U.PORTO

FEP FACULDADE DE ECONOMIA UNIVERSIDADE DO PORTO

SYMBOLIC AND COMPOSITIONAL ANALYSIS OF TEXTUAL DATA

Tânia Manuela Costa da Silva

Dissertation

Master in Modeling, Data Analysis and Decision Support Systems

Supervised by
Prof. Maria Paula de Pinho de Brito Duarte Silva

2021

Dedicated to my Parents

Biography

Tânia Silva was born on August 20th, 1998 in Vila Nova de Famalicão, Braga, Portugal.

In 2016 she started her bachelor degree in Economics at University of Porto. During her bachelor she did her first summer internship at Department of Statistics of Bank of Portugal. After graduating in 2019 in Economics she did her second internship at Bank of Portugal, for a period of six months. In September 2019 she joined the Master in Modeling, Data Analysis and Decision Support Systems.

Currently and since September 2020, she is working in Audit at Ernest & Young. This work allows her to have insights of different types of businesses and understand how things work.

The motivation for writing this thesis comes from believing in the potential of data and the added value it can bring to any analysis, company and sector of activity.

Acknowledgments

I would like to thank all the persons who made this possible and in one way or another motivated me to complete this dissertation.

First of all, I would like to express my gratitude to professora Maria Paula Brito, for the motivation, patience, the guidance and advices, and all the transmitted knowledge in this theme. I would like to genuinely thank for the dedication, commitment and for the many revised versions.

I must thank all of my professors during FEP' journey (bachelor and master) for the share of useful insights and knowledge.

I would also like to thank my friends for always reassuring me, believe when I didn't believe that this would be possible and for their patience.

And finally, I am so grateful for my parents, that made this journey easier , have supported me and encouraged me to end this life's chapter.

Abstract

The exponential growth of textual data production makes it central to develop new analysis techniques for this unstructured kind of data. Natural Language Processing (NLP) combines the field of linguistics and computer science to decipher language structure and to design models which can understand and separate significant details from text and speech. Symbolic Data Analysis appears as an extension of traditional approaches to deal with the analysis of complex data and new data types. An alternative is Compositional Data Analysis that is defined for random vectors with strictly positive components and a constant sum. The standard statistical techniques have no applicability in compositional data, so new techniques and transformations are needed.

In this dissertation it will be applied *Topic Modeling* to describe textual data by a distribution on some topics and then compare the use of symbolic and compositional clustering approaches to analyse the resulting distributional data. As a complement, Discriminant Analysis is also applied based on the analysis of the predictor variables.

It is highlighted the importance of data pre-processing and the fact that the different approaches lead to very different solutions. Based on the data set used, it was concluded that in the symbolic clustering analysis, texts related to the same topic appear more frequently in the same cluster than by using a compositional approach.

Key Words: Text, Text Mining, Compositional, Symbolic, Discriminant, Speeches, Barack Obama, Donald Trump.

Resumo

O aumento exponencial da produção de dados textuais torna fulcral o desenvolvimento de técnicas de análise para este tipo de dados não estruturado. O processamento de linguagem natural (PLN) combina as áreas da linguística e da ciência da computação para decifrar a estrutura da linguagem e desenhar modelos que possam compreendê-la e identificar detalhes significativos do texto e discurso. A análise de dados simbólicos surge como uma extensão das abordagens tradicionais para analisar dados mais complexos e novos tipos de dados. Uma alternativa a esta abordagem é a análise de dados composicionais, onde composições são vetores cujas componentes são positivas e têm uma soma constante. Os métodos estatísticos clássicos não têm aplicabilidade em dados composicionais, pelo que novas técnicas e transformações são necessárias.

Nesta dissertação será aplicado *Topic Modeling* para descrever dados textuais sob a forma de distribuições em tópicos e após isso comparar o uso de métodos de classificação simbólicos e composicionais. Complementarmente, foi aplicada análise discriminante com base na análise das variáveis independentes. É destacada a importância do pré-processamento e do facto dos diferentes métodos conduzirem a soluções muito diferentes. Tendo em conta o conjunto de dados usado, foi concluído que na análise classificatória de dados simbólicos textos relacionados com o mesmo tópico tendem mais frequentemente a agregar-se na mesma classe do que na análise classificatória de dados composicionais.

Palavras-Chave: Texto, *Text Mining*, Composicional, Simbólico, Discriminante, Discursos, Barack Obama, Donald Trump.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter we give an overview of the theme that will be developed during the present dissertation. This overview includes the motivation and description of the problem studied, the objectives to be achieved and the thesis structure and organization.

1.1 Motivation

Text Mining (TM) has reached considerable interest in recent years due to the daily exponential increase of textual data and the expectation of a tremendous growth in the next few years. TM is quite similar to data mining, which refers to the extraction or mining of knowledge from large amounts of data. However, in TM the data comes from unstructured sources of information and despite being easily understood by individuals, it is hardly perceived by machines. Fan et al. (2006) also refer that there is a trade-off between the humans' capabilities to comprehend unstructured data and the ability of computers to process text in large volumes at high speed. The better answer of text mining is creating technology that combines human capabilities with the speed and accuracy of a computer.

The rise of the field of text mining which aims at discovering, extracting and accessing information contained in textual data required the development of approaches to perform the exploratory analysis – as correspondence analysis and

clustering. According to Petrović et al. (2009) correspondence analysis is an unsupervised approach that allows the construction of a low–dimensional projection space with simultaneous placement of both documents and features. It is mostly applied to contingency tables (Široki et al., 2019). The use of correspondence analysis in a linguistic context i.e. textual data analysis was first proposed by Benzécri (1992).

According to Allahyari et al. (2017) clustering methods can be applied at different levels as documents, paragraphs, sentences and terms and allow finding groups of similar documents. Clustering is used to organize documents in order to enhance retrieval and support browsing. Many clustering algorithms that can be applied to textual data have been proposed. MacQueen (1967) proposed K-means that is a partition-based clustering algorithm. Other types of clustering algorithms proposed are density-based e.g. DBSCAN proposed by Ester et al. (1996), hierarchy-based e.g. BIRCH proposed by Wang et al. (2007), grid-based e.g. STING proposed by Wang et al. (1997) and model-based as neural networks proposed by Kohonen et al. (2000). Other algorithms were developed in other fields, in order to enhance clustering performance such as spectral clustering in physics, non-regression matrix factorization in mathematics and LDCC based on Latent Dirichlet Allocation from Topic Models.

1.2 Objectives

In this dissertation we want to describe textual data, by means of discrete distributions and then use symbolic and compositional clustering methods to organize the texts in classes. The textual data analysed consist of 83 presidential speeches of the United States of America (USA) from 2008 to 2020, where 49 speeches were made by Barack Obama and 34 by Donald Trump. These speeches were obtained from UVA's Miller Center[1].

The first step of TM is the pre-processing of data. After this cleaning process the aim is to describe the speeches by means of discrete distributions, and then use

[1] millercenter.org/the-presidency/presidential-speeches retrieved november 15, 2020

clustering techniques adapted to distributional data for their analysis. It will be applied topic modeling, aimed at describing each speech by a distribution on those topics, using symbolic and compositional clustering techniques to organize the speeches in classes. The two main topic models are Probabilistic Latent Semantic Analysis (pDSA) and Latent Dirichlet Allocation (LDA). We shall use LDA in order to identify the topics by words probabilities, i.e. we will obtain the weight of each topic/subtopic in a certain document/text. This unsupervised technique enables to extract thematic information (topics) from a collection of documents (Mcauliffe and Blei, 2007). As a complement, it will be applied Discriminant Analysis to predict the author of the speech based on the topics of the speech.

As an alternative approach to Topic Modeling we could have used Correspondence Analysis, in order to identify topics from the factorial representation.

1.3 Dissertation Structure

This dissertation is organized in five chapters. The first chapter is an introduction to the theme, presenting the motivation and problem studied and the main objectives. The second chapter presents a literature review and state-of-art analysis. It is subdivided into six subsections: textual data analysis and text mining, correspondence analysis, topic modeling, cluster analysis, symbolic data analysis, compositional data analysis and discriminant analysis. The third chapter is about the data and software used. In the fourth section we describe and comment the results obtained using topic modeling to describe each speech by a distribution on topics and then by the application of compositional and symbolic clustering techniques. Finally, in the fifth section we draw the main conclusions and make final considerations regarding the theme of this dissertation.

Chapter 2

Literature Review

As established in Chapter 1, this dissertation is about text mining. The approach followed includes different methods: topic modeling, cluster analysis, symbolic data analysis, compositional data analysis and discriminant analysis.

2.1 Textual Data Analysis and Text Mining

In recent years, we have seen an increase in the quantities of available textual data. Contrasting with textual data analysis that starts from the postulate that the internal organization of a speech "memorizes" the external processes which led to its production (Reinert, 1993), TM is useful at extracting knowledge for decision-making from a large volume of semi-structured and unstructured textual data. The objective in TM focus more on modeling and predictive analysis than textual analysis. According to Tufféry (2005) TM is an ensemble of techniques and methods that deal with natural language in large quantities, in order to identify and structure the content and themes, discovering hidden information and/or enabling automatic decision making. TM is an extension of Data Mining developed around lexicometry[1] or lexical statistics (Benzécri, 1981), so it can also be classified into two approaches: exploring textual data and its content – descriptive TM – and/or using this information to optimize decision-making and business processes organization - predictive TM.

[1]Lexicometry refers to the measurement of the frequency with which words occur in text.

Text mining is widely used in diverse fields. Its application is very popular in enterprises, which wish to learn more about their customers' preferences using feedback and information collected through their website, email or call centers. Talib et al. (2016) mentioned the following fields of application: academic research where text mining is used to find and classify research papers, in life sciences highlighting the opportunity they give to infer relationships among diseases, species, symptoms, course of treatments and genes, in social media it helps analyzing posts, likes and followers.

The textual data analysis and text mining framework generally includes some first steps: acquiring text data, data cleaning and pre-processing, data normalization, conversion of text to machine readable vectors, dimensionality reduction techniques, feature selection and, finally, application of NLP and machine learning algorithms.

In the context of textual data, the two main unsupervised learning methods commonly used, according to Aggarwal and Zhai (2012), are clustering and topic modeling. It will be applied topic modeling that can be considered as the process of clustering with a generative probabilistic model. A topic can be considered to be analogous to a cluster and the membership of a document to a cluster is probabilistic in nature. Also, since the documents may be represented as a linear probabilistic combination of these different topics, topic modeling provides an extremely general framework, which relates to both the clustering and dimension reduction problems.

