Clustering Techniques. Em *KDD Workshop on Text Mining*, volume 400, pages 525–526. Boston.

Tan, A.-H. (1999). Text Mining: The State of the Art and the Challenges. Em *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, pages 65–70.

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tekaia, F., Yeramian, E., e Dujon, B. (2002). Amino Acid Composition of Genomes, Lifestyles of Organisms, and Evolutionary Trends: A Global Picture with Correspondence Analysis. *Gene*, 297(1):51–60. Elsevier.

Willett, P. (1988). Recent Trends in Hierarchic Document Clustering: A Critical Review. *Information Processing & Management*, 24(5):577–597. Elsevier.

Zhao, Y., Karypis, G., and Fayyad, U. (2005). Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, 10(2):141–168. Springer.

Anexo A

Dtm-Vic — *Data and Text Mining*: Visualização, Inferência, Classificação

O *software* Dtm-Vic cujo menu principal está representado na Figura A.1 foi desenvolvido por Ludovic Lebart para analisar dados complexos, tanto numéricos como textuais.

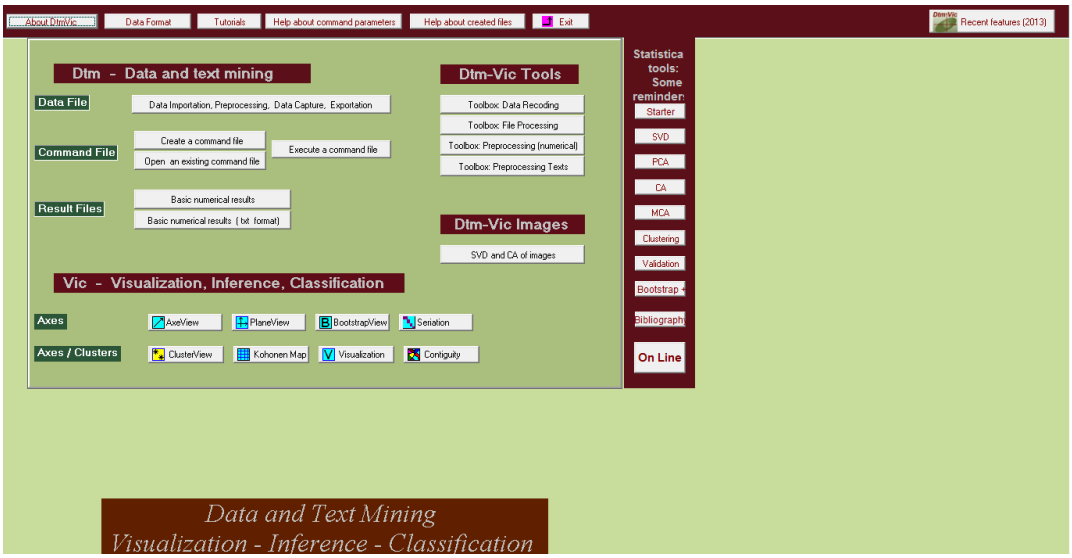

Figura A.1: Menu principal do *software* Dtm-Vic.

O programa tem à sua disposição técnicas de visualização (análise de componentes principais, análise de correspondências simples e múltiplas) e de classificação (método híbrido, combinando classificação hierárquica - critério de Ward - e K-médias; mapas auto-organizados de Kohonen); validação de técnicas de visuali-

zação: re-amostragem (*bootstrap*, *bootstrap* parcial, *bootstrap* total, *bootstrap* sobre variáveis) e análise de contiguidade e métodos afins.

Este *software* oferece ainda um tutorial dividido em quatro temas - tutorial A: 'An Introduction to Dtm-Vic', tutorial B: 'Dtm-Vic and Textual Data', tutorial C: 'Dtm-Vic and Numerical Data', tutorial D: 'Data and Text importation'. No tutorial A estão disponíveis seis aplicações introdutórias. Os três primeiros exemplos dizem respeito a dados numéricos (Análise de Componentes Principais, Análise de Correspondências e Análise de Correspondências Múltiplas). Os restantes exemplos tratam de dados numéricos e textuais. Três aplicações mais avançadas relacionadas com dados textuais constituem o tutorial B, enquanto que o tutorial C inclui aplicações avançadas utilizando dados numéricos. Além disto, quatro exemplos de importação de dados estão disponíveis no tutorial D. Estes tutoriais podem ser lidos diretamente do menu principal do programa.

Para utilizar os comandos do programa basta clicar em '*Create a command file*' e a janela da Figura A.2 aparecerá. O utilizador pode 'correr' o programa com os seus próprios dados, mudando alguns parâmetros e respeitando os formatos de entrada dos dados. A informação sobre cada parâmetro pode ser acedida no menu principal em '*Help about command parameters*'.

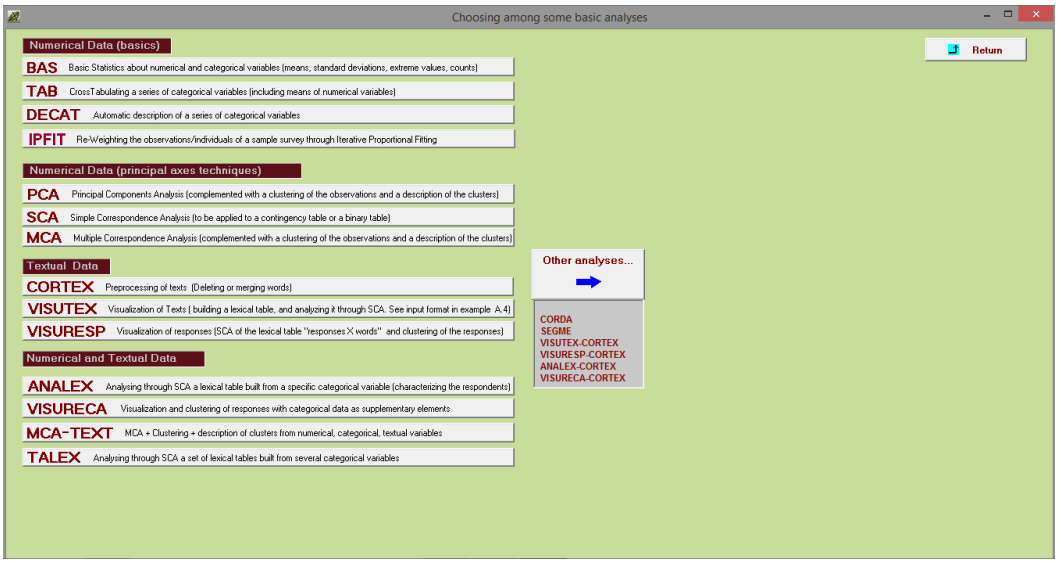

Figura A.2: Comandos do *software* Dtm-Vic.

Os resultados obtidos podem ser vistos em '*Basic Numerical Results*'. Vários ficheiros com os resultados são gerados à parte e a informação do que cada um contém pode ser consultada em '*Help about created files*'. O programa ainda disponibiliza algumas explicações sobre alguns métodos, nomeadamente, SVD, PCA, CA, MCA, *Clustering*, *Validation* e *Bootstrap*. Este *software* pode ser obtido gratuitamente na Internet em $http://www.dtmvic.com/05_software E.html$.

Anexo B

Dados notícias - palavras retidas

Tabela B.1: Palavras retidas e frequências

Palavras	Frequência	Palavras	Frequência
Brasil	61	fazer	54
Costa	36	fim	124
Estado	73	final	42
Gbagbo	40	grande	37
Governo	116	grupo	38
Itália	36	hoje	241
Lisboa	59	início	47
Lusa	91	janeiro	76
Luís	37	maior	35
Para	47	melhor	42
Porto	45	menos	38
Portugal	88	mil	74
Presidente	49	milhões	91
República	42	ministro	35
Segundo	61	novo	48
Segurança	51	nível	35
Silva	65	onde	53
Social	56	partir	47
acordo	52	país	132
afirmou	51	países	51
agora	35	pessoas	71
agência	46	poder	40
ainda	110	pontos	39
ano	174	portuguesa	55
anos	136	português	36
anunciou	35	presidente	84

apenas 36
aumento 36
cento 157
cerca 57
comunicado 35
contra 42
crise 56
decisão 53
devido 37
dezembro 44
dia 77
disse 132
dois 82
durante 35
empresa 37
enquanto 35
equipa 42
euros 131
primeira 38
primeiro 36
quatro 60
quinta-feira 47
segundo 80
ser 124
situação 53
sábado 36
ter 81
todos 64
trabalhadores 53
trabalho 50
três 50
têm 39
vai 88
vez 42
vão 36

Anexo C

Análise de Correspondências dos dados notícias — Histograma com os valores próprios (*output* parcial).

number	Eigen value	percent.	cumulat. percent.	
1	.3893	6.01	6.01	**********
2	.3465	5.35	11.35	*********
3	.3255	5.02	16.37	*********
4	.3023	4.66	21.04	********
5	.2308	3.56	24.60	******
6	.2051	3.16	27.76	*****
7	.1783	2.75	30.51	*****
8	.1654	2.55	33.06	****
9	.1531	2.36	35.43	****
10	.1483	2.29	37.71	****
11	.1460	2.25	39.97	****
12	.1358	2.09	42.06	****
13	.1290	1.99	44.05	****
14	.1273	1.96	46.02	****
15	.1237	1.91	47.92	****
16	.1157	1.78	49.71	***
17	.1136	1.75	51.46	***
18	.1121	1.73	53.19	***
19	.1081	1.67	54.86	***
20	.1036	1.60	56.46	***
21	.1006	1.55	58.01	***
22	.0979	1.51	59.52	***
23	.0964	1.49	61.01	***
24	.0935	1.44	62.45	***
25	.0888	1.37	63.82	***
26	.0868	1.34	65.16	***
27	.0851	1.31	66.47	***
28	.0805	1.24	67.71	***
29	.0783	1.21	68.92	***
30	.0767	1.18	70.11	***
31	.0757	1.17	71.27	***
32	.0740	1.14	72.42	***
33	.0704	1.09	73.50	***
34	.0672	1.04	74.54	**
35	.0660	1.02	75.56	**
36	.0624	.96	76.52	**
37	.0607	.94	77.46	**
38	.0590	.91	78.37	**
39	.0580	.90	79.26	**
40	.0572	.88	80.14	**
41	.0546	.84	80.99	**
42	.0531	.82	81.81	**
43	.0514	.79	82.60	**

Figura C.1: Histograma com os primeiros 43 valores próprios da AC do conjunto de dados notícias.

Anexo D

Classificação Hierárquica - dados notícias

Tabela D.1: Classes formadas através da aplicação da Classificação Hierárquica às 30 coordenadas fatoriais das 227 notícias — partição em 3 classes.

Classe	Notícias				
Classe 1	1	45	85	136	180
	2	46	86	138	182
	4	47	87	139	184
	5	48	90	140	185
	6	49	92	141	186
	7	50	93	142	187
	8	51	95	143	188
	9	52	97	144	189
	11	53	98	145	190
	12	54	102	147	191
	13	55	103	149	194
	14	56	104	150	196
	15	57	105	151	197
	16	58	106	152	198
	17	59	107	153	199
	19	60	108	154	200
	21	61	109	155	202
	22	62	110	156	203
	23	63	111	157	204
	24	65	112	158	205
	25	68	113	159	206
	26	69	114	161	207
	27	70	115	163	208

	28	71	119	164	209
	29	72	121	165	210
	30	73	122	166	211
	32	74	123	167	212
	33	75	124	168	213
	34	76	125	170	215
	35	78	127	171	218
	36	79	128	172	219
	37	80	130	173	220
	39	81	131	174	221
	41	82	132	175	224
	42	83	134	176	225
	43	84	135	178	227
Classe 2	3	88	117	162	195
	10	91	118	169	201
	18	94	120	177	214
	38	96	126	179	216
	44	99	129	181	217
	64	100	133	183	222
	67	101	146	192	
	77	116	160	193	
Classe 3	20	40	89	148	226
	31	66	137	223	

Tabela D.2: Classes formadas através da aplicação da Classificação Hierárquica às 30 coordenadas fatoriais das 227 notícias — partição em 23 classes.