2.2 Correspondence Analysis

Concepts and Definitions

Correspondence analysis (CA) is an extension of Principal Component Analysis (PCA) that allows summarizing the information contained in contingency tables and putting in evidence the main factors or properties of the data. PCA enables to determine new variables, linear combinations of the original ones, which maximize variance and are non-correlated reducing the dimension, which means that we will

rely on a smaller number of variables. The techniques are similar, CA is a variant of PCA aimed primarily at categorical data.

The two pioneering papers in correspondence analysis were from Richardson and Kuder (1933) and Hirschfeld (1935). The version adapted to textual data was later developed by Benzécri et al. (1973). CA is a technique that allows finding a multidimensional representation of the dependencies between rows (observations) and columns (variables) in a low dimensional space. In the context of textual data analysis, it can be applied to specific contingency tables indicating which words (columns) appear in which document (rows) and the respective frequency. Thus, it allows investigating the possible relations between two qualitative variables, exploiting distances between documents and terms. The proximity between two documents can be explained by their use of specific words. The proximity between two words can be explained by their similar distribution over the documents. Consider a contingency Table 2.1 with r rows and s columns and suppose you wish to exploit the relation between A (A_1, A_2, ..., A_r) and B(B_1, B_2, ..., B_s):

Table 2.1: Contingency Table.

	B_1	...	B_j	...	B_s	Total
A_1	n_{11}	...	n_{1j}	...	n_{1s}	$n_{1.}$
...	...	...	...	...	...	...
A_i	n_{i1}	...	n_{ij}	...	n_{is}	$n_{i.}$
...	...	...	...	...	...	...
A_r	n_{r1}	...	n_{rj}	...	n_{rs}	$n_{r.}$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{.s}$	n

Each cell records the absolute frequency, n_{ij}, and n represents the total number of observations. The relative frequency of cell ij, corresponding to category A_i of variable A and category B_j of variable B , is $f_{ij} = \frac{n_{ij}}{n}$. The marginal relative frequencies are represented by $f_{i.} = \sum_{j=1}^{s} f_{ij}$ and $f_{.j} = \sum_{i=1}^{r} f_{ij}$.

The row-profiles and column-profiles provide an estimate of the conditional probabilities of observing a category of one variable, knowing the observed category of the other variable (conditional probability). The row profiles ($\frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{fi.}$) correspond to the proportion of observations that verify the category B_j knowing

that they verify the category A_i, analogously for the column profiles.

The distance between row-profiles is computed using the chi-square distance: $d^2(i,i^{'}) = \sum_{j=1}^{s} \frac{1}{f_{.j}}(\frac{f_{ij}}{f_{i.}} - \frac{f_{i^{'}j}}{f_{i^{'}.}})^2$, analogously for the column profiles $d^2(j,j^{'}) = \sum_{i=1}^{r} \frac{1}{f_{i.}}(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij^{'}}}{f_{.j^{'}}})^2$.

Applications

According to Kamalja and Khangar (2017) there are a variety of research areas where CA can be applied such as social sciences, medical research, engineering, market research, software development, etc. Considering textual visualization tasks, Petrović et al. (2009) applied correspondence analysis to explore the use of three textual features (letter n-grams, words and word digrams) in order to visualize a Croatian-English corpus. Morin (2004) has used CA to improve the information retrieval on abstracts of internal reports from a research center in France.

Sadika R. (2014) applied CA to closed and open-ended questions of a survey about the Tunisian Revolution in 2011. It was also performed a statistical comparison of clustering results with and without some pre-processing techniques as lemmatization.

Dias (2015) applied CA and Clustering to three data sets: the first consisting of 227 news items published by the Lusa agency; the second set of textual data consists of the list of entities cited in the first data set; at last, it was used the book 'Segredos da Maçonaria Portuguesa'. For the three data sets, it was applied correspondence analysis putting in evidence the main factors. Then, clustering methods as hierarchical ascending classification, a Kohonen map and the K-means algorithm were applied, to group the texts by topics. One of the main conclusions is that clustering analysis is a great complement to correspondence analysis.

Summa and Brito (2018) proposed factorial analysis to obtain typologies of overall reacting and sentiment as a consequence of populist speeches from EU leaders and commissioners pro and against populism.

2.3 Topic Modeling

Topic modeling is a clustering algorithm that creates a probabilistic generative model for the corpus of text documents and can be used for various text mining tasks such as summarization, document classification, novelty detection, etc. (Aggarwal and Zhai, 2012).

Topic Modeling can model objects as latent topics that can reflect meaning of the collection of documents. Thus, its main significance is to find the structure of word use and how to link documents that share the same structure. Documents are represented as a mixture of topics, where a topic is a probability distribution over words. So, a topic is a collection of words that are likely to appear in the same context.

Topic models are applied in various fields including medical sciences, software engineering, geography, political science etc.

There are two main topic modeling methods: Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). LSA uses singular value decomposition (SVD) and allows comparisons that capture the semantic structure in the documents resulting from the cooccurrence of words across the collection. Basically, SVD is used to perform dimensionality reduction on the *tf-idf* vectors. In *tf-idf* the term frequency count is compared to an inverse document frequency count, which measures the number of occurrences of a word in the entire corpus (Salton and McGill, 1983). Probabilistic Latent Semantic Analysis (PLSA) models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of "topics". Each word is generated from a single topic, and different words in a document may be generated from different topics (Hofmann, 1999). PLSA which was introduced by Hofmann (1999), does not provide any probabilistic model at the document level which makes the generalization to new unseen documents difficult. Later, it was generalized by Blei et al. (2003) with the introduction of a Dirichlet prior on mixture weights of topics per documents. Thus, LDA overcomes much of the difficulties that appeared with PLSA by using a Bayesian

approach. In this dissertation we shall focus on LDA.

2.3.1 Latent Dirichlet Allocation

Blei et al. (2003) describe LDA as a three-level hierarchical Bayesian model, it is a generative probabilistic model for collections of discrete data such as textual data. The main task of LDA consists in finding the latent variable in a text document that is the topic. It is necessary to define some notation. The word (w) is referred as a term. A document (d) is a collection of words and a collection of documents is called a corpus (D). Vocabulary is the collection of all terms in the corpus. LDA has three main constructs: word-topic-document. This probabilistic model is used to discover the topic (Z_i) characterized by word distribution.

In LDA, to generate a document, the first step is to randomly choose a distribution over topics. Next, for each word in the document, it is randomly chosen a topic from the distribution over topics and it is randomly chosen a word from the corresponding topic (distribution over the vocabulary). Blei et al. (2003) represent this generative process as follows:

1. Generate a set of multinomial topic distributions β_k from a Dirichlet distribution.
2. Choose size of document N~ Poisson distribution
    a. Sample a topic distribution $\theta_d \sim Dir(\alpha)$
    b. For each of the N words w_n
        i. Sample a topic $z_n \sim Mult\ (\theta_d)$
        ii. Sample a word w_n from $p(w_n|Z_n, \beta)$, a multinomial probability conditioned on the topic Z_n.

Figure 2.1: Generative process of LDA.

(Adapted from Blei et al. (2003)).

Figure 2.2 represents a probabilistic graphical model for LDA (Blei et al., 2003). Nodes are random variables, edges indicate dependence, shaded nodes are observed, unshaded nodes are latent variables and plates (boxes) indicate replicated variables i.e. repetitions of sampling steps, with the lower right cor-

ner variable indicating the number of samples. K denotes the number of topics and $\varphi_1, ..., \varphi_k$ are V-dimensional vectors storing the parameters of the Dirichlet-distributed topic-word distributions (V is the number of words in the vocabulary). The parameters α and β are corpus-level parameters. θ_d are document-level variables, sampled once per document. The variables $z_{d,n}$ and $w_{d,n}$ are word/term-level variables and are sampled once for each word in each document.

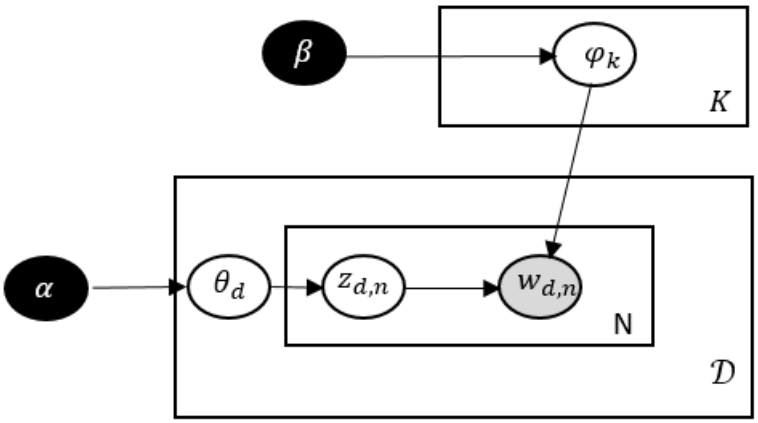


Figure 2.2: Graphical model representation of smoothed LDA.

(Adapted from Blei et al. (2003)).

Summarizing, the LDA process starts by sampling the term probabilities of each topic from a Dirichlet distribution. The following step is sampling the topic probabilities of the current document from another Dirichlet. Next, choose a topic. Finally, training an LDA model involves finding the optimal set of parameters under which the probability of generating the training set is maximized.

2.4 Cluster Analysis

Cluster analysis aims at grouping a set of objects into groups/clusters such that individuals from the same cluster have a high degree of similarity, and well separating the groups that should be "relatively distinct" from each other.

The clustering methods can be divided into hierarchical and non-hierarchical (also called partitioning methods). The hierarchical methods are divided into

agglomerative methods that follow a "bottom-up" approach which progressively merge the objects, and divisive methods that use a "top-down" approach where all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy (Gowda and Krishna, 1978). Non-hierarchical methods aim at determining partitions P = $\{C_1, ..., C_k\}$, i.e. families of classes which do not intersect and that jointly cover the whole set of observations, Ω.

We have to use comparison measures to evaluate the similarity or dissimilarity between the elements of the set to be clustered. There are two types of measures: similiraty measures, where high values indicate an important similarity between the elements and dissimilarity measures, where high values indicate that the elements are quite different. A non exaustive list of comparison measures that can be used is: Euclidean distance, "Manhattan" or "City Block" distances, Chebyshev distance and Minkowski distances, that can be computed as follows:

$$d(I_h, I_i) = \sqrt[q]{\sum_{j=i}^{p} (x_{hj} - x_{ij})^q} \tag{2.1}$$

where x_{ij} is the value of variable j for individual I_i.