Classe	Notícias				
Classe 1	1	12	33	136	206
Classe 2	2	39	83	144	200
	14	53	90	164	213
	15	54	114	166	218
	16	68	115	174	225
	17	71	124	175	
	19	79	128	178	
	22	81	141	197	
	28	82	143	199	
Classe 3	3	162	181	193	
Classe 4	4	61	127	170	191
	11	62	140	172	202
	21	70	142	173	207
	26	72	145	182	208

	30	74	149	184	215
	35	78	150	185	221
	41	84	153	187	
	46	97	163	188	
	48	103	165	189	
	52	113	167	190	
Classe 5	5	73	93		
	25	87	171		
Classe 6	6	50	51	130	
Classe 7	7	111	161	198	
	42	131	176	211	
	102	158	194		
Classe 8	8	104	151	205	
	23	112	152	209	
	76	147	154	210	
Classe 9	9	24	203		
	13	49	220		
Classe 10	10	44	118	183	217
Classe 11	18	94	100	117	
	88	99	101	201	216
Classe 12	20	40	89	148	226
	31	66	137	223	
Classe 13	27	75	122		
Classe 14	29	63	119	168	212
	43	69	125	180	219
	45	98	132	186	227
	55	107	155	196	
Classe 15	32	59	86	92	157
Classe 16	34	106	135		
Classe 17	36	47	58	60	65
Classe 18	37	121	159	204	
Classe 19	38	91	129	192	
	64	120	160	195	
	77	126	169	214	
Classe 20	56	57	108		
Classe 21	67	116	146	179	
	96	133	177	222	
Classe 22	80	95	134		
	85	123	156		
Classe 23	105	110	139		
	109	138	224		

Anexo E

Classificação Não Hierárquica - dados notícias

Tabela E.1: Classes formadas através da aplicação do algoritmo K-médias às 30 coordenadas fatoriais das 227 notícias — partição em 27 classes.

Classe	Notícias				
Classe 1	1	41	136	206	
	9	132	186	220	
Classe 2	105	110	139		
	109	138	224		
Classe 3	3	162	169	181	193
Classe 4	18	94	100	117	201
	88	99	101	120	216
Classe 5	5	25	171		
Classe 6	10	118	183		
	44	129	217		
Classe 7	135				
Classe 8	8				
Classe 9	56	57	108		
Classe 10	47				
Classe 11	2	55	114	164	197
	11	61	115	166	198
	14	63	119	167	199
	15	64	126	168	200
	16	68	127	170	202
	17	70	128	173	203
	19	71	131	174	205
	22	74	140	175	207
	26	77	141	176	208

	28	78	142	178	209
	30	79	143	180	210
	38	81	144	182	211
	39	82	147	184	213
	42	83	149	185	214
	43	84	153	189	215
	45	91	155	191	218
	46	98	158	192	219
	48	107	160	194	221
	53	111	161	195	225
	54	113	163	196	227
Classe 12	12				
Classe 13	13				
Classe 14	124				
Classe 15	36				
Classe 16	4	35	125	172	
	7	62	145	187	
	21	97	150	188	
	33	102	165	190	
Classe 17	51	121	130	159	204
Classe 18	67	116	146	179	
	96	133	177	222	
Classe 19	32	59	86	92	157
Classe 20	20	89	226		
	31	137			
	40	148			
	66	223			
Classe 21	6	50			
Classe 22	34	106			
Classe 23	23	112	152	212	
	104	151	154		
Classe 24	29	65	76	90	123
	52	69	80	93	134
	58	72	85	95	156
	60	73	87	103	
Classe 25	49				
Classe 26	37				
Classe 27	27	75	122		

Anexo F

Mapas de Kohonen - dados notícias.

Figura F.1: Mapa de Kohonen (4 x 4) representando as 227 notícias e as 87 entidades.

Figura F.2: Mapa de Kohonen (5 x 5) representando as 227 notícias e as 87 entidades.

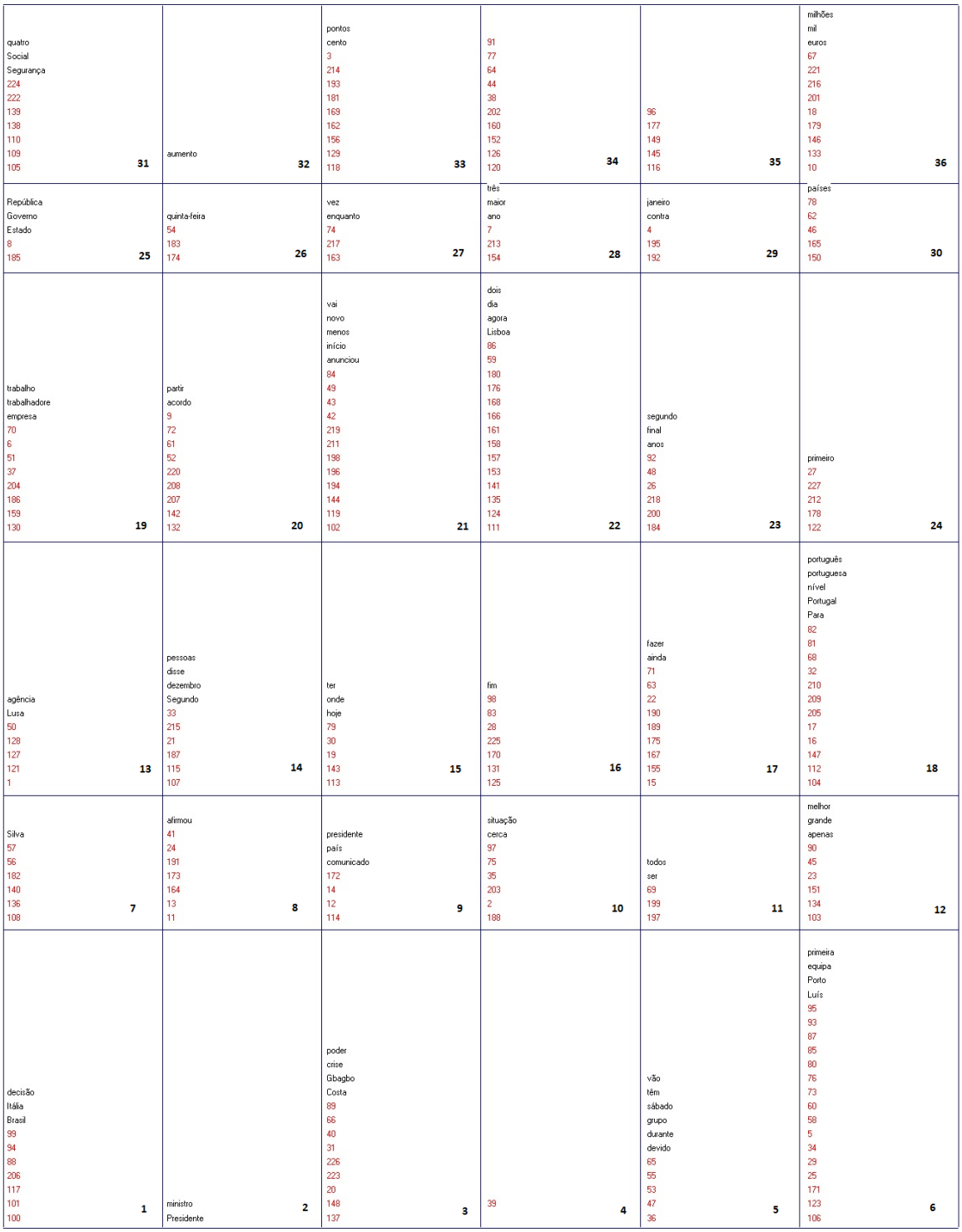

Figura F.3: Mapa de Kohonen (6 x 6) representando as 227 notícias e as 87 entidades.

Anexo G

Tabela de Contingência — dados notícias e entidades

Entidades	Notícia (frequência)									
África*do*Sul	11(5)	35(1)	46(3)	92(1)	175(10)					
Agência*Brasil	88(4)	100(7)	117(7)	149(1)						
Agência*Lusa	29(2)	56(1)	57(1)	98(1)	104(1)	174(1)	188(3)	189(3)		
Alassane*Ouattar	20(1)	31(2)	39(1)	40(2)	66(1)	89(1)	137(1)	148(1)	223(1)	226(1)
Ano*Novo	2(2)	4(1)	7(1)	22(1)	24(2)	30(2)	70(1)	97(1)	125(1)	161(1)
	170(3)	172(1)	175(1)	221(2)						
BPN	56(5)	57(5)	108(14)	184(4)						
Benfica	29(6)	61(3)	80(3)	85(1)	90(1)	95(2)	171(4)	205(1)		
Brasil	11(4)	26(1)	41(2)	61(1)	94(2)	99(6)	101(6)	136(1)	140(1)	173(1)
	182(1)	191(3)	201(2)	206(3)	214(1)	216(1)				
Brasília	6(3)	41(1)	51(3)	88(1)	94(1)	99(1)	100(1)	101(1)	117(1)	130(2)
	182(1)									
Caixa*Geral*de*Aposentações	133(8)	186(4)								
Caixa*Geral*de*Depósitos	10(2)	56(1)	57(1)	71(1)	108(3)	127(1)	183(1)	184(2)	202(1)	
Cavaco*Silva	56(7)	57(7)	71(1)	108(10)						
Cesare*Battisti	88(5)	94(3)	99(6)	100(7)	101(6)	117(7)				
China	11(2)	169(2)	170(1)	201(10)	213(1)	214(5)	216(4)			
Coimbra	28(1)	98(1)	131(5)	158(3)	176(2)	180(2)	211(1)	212(1)		
Costa*do*Marfim	39(2)	66(4)	89(2)	137(3)	148(7)	172(2)	173(1)	223(2)	226(1)	
Diário*da*República	105(1)	109(1)	110(1)	119(1)	126(2)	129(2)	132(1)	133(3)	138(1)	139(1)
	153(1)	163(1)	177(1)	186(2)	222(1)	224(1)				
Espanha	47(1)	59(1)	64(1)	90(1)	92(1)	104(1)	147(1)	150(1)	165(1)	183(1)
	209(4)	210(4)	212(1)	227(1)						
Estado	1(1)	10(3)	20(1)	24(1)	38(2)	41(3)	77(1)	78(1)	109(1)	110(1)
	114(1)	127(2)	133(2)	138(1)	139(1)	148(1)	167(1)	173(1)	178(1)	182(1)
	186(2)	188(1)	189(2)	190(1)	192(1)	206(1)				
Europa	3(1)	16(1)	17(1)	22(2)	27(2)	29(1)	54(1)	55(1)	120(1)	165(1)
	195(2)	209(1)	210(1)							
ex-ativista	88(2)	99(4)	100(3)	101(4)	117(3)					
FC*Porto	73(1)	87(4)	93(3)	151(3)	171(6)					
França	22(1)	75(1)	99(1)	100(1)	101(1)	117(1)	134(1)	148(1)	196(1)	203(2)
	223(1)									
Governo	4(1)	11(2)	41(2)	52(2)	66(1)	78(1)	88(2)	89(1)	94(1)	99(1)
	101(1)	104(2)	105(3)	109(5)	110(5)	119(1)	122(1)	126(1)	129(1)	131(4)
	138(6)	139(6)	140(3)	148(1)	149(1)	153(1)	158(1)	167(3)	169(3)	170(5)
	174(1)	175(1)	176(1)	177(3)	182(3)	183(1)	184(2)	187(1)	188(1)	190(1)
	195(1)	196(1)	200(1)	202(1)	203(1)	211(1)	217(1)	221(2)	227(3)	

Guarda	93(1)	128(2)	142(2)	194(1)	207(3)	208(3)				
Itália	15(1)	88(6)	94(3)	99(5)	100(5)	101(5)	117(5)	150(1)		
Laurent*Gbagbo	20(3)	31(1)	39(1)	40(7)	66(5)	89(4)	137(4)	148(2)	223(8)	226(2)
Lisboa	4(1)	29(3)	30(1)	45(3)	68(1)	69(2)	70(1)	82(1)	84(1)	95(1)
	108(1)	118(2)	119(1)	131(1)	158(1)	161(1)	176(1)	180(3)	181(1)	182(1)
	193(1)	194(2)	211(1)	212(4)						
Lousã	131(5)	158(3)	176(3)	211(1)						
Lusa	5(1)	6(2)	14(1)	15(1)	28(1)	30(1)	44(1)	47(1)	50(3)	51(1)
	58(1)	60(1)	61(1)	65(2)	68(1)	70(2)	78(2)	81(1)	82(1)	84(1)
	91(1)	104(2)	105(1)	108(4)	119(1)	121(1)	125(1)	128(1)	130(1)	138(1)
	139(1)	142(1)	143(1)	147(1)	155(2)	175(1)	179(1)	183(2)	184(1)	185(2)
	187(2)	190(2)	202(1)	209(1)	210(1)	212(1)	215(1)	219(1)	222(1)	224(1)
	225(1)	227(1)								
Moçambique	15(1)	35(3)	122(1)	127(3)	175(3)	212(1)				
ONU	20(1)	66(1)	69(2)	89(2)	136(2)	137(3)	227(1)			
PS	56(1)	57(1)	65(2)	78(1)	108(1)	129(1)	157(1)	158(1)	176(1)	188(3)
PSD	52(2)	65(1)	78(1)	109(1)	110(1)	129(1)	158(2)	176(2)	188(2)	
PSI	91(8)	120(2)	181(2)	193(2)						
Porto	6(1)	45(1)	51(1)	55(1)	80(1)	82(1)	84(1)	130(1)	180(1)	188(4)
	194(1)	212(1)								
Portugal	15(1)	29(1)	47(1)	53(1)	55(1)	58(2)	60(2)	68(3)	69(1)	72(1)
	76(1)	90(1)	103(3)	104(2)	107(2)	118(1)	122(2)	128(1)	147(1)	151(5)
	152(1)	155(1)	160(3)	163(1)	165(1)	184(1)	187(2)	188(2)	189(1)	190(2)
	191(1)	194(1)	195(2)	197(1)	201(2)	205(1)	207(1)	208(1)	209(1)	210(3)
	212(4)									
Presidente	20(1)	24(1)	31(3)	33(1)	39(1)	53(1)	65(1)	89(1)	100(1)	114(2)
	117(1)	136(1)	169(1)	173(6)	223(2)	226(2)	227(1)			
Presidente*Lula*da*Silva	41(1)	99(4)	101(4)	140(4)						
Presidente*da*República	36(2)	40(1)	56(6)	57(6)	107(1)	108(2)	129(1)	172(4)	186(1)	
primeiro-ministro	4(1)	30(1)	40(2)	88(1)	94(1)	97(2)	114(1)	169(1)	182(1)	221(2)
RN	1(1)	2(1)	3(1)	7(1)	8(1)	9(1)	12(1)	13(1)	19(1)	22(1)
	24(1)	27(1)	30(1)	223(1)	226(1)					
Reino*Unido	58(3)	60(3)	66(3)	89(1)	97(3)	152(1)	221(3)			
Rússia	11(2)	15(1)	24(2)	114(4)	136(1)	178(1)	212(1)	214(1)		
SNGB	6(3)	51(5)	130(5)							
Sara*Moreira	16(6)	17(6)	205(1)							
Segurança*Social	105(2)	107(1)	109(3)	110(3)	138(6)	139(6)				
Supremo*Tribunal*Federal	88(2)	99(3)	100(5)	101(3)	117(5)					
União*Europeia	4(1)	31(2)	66(2)	68(4)	75(5)	112(1)	114(3)	136(2)	145(2)	160(1)
	195(1)	212(1)	213(1)							
Varzim*Sol	6(4)	51(5)	130(3)							

Figura G.1: Notícia e frequência das 50 entidades retidas.

Anexo H

Análise de Correspondências dos dados notícias e entidades — Histograma com os valores próprios (*output* parcial).

number	Eigen value	percent.	cumulat. percent.	
1	.8706	5.17	5.17	**********
2	.8504	5.05	10.23	**********
3	.8121	4.83	15.05	**********
4	.8018	4.76	19.81	**********
5	.7902	4.70	24.51	**********
6	.7486	4.45	28.96	*********
7	.7475	4.44	33.40	*********
8	.7106	4.22	37.62	*********
9	.6661	3.96	41.58	********
10	.6344	3.77	45.35	********
11	.6321	3.76	49.10	********
12	.6042	3.59	52.69	********
13	.5585	3.32	56.01	*******
14	.5391	3.20	59.22	*******
15	.4885	2.90	62.12	*******
16	.4735	2.81	64.93	*******
17	.4632	2.75	67.68	*******
18	.3887	2.31	69.99	******
19	.3773	2.24	72.24	******
20	.3441	2.04	74.28	*****
21	.3296	1.96	76.24	*****
22	.3214	1.91	78.15	*****
23	.3163	1.88	80.03	*****
24	.2851	1.69	81.72	****
25	.2654	1.58	83.30	****

Figura H.1: Histograma com os primeiros 25 valores próprios.

Anexo I

Análise de Correspondências - Notícias e entidades

Tabela I.1: Coordenadas, contribuições absolutas e relativas das 50 entidades retidas para o eixo 1.

Entidades	Coordenadas	CTA	CTR
Agência*Brasil	**1,91**	**7,0**	**0,28**
Agência*Lusa	-0,61	0,5	0,02
Alassane*Ouattar	0,19	0,0	0,00
Ano*Novo	-0,15	0,0	0,00
BPN	-0,82	1,9	0,05
Benfica	**-1,05**	**2,4**	*0,04*
Brasil	**1,03**	**3,9**	**0,12**
Brasília	1,02	1,7	0,12
Caixa*Geral*de*Aposentações	-0,22	0,1	0,00
Caixa*Geral*de*Depósitos	-0,60	0,5	0,02
Cavaco*Silva	-0,86	1,9	0,05
Cesare*Battisti	**1,83**	**11,6**	**0,52**
China	0,54	0,7	0,01
Coimbra	-0,41	0,3	0,01
Costa*do*Marfim	0,14	0,1	0,00
Diário*da*República	-0,19	0,1	0,00
Espanha	-0,50	0,5	0,01
Estado	-0,01	0,0	0,00
Europa	**-1,40**	**3,2**	**0,08**
FC*Porto	**-1,20**	**2,5**	*0,03*
França	0,87	0,9	0,03
Governo	0,07	0,0	0,00
Guarda	-0,79	0,8	0,01

Itália	**1,73**	**9,4**	**0,46**
Laurent*Gbagbo	0,18	0,1	0,00
Lisboa	-0,57	1,2	0,03
Lousã	-0,37	0,2	0,00
Lusa	-0,37	1,0	0,02
Moçambique	-0,04	0,0	0,00
ONU	0,08	0,0	0,00
PS	-0,51	0,3	0,01
PSD	-0,29	0,1	0,01
PSI	**-1,88**	**5,0**	*0,06*
Porto	-0,34	0,2	0,01
Portugal	-0,51	1,8	0,04
Presidente	0,36	0,3	0,01
Presidente*Lula*da*Silva	**1,41**	**2,6**	**0,10**
Presidente*da*República	-0,68	1,1	0,03
RN	-0,36	0,2	0,00
Reino*Unido	-0,09	0,0	0,00
Rússia	0,21	0,1	0,00
SNGB	0,59	0,5	0,01
Sara*Moreira	**-4,17**	**22,9**	**0,25**
Segurança*Social	-0,07	0,0	0,00
Supremo*Tribunal*Federal	**1,88**	**6,4**	**0,47**
União*Europeia	0,00	0,0	0,00
Varzim*Sol	0,59	0,4	0,01
ex-ativista	**1,83**	**5,4**	**0,48**
primeiro-ministro	0,34	0,1	0,01
África*do*Sul	0,22	0,1	0,00

Tabela I.2: Coordenadas, contribuições absolutas e relativas das 227 notícias para o eixo 1.

Notícias	Coordenadas	CTA	CTR		Notícias	Coordenadas	CTA	CTR
1	-0,20	0,0	0,00		115	0,00	0,0	0,00
2	-0,24	0,0	0,00		116	0,00	0,0	0,00
3	-0,94	0,2	0,02		**117**	**1,86**	**10,5**	**0,44**
4	-0,07	0,0	0,00		118	-0,59	0,1	0,02
5	-0,40	0,0	0,01		119	-0,28	0,0	0,01
6	0,50	0,3	0,01		**120**	**-1,84**	**1,0**	**0,08**
7	-0,28	0,0	0,00		121	-0,40	0,0	0,01
8	-0,39	0,0	0,00		122	-0,26	0,0	0,01
9	-0,39	0,0	0,00		123	0,00	0,0	0,00
10	-0,26	0,0	0,00		124	0,00	0,0	0,00

11	0,49	0,4	0,02	125	-0,28	0,0	0,00
12	-0,39	0,0	0,00	126	-0,11	0,0	0,00
13	-0,39	0,0	0,00	127	-0,13	0,0	0,00
14	-0,40	0,0	0,01	128	-0,66	0,2	0,02
15	0,22	0,0	0,01	129	-0,32	0,1	0,01
16	**-4,04**	**11,6**	**0,25**	130	0,54	0,4	0,01
17	**-4,04**	**11,6**	**0,25**	131	-0,30	0,1	0,00
18	0,00	0,0	0,00	132	-0,20	0,0	0,00
19	-0,39	0,0	0,00	133	-0,19	0,0	0,00
20	0,18	0,0	0,00	134	0,93	0,1	0,01
21	0,00	0,0	0,00	135	0,00	0,0	0,00
22	-0,52	0,1	0,01	136	0,27	0,1	0,01
23	0,00	0,0	0,00	137	0,15	0,0	0,00
24	0,01	0,0	0,00	138	-0,04	0,0	0,00
25	0,00	0,0	0,00	139	-0,04	0,0	0,00
26	1,10	0,1	0,04	**140**	**0,92**	**0,7**	**0,04**
27	-1,13	0,4	0,03	141	0,00	0,0	0,00
28	-0,42	0,0	0,01	142	-0,70	0,1	0,01
29	**-0,92**	**1,1**	**0,06**	143	-0,40	0,0	0,01
30	-0,23	0,0	0,00	144	0,00	0,0	0,00
31	0,22	0,0	0,00	145	0,01	0,0	0,00
32	0,00	0,0	0,00	146	0,00	0,0	0,00
33	0,38	0,0	0,00	147	-0,49	0,1	0,03
34	0,00	0,0	0,00	148	0,21	0,1	0,00
35	0,03	0,0	0,00	149	1,06	0,2	0,07
36	-0,73	0,1	0,01	150	0,66	0,1	0,02
37	0,00	0,0	0,00	**151**	**-0,82**	**0,5**	**0,05**
38	-0,01	0,0	0,00	152	-0,32	0,0	0,01
39	0,22	0,0	0,00	153	-0,06	0,0	0,00
40	0,15	0,0	0,00	154	0,00	0,0	0,00
41	0,55	0,3	0,04	155	-0,45	0,1	0,02
42	0,00	0,0	0,00	156	0,00	0,0	0,00
43	0,00	0,0	0,00	157	-0,55	0,0	0,00
44	-0,40	0,0	0,01	158	-0,39	0,2	0,01
45	-0,55	0,1	0,01	159	0,00	0,0	0,00
46	0,24	0,0	0,00	160	-0,41	0,1	0,01
47	-0,49	0,1	0,03	161	-0,39	0,0	0,01
48	0,00	0,0	0,00	162	0,00	0,0	0,00
49	0,00	0,0	0,00	163	-0,37	0,0	0,01
50	-0,40	0,0	0,01	164	0,00	0,0	0,00
51	**0,59**	**0,5**	*0,02*	165	-0,86	0,2	0,05
52	-0,12	0,0	0,00	166	0,00	0,0	0,00