The other measures are defined as particular cases of Minkowski distance. When q=2 we have the Euclidean distance; when q=1 we have the "Manhattan" distance; At last, when $q = \infty$ we have the Chebyshev distance.

In hierarchical agglomerative clustering each data point starts as a single cluster. Several methods have been proposed in literature to choose classes to be merged at each step, we focus on Single Linkage, Complete Linkage, Average Linkage and Ward's Linkage. Single Linkage is the measure between the pair of individuals, one in each cluster, X and Y, which are the closest among all possible pairs and is expressed as follows:

$$d(X, Y) = \min_{x \in X, y \in Y} d(x,y) \tag{2.2}$$

Complete Linkage is the distance between the pair of individuals, one in each cluster, which are most distant from all possible pairs and can be expressed as

follows:

$$d(X,Y) = \max_{x \in X, y \in Y} d(x,y) \tag{2.3}$$

Average linkage between groups is the average value of the dissimilarity values between elements of each cluster (X and Y) and is expressed as follows:

$$d(X,Y) = \frac{1}{|X||Y|} \sum_{x \in X} \sum_{y \in Y} d(x,y) \tag{2.4}$$

Another method is the Ward's index that defines dissimilarity between two classes X and Y as the increase in dispersion (measured by sum of squares of distances to the centroid) when X and Y are merged to form X∪Y and that may be obtained by:

$$d(X,Y) = \frac{|X||Y|}{|X|+|Y|} d^2(g_X, g_Y) \tag{2.5}$$

where g_x is the centroid (center of gravity) of the set X.

The main application areas of cluster analysis concern providing objects' taxonomies, data reduction i.e. grouping the objects (or variables) for further analysis, analysis of similarities or dissimilarities between objects, etc.

Ester et al. (1996) presented the clustering algorithm DBSCAN which relies on a density-based notion of clusters. Two parameters are used in the algorithm: *Eps* that specifies how close points should be to be considered a part of a cluster, and *MinPts* that defines the minimum number of points to form a dense region. The method is able to separate "noise" from clusters of points, where "noise" consists of points in low density regions. Wang et al. (1997) introduced a STatistical INformation Grid-method (STING) to efficiently process many common "region oriented" queries on a set of points. Another clustering algorithm for large data sets is called BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) and was proposed by Wang et al. (2007). Kohonen et al. (2000) proposed the implementation of a system that is able to organize vast document collections according to textual similarities and is based on the Self-Organizing Map (SOM) algorithm.

Other data specific clustering approaches will be presented in Sections 2.5.1

and 2.6.1.

Determination of the number of clusters

In the case of non-hierarchical methods, we need to fix in advance the number of clusters. The most popular indices to determine the number of clusters are Calinski and Harabasz index and Silhouette index.

In detail, Caliński and Harabasz (1974) evaluates the clustering effect by the tightness of the cluster and the tightness between the clusters and is computed as follows:

$$CH(k) = \frac{\frac{B(k)}{k-1}}{\frac{W(k)}{n-k}} \tag{2.6}$$

where k is the number of clusters, n the number of observations, W(k) is the within-class dispersion ($W(k) = \frac{1}{n}\sum_{h=1}^{k}\sum_{I \in C_h} d^2(I, g_h)$) and B(k) is the between-class dispersion ($B(k) = \frac{1}{n}\sum_{h=1}^{k} n_h d^2(g_h, g)$).

We can also consider the Silhouette index for the determination of the number of clusters. The Silhouette index obtains the optimal clustering number by analysing the difference between the average distance within the cluster and the minimum distance between the clusters and was introduced by Rousseeuw (1987). It is computed as follows:

$$S_i = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2.7}$$

where a(i) represents the average distance of element i to other elements in the same cluster and b(i) represents the minimum average distance of the element i to the other clusters.

Hardy and Lallemand (2004) investigated the problem of the determination of the number of clusters for symbolic objects described by multi-valued and modal variables. The five best stopping rules were identified: Calinski and Harabasz index, Duda and Hart rule, the C-index, the τ-index and the Beale test. The Duda and Hart rule and the Beale test are statistical hypothesis tests. Duda and Hart rule is defined as the ratio of the square error of a two cluster solution over that for one solution so that the process of division stops if the ratio is not small

enough or a merger proceeds if this is the case (Duda et al., 1973). Beale test is a statistical test that is based on the increase in the mean square deviation from the cluster centroids (MSD) as one moves from two to one cluster against the MSD when two clusters were present (Beale, 1969). The analysis of this indices should provide the "best" number of clusters. The Calinski and Harabasz method, the C-index and the τ-index use various forms of sum of squares within and between clusters.

The C-index is computed as follows:

$$C - index = \frac{D_u - (r - D_{min})}{(r \times D_{max}) - (r \times D_{min})} \tag{2.8}$$

where $D_{min} \neq D_{max}$, C-index $\in (0, 1)$. D_u is the sum of all within-cluster dissimilarities, r is their number, D_{min} is the smallest and D_{max} is the largest within-cluster dissimilarities, respectively (Hubert and Levin, 1976).

The τ-index is defined as follows:

$$\tau = \frac{s(+) - s(-)}{s(+) + s(-)} \tag{2.9}$$

where s(+) is the number of concordant comparisons and s(-) is the number of discordant comparisons. Comparisons are made between all within-cluster dissimilarities and all between-cluster dissimilarities. A comparison is considered to be concordant (s(+)) if a within-cluster dissimilarity is strictly less than a between-cluster dissimilarity. A comparison is considered to be discordant (s(-)) if a within-cluster dissimilarity is strictly greater than a between-cluster dissimilarity (Baker and Hubert, 1975).

2.5 Symbolic Data Analysis

Information retrieval from large data sets appear as a challenge, because traditional/classic methods are not capable to handle it. One approach to solve this issue is Symbolic Data Analysis that allows summarizing large data sets in such a way that the resulting summary data set is of a manageable size. Symbolic

Data Analysis has been introduced by Diday (1988) and enables to consider data that contain information which cannot be represented within the classical data models. Additionally, the results obtained are directly interpretable in terms of the input descriptive variables.

Symbolic data contains internal variability and is different from classic data (categorical, quantitative). In the case of classical statistics, data is usually represented in a $n \times p$ data array where one single value is recorded for each variable (p) and for each observation (n). Symbolic data, with intrinsic variability and where new variables types have been introduced (e.g. in the form of sets, intervals or distributions over a given domain), require specific methods and approaches - Symbolic Data Analysis (SDA).

As in classical statistics, the new variables can be distinguished between numerical and categorical variables (Brito, 2014). According to Brito (2014) there are three main new types of symbolic data variables: multi-valued (quantitative or categorical), histogram-valued variables (includes the particular case: interval-valued variables) and categorical modal variables. In the first case, a quantitive multi-valued variables, Y, is defined by an application:

$$Y : S \to B \tag{2.10}$$

such that,

$$s_i \mapsto Y(s_i) = \{c_{i1}, ..., c_{in_i}\} \tag{2.11}$$

where S is a given set of units. For quantitative variables, B is the power set of an underlying set O $\subseteq \mathbb{R}$ (excepting, $\emptyset$, empty set) and $Y(s_i)$ is a non-empty set of real numbers. In the case of categorical multi-valued variables, B is the set of non-empty subsets of a set of categories O $= \{m_1, ..., m_k\}$ and the values of $Y(s_i)$ are finite sets of categories.

In the case of histogram-valued variables we define subintervals between the global lower and upper bounds and compute frequencies for these intervals. Thus, given $S = \{s_1, ..., s_n\}$, a histogram-valued is defined by an application:

$$Y : S \to B \tag{2.12}$$

such that,

$$s_i \mapsto Y(s_i) = \{[\underline{I}_{i1}, \overline{I}_{i1}], p_{i1}; [\underline{I}_{i2}, \overline{I}_{i2}], p_{i2}; ...; [\underline{I}_{ik_i}, \overline{I}_{ik_i}], p_{ik_i}\} \tag{2.13}$$

where $I_{il} = [\underline{I}_{il}, \overline{I}_{il}]$, l=1,...,$k_i$ are the subintervals considered for observation $s_i, p_{i1} + ... + p_{ik_i} = 1$; B is now the set of frequency distributions over $I_{i1}, ..., I_{ik_i}$. For each unit s_i values are assumed to be uniformly distributed within each subinterval. For different observations, the number and length of subintervals of the histograms may naturally be different.

Interval-valued variables may be considered as a particular case of histogram-valued variables (when k=1).

A categorical modal variable Y is expressed as:

$$Y = \{m_1(p_1), ..., m_k(p_k)\} \tag{2.14}$$

where Y has a finite domain $O = \{m_1, ..., m_k\}$. For each element, we are given a category set and for each category m_l a measure, p_l, indicates how frequent or likely that category is for this element. The measures are often weights, probabilities or relative frequencies. Now, B is the set of distributions (probability, frequency, etc.) over O.

Next, we will exemplify each variable type presented above. Consider a dataset containing information about three special products that a company produces. Table 2.2 presents data of these three products. For product type B, for instance, the expected demand ranges from 1500 to 3000, the number of special requirements is 1 or 4 and production time is between 10 and 20 minutes for 15% of the products, between 20 and 30 minutes for 70% and between 30 and 50 minutes for the remaining 15%. Thus, the expected demand is an interval-valued variable, the number of special requirements is a multi-valued quantitative variable and production time is a histogram-valued variable.

Additionaly, we have in Table 2.3 the information about the market shares by product producers. The variable presented is a categorical modal variable.

Table 2.2: Data for products.

Product	Expected demand	No. of req.	Production time(min)
A	[1000,2000]	{1,2,3}	{[10,20[,0.05; [20,30[,0.65; [30,50[,0.30}
B	[1500,3000]	{1,4}	{[10,20[,0.15; [20,30[,0.70; [30,50[,0.15}
C	[100,200]	{2,3}	{[10,20[,0.02; [20,30[,0.50; [30,50[,0.48}

Table 2.3: Data for market share.

Product	Main producers
A	{Company ABC (0.25), Company BCD (0.60); Company CDE (0.15)}
B	{Company ABC (0.35), Company CDE (0.50); Company EFG (0.15)}
C	{Company BCD (0.80); Company EFG (0.20)}

Brito (2014) also refers other types of symbolic data: taxonomic variables – whose values are organized in a tree with several levels of generality and constrained variables – where the variable is hierarchically dependent from another one if its application is constrained by the values taken by it.

Symbolic data are complex data as they cannot be reduced to standard data without losing much information.