53	-0,08	0,0	0,00	167	0,05	0,0	0,00
54	-1,50	0,2	0,03	168	0,00	0,0	0,00
55	-0,80	0,2	0,04	169	0,30	0,1	0,01
56	**-0,81**	**1,4**	**0,06**	170	0,05	0,0	0,00
57	**-0,81**	**1,4**	**0,06**	**171**	**-1,22**	**1,5**	**0,05**
58	-0,30	0,1	0,00	172	-0,40	0,1	0,01
59	-0,54	0,0	0,01	173	0,39	0,1	0,01
60	-0,30	0,1	0,00	174	-0,29	0,0	0,00
61	-0,54	0,1	0,01	175	0,11	0,0	0,00
62	0,00	0,0	0,00	176	-0,38	0,1	0,01
63	0,00	0,0	0,00	177	0,00	0,0	0,00
64	-0,54	0,0	0,01	178	0,11	0,0	0,00
65	-0,31	0,1	0,01	179	-0,40	0,0	0,01
66	0,10	0,0	0,00	180	-0,51	0,2	0,02
67	0,00	0,0	0,00	**181**	**-1,55**	**0,7**	**0,06**
68	-0,29	0,1	0,01	182	0,27	0,1	0,02
69	-0,32	0,1	0,01	183	-0,38	0,1	0,02
70	-0,39	0,1	0,02	184	-0,56	0,3	0,03
71	-0,78	0,1	0,02	185	-0,40	0,0	0,01
72	-0,54	0,0	0,02	186	-0,23	0,0	0,00
73	-1,28	0,2	0,02	187	-0,36	0,1	0,03
74	0,00	0,0	0,00	188	-0,42	0,3	0,02
75	0,16	0,0	0,00	189	-0,42	0,1	0,01
76	-0,54	0,0	0,02	190	-0,30	0,1	0,02
77	-0,01	0,0	0,00	191	0,69	0,2	0,03
78	-0,27	0,0	0,01	192	-0,01	0,0	0,00
79	0,00	0,0	0,00	**193**	**-1,55**	**0,7**	**0,06**
80	-0,94	0,4	0,03	194	-0,60	0,2	0,03
81	-0,40	0,0	0,01	195	-0,67	0,3	0,04
82	-0,46	0,1	0,02	196	0,50	0,1	0,01
83	0,00	0,0	0,00	197	-0,54	0,0	0,02
84	-0,46	0,1	0,02	198	0,00	0,0	0,00
85	-1,13	0,1	0,02	199	0,00	0,0	0,00
86	0,00	0,0	0,00	200	0,07	0,0	0,00
87	**-1,28**	**0,7**	*0,02*	201	0,50	0,3	0,01
88	**1,68**	**6,6**	**0,45**	202	-0,32	0,0	0,01
89	0,15	0,0	0,00	203	0,64	0,1	0,01
90	-0,74	0,2	0,04	204	0,00	0,0	0,00
91	**-1,83**	**3,1**	**0,05**	**205**	**-2,05**	**1,3**	**0,25**
92	-0,15	0,0	0,00	206	0,83	0,3	0,04
93	-0,92	0,4	0,03	207	-0,77	0,2	0,01
94	**1,51**	**2,3**	**0,28**	208	-0,77	0,2	0,01

95	-0,96	0,3	0,03	209	-0,63	0,4	0,03
96	0,00	0,0	0,00	210	-0,63	0,4	0,03
97	0,04	0,0	0,00	211	-0,35	0,0	0,01
98	-0,55	0,1	0,01	212	-0,41	0,3	0,04
99	**1,60**	**8,1**	**0,44**	213	0,29	0,0	0,00
100	**1,86**	**10,5**	**0,44**	214	0,60	0,3	0,01
101	**1,60**	**8,1**	**0,44**	215	-0,40	0,0	0,01
102	0,00	0,0	0,00	216	0,69	0,2	0,02
103	-0,57	0,1	0,02	217	0,07	0,0	0,00
104	-0,33	0,1	0,03	218	0,01	0,0	0,00
105	-0,08	0,0	0,00	219	-0,40	0,0	0,01
106	0,00	0,0	0,00	220	-0,04	0,0	0,00
107	-0,47	0,1	0,02	221	0,03	0,0	0,00
108	**-0,79**	**2,2**	**0,06**	222	-0,30	0,0	0,01
109	-0,03	0,0	0,00	223	0,22	0,1	0,00
110	-0,03	0,0	0,00	224	-0,30	0,0	0,01
111	0,00	0,0	0,00	225	-0,40	0,0	0,01
112	0,01	0,0	0,00	226	0,16	0,0	0,00
113	0,00	0,0	0,00	227	-0,04	0,0	0,00
114	0,18	0,0	0,00				

Tabela I.3: Coordenadas, contribuições absolutas e relativas das 50 entidades retidas para o eixo 2.

Entidades	Coordenadas	CTA	CTR
Agência*Brasil	**-1,05**	**2,2**	**0,09**
Agência*Lusa	0,53	0,4	0,01
Alassane*Ouattar	0,28	0,1	0,01
Ano*Novo	0,11	0,0	0,00
BPN	**1,53**	**6,8**	**0,18**
Benfica	-0,39	0,3	0,01
Brasil	-0,51	1,0	0,03
Brasília	-0,52	0,5	0,03
Caixa*Geral*de*Aposentações	0,87	0,9	0,01
Caixa*Geral*de*Depósitos	1,16	1,8	0,08
Cavaco*Silva	**1,62**	**6,8**	**0,19**
Cesare*Battisti	**-1,01**	**3,6**	**0,16**
China	-0,21	0,1	0,00
Coimbra	0,44	0,3	0,01
Costa*do*Marfim	0,35	0,3	0,01
Diário*da*República	0,59	0,8	0,02
Espanha	-0,11	0,0	0,00

Estado	0,38	0,5	0,01
Europa	**-1,51**	**3,8**	**0,09**
FC*Porto	-0,23	0,1	0,00
França	-0,48	0,3	0,01
Governo	0,22	0,5	0,01
Guarda	0,14	0,0	0,00
Itália	**-0,96**	**3,0**	**0,14**
Laurent*Gbagbo	0,30	0,3	0,01
Lisboa	0,18	0,1	0,00
Lousã	0,49	0,3	0,01
Lusa	0,33	0,8	0,02
Moçambique	0,24	0,1	0,00
ONU	0,24	0,1	0,00
PS	0,82	0,9	0,04
PSD	0,51	0,3	0,01
PSI	-0,59	0,5	0,01
Porto	0,08	0,0	0,00
Portugal	0,02	0,0	0,00
Presidente	0,06	0,0	0,00
Presidente*Lula*da*Silva	-0,70	0,7	0,02
Presidente*da*República	**1,39**	**4,8**	**0,12**
RN	-0,31	0,1	0,00
Reino*Unido	0,20	0,1	0,00
Rússia	0,02	0,0	0,00
SNGB	-0,30	0,1	0,00
Sara*Moreira	**-6,23**	**52,4**	**0,55**
Segurança*Social	0,41	0,4	0,01
Supremo*Tribunal*Federal	-1,04	2,0	0,15
União*Europeia	0,05	0,0	0,00
Varzim*Sol	-0,30	0,1	0,00
ex-ativista	-1,01	1,7	0,15
primeiro-ministro	-0,01	0,0	0,00
África*do*Sul	0,06	0,0	0,00

Tabela I.4: Coordenadas, contribuições absolutas e relativas das 227 notícias para o eixo 2.

Notícias	Coordenadas	CTA	CTR		Notícias	Coordenadas	CTA	CTR
1	0,04	0,0	0,00		115	0,00	0,0	0,00
2	-0,03	0,0	0,00		116	0,00	0,0	0,00
3	-0,99	0,2	0,03		**117**	**-1,03**	**3,3**	**0,14**
4	0,12	0,0	0,00		118	0,14	0,0	0,00

5	0,36	0,0	0,01	119	0,36	0,1	0,02
6	-0,24	0,1	0,00	120	-0,98	0,3	0,02
7	-0,11	0,0	0,00	121	0,36	0,0	0,01
8	-0,34	0,0	0,00	122	0,14	0,0	0,00
9	-0,34	0,0	0,00	123	0,00	0,0	0,00
10	0,75	0,3	0,02	124	0,00	0,0	0,00
11	-0,12	0,0	0,00	125	0,24	0,0	0,00
12	-0,34	0,0	0,00	126	0,51	0,1	0,01
13	-0,34	0,0	0,00	127	0,48	0,1	0,01
14	0,36	0,0	0,01	128	0,17	0,0	0,00
15	-0,07	0,0	0,00	129	0,75	0,3	0,05
16	**-6,03**	**26,4**	**0,56**	130	-0,27	0,1	0,00
17	**-6,03**	**26,4**	**0,56**	131	0,41	0,3	0,01
18	0,00	0,0	0,00	132	0,64	0,0	0,01
19	-0,34	0,0	0,00	**133**	**0,79**	**0,8**	*0,02*
20	0,29	0,1	0,01	134	-0,53	0,0	0,00
21	0,00	0,0	0,00	135	0,00	0,0	0,00
22	-0,80	0,3	0,03	136	0,02	0,0	0,00
23	0,00	0,0	0,00	137	0,32	0,1	0,01
24	0,06	0,0	0,00	138	0,36	0,2	0,01
25	0,00	0,0	0,00	139	0,36	0,2	0,01
26	-0,55	0,0	0,01	140	-0,36	0,1	0,01
27	**-1,20**	**0,5**	**0,04**	141	0,00	0,0	0,00
28	0,42	0,0	0,01	142	0,22	0,0	0,00
29	-0,19	0,0	0,00	143	0,36	0,0	0,01
30	0,08	0,0	0,00	144	0,00	0,0	0,00
31	0,15	0,0	0,00	145	0,05	0,0	0,00
32	0,00	0,0	0,00	146	0,00	0,0	0,00
33	0,06	0,0	0,00	147	0,09	0,0	0,00
34	0,00	0,0	0,00	148	0,29	0,1	0,01
35	0,21	0,0	0,00	149	-0,45	0,0	0,01
36	**1,51**	**0,5**	**0,05**	150	-0,58	0,1	0,02
37	0,00	0,0	0,00	151	-0,08	0,0	0,00
38	0,41	0,0	0,01	152	0,12	0,0	0,00
39	0,29	0,0	0,01	153	0,44	0,0	0,01
40	0,36	0,2	0,01	154	0,00	0,0	0,00
41	-0,08	0,0	0,00	155	0,25	0,0	0,01
42	0,00	0,0	0,00	156	0,00	0,0	0,00
43	0,00	0,0	0,00	157	0,89	0,1	0,01
44	0,36	0,0	0,01	158	0,49	0,3	0,02
45	0,17	0,0	0,00	159	0,00	0,0	0,00
46	0,07	0,0	0,00	160	0,03	0,0	0,00

47	0,09	0,0	0,00	161	0,16	0,0	0,00
48	0,00	0,0	0,00	162	0,00	0,0	0,00
49	0,00	0,0	0,00	163	0,33	0,0	0,01
50	0,36	0,0	0,01	164	0,00	0,0	0,00
51	-0,30	0,1	0,00	165	-0,58	0,1	0,02
52	0,39	0,1	0,01	166	0,00	0,0	0,00
53	0,04	0,0	0,00	167	0,28	0,0	0,01
54	-1,64	0,3	0,04	168	0,00	0,0	0,00
55	-0,51	0,1	0,02	169	0,04	0,0	0,00
56	**1,54**	**5,2**	**0,22**	170	0,15	0,0	0,00
57	**1,54**	**5,2**	**0,22**	171	-0,32	0,1	0,00
58	0,18	0,0	0,00	**172**	**0,99**	**0,7**	**0,05**
59	-0,12	0,0	0,00	173	0,07	0,0	0,00
60	0,18	0,0	0,00	174	0,40	0,0	0,01
61	-0,29	0,0	0,00	175	0,14	0,0	0,00
62	0,00	0,0	0,00	176	0,50	0,3	0,02
63	0,00	0,0	0,00	177	0,34	0,0	0,01
64	-0,12	0,0	0,00	178	0,22	0,0	0,00
65	0,52	0,2	0,02	179	0,36	0,0	0,01
66	0,28	0,1	0,01	180	0,27	0,0	0,00
67	0,00	0,0	0,00	181	-0,36	0,0	0,00
68	0,09	0,0	0,00	182	0,02	0,0	0,00
69	0,19	0,0	0,00	183	0,42	0,1	0,02
70	0,26	0,0	0,01	**184**	**1,00**	**1,0**	**0,10**
71	**1,51**	**0,5**	**0,07**	185	0,36	0,0	0,01
72	0,02	0,0	0,00	**186**	**0,82**	**0,6**	**0,03**
73	-0,25	0,0	0,00	187	0,20	0,0	0,01
74	0,00	0,0	0,00	188	0,41	0,3	0,01
75	-0,05	0,0	0,00	189	0,43	0,1	0,01
76	0,02	0,0	0,00	190	0,24	0,0	0,01
77	0,41	0,0	0,01	191	-0,41	0,1	0,01
78	0,47	0,1	0,03	192	0,41	0,0	0,01
79	0,00	0,0	0,00	193	-0,36	0,0	0,00
80	-0,30	0,0	0,00	194	0,13	0,0	0,00
81	0,36	0,0	0,01	195	-0,49	0,2	0,02
82	0,22	0,0	0,00	196	-0,15	0,0	0,00
83	0,00	0,0	0,00	197	0,02	0,0	0,00
84	0,22	0,0	0,00	198	0,00	0,0	0,00
85	-0,43	0,0	0,00	199	0,00	0,0	0,00
86	0,00	0,0	0,00	200	0,23	0,0	0,01
87	-0,25	0,0	0,00	201	-0,24	0,1	0,00
88	**-0,91**	**2,0**	**0,13**	202	0,62	0,1	0,03

89	0,28	0,1	0,01	203	-0,27	0,0	0,00
90	-0,17	0,0	0,00	204	0,00	0,0	0,00
91	-0,53	0,3	0,00	**205**	**-2,39**	**1,8**	**0,34**
92	-0,02	0,0	0,00	206	-0,31	0,0	0,01
93	-0,07	0,0	0,00	207	0,12	0,0	0,00
94	**-0,81**	**0,7**	**0,08**	208	0,12	0,0	0,00
95	-0,22	0,0	0,00	209	-0,19	0,0	0,00
96	0,00	0,0	0,00	210	-0,19	0,0	0,00
97	0,12	0,0	0,00	211	0,36	0,1	0,01
98	0,52	0,1	0,01	212	0,13	0,0	0,00
99	**-0,86**	**2,4**	**0,13**	213	-0,09	0,0	0,00
100	**-1,03**	**3,3**	**0,14**	214	-0,24	0,0	0,00
101	**-0,86**	**2,4**	**0,13**	215	0,36	0,0	0,01
102	0,00	0,0	0,00	216	-0,29	0,0	0,00
103	0,16	0,0	0,00	217	0,23	0,0	0,01
104	0,16	0,0	0,01	218	0,00	0,0	0,00
105	0,37	0,1	0,02	219	0,36	0,0	0,01
106	0,00	0,0	0,00	220	0,04	0,0	0,00
107	0,50	0,1	0,03	221	0,15	0,0	0,00
108	**1,43**	**7,4**	**0,20**	222	0,50	0,1	0,02
109	0,37	0,2	0,02	223	0,19	0,1	0,00
110	0,37	0,2	0,02	224	0,50	0,1	0,02
111	0,00	0,0	0,00	225	0,36	0,0	0,01
112	0,05	0,0	0,00	226	0,16	0,0	0,00
113	0,00	0,0	0,00	227	0,18	0,0	0,01
114	0,07	0,0	0,00				

Tabela I.5: Coordenadas, contribuições absolutas e relativas das 50 entidades retidas para o eixo 3.

Entidades	Coordenadas	CTA	CTR
Agência*Brasil	-0,40	0,3	0,01
Agência*Lusa	-0,28	0,1	0,00
Alassane*Ouattar	**1,90**	**4,7**	**0,29**
Ano*Novo	0,45	0,4	0,01
BPN	0,08	0,0	0,00
Benfica	**-0,99**	**2,2**	*0,03*
Brasil	-0,11	0,0	0,00
Brasília	-0,95	1,6	0,11
Caixa*Geral*de*Aposentações	0,30	0,1	0,00
Caixa*Geral*de*Depósitos	0,06	0,0	0,00
Cavaco*Silva	0,12	0,0	0,00

Cesare*Battisti	-0,39	0,6	0,02
China	0,18	0,1	0,00
Coimbra	-0,47	0,4	0,01
Costa*do*Marfim	**1,81**	**8,6**	**0,23**
Diário*da*República	0,04	0,0	0,00
Espanha	-0,28	0,2	0,00
Estado	0,30	0,3	0,01
Europa	0,16	0,0	0,00
FC*Porto	**-1,46**	**3,9**	**0,05**
França	0,45	0,3	0,01
Governo	0,02	0,0	0,00
Guarda	-0,91	1,1	0,01
Itália	-0,38	0,5	0,02
Laurent*Gbagbo	**1,94**	**15,1**	**0,33**
Lisboa	-0,63	1,5	0,04
Lousã	-0,43	0,2	0,01
Lusa	-0,43	1,4	0,03
Moçambique	-0,01	0,0	0,00
ONU	**1,41**	**2,6**	**0,09**
PS	-0,16	0,0	0,00
PSD	-0,23	0,1	0,00
PSI	**-4,31**	**28,3**	**0,29**
Porto	-0,81	1,1	0,04
Portugal	-0,22	0,4	0,01
Presidente	**1,17**	**4,0**	**0,10**
Presidente*Lula*da*Silva	-0,29	0,1	0,00
Presidente*da*República	0,40	0,4	0,01
RN	1,09	1,9	0,03
Reino*Unido	0,81	1,2	0,03
Rússia	0,49	0,3	0,01
SNGB	**-1,61**	**3,7**	**0,09**
Sara*Moreira	**2,05**	**5,9**	**0,06**
Segurança*Social	0,00	0,0	0,00
Supremo*Tribunal*Federal	-0,40	0,3	0,02
União*Europeia	0,78	1,7	0,03
Varzim*Sol	**-1,60**	**3,3**	**0,09**
ex-ativista	-0,39	0,3	0,02
primeiro-ministro	0,60	0,5	0,02
África*do*Sul	0,04	0,0	0,00

Tabela I.6: Coordenadas, contribuições absolutas e relativas das 227 notícias para o eixo 3.

Notícias	Coordenadas	CTA	CTR		Notícias	Coordenadas	CTA	CTR
1	0,77	0,1	0,02		115	0,00	0,0	0,00
2	0,73	0,2	0,02		116	0,00	0,0	0,00
3	0,69	0,1	0,01		117	-0,37	0,4	0,02
4	0,27	0,0	0,01		118	-0,55	0,1	0,02
5	-0,48	0,0	0,01		119	-0,28	0,0	0,01
6	**-1,34**	**2,6**	0,11		**120**	**-3,13**	**3,2**	**0,23**
7	0,85	0,2	0,02		121	-0,48	0,0	0,01
8	1,20	0,2	0,02		122	-0,12	0,0	0,00
9	1,20	0,2	0,02		123	0,00	0,0	0,00
10	0,22	0,0	0,00		124	0,00	0,0	0,00
11	0,08	0,0	0,00		125	0,01	0,0	0,00
12	1,20	0,2	0,02		126	0,04	0,0	0,00
13	1,20	0,2	0,02		127	0,11	0,0	0,00
14	-0,48	0,0	0,01		128	-0,69	0,2	0,02
15	-0,13	0,0	0,00		129	0,02	0,0	0,00
16	**1,97**	**3,0**	**0,06**		**130**	**-1,48**	**2,9**	**0,10**
17	**1,97**	**3,0**	**0,06**		131	-0,37	0,2	0,01
18	0,00	0,0	0,00		132	0,04	0,0	0,00
19	1,20	0,2	0,02		133	0,27	0,1	0,00
20	**1,68**	**2,1**	**0,28**		134	0,49	0,0	0,00
21	0,00	0,0	0,00		135	0,00	0,0	0,00
22	0,51	0,1	0,01		**136**	**0,94**	**0,7**	**0,07**
23	0,00	0,0	0,00		**137**	**1,95**	**4,5**	**0,26**
24	0,70	0,4	0,04		138	0,00	0,0	0,00
25	0,00	0,0	0,00		139	0,00	0,0	0,00
26	-0,12	0,0	0,00		140	-0,17	0,0	0,00
27	0,52	0,1	0,01		141	0,00	0,0	0,00
28	-0,50	0,1	0,01		142	-0,83	0,2	0,02
29	**-0,72**	**0,7**	*0,03*		143	-0,48	0,0	0,01
30	0,28	0,1	0,01		144	0,00	0,0	0,00
31	**1,50**	**2,0**	**0,16**		145	0,87	0,2	0,02
32	0,00	0,0	0,00		146	0,00	0,0	0,00
33	1,30	0,2	0,04		147	-0,35	0,0	0,01
34	0,00	0,0	0,00		**148**	**1,64**	**3,8**	**0,18**
35	0,00	0,0	0,00		149	-0,21	0,0	0,00
36	0,44	0,0	0,00		150	-0,37	0,0	0,01
37	0,00	0,0	0,00		**151**	**-0,76**	**0,5**	**0,04**
38	0,33	0,0	0,00		152	0,33	0,0	0,01