Descriptive statistics have been developed for these new types of variables, principal components methods for interval-valued exist, regression methods for interval-valued and histogram-valued variables exist, and a lot of research on (dis)similarity measures and subsequent clustering techniques has been developed. Despite all those developments, there are still many opportunities for methodological improvement (Beranger et al., 2018).

The text analyzed - the speeches - will be represented as categorical modal symbolic data. After this, clustering techniques developed for this type of symbolic data will be applied. In the next subsection will be described some of these clustering techniques.

2.5.1 Clustering of Symbolic Data

Methods and concepts

Many different clustering approaches, hierarchical and non-hierarchical, have been developed in the context of Symbolic Data Analysis. Some approaches will be mentioned in this subchapter, however this is a field that is currently developing, so that the list is not exhaustive.

K-means clustering extension approaches have been developed, Ralambondrainy (1995) proposes a hybrid numeric-symbolic method that integrates an extended version of the K-means for cluster determination and a complementary characterization algorithm (GENER) for cluster description. Fuzzy extensions have been developed in order to apply the concept of fuzziness on a symbolic dataset, describing the clustering problem in a partitioning approach (El-Sonbaty and Ismail, 1998). Brito and Diday (1990) had developed symbolic pyramidal clustering by coupling the pyramidal model with symbolic criteria to form clusters. A method for symbolic hierarchical clustering has been developed to deal with multi-valued data by Brito (1991, 1994) and was further developed for modal variables (Brito, 1998).

SCLUST is a non hierarchical clustering method that can be applied to a set of symbolic objects in order to generate partitions (Verde et al., 2000). This method is a generalization to symbolic objects of the dynamic clustering method (Celeux et al., 1989). The author defines prototype as a model of a class, and its representation can be an element of the same space of representation of the concepts to be clustered which generalizes the characteristics of the elements belonging to the class. The method starts from a partition on a prefixed number of clusters and alternates an assignment step and a representation step until convergence is achieved (or the limit e.g. maximum number of iterations is reached). The assignment step is based on minimum distance to cluster prototypes and the representation step enables the determination of new prototypes in each cluster. The method determines a series of partitions that improves at each step a mathematical criterion.

Korenjak-Černe et al. (2011) have developed two other clustering methods based on the data descriptions with discrete distributions: the adapted leaders method and the adapted agglomerative hierarchical clustering Ward's method. Units are described by discrete distributions, because such a description enables to deal with all types of variables (numerical, ordinal and nominal) and more detailed information about the raw data than the mean value is preserved. Despite the fact that each of the methods can be used separately, they can also be used to perform clustering in two stages. The first stage of this combined method consists in efficiently cluster a large data set with the adapted leaders method. In the second stage it is applied an agglomerative hierarchical method to clusters' leaders to reveal the internal structure among them and to determine the appropriate number of clusters. The adapted version of the leaders method (first stage) is a variant of a dynamic clustering method (Diday, 1972) and is described as follows:

Determine an initial clustering
Repeat
    Determine leaders of the clusters in the current clustering
    Assign each unit to the nearest new leader – producing a new clustering
Until the leaders stabilize

Figure 2.3: Leaders algorithm.

(Adapted from Korenjak-Černe and Batagelj (2002)).

The leaders method is a local optimization method. In the second stage the method used is also very popular, it allows a tree-representation of the results - a dendogram. An adapted agglomerative clustering method based on the descriptions of units, cluster and cluster's leaders by discrete distributions is described with the following procedure:

Each unit forms a cluster: $C_n = \{\{X\}: X \in U\}$;

They are at level 0 : h({X})=0, $X \in U$;

For k=n-1 to 1 **do**

Determine the closest pair of clusters

$$(p, q) = argmin_{i,j;\, i \neq j}\{D(C_i, C_j): C_i, C_j \in C_{k+1}\}$$

Join the closest pair of clusters $C_{(pq)}$=$C_p \cup C_q$

$$C_k = (C_{k+1} \backslash \{C_q; C_p\}) \cup \{C_{(pq)}\}$$

h($C_{(pq)}$)=D(C_p, C_q)

determine the dissimilarities D($C_{(pq)}, C_s$), $C_s \epsilon C_k$

endfor

Figure 2.4: Agglomerative hierarchical clustering method.

(Retrieved from Korenjak-Černe and Batagelj (2002)).

Considering that C_k is a partition of the finite set of units U into k clusters. The level h of the cluster $C_{pq} = C_p \cup C_q$ is determined by the dissimilarity between the join clusters C_p and C_q.

In this approach, the leaders of the clusters are taken as units for hierarchical clustering from the clustering obtained by the leaders method (Korenjak-Černe et al., 2011).

It was also developed by Kejžar et al. (2021) an adaptation of the method described above - agglomerative hierarchical and leaders method - that can be used with alternative dissimilarity measures and that allows the use of weights for each object to consider its size (counts/frequencies).

2.6 Compositional Data Analysis

Compositional Data Analysis refers to the analysis of compositional data, that are defined as random vectors with strictly positive components and a constant sum. This kind of data is measured in proportions, percentages, parts per million, or similar. Different applications of compositional data include geology, economy, chemistry, genetics, sociology, etc.

A composition is a vector x defined on the (D-1) – dimensional simplex space

$$S^D = \{x = [x_1, x_2, ..., x_D] : x_1 > 0; x_2 > 0, ..., x_D > 0; \sum_{i=1}^{D} x_i = k\} \tag{2.15}$$

where k is an arbitrary positive constant - usually 1 (parts per one), 100 (percentages) or 10^6.

Consider the example detailed in Carson et al. (2016) of physical activity behavior data. Body sensors enable measuring precisely the time spent sleeping, in sedentary behavior (SB), in light activity (LIPA) or in moderate and vigorous activity (MVPA) over 24 hours. Hence the sum of the time spent in each behavior will be 24 hours, apart from slight measurement or rounding-off errors, and in percentage it will sum up to 100% of the day. Thus, a four - part composition consisting of sleep, SB, LIPA and MPVA times over a day would satisfy:

$$t_{sleep} + t_{SB} + t_{LIPA} + t_{MPVA} = 24hours \tag{2.16}$$

Data analysis involving compositions should fulfil two main principles: scale invariance (compositions provide information only about the relative magnitudes of their components, and therefore the closure constant k is irrelevant) and sub compositional coherence. Working with ratios, or equivalently logratios, involves not only scale invariance, but automatically subcompositional coherence, since ratios within a subcomposition are equal to the corresponding ratios within the full composition.

In the 1980's appropriate methodology, taking account of some logically necessary principles of compositional data analysis and the special nature of compositional sample spaces, come from Atchison and Shen (1980) and Aitchison (1982, 1983, 1985), culminating in the methodological monograph Aitchison (1986).

In Aitchison (1986) three main transformations were proposed in order to transform data from the simplex space to the standard real space - the additive

log-ratio transformation:

$$alr(x) = [\log(\frac{x_1}{x_D}), ..., \log(\frac{x_{D-1}}{x_D})] \tag{2.17}$$

the centered log-ratio transformation:

$$clr(x) = [\log(\frac{x_1}{g(x)}), ..., \log(\frac{x_D}{g(x)})] \tag{2.18}$$

and the isometric log-ratio transformation:

$$ilr(x) = [< x, e_1 >, ..., < x, e_{D-1} >] \tag{2.19}$$

where x represents the composition vector, g(x) is the geometric mean of the composition x, $< , >$ indicates the ordinary Euclidean inner product [2] and $e_1, ..., e_{D-1}$ forms an orthonormal basis in the Simplex.

Blasco-Duatis and Coenders (2020) applied compositional analysis and visualization tools as compositional biplot to provide a sentiment analysis of the political parties' discussion on Twiter about the motion of no confidence in the Spanish government. Kim et al. (2020) have used topic modeling to identify the fundamental dimensions or building blocks of religion & spirituality. The data used consisted of 255 self-report inventories of religion & spirituality published from 1929 to 2017. It was also noticed that as the topic proportional sum in each document is always 1 or 100 percent, this is a type of compositional data.

Our data - speeches - will be considered as compositions, in the context of Compositional Data Analysis. Afterwards, transformations and appropriate methods will be applied to obtain classes of speeches. In the next subsection will be detailed some clustering approaches appropriate for compositional data.

2.6.1 Clustering of Compositional Data

In order to apply any hierarchical method of classification, it is necessary to establish in advance which are the measures of difference, central tendency and

[2] The inner product is computed as follows: $< x, y >= \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} log\frac{x_i}{x_j} log\frac{y_i}{y_j} \; \forall x, y \in S^D$

dispersion, to be used in accordance with the nature of data to be classified. When classifying a compositional data set, appropriate measures of difference have to be used, as Aitchison's or the Mahalanobis (clr). To calculate the matrix of differences associated with hierarchical methods, as Single Linkage, Complete Linkage and Average Linkage applied to a compositional data set, only Aitchison's distance will be suitable. The Aitchison distance between x and y $\in S^D$ is defined as:

$$d_{At} = d_A(x,y) = \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}(ln\frac{x_i}{x_j} - ln\frac{y_i}{y_j})^2} \tag{2.20}$$

Any clustering method which reduces the measure of difference from a composition to a cluster C of compositions to the difference between the composition and the "center" of the group, would have to take into account that the arithmetic mean $\overline{C}$ of the data set is usually not representative of the "center" of the set, and neither is compatible with the group of perturbations. The concept of group of pertubations was introduced by Aitchison as a means to characterize the "difference" between two compositions. Perturbing a vector $x = [x_1, x_2, ..., x_D]$ in S^D by a vector $y = [y_1, y_2, ..., y_D]$ (also in S^D) results in a new vector defined by $x \oplus y = C[x_1 \times y_1, x_2 \times y_2, ..., x_D \times y_D]$, where C is the closure operation. The closure operation is defined as:

$$C(x) = [\frac{k \times x_1}{\sum_{i=1}^{D} x_i}, \frac{k \times x_2}{\sum_{i=1}^{D} x_i}, ..., \frac{k \times x_D}{\sum_{i=1}^{D} x_i}] \tag{2.21}$$

where k depends on the units of measurement: usual values are 1 (proportions), 100 (percentages), 10^6 (ppm) and 10^9 (ppb).

Aitchison proposed the geometric mean center(C) as a more representative point of the central tendency of a compositional dataset C in S^D, that is defined as:

$$cen(C) = \frac{(g_1, g_2, ..., g_D)}{g_1 + g_2 + ... + g_D} \tag{2.22}$$

where $g_j = (\prod_{i=1}^{N} x_{ij})^{\frac{1}{N}}$ is the geometric mean of the jth component of the compositions $x_1, x_2, ..., x_n$ in C for a data set of size N.

Thus, this definition of the center of a set of compositions should be used, in

addition to Aitchison's distance.

However, other methods use the measure of dispersion to classify the observations of a data set as the Ward method. This method is based on the concept of variability on a cluster C. When the data set is compositional the variability should be defined as follows:

$$\sum_{x \in C} d_{at}^2(x, cen(C)) \tag{2.23}$$

where d_{at} denotes the Aitchison's distance.

This kind of adaptations are introduced to make the standard hierarchical clustering methods compatible with the compositional nature of a data set X. If these methods were applied to the transformed data set clr(X), these adaptations can be omitted (Martín-Fernández et al., 1998).

Palarea-Albaladejo et al. (2012) have shown that the most commonly-used distances for compositions do not agree with the principles of the simplex space mentioned before: scale invariance and subcompositional coherence. The only two measures that satisfy them are the Aitchison distance and the C-KL dissimilarity. Therefore, the adequacy of different dissimilarities in the simplex, together with the behavior of the common log-ratio transformations, is discussed in the basis of compositional principles. After that, a coherent framework for applying the widely-used C-means (FCM) algorithm is introduced.

Other authors have described a strategy for clustering compositional data with K-means algorithm and several adapted transformations. For instance, Godichon-Baggioni et al. (2019) refer that the choice of an appropriate transformation in practice depends strongly on the type of cluster profiles that are of interest in a given context.

2.7 Discriminant Analysis

Discriminant Analysis is a statistical method that allows understanding what distinguishes two or more groups - descriptive purpose - and building a rule to

predict or decide to which group a new case belongs to - classification purpose. This method of classification, in contrast with clustering as mentioned in the previous sections, assumes prior knowledge of the class membership of the multivariate data. The prediction is based on a discriminant rule which is obtained through training data. Typically, then for the test set we have only the multivariate data without class membership information. The aim is to predict the class of membership for this set of observations.

There are different methods of discriminant analysis such as Support Vector Machines, Neural Networks, *Fisher*'s linear discriminant analysis, K-Nearest neighbours discriminant analysis, etc. We will focus on Linear Discriminant Analysis, with Fisher (1936) method as a specific approach and Quadratic Discriminant Analysis (QDA).

The assumptions for Linear Discriminant Analysis include normality of the independent variables, homoscedasticity that corresponds to a situation in which variances and covariances of the independent variables are the same across groups, no multicolinearity that corresponds to the non-existence of independent variables substantially correlated amongst each other and independence that refers to the fact that the data points should consist of a independent and identically distributed sample. Fisher (1936) does not make some of the assumptions of Linear Discriminant Analysis such as normal distributed variables or equal covariance across classes. However, only under certain assumptions we obtain "optimality" of the classification rule in the sense of a minimal misclassification error, considering the training data.

The idea behind *Fisher's* method consist in searching for projection directions which allow for a maximum separation of the group means with in the projected data.

Consider that n observations are given from a training data set. Also consider a two-group case (g=2), one considers a projection direction a $\in \mathbb{R}$, with a $\neq$ 0 and that group means are μ_1 and μ_2, then the projected group means are denoted as $\mu_{1,y} = a'\mu_1$ and $\mu_{2,y} = a'\mu_2$. Denoting by B and W, the matrices of sums of squares and cross products between and within groups, respectively, the

maximization problem can be expressed as:

$$\frac{a'B_a}{a'W_a} \tag{2.24}$$

The discriminant function, hereinafter denoted by Z, is a linear combination of the original variables:

$$Z = Y\gamma = \gamma_1 Y_1 + \gamma_2 Y_2 + ... + \gamma_p Y_p \tag{2.25}$$

where Y are the centered original variables and γ is the p x 1 vector of coefficients.

To lead to the best separation, the mean-values of Z for the two groups should be as different as possible - maximum sum of squares between-groups - and the values of Z within each group should be as similar as possible - minimum sum of squares within-groups.

QDA represents an extension of Linear Discriminant Analysis for nonlinear class separations. QDA assumes importance when there is prior knowledge that individual classes exhibit distinct covariances, so this extension proposes the estimation of a covariance matrix for every group of observations.

Some of the main applications of discriminant analysis include bankrupcy prediction (Altman, 1968), face recognition (Etemad and Chellappa, 1997), marketing (Dahiya and Sachar, 2021), biomedical studies as in medicine with the assessment of severity state of a patient and prognosis of disease outcome (Hughes et al., 1963) and earth science (e.g. discriminant analysis for finding patterns and to classify various zones as developed by Tahmasebi et al. (2010)). Kitchens and Powell (1975) have also applied discriminant analysis to isolate influence variables in a political campaign.

2.7.1 Discriminant Analysis for Compositional Data

According to Filzmoser et al. (2012), since compositional data include only relative information, some transformations should be made before applying methods that are based on the standard Euclidean geometry. Rather than working in

the simplex space, compositional data are usually transformed to the Euclidean space.

Despite the fact that results obtained by the *alr* approach are usually of intuitive interpretation, the *alr* transformation is not isometric and the application should be avoided, because the corresponding basis on the simplex is not orthonomal with respect to the Aitchison geometry. Isometric logratio (*ilr*) transformation fulfill the orthonormality and is defined as:

$$z = (z_1, ..., z_{D-1})^t, z_i = \sqrt{\frac{i}{i+1}} ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}, \text{for i =1,...,D-1.} \tag{2.26}$$

The ilr transformation moves the whole Aitchison geometry on the simplex to the standard Euclidian geometry.

The authors have also concluded that in presence of outliers the robust version of discriminant analysis is preferable to the classical one. Robust estimator is needed to make discriminant analysis work optimally within the classification though in the condition of data which contains outliers i.e. case with such an extreme value on one or more variables that distorts the statistics. This consists on replacing mean vectors and covariance matrices in discriminant analysis by robust counteparts. However, robust discriminant analysis will not necessarily improve the misclassification rates, which depend on the position of the outliers on the space.

In this dissertation we have used discriminant analysis, particularly *Fisher* method, to predict the speech author.

Chapter 3

Data and Software

In this chapter we will present the data and software to be used.

Data

The data consist of 83 presidential american speeches from 2008 to 2020, where 49 speeches were made by Barack Obama and 34 by Donald Trump. These speeches were retrieved from UVA's Miller Center[1] and have different sizes. Each speech is considered as a document. The words taken from all documents after pre-processing are referred as terms.

Pre-processing

Given the corpus of raw data (in our case Donald Trump and Barack Obama presidential speeches), the first step of TM is the pre-processing. This phase is particularly important as it involves a sequence of techniques which should improve the next phases of a TM process, leading to better performances.

The software used was KNIME and in Figure A1 of the *Appendix*, is represented the general view of KNIME and the specific pre-processing techniques applied.

After concatenating the two collections of documents (classified as DT and BO, where DT corresponds to Donald Trump speeches and BO corresponds to Barack Obama speeches), we obtained a table with 83 rows, containing noisy and

[1] millercenter.org/the-presidency/presidential-speeches retrieved november 15, 2020

some uninformative data where each row corresponds to one speech retrieved from UVA's Miller Center.

Some of the steps needed to pre-process the data include removal of punctuation and symbols (€) and the elimination of short words, with less than two characters, and stopwords. Additionally, we eliminated numbers from the documents (using the *Number Filter* node).

In particular, for the stopword technique we created a list of words, that in this context were considered irrelevant. Some examples of words included in that list[2] were "the", "donald", "trump", "applause", "america" and "thank". In fact, first we have checked the frequency of all words and concluded that the more frequent words were "united", "states", "america", "thank" and "applause". It is assumed that words appearing often in the documents may be more relevant for identification of the class than the words appearing rarely. However, it is also assumed that a token that appears in many documents is probably irrelevant (e.g. "america", "thank", "applause", etc.).

Another important step was the conversion of all alphanumeric characters to lowercase, to be able to identify, across all documents, the same word written in different ways. It was also applied lemmatization to the data with the Standford Core NLP library[3]. As a result, we obtained the lemma of a token by removing inflections, such as plurals, pronoun cases and verb endings. An alternative of lemmatization is stemming tokens e.g using Snowball stemming library[4]. The difference between these two feature reduction techniques is that the latter one consists in cutting off the end or the beginning of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word, while lemmatization takes into consideration the morphological analysis of the words.

Table 3.1 presents a comparison before and after pre-processing of an extract from a speech.

[2]The complete list of stopwords is represented in Table C1 of the *Appendix*

[3]More details about the Stanford Core NLP library may be checked in this site - https://stanfordnlp.github.io/CoreNLP/index.html

[4]More details about the Snowball stemming library may be checked in this site - http://snowball.tartarus.org/

Table 3.1: Example - extract from a speech - before and after the pre-processing techniques.

Before Pre-processing	After Pre-processing
"President: Barack Obama, Madam Speaker, Vice President Biden, Members of Congress, and the American people: When I spoke here last winter, this Nation was facing the worst economic crisis since the Great Depression..."	"madam speaker vice members congress spoke winter nation facing worst economic crisis since depression"

After this cleansing process, we created a Bag of Words (BoW) with the tokens occurring in the 83 pre-processed documents. The output of the BoW node is a table of two columns: one containing the tokens extracted from the pre-processed document and the other with the corresponding document from where the token was extracted. The number of rows in this table -78758 words- corresponds to the number of words of the "bag". Then, in order to eliminate the words with lowest frequencies, we computed the absolute and relative frequency of each pre-processed token. As such words which only appeared once in the "bag" were removed from the table. This method of removing words with low frequencies is called TF-IDF.

Based on the "bag" with the pre-processed tokens and after all the pre-processing techniques and with the usage of the *Tag Cloud* node, as represented in Figure 3.1, it is possible to visually represent the most frequent words of our dataset.

Figure 3.1: Tag Cloud after pre-processing.

We have then applied Topic Modeling, namely Latent Dirichlet Allocation (LDA) by using the *Topic extractor* (Parallel LDA) node. This was applied to the pre-processed documents, i.e. filtered, lemmatized, etc. The number of topics defined was 10, as well as the number of top words to extract per topic. The Alpha parameter (α) that defines the prior weight of each topic k in a document, was set at 0.1. The Beta parameter (β) that defines the prior weight of each word w in a topic, was also set at 0.1. It was defined 10 iterations and a number of threads of 8.