39	**1,92**	**2,0**	**0,28**	153	0,03	0,0	0,00
40	**1,75**	**4,0**	**0,21**	154	0,00	0,0	0,00
41	-0,07	0,0	0,00	155	-0,40	0,1	0,02
42	0,00	0,0	0,00	156	0,00	0,0	0,00
43	0,00	0,0	0,00	157	-0,18	0,0	0,00
44	-0,48	0,0	0,01	158	-0,40	0,2	0,01
45	-0,75	0,2	0,03	159	0,00	0,0	0,00
46	0,04	0,0	0,00	160	0,03	0,0	0,00
47	-0,35	0,0	0,01	161	-0,10	0,0	0,00
48	0,00	0,0	0,00	162	0,00	0,0	0,00
49	0,00	0,0	0,00	163	-0,10	0,0	0,00
50	-0,48	0,1	0,01	164	0,00	0,0	0,00
51	**-1,49**	**3,6**	**0,10**	165	-0,13	0,0	0,00
52	-0,12	0,0	0,00	166	0,00	0,0	0,00
53	0,52	0,1	0,02	167	0,10	0,0	0,00
54	0,18	0,0	0,00	168	0,00	0,0	0,00
55	-0,32	0,0	0,01	169	0,35	0,1	0,02
56	0,17	0,1	0,00	170	0,20	0,0	0,00
57	0,17	0,1	0,00	**171**	**-1,41**	**2,2**	**0,06**
58	0,29	0,1	0,00	**172**	**0,90**	**0,6**	**0,04**
59	-0,31	0,0	0,00	**173**	**1,11**	**1,2**	**0,07**
60	0,29	0,1	0,00	174	-0,14	0,0	0,00
61	-0,78	0,3	0,03	175	0,02	0,0	0,00
62	0,00	0,0	0,00	176	-0,38	0,2	0,01
63	0,00	0,0	0,00	177	0,03	0,0	0,00
64	-0,31	0,0	0,00	178	0,44	0,0	0,01
65	-0,05	0,0	0,00	179	-0,48	0,0	0,01
66	**1,58**	**4,6**	**0,33**	180	-0,67	0,3	0,03
67	0,00	0,0	0,00	**181**	**-3,42**	**3,8**	**0,30**
68	0,17	0,0	0,00	182	-0,10	0,0	0,00
69	0,30	0,0	0,00	183	-0,24	0,0	0,01
70	-0,29	0,0	0,01	184	-0,02	0,0	0,00
71	0,10	0,0	0,00	185	-0,48	0,1	0,01
72	-0,25	0,0	0,00	186	0,28	0,1	0,00
73	-1,62	0,3	0,04	187	-0,29	0,0	0,02
74	0,00	0,0	0,00	188	-0,36	0,2	0,01
75	0,80	0,4	0,02	189	-0,08	0,0	0,00
76	-0,25	0,0	0,00	190	-0,19	0,0	0,01
77	0,33	0,0	0,00	191	-0,15	0,0	0,00
78	-0,17	0,0	0,00	192	0,33	0,0	0,00
79	0,00	0,0	0,00	**193**	**-3,42**	**3,8**	**0,30**
80	**-1,04**	**0,5**	*0,03*	194	-0,71	0,3	0,04

81	-0,48	0,0	0,01	195	0,12	0,0	0,00
82	-0,69	0,2	0,04	196	0,26	0,0	0,00
83	0,00	0,0	0,00	197	-0,25	0,0	0,00
84	-0,69	0,2	0,04	198	0,00	0,0	0,00
85	-1,09	0,1	0,02	199	0,00	0,0	0,00
86	0,00	0,0	0,00	200	0,02	0,0	0,00
87	**-1,62**	**1,1**	**0,04**	201	0,09	0,0	0,00
88	-0,37	0,3	0,02	202	-0,13	0,0	0,00
89	**1,67**	**3,7**	**0,36**	203	0,34	0,0	0,00
90	-0,55	0,1	0,02	204	0,02	0,0	0,00
91	**-4,31**	**18,2**	**0,29**	205	0,31	0,0	0,01
92	-0,13	0,0	0,00	206	-0,01	0,0	0,00
93	**-1,17**	**0,7**	**0,05**	207	-0,82	0,3	0,01
94	-0,32	0,1	0,01	208	-0,82	0,3	0,01
95	-0,96	0,3	0,03	209	-0,25	0,1	0,00
96	0,00	0,0	0,00	210	-0,25	0,1	0,00
97	0,75	0,4	0,02	211	-0,42	0,1	0,01
98	-0,41	0,0	0,00	212	-0,31	0,2	0,02
99	-0,33	0,4	0,02	213	0,53	0,1	0,01
100	-0,37	0,4	0,02	214	0,20	0,0	0,00
101	-0,33	0,4	0,02	215	-0,48	0,0	0,01
102	0,00	0,0	0,00	216	0,13	0,0	0,00
103	-0,26	0,0	0,01	217	0,02	0,0	0,00
104	-0,25	0,0	0,02	218	0,00	0,0	0,00
105	-0,05	0,0	0,00	219	-0,48	0,0	0,01
106	0,00	0,0	0,00	220	-0,02	0,0	0,00
107	-0,01	0,0	0,00	221	0,56	0,3	0,02
108	0,02	0,0	0,00	222	-0,22	0,0	0,00
109	0,02	0,0	0,00	**223**	**1,84**	**5,5**	**0,32**
110	0,02	0,0	0,00	224	-0,22	0,0	0,00
111	0,00	0,0	0,00	225	-0,48	0,0	0,01
112	0,87	0,1	0,02	**226**	**1,75**	**2,3**	**0,33**
113	0,00	0,0	0,00	227	0,31	0,1	0,02
114	**0,76**	**0,7**	**0,04**				

Anexo J

Classificação Hierárquica - dados entidades e notícias

Tabela J.1: Classes formadas através da aplicação da Classificação Hierárquica às 24 coordenadas fatoriais das 227 notícias — partição em 2 classes.

Classe	Notícias				
Classe 1	1	51	96	141	186
	2	52	97	142	187
	3	53	98	143	188
	4	54	99	144	189
	5	55	100	145	190
	6	56	101	146	191
	7	57	102	147	192
	10	58	103	148	193
	11	59	104	149	194
	14	60	105	150	195
	15	61	106	151	196
	16	62	107	152	197
	17	63	108	153	198
	18	64	109	154	199
	20	65	110	155	200
	21	66	111	156	201
	22	67	112	157	202
	23	68	113	158	203
	24	69	114	159	204
	25	70	115	160	205
	26	71	116	161	206
	27	72	117	162	207
	28	73	118	163	208

	29	74	119	164	209
	30	75	120	165	210
	31	76	121	166	211
	32	77	122	167	212
	33	78	123	168	213
	34	79	124	169	214
	35	80	125	170	215
	36	81	126	171	216
	37	82	127	172	217
	38	83	128	173	218
	39	84	129	174	219
	40	85	130	175	220
	41	86	131	176	221
	42	87	132	177	222
	43	88	133	178	223
	44	89	134	179	224
	45	90	135	180	225
	46	91	136	181	226
	47	92	137	182	227
	48	93	138	183	
	49	94	139	184	
	50	95	140	185	
Classe 2	8	9	12	13	19

Tabela J.2: Classes formadas através da aplicação da Classificação Hierárquica às 24 coordenadas fatoriais das 227 notícias — partição em 12 classes.

Classe	Notícias				
Classe 1	1	7	30	125	
	2	22	54	161	
	3	27	70		
Classe 2	4	65	108	153	190
	6	67	109	154	191
	15	68	110	155	194
	18	69	111	156	195
	21	71	113	157	197
	23	72	115	158	198
	25	74	116	159	199
	26	76	117	160	200
	28	78	118	162	201
	32	79	119	163	202
	34	82	122	164	204

	36	83	123	166	205
	37	84	124	167	206
	41	86	126	168	211
	42	88	129	169	212
	43	94	130	170	213
	45	96	131	172	214
	48	97	132	174	216
	49	98	133	176	217
	51	99	135	177	218
	52	100	138	180	220
	55	101	139	182	221
	56	102	140	183	222
	57	103	141	184	224
	58	104	144	186	227
	60	105	146	187	
	62	106	149	188	
	63	107	152	189	
Classe 3	5	143	215		
	14	179	219		
	44	185	225		
Classe 4	8	9	12	13	19
Classe 5	10	35	66	136	192
	20	38	77	137	223
	24	39	89	148	226
	31	40	114	173	
	33	53	127	178	
Classe 6	11	59	92	165	210
	46	64	147	175	
	47	90	150	209	
Classe 7	16	17			
Classe 8	29	61	80	85	95
Classe 9	73	87	93	151	171
Classe 10	75	134	196		
	112	145	203		
Classe 11	91	120	181	193	
Classe 12	128	142	207	208	

Tabela J.3: Classes formadas através da aplicação da Classificação Hierárquica às 24 coordenadas fatoriais das 227 notícias — partição em 25 classes.

Classe	Notícias				
Classe 1	1	7	70	161	
	2	30	125		
Classe 2	3	22	27	54	
Classe 3	4	62	106	149	190
	15	63	111	154	191
	18	67	113	155	198
	21	74	115	156	199
	23	79	116	159	202
	25	83	117	162	204
	26	86	119	164	205
	32	88	122	166	206
	34	94	123	167	218
	37	96	124	168	220
	41	99	135	169	227
	42	100	140	170	
	43	101	141	182	
	48	102	144	183	
	49	104	146	187	
Classe 4	5	50	143	215	
	14	81	179	219	
	44	121	185	225	
Classe 5	6	51	130		
Classe 6	8	9	12	13	19
Classe 7	10	38	127		
	35	77	192		
Classe 8	11	46	92	175	
Classe 9	16	17			
Classe 10	20	39	89	148	226
	24	40	114	173	
	31	53	136	178	
	33	66	137	223	
Classe 11	28	131	158	176	211
Classe 12	29	61	80	85	95
Classe 13	36	57	108	184	
	56	71	172		
Classe 14	45	72	103	163	197
	55	76	107	180	212

	68	82	118	194	
	69	84	160	195	
Classe 15	47	64	147	165	210
	59	90	150	209	
Classe 16	52	78	129	174	189
	65	98	157	188	
Classe 17	58	60	97	152	221
Classe 18	73	87	93	151	171
Classe 19	75	112	145		
Classe 20	91	120	181	193	
Classe 21	105	126	139	200	224
	109	132	153	217	
	110	138	177	222	
Classe 22	128	142	207	208	
Classe 23	133	186			
Classe 24	134	196	203		
Classe 25	201	213	214	216	

Anexo K

Classificação Não Hierárquica - dados entidades e notícias

Tabela K.1: Classes formadas através da aplicação do algoritmo K-médias às 24 coordenadas fatoriais das 227 notícias — partição em 19 classes.

Classe	Notícias				
Classe 1	128	142	207	208	
Classe 2	134	196	203		
Classe 3	3	22	27	54	
Classe 4	35	127			
Classe 5	126	132	133	186	
Classe 6	6	51	130		
Classe 7	11	46	175		
Classe 8	1	8	12	19	
	7	9	13		
Classe 9	73	87	93	151	171
Classe 10	91	120	181	193	
Classe 11	2	60	152		
	58	97	221		
Classe 12	4	63	105	141	177
	15	67	106	144	180
	18	69	107	146	182
	21	71	108	149	184
	23	72	109	153	188
	25	74	110	154	190
	26	76	111	156	191
	30	79	113	158	194
	32	82	115	159	195

	34	83	116	160	197
	36	84	117	161	198
	37	86	118	162	199
	42	88	119	163	200
	43	94	122	164	204
	45	96	123	166	205
	48	98	124	167	211
	49	99	129	168	212
	52	100	131	169	217
	55	101	135	170	218
	56	102	138	172	220
	57	103	139	174	227
	62	104	140	176	
Classe 13	47	64	92	150	209
	59	90	147	165	210
Classe 14	201	213	214	216	
Classe 15	68	75	112	145	
Classe 16	16	17			
Classe 17	29	61	80	85	95
Classe 18	10	38	66	137	192
	20	39	77	148	206
	24	40	89	173	223
	31	41	114	178	226
	33	53	136	189	
Classe 19	5	65	125	183	219
	14	70	143	185	222
	28	78	155	187	224
	44	81	157	202	225
	50	121	179	215	