The ten topics defined were as follows: *Coronavirus*, *Energy & Oil*, *Family/Children*, *Security(War)*, *Jobs/Work*, *Economy*, *External Relations*, *Tax*, *System*, *Human Rights*. As mentioned before the number of top words [5] per topic is also 10. For instance, for topic*Coronavirus* the top words that define it are "country", "virus", "testing", "governors", "working", "cases", "secretary", "question", "tests" and "health".

Table 3.2 presents a partial view of the output obtained after pre-processing and the application of the method Latent Dirichlet Allocation. This representation of the data is the basis for clustering and discriminant analysis.

Table 3.2: Partial view of the output after pre-processing and LDA.

Speech	Coronavirus	Energy&Oil	Family/Children	Security(War)	Jobs/work	Economy	External Relations	Tax	System	Human Rights	Assigned topic
1	0.0000	0.0000	0.2593	0.1437	0.0278	0.5464	0.0000	0.0000	0.0012	0.0211	Economy
2	0.0000	0.0013	0.1795	0.0064	0.0014	0.5974	0.1229	0.0037	0.0868	0.0002	Economy
3	0.0017	0.0001	0.7222	0.0881	0.0036	0.1575	0.0036	0.0001	0.0155	0.0071	Family/Children
4	0.0000	0.0000	0.0364	0.2746	0.0004	0.0763	0.0004	0.0000	0.0054	0.6060	Human Rights
5	0.0000	0.0000	0.0002	0.7688	0.0038	0.0725	0.0002	0.0001	0.0002	0.1537	Security(war)
...	...	...	...	...	...	...	...	...	...	...	...
82	0.7257	0.0000	0.0011	0.0019	0.0735	0.0300	0.0001	0.0567	0.1040	0.0065	Coronavirus
83	0.6922	0.0045	0.0000	0.0205	0.0535	0.0630	0.0379	0.0376	0.0904	0.0000	Coronavirus

[5] The top words for the ten topics are represented in Table D1 of the Appendix

Note that even though these techniques are of extreme importance to the quality of the results, there are some limitations. These limitations include lacks of contextual information and the fact that order and semantic relationships between words are ignored. As such, the results of the pre-processed method should be dealt with caution.

Software

As mentioned in the previous subsection for pre-processing we used KNIME. It allows the assembly of nodes blending different data sources, including pre-processing for modeling, data analysis and visualization. For the descriptive statistics we used CoDaPack also referred as Compositional Data Package which implements the most elementary statistical methods applicable to compositional data. It allows making transformations between the real space to the simplex and vice versa, operations, 2D and 3D graphical outputs and Compositional Descriptive Statistics. Finally, we used R to apply clustering methods and to perform the discriminant analysis.

Chapter 4

Analysis of the results

4.1 Introduction

In this chapter we present descriptive statistics of the analysed data, and the clustering results using compositional and symbolic approaches. For the descriptive statistics we used essentially CoDaPack. For the application of clustering methods we used R, namely the package *Compositions* and *Clamix*.

In order to evaluate the quality of our results we present here below a brief historical background of the tenures of both presidents.

Barack Obama is an american lawyer and politician who served as the 44^{th} president of the United States of America from 2009 to 2017. During the first tenure the main reforms included economic stimulus activities in response to the Great Recession and the financial crisis. Other reforms that were passed include the Affordable Care Act (commonly referred as *Obamacare*). Obama ordered the end to US involvement in the Iraq War, increased the number of troops in Afeganistan, reduced nuclear weapons with Russia, authorized an armed intervention in the Lybian Civil War and ordered a military operation in the Pakistan that resulted in the death of Bin Laden.

Donald Trump is an american businessman, television personality and politician who served as the 45^{th} president of the United States from 2017 to 2021. During his presidency, Trump signed an executive order banning citizens from seven Muslim-majority countries. Trump signed massive tax cuts, revoked provi-

sions that had been created due to the economic crisis, tried to undo *Obamacare* and also made several revisions to environmental regulations to allow for the expansion of fossil fuel exploration. Trump withdrew the US from negotiations on the Trans-Pacific Partnership and the Paris Agreement on climate change and also withdrew US from the nuclear agreement with Iran. Donald Trump has also imposed tariffs on several imported products which triggered a trade war with China. His reaction to the pandemic (Covid-19) was slow and ignored expert recommendations.

4.2 Descriptive Statistics

Standard descriptive statistics are somewhat useless when dealing with compositional data. In Aitchison geometry, the arithmetic mean, the variance and standard deviation cannot be seen as measures of central tendency and dispersion. Instead, some alternatives have been introduced in Section 2.6.1 as the concept of centre, variation matrix and total variance.

In Table 4.1 are represented the center as well as the quartiles of the speeches' data set. The center of the data set is represented as:

$$\hat{E} = C[g_1, g_2, ..., g_D] \tag{4.1}$$

where, $g_i = (\prod_{k=1}^{N} x_{ki})^{1/N}$ stands for the geometric mean of part X_i in the data set X.

Table 4.1: Center and quartiles of speeches' data set.

Topic	Center	0	25	50	75	100
Coronavirus	0.0017	0.0000	0.0000	0.0001	0.0088	0.7667
Energy&Oil	0.0027	0.0000	0.0000	0.0001	0.0133	0.5491
Family/Children	0.0416	0.0001	0.0003	0.0286	0.1190	0.7749
Security(War)	0.2335	0.0001	0.0204	0.0964	0.2133	0.8526
Jobs/work	0.0262	0.0001	0.0005	0.0071	0.0614	0.9060
Economy	0.5758	0.0008	0.0753	0.1460	0.4445	0.8974
External Relations	0.0231	0.0000	0.0002	0.0042	0.1728	0.5205
Tax	0.0027	0.0000	0.0001	0.0002	0.0115	0.9984
System	0.0425	0.0001	0.0010	0.0177	0.0817	0.5775
Human Rights	0.0503	0.0001	0.0010	0.0220	0.1058	0.6432

We also have computed these measures by group. The centre and the quartiles of the speeches made by Donald Trump are represented in Table 4.2. The same output is represented in Table 4.3 for the 49 Barack Obama speeches.

According to the summary of compositional statistics, we can conclude that the topics most addressed by Donald Trump are *External Relations* (30,05%), *Economy* (24,71%) and *Security (War)* (24,75%), while the topics most addressed by Barack Obama are *Economy* (67,08%) followed by *Security(War)* (14,69%). Therefore, we notice that the topic *Economy* in the speeches from BO has more than twice the weight of the same topic in speeches whose author is DT. Additionally, we also notice that the weight of the topic *Human Rights* is higher in speeches from Barack Obama (7,03% comparing with 1,66% in the speeches from Donald Trump). Note that the topic *Coronavirus* is only addressed by Donald Trump in accordance with the chronology of the speeches.

Table 4.3: Center and quartiles of speeches made by Barack Obama.

Topic	Center	0	25	50	75	100
Coronavirus	0.0005	0.0000	0.0000	0.0000	0.0009	0.0613
Energy&Oil	0.0027	0.0000	0.0001	0.0001	0.0259	0.5491
Family/Children	0.0432	0.0001	0.0027	0.0478	0.1795	0.7749
Security(War)	0.1469	0.0002	0.0108	0.0964	0.1781	0.8526
Jobs/work	0.0052	0.0001	0.0002	0.0027	0.0090	0.0889
Economy	0.6708	0.0119	0.1106	0.3060	0.6698	0.8974
External Relations	0.0025	0.0001	0.0002	0.0006	0.0042	0.1229
Tax	0.0011	0.0000	0.0001	0.0001	0.0034	0.0477
System	0.0567	0.0003	0.0042	0.0386	0.0869	0.5775
Human Rights	0.0703	0.0001	0.0071	0.0452	0.1340	0.6432

Table 4.2: Center and quartiles of speeches made by Donald Trump.

Topic	Center	0	25	50	75	100
Coronavirus	0.0054	0.0000	0.0000	0.0003	0.0299	0.7667
Energy&Oil	0.0015	0.0000	0.0000	0.0001	0.0029	0.2048
Family/Children	0.0210	0.0001	0.0002	0.0028	0.0696	0.6416
Security(War)	0.2435	0.0001	0.0247	0.1091	0.2239	0.7146
Jobs/work	0.1442	0.0001	0.0080	0.0564	0.2087	0.9060
Economy	0.2471	0.0008	0.0300	0.0890	0.2124	0.3049
External Relations	0.3005	0.0000	0.0523	0.2604	0.4175	0.5205
Tax	0.0052	0.0000	0.0001	0.0004	0.0218	0.9984
System	0.0150	0.0001	0.0002	0.0054	0.0552	0.1366
Human Rights	0.0166	0.0001	0.0004	0.0057	0.0511	0.179

Following the compositional statistics approach, we also have obtained the variation array of the speeches that is represented in Table 4.4. This matrix is composed by an upper diagonal that contains the log-ratio variances and a lower diagonal that contains the log-ratio means. The ij^{th} component of the upper diagonal corresponds to $var(ln(\frac{x_i}{x_j}))$ and the ij^{th} component of the lower diagonal corresponds to $E(ln(\frac{x_i}{x_j}))$, where i,j=1,2,...,D. Table 4.4 indicates that the lowest variability is related with the topic *Economy*, as the simple log-ratios are the lowest. The *clr*-biplot represented in Figure 4.1 reflects the variability pattern of the variation array. The smallest ray is the one corresponding to *Economy*, which is expected given the concentration of smallest variances in the array. The largest rays corresponds to *External Relations* and *Human Rights*, which is corroborative with the concentration of the largest variances on those topics as represented in

Table 4.4. Table 4.4 also shows that the total variance is 69.7659, that corresponds to the sum of the values in the last column.

Table 4.4: Variation Array.

Xi/Xj	Coronavirus	Energy&Oil	Family/Children	Security(War)	Jobs/work	Economy	External Relations	Tax	System	Human Rights	clr variances
Coronavirus		17.6456	21.4272	17.1394	15.0168	15.3448	20.5275	14.1829	16.1046	23.0723	9.0695
Energy&Oil	0.4964		19.5431	13.5179	19.9391	10.7271	20.3917	15.7894	14.4745	15.5606	7.7823
Family/Children	3.2180	2.7215		13.7269	14.8236	12.4058	19.6392	22.6252	18.1572	11.4743	8.4057
Security(War)	4.9439	4.4475	1.7259		15.0278	7.3788	12.5013	14.1662	16.2649	7.8203	4.7777
Jobs/work	2.7568	2.2603	-0.4612	-2.1871		10.7642	11.6345	14.6261	13.5439	20.7074	6.6318
Economy	5.8464	5.3499	2.6284	0.9025	3.0896		14.6402	9.1878	6.7102	9.3661	2.6759
External Relations	2.6295	2.1331	-0.5884	-2.3144	-0.1272	-3.2168		18.7103	22.5444	18.9408	8.9764
Tax	0.4933	-0.0032	-2.7247	-4.4506	-2.2635	-5.3531	-2.1362		12.8441	20.4044	7.2771
System	3.2394	2.7429	0.0214	-1.7045	0.4826	-2.6070	0.6098	2.7461		16.6189	6.7497
Human Rights	3.4088	2.9123	0.1908	-1.5351	0.6520	-2.4376	0.7792	2.9155	0.1694		7.4199

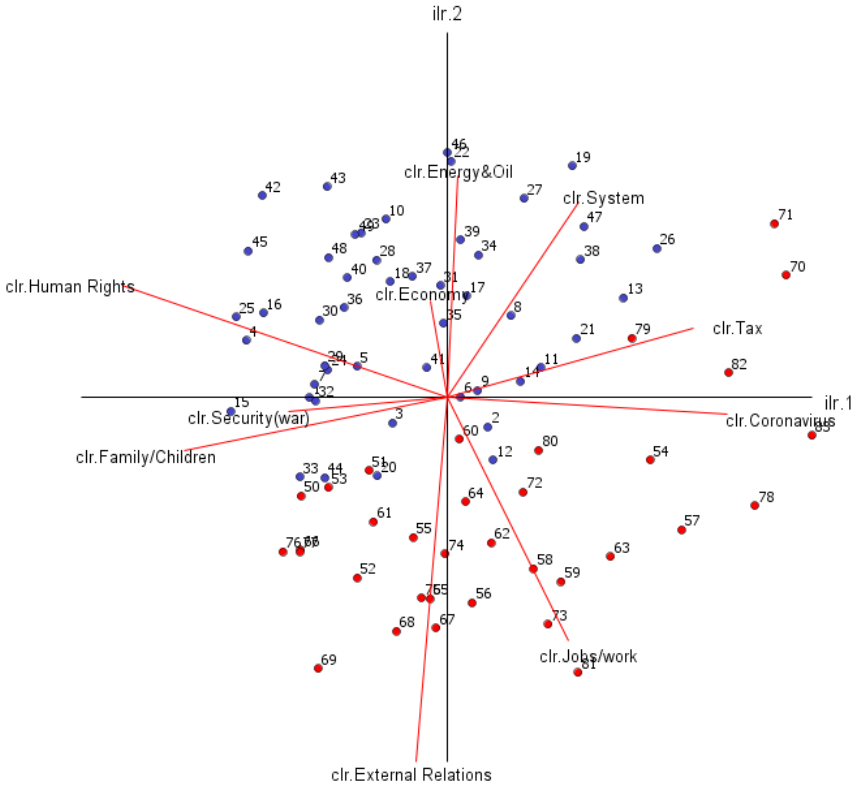


Figure 4.1: Compositional biplot.

Additionally, we perfomed some ternary diagrams by group. Ternary/Quaternary Diagrams display closed three/four-part subcompositions. For instante, the subcomposition displayed in Figure 4.2 is composed by the topics *External Relations*, *Security (War)* and *Human Rights*. As concluded before, this ternary diagram shows that Donald Trump speeches focus more on *External Relations* topics and that Barack Obama speeches include more about *Human Rights* and *Security (War)* topics.

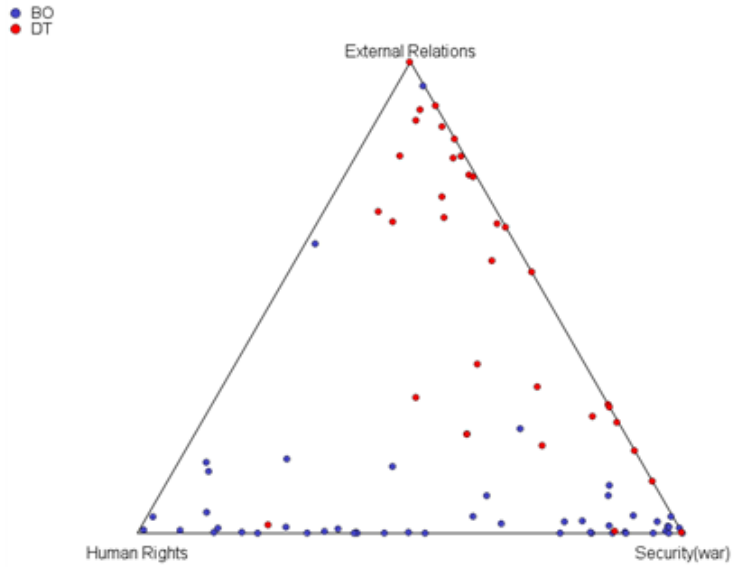


Figure 4.2: Ternary Diagram: *External Relations*, *Security (War)* and *Human Rights*.

The quaternary diagram represented in Figure 4.3 is composed by the topics *Family/Children*, *Economy*, *Coronavirus* and *External Relations*. It can be concluded that the topics of *Family/Children* and *Economy* are more likely to be included in speeches from Barack Obama and the topics of *External Relations* and *Coronovirus* are more likely to be included in speeches from Donald Trump.

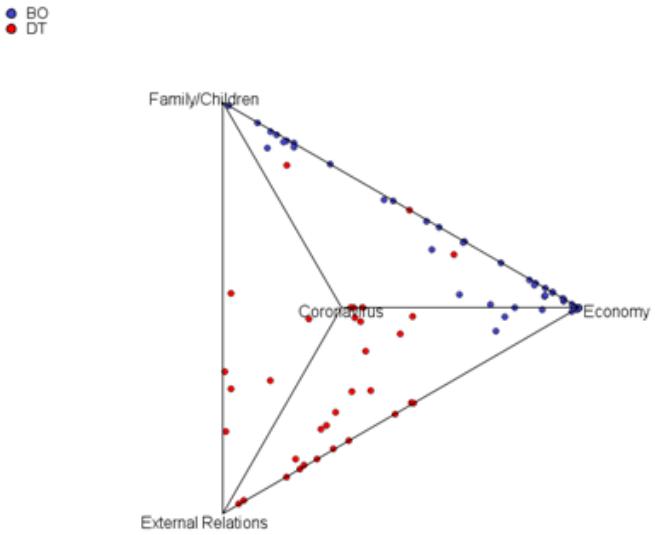


Figure 4.3: Quaternary Diagram: *Economy*, *External Relations*, *Coronavirus* and *Family/Children*.

4.3 Clustering techniques applied to compositional data

In this section we describe the results of applying clustering techniques suitable to compositional data. Particularly, we have applied three different aggregation methods - single, complete and Ward Linkage using Aitchison's distance as comparison measure.

The first clustering method applied was hierarchical clustering with average linkage. To identify the optimal number of clusters we have used the Silhouette and the Calinski Harabaz indices. As mentioned before the Silhouette index determines the optimal number of clusters by analysing the difference between the average distance within the cluster and the minimum distance between clusters. The Calinski Harabaz index is a measure of how similar each element is to its own cluster (cohesion) compared with the others clusters (separation). It was determined as represented in Figure E1 of the *Appendix* that the optimal number of clusters is 10, where the Calinski Harabaz index is maximum and the Silhouette index takes a high value.

The partition obtained using average linkage is represented in Figure 4.4.

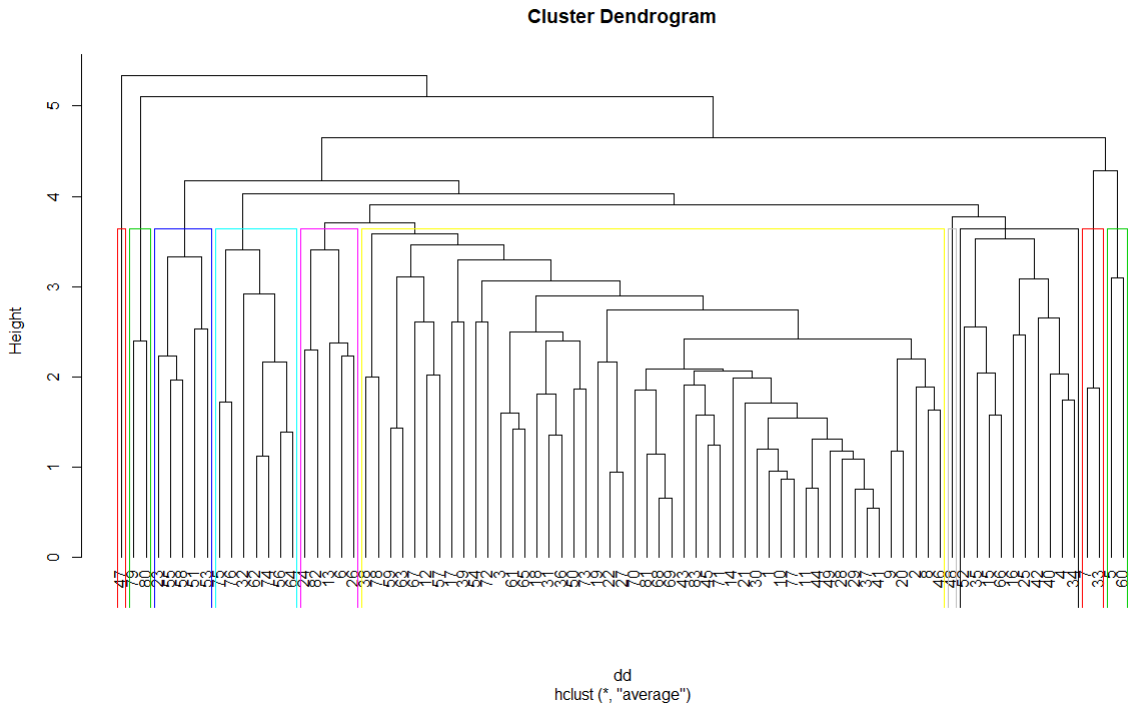


Figure 4.4: Cluster dendogram – Compositional data - Average linkage.