Anexo L

Mapas de Kohonen - dados notícias e entidades

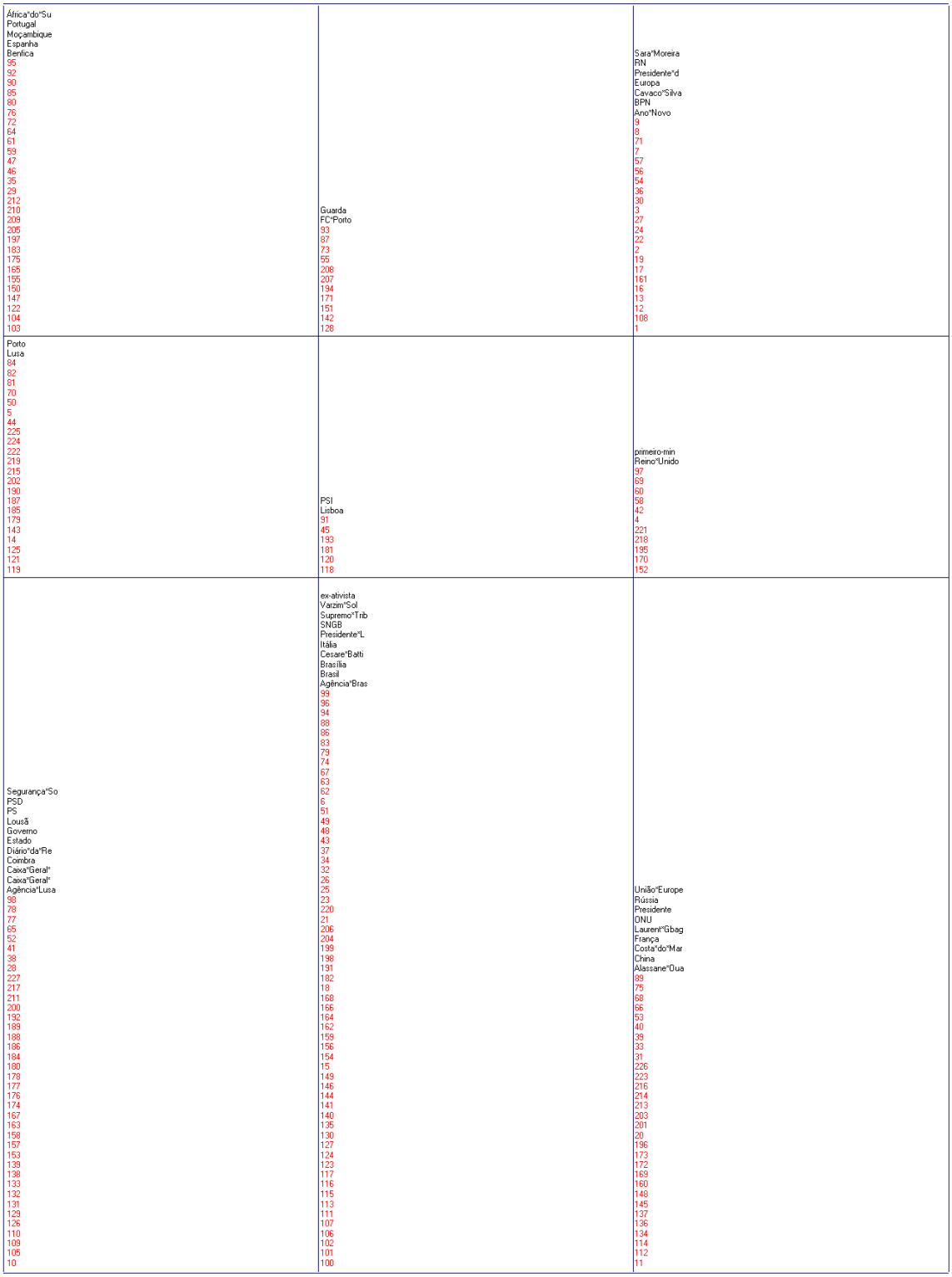

Figura L.1: Mapa de Kohonen (3 x 3) representando as 227 notícias e as 50 entidades.

Figura L.2: Mapa de Kohonen (4 x 4) representando as 227 notícias e as 50 entidades.

Anexo M

Dados Livro - Valores próprios e inércia para os 38 primeiros eixos.

Tabela M.1: Valores próprios, inércia e inércia acumulada para os 38 primeiros eixos.

Eixo	λ	Inércia (%)	% acumulada	Eixo	λ	Inércia (%)	% acumulada
1	0,8600	3,30	3,30	20	0,5425	2,08	50,70
2	0,8202	3,14	6,44	21	0,5305	2,03	52,74
3	0,7725	2,96	9,40	22	0,5213	2,00	54,73
4	0,7613	2,92	12,32	23	0,5067	1,94	56,68
5	0,7414	2,84	15,16	24	0,5035	1,93	58,61
6	0,7235	2,77	17,93	25	0,4974	1,91	60,51
7	0,7048	2,70	20,63	26	0,4822	1,85	62,36
8	0,6853	2,63	23,26	27	0,4769	1,83	64,19
9	0,6698	2,57	25,83	28	0,4730	1,81	66,00
10	0,6570	2,52	28,34	29	0,4554	1,75	67,75
11	0,6248	2,39	30,74	30	0,4429	1,70	69,44
12	0,6153	2,36	33,10	31	0,4368	1,67	71,12
13	0,6106	2,34	35,44	32	0,4276	1,64	72,76
14	0,5947	2,28	37,72	33	0,4174	1,60	74,36
15	0,5843	2,24	39,95	34	0,4097	1,57	75,93
16	0,5807	2,23	42,18	35	0,4034	1,55	77,47
17	0,5703	2,19	44,37	36	0,3956	1,52	78,99
18	0,5604	2,15	46,51	37	0,3908	1,50	80,49
19	0,5509	2,11	48,62	38	0,3801	1,46	81,94

Anexo N

Análise de Correspondências - Livro

Tabela N.1: Coordenadas, contribuições absolutas e relativas das 56 entidades retidas para o eixo 1.

Entidades	Coordenadas	CTA	CTR
Abel*Pinheiro	-0,37	0,2	0,00
António*Arnaut	0,25	0,0	0,00
António*José*Vilela	-0,70	1,1	0,02
António*Reis	-0,04	0,0	0,00
Bairro*Alto	0,11	0,0	0,00
CO	-0,18	0,0	0,00
Carbonária	0,57	0,4	0,00
Cf	-0,47	0,8	0,02
Coimbra	0,28	0,1	0,00
Conselho*da*Ordem	0,28	0,1	0,00
EUA	-0,51	0,2	0,00
GLLP	-0,41	1,1	0,03
GLRP	**-0,44**	**1,9**	**0,04**
GOL	0,06	0,0	0,00
Governo	-0,21	0,1	0,00
Grande*Dieta	**2,16**	**20,5**	**0,39**
Grande*Loja	-0,32	0,1	0,00
Grande*Loja*Legal*de*Portugal	-0,44	0,1	0,00
Grande*Loja*Regular*de*Portugal	-0,44	0,2	0,00
Grande*Oriente*Lusitano	0,05	0,0	0,00
Grão	-0,37	0,3	0,00
Irmão	-0,30	0,1	0,00
Irmãos	-0,18	0,2	0,00
Isaltino*Morais	-0,31	0,1	0,00
Jorge*Silva*Carvalho	-0,64	0,7	0,02

José*Moreno	-0,37	0,1	0,00
Justiça	0,29	0,1	0,00
Lisboa	0,14	0,1	0,00
Loja*Mercúrio	-0,48	0,2	0,00
Loja*Universalis	-0,20	0,0	0,00
Maçonaria	-0,13	0,0	0,00
Mercúrio	-0,42	0,1	0,00
Mário*Martin*Guia	-0,51	0,2	0,00
NUIPC	**-1,75**	**2,4**	**0,05**
Nuno*Vasconcellos	-0,80	0,9	0,02
Ongoing	-0,86	0,7	0,02
PS	-0,12	0,0	0,00
PSD	-0,36	0,2	0,00
Paulo*Portas	-0,53	0,2	0,00
País	0,01	0,0	0,00
Porto	0,10	0,0	0,00
Portugal	-0,18	0,1	0,00
Presidente	0,43	0,2	0,00
Público	-0,52	0,2	0,00
Representante	**4,46**	**58,2**	**0,71**
SIED	-0,70	0,4	0,01
SIS	-0,58	0,3	0,01
Silva*Carvalho	-0,98	1,0	0,02
Sábado	-0,70	1,0	0,02
TDLSB	**-1,85**	**2,8**	**0,05**
Venerável	0,57	0,6	0,01
grão-mestre	0,01	0,0	0,00
grão-mestre*do*GOL	0,08	0,0	0,00
presidente	1,19	1,4	0,02
secretário	-0,08	0,0	0,00
secretário*de*Estado	-0,33	0,1	0,00

Tabela N.2: Coordenadas, contribuições absolutas e relativas das 56 entidades retidas para o eixo 2.

Entidades	Coordenadas	CTA	CTR
Abel*Pinheiro	0,09	0,0	0,00
António*Arnaut	-0,24	0,0	0,00
António*José*Vilela	0,60	0,8	0,02
António*Reis	-0,37	0,3	0,01
Bairro*Alto	-0,23	0,0	0,00
CO	-0,58	0,4	0,01

Carbonária	-0,59	0,4	0,01
Cf	0,55	1,2	0,02
Coimbra	-0,22	0,1	0,00
Conselho*da*Ordem	-0,38	0,2	0,00
EUA	0,23	0,0	0,00
GLLP	**-0,56**	**2,1**	**0,06**
GLRP	**-0,60**	**3,7**	**0,07**
GOL	-0,21	0,6	0,01
Governo	0,03	0,0	0,00
Grande*Dieta	0,48	1,0	0,02
Grande*Loja	-0,64	0,5	0,01
Grande*Loja*Legal*de*Portugal	-0,52	0,2	0,00
Grande*Loja*Regular*de*Portugal	-0,43	0,2	0,00
Grande*Oriente*Lusitano	-0,33	0,4	0,01
Grão	-0,84	1,5	0,02
Irmão	-0,15	0,0	0,00
Irmãos	-0,19	0,2	0,00
Isaltino*Morais	-0,04	0,0	0,00
Jorge*Silva*Carvalho	0,67	0,8	0,02
José*Moreno	-0,44	0,2	0,00
Justiça	-0,26	0,1	0,00
Lisboa	-0,22	0,3	0,01
Loja*Mercúrio	0,12	0,0	0,00
Loja*Universalis	0,23	0,0	0,00
Maçonaria	-0,17	0,0	0,00
Mercúrio	-0,15	0,0	0,00
Mário*Martin*Guia	-0,69	0,4	0,01
NUIPC	**6,06**	**30,5**	**0,59**
Nuno*Vasconcellos	**1,17**	**2,1**	**0,04**
Ongoing	**1,40**	**1,9**	**0,04**
PS	0,03	0,0	0,00
PSD	0,26	0,1	0,00
Paulo*Portas	0,14	0,0	0,00
País	-0,27	0,1	0,00
Porto	-0,24	0,1	0,00
Portugal	-0,21	0,1	0,00
Presidente	-0,36	0,2	0,00
Público	0,61	0,3	0,01
Representante	**1,41**	**6,1**	**0,07**
SIED	0,81	0,6	0,02
SIS	0,55	0,3	0,01
Silva*Carvalho	**1,52**	**2,4**	**0,04**

Sábado	0,61	0,8	0,02
TDLSB	**6,53**	**36,8**	**0,61**
Venerável	-0,29	0,2	0,00
grão-mestre	-0,39	0,7	0,01
grão-mestre*do*GOL	-0,28	0,1	0,00
presidente	-0,23	0,1	0,00
secretário	-0,61	0,6	0,01
secretário*de*Estado	0,14	0,0	0,00

Tabela N.3: Coordenadas, contribuições absolutas e relativas das 56 entidades retidas para o eixo 3.

Entidades	Coordenadas	CTA	CTR
Abel*Pinheiro	1,0	1,3	0,02
António*Arnaut	0,8	0,5	0,01
António*José*Vilela	0,9	1,8	0,03
António*Reis	0,4	0,4	0,01
Bairro*Alto	0,3	0,1	0,00
CO	0,2	0,1	0,00
Carbonária	**3,2**	**13,3**	**0,15**
Cf	0,2	0,2	0,00
Coimbra	0,8	0,6	0,01
Conselho*da*Ordem	0,5	0,4	0,01
EUA	0,6	0,3	0,01
GLLP	**-1,3**	**12,1**	**0,31**
GLRP	**-1,4**	**22,2**	**0,42**
GOL	0,3	1,3	0,02
Governo	0,5	0,3	0,00
Grande*Dieta	-0,2	0,2	0,00
Grande*Loja	**-1,6**	**3,3**	**0,04**
Grande*Loja*Legal*de*Portugal	-1,1	1,1	0,02
Grande*Loja*Regular*de*Portugal	-1,0	1,2	0,02
Grande*Oriente*Lusitano	0,4	0,6	0,01
Grão	**-1,3**	**3,5**	**0,05**
Irmão	0,3	0,1	0,00
Irmãos	-0,2	0,3	0,01
Isaltino*Morais	-0,1	0	0,00
Jorge*Silva*Carvalho	0,1	0	0,00
José*Moreno	-0,8	0,8	0,01
Justiça	0,7	0,5	0,01
Lisboa	**0,7**	**2,7**	**0,05**
Loja*Mercúrio	-0,5	0,2	0,00

Loja*Universalis	0,3	0,1	0,00
Maçonaria	0,9	1,3	0,02
Mercúrio	-0,5	0,2	0,00
Mário*Martin*Guia	**-1,8**	**2,8**	**0,05**
NUIPC	-0,8	0,6	0,01
Nuno*Vasconcellos	-0,2	0,1	0,00
Ongoing	0,0	0	0,00
PS	0,3	0,2	0,00
PSD	0,4	0,3	0,01
Paulo*Portas	1,1	1	0,02
País	0,3	0,1	0,00
Porto	0,6	0,6	0,01
Portugal	0,5	0,7	0,01
Presidente	**1,9**	**4,9**	**0,06**
Público	0,6	0,3	0,00
Representante	-1,3	5	0,05
SIED	0,3	0,1	0,00
SIS	0,5	0,3	0,01
Silva*Carvalho	0,3	0,1	0,00
Sábado	0,9	1,8	0,03
TDLSB	-0,9	0,8	0,01
Venerável	**1,5**	**4,4**	**0,06**
grão-mestre	0,0	0	0,00
grão-mestre*do*GOL	0,9	0,9	0,01
presidente	**1,7**	**3,3**	**0,04**
secretário	-0,4	0,3	0,00
secretário*de*Estado	0,5	0,2	0,01

Anexo O

Classificação Hierárquica - livro

Tabela O.1: Classes formadas através da aplicação da Classificação Hierárquica às 30 coordenadas fatorias das 56 entidades retidas — partição em 15 classes.

Classe	Entidades	
Classe 1	Abel*Pinheiro	Paulo*Portas
	EUA	
Classe 2	António*Arnaut	
Classe 3	António*José*Vilela	Lisboa
	António*Reis	Loja*Universalis
	Bairro*Alto	Maçonaria
	Cf	Mário*Martin*Guia
	Coimbra	PS
	Conselho*da*Ordem	PSD
	GLLP	País
	GLRP	Porto
	GOL	Portugal
	Governo	Representante
	Grande*Dieta	Sábado
	Grande*Oriente*Lusitano	Venerável
	Grão	grão-mestre
	Irmão	secretário
	Irmãos	secretário*de*Estado
	Justiça	
Classe 4	CO	
Classe 5	Carbonária	
Classe 6	Grande*Loja	
Classe 7	Grande*Loja*Legal*de*Portugal	Grande*Loja*Regular*de*Portugal
Classe 8	Isaltino*Morais	Loja*Mercúrio

	José*Moreno	Mercúrio
Classe 9	Jorge*Silva*Carvalho SIED	SIS
Classe 10	NUIPC	TDLSB
Classe 11	Nuno*Vasconcellos Ongoing	Silva*Carvalho
Classe 12	Presidente	
Classe 13	Público	
Classe 14	grão-mestre*do*GOL	
Classe 15	presidente	

Tabela O.2: Classes formadas através da aplicação da Classificação Hierárquica às 30 coordenadas fatorias das 56 entidades retidas — partição em 4 classes.

Classe	Entidades	
Classe 1	Abel*Pinheiro	Paulo*Portas
	EUA	
Classe 2	António*Arnaut	Loja*Mercúrio
	António*José*Vilela	Loja*Universalis
	António*Reis	Maçonaria
	Bairro*Alto	Mercúrio
	CO	Mário*Martin*Guia
	Cf	Nuno*Vasconcellos
	Coimbra	Ongoing
	Conselho*da*Ordem	PS
	GLLP	PSD
	GLRP	País
	GOL	Porto
	Governo	Portugal
	Grande*Dieta	Público
	Grande*Loja	Representante
	Grande*Oriente*Lusitano	SIED
	Grão	SIS
	Irmão	Silva*Carvalho
	Irmãos	Sábado
	Isaltino*Morais	Venerável
	Jorge*Silva*Carvalho	grão-mestre
	José*Moreno	grão-mestre*do*GOL
	Justiça	secretário
	Lisboa	secretário*de*Estado
Classe 3	Carbonária	presidente

	Presidente	
Classe 4	Grande*Loja*Legal*de*Portugal	Grande*Loja*Regular*de*Portugal

Anexo P

Classificação Não Hierárquica - livro

Tabela P.1: Classes formadas a partir da aplicação do algoritmo K-médias às 30 coordenadas fatoriais das 56 entidades retidas — partição em 19 classes.

Classe	Entidades	
Classe 1	Abel*Pinheiro	Paulo*Portas
	EUA	
Classe 2	António*Arnaut	
Classe 3	Público	
Classe 4	Maçonaria	
Classe 5	Presidente	
Classe 6	CO	
Classe 7	Carbonária	
Classe 8	Isaltino*Morais	
Classe 9	Venerável	
Classe 10	secretário	
Classe 11	presidente	
Classe 12	NUIPC	TDLSB
Classe 13	Silva*Carvalho	
Classe 14	Irmão	
Classe 15	grão-mestre*do*GOL	
Classe 16	Justiça	
Classe 17	Grande*Loja	
Classe 18	Grande*Loja*Legal*de*Portugal	Grande*Loja*Regular*de*Portugal
Classe 19	António*José*Vilela	Loja*Mercúrio
	António*Reis	Loja*Universalis
	Bairro*Alto	Mercúrio
	Cf	Mário*Martin*Guia
	Coimbra	Nuno*Vasconcellos

Conselho*da*Ordem	Ongoing
GLLP	PS
GLRP	PSD
GOL	País
Governo	Porto
Grande*Dieta	Portugal
Grande*Oriente*Lusitano	Representante
Grão	SIED
Irmãos	SIS
Jorge*Silva*Carvalho	Sábado
José*Moreno	grão-mestre
Lisboa	secretário*de*Estado

Tabela P.2: Classes formadas a partir da aplicação do algoritmo K-médias às 30 coordenadas fatoriais das 56 entidades retidas — partição em 21 classes.

Classe	Entidades	
Classe 1	Abel*Pinheiro	Paulo*Portas
Classe 2	António*Arnaut	
Classe 3	grão-mestre*do*GOL	
Classe 4	Silva*Carvalho	
Classe 5	Venerável	
Classe 6	CO	
Classe 7	Carbonária	
Classe 8	NUIPC	TDLSB
Classe 9	presidente	
Classe 10	Conselho*da*Ordem	
Classe 11	António*José*Vilela	Loja*Universalis
	António*Reis	Mercúrio
	Bairro*Alto	Mário*Martin*Guia
	Cf	Nuno*Vasconcellos
	Coimbra	Ongoing
	EUA	PS
	GLLP	PSD
	GLRP	País
	GOL	Porto
	Governo	Portugal
	Grande*Dieta	Representante
	Grande*Oriente*Lusitano	SIED
	Irmãos	SIS
	Jorge*Silva*Carvalho	Sábado

	José*Moreno	grão-mestre
	Lisboa	secretário*de*Estado
	Loja*Mercúrio	
Classe 12	Irmão	
Classe 13	Público	
Classe 14	Isaltino*Morais	
Classe 15	secretário	
Classe 16	Justiça	
Classe 17	Grande*Loja	
Classe 18	Grande*Loja*Legal*de*Portugal	Grande*Loja*Regular*de*Portugal
Classe 19	Maçonaria	
Classe 20	Presidente	
Classe 21	Grão	

Tabela P.3: Classes formadas a partir da aplicação do algoritmo K-médias às 30 coordenadas fatoriais das 56 entidades retidas — partição em 4 classes.

Classe	Entidades	
Classe 1	NUIPC	TDLSB
Classe 2	António*Arnaut	
Classe 3	Abel*Pinheiro	Lisboa
	António*José*Vilela	Loja*Mercúrio
	António*Reis	Loja*Universalis
	Bairro*Alto	Maçonaria
	CO	Mercúrio
	Carbonária	Mário*Martin*Guia
	Cf	Nuno*Vasconcellos
	Coimbra	Ongoing
	Conselho*da*Ordem	PS
	EUA	PSD
	GLLP	Paulo*Portas
	GLRP	País
	GOL	Porto
	Governo	Portugal
	Grande*Dieta	Presidente
	Grande*Loja	Público
	Grande*Loja*Legal*de*Portugal	Representante
	Grande*Loja*Regular*de*Portugal	SIED
	Grande*Oriente*Lusitano	SIS
	Grão	Silva*Carvalho
	Irmão	Sábado

	Irmãos	Venerável
	Isaltino*Morais	grão-mestre
	Jorge*Silva*Carvalho	grão-mestre*do*GOL
	José*Moreno	secretário
	Justiça	secretário*de*Estado
Classe 4	presidente	

Anexo Q

Mapas de Kohonen - livro

Carbonária	TDLSB NUIPC		
presidente Presidente	grão-mestre* Conselho*da*	PS Isaltino*Mor Cf António*Arna	Paulo*Portas Irmão EUA Abel*Pinheir
Venerável Porto Lisboa Justiça Coimbra	secretário grão-mestre Representant Portugal País Mário*Martin Maçonaria José*Moreno Irmãos Grão Grande*Orien Grande*Loja* Grande*Loja* Grande*Loja Grande*Dieta GOL GLRP GLLP CO Bairro*Alto António*Reis	Sábado Mercúrio Loja*Univers Loja*Mercúri António*José	secretário*d Silva*Carval SIS SIED Público PSD Ongoing Nuno*Vasconc Jorge*Silva* Governo

Figura Q.1: Mapa de Kohonen (4 x 4) representando as 56 entidades do livro.

Figura Q.2: Mapa de Kohonen (5 x 5) representando as 56 entidades do livro.

Sábado António*José **20**	Maçonaria GOL Bairro*Alto **21**	Público Irmãos **22**	Ongoing Nuno*Vasconc **23**	Paulo*Portas EUA Abel*Pinheir **24**	Loja*Mercúri Isaltino*Mor **25**
Grão CO António*Reis **16**	Portugal Grande*Orien Cf **17**	TDLSB NUIPC **18**			José*Moreno **19**
				grão-mestre Representant Mário*Martin Mercúrio Grande*Dieta GLRP GLLP **14**	Justiça Conselho*da* **15**
País PS Governo **8**	secretário*d PSD **9**	SIS SIED Jorge*Silva* **10**	Silva*Carval **11**	Venerável Loja*Univers Lisboa **12**	Irmão **13**
Porto Coimbra **6**					grão-mestre* **7**
Presidente António*Arna **1**	Carbonária **2**	presidente Grande*Loja **3**	secretário **4**		Grande*Loja* Grande*Loja* **5**

Figura Q.3: Mapa de Kohonen (6 x 6) representando as 56 entidade do livro.