The clusters are organized as follows:

Red={'47'}
Green={'79', '80'}
Purple={'23', '55', '58', '51', '53'}
Blue={'75', '76', '32', '62', '74', '56', '64'}
Pink={'24', '82', '13', '6', '26'}
Yellow={'38', '78', '59', '63', '67', '12', '57', '17', '39', '54', '72', '3', '61', '65', '18', '31', '36', '50', '73', '19', '22', '27', '70', '81', '68', '69', '43', '83', '45', '71', '14', '21', '30', '1', '10', '77', '11', '44', '49', '28', '29', '37', '41', '9', '20', '2', '8', '46'}
Grey={'48'}
Black={'52', '35', '15', '66', '16', '25', '42', '40', '4', '34'}
Red2={'7', '33'}
Green2={'5', '60'}

In Table 4.5 is represented the center of each cluster. The main characteristics of each cluster are summarized in Table 4.6. For instance, the first cluster (in *Red*) is composed by a single speech from Barack Obama (BO) with the main topic *System*. Note that, main topic represents the topic with highest probability to

appear in the speech [1]. The second cluster is composed by two speeches both from Donald Trump (DT) and the main topics are *Tax* and *Jobs/Work*. The center of the cluster represented in Table 4.5 reinforces this conclusion, with an average probability of 0.4992 for *System* and 0.4530 for *Jobs/Work*. The description of the composition of the remaining clusters is detailed in Table 4.6.

This method classifies speeches '79' and '80' as similar, since it joins them in *Green* cluster. Although the two speeches have similar distributions - the main topic has a weight of more than 90% - they are about very different topics. The *Red2* cluster is composed by speches whose probability of the topic *Family/Children* occuring is more than 35% jointly with a probability of the topic *Economy* occur higher than 20%. The *Grey* cluster isolates the speech whose probability of the topics *Security (War)* and *Human Rights* are both higher than 40%.

The *Yellow* cluster is composed by a large number of speeches and has a main topic distribution as represented in Figure 4.5. Most of the spechees in this cluster are from BO.

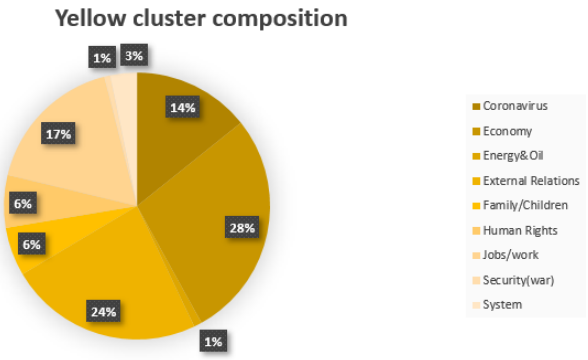


Figure 4.5: Yellow cluster main topics' distribution.

[1] The detail about each speech main topic, date and author is represented in Table B1 of the *Appendix*.

Table 4.5: Center of each cluster - Average linkage solution.

	Coronavirus	*Energy& Oil*	*Family Children*	*Security (War)*	*Jobs work*	*Economy*	*External Relations*	*Tax*	*System*	*Human Rights*
Average										
red	0.0000	0.0000	0.0001	0.0072	0.0001	0.4265	0.0000	0.0012	0.5644	0.0001
green	0.0000	0.0000	0.0123	0.0001	0.4530	0.0137	0.0000	0.4992	0.0212	0.0000
purple	0.0000	0.0011	0.0001	0.4847	0.0918	0.0799	0.2126	0.0020	0.0234	0.1040
blue	0.0096	0.0148	0.1130	0.4713	0.0472	0.0067	0.2901	0.0003	0.0023	0.0440
pink	0.1503	0.0052	0.2677	0.0089	0.0356	0.3833	0.0002	0.0182	0.1018	0.0283
yellow	0.0623	0.0330	0.1085	0.0978	0.0849	0.3561	0.0965	0.0121	0.0848	0.0634
grey	0.0000	0.0000	0.0001	0.4553	0.0001	0.1330	0.0000	0.0000	0.0035	0.4075
black	0.0106	0.0490	0.1135	0.3491	0.0004	0.1371	0.0730	0.0002	0.0156	0.2510
red2	0.0000	0.0001	0.4396	0.0909	0.0666	0.2836	0.0008	0.0001	0.0039	0.1139
green2	0.0000	0.1024	0.0015	0.4096	0.1062	0.1614	0.1106	0.0001	0.0277	0.0801

Table 4.6: Characteristics of each cluster - Average linkage solution.

Cluster	Dimension	Topics distribution	Author distribution
red	1	{System, (100%)}	BO (100%)
green	2	{Tax, (50%); Jobs/Work, (50%)}	DT (100%)
purple	5	{Security (War), (80%); Jobs/Work, (20%)}	DT (80%); BO (20%)
blue	7	{Security (War), (57%); External Relations, (43%)}	DT (86%); BO (14%)
pink	5	{Family/Children, (40%); Economy, (40%); Coronavirus, (20%)}	DT (20%); BO (80%)
yellow	48	Various topics	DT (37,5%); BO (62,5%)
grey	1	{Security (War), (100%)}	BO (100%)
black	10	{External Relations, (10%); EnergyOil, (10%); Human Rights, (40%); Security(War), (30%); Family/Children, (10%)}	DT (20%); BO (80%)
red2	2	{Family/Children, (100%)}	BO (100%)
green2	2	{Security(War), (50%); Economy, (50%)}	DT (50%); BO (50%)

The next clustering method applied was Ward's method. The optimal number of clusters was obtained through the Silhouette and the Calinski Harabaz indices. It was determined as represented in Figure E2 of the *Appendix* that the optimal number of clusters is ten, where the Silhouette index is maximum and the Calinski Harabaz index takes an high value.

The partition obtained using Ward linkage is represented in Figure 4.6.

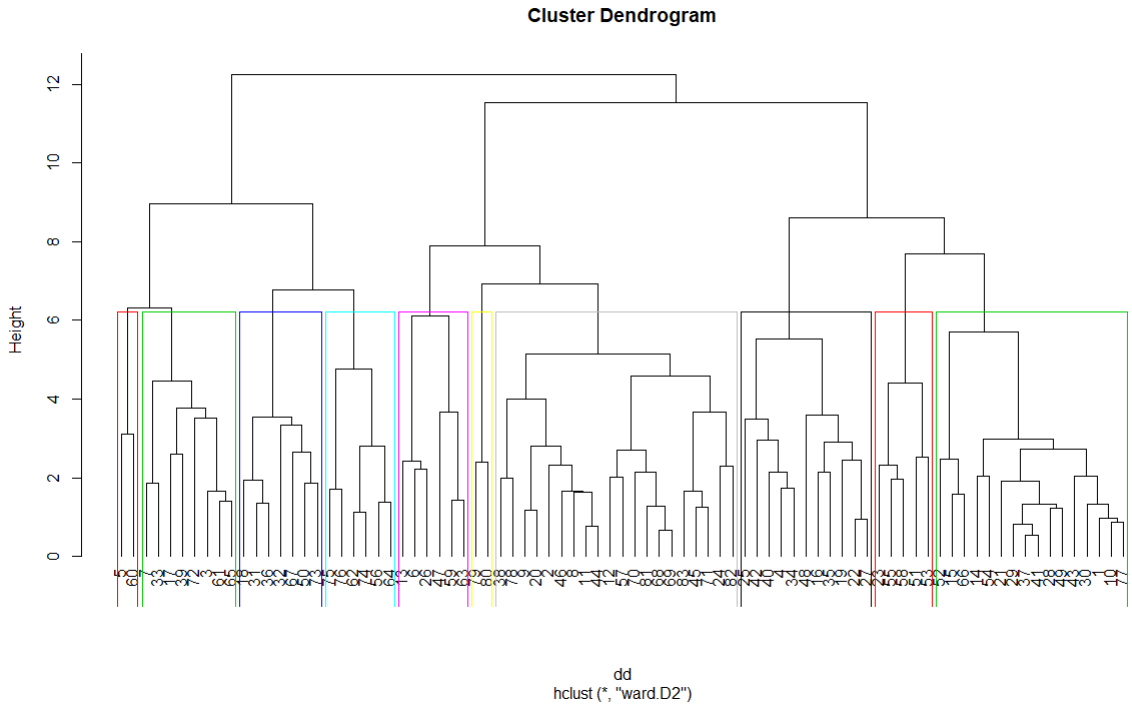


Figure 4.6: Cluster dendogram – Compositional data - Ward linkage.

The clusters are organized as follows:

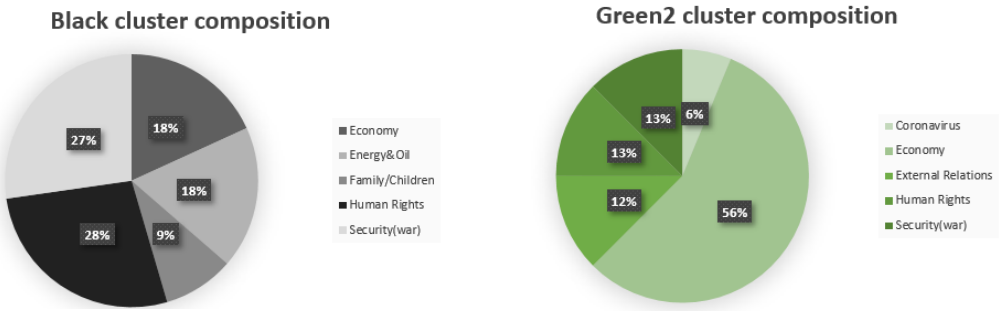
Red={'5', '60'}
Green={'7', '33', '17', '39', '72', '3', '61','65'}
Purple={'18', '31', '36', '32', '67', '50', '73'}
Blue={'75', '76', '62', '74', '56', '64'}
Pink={'13', '6', '26', '47', '59','63'}
Yellow={'79', '80'}
Grey={'38', '78', '9', '20', '2', '46', '8', '11', '44', '12', '57', '70', '81', '68', '69', '83', '45', '71', '24', '82' }
Black={'25', '42', '40', '4', '34', '48','16', '35', '19', '22', '27'}
Red2={'23', '55', '58', '51', '53'}

Green2={'52', '15', '66', '14', '54', '21', '29', '37', '41', '28', '49', '43', '30', '1', '10', '77'}

The center of each cluster is detailed in Table 4.7. In Table 4.8 we have the main characteristics of each cluster summarized, where, as seen before for the Average linkage solution analysis, we have information about the main topics representation and author distribution in each cluster.

Comparing this solution with the previous solution obtained using average linkage we notice that only two cluster remain the same - *Yellow* and *Red* clusters that correspond to the *Green* and *Green2* clusters in solution obtained through Average linkage.

The *Black* cluster is composed by eleven speeches all from BO that have a main topic distribution as represented in Figure 4.7a. The main topics with greatest weight in the cluster are *Human Rights* and *Security (War)*. The *Green2* cluster is mostly composed by speeches from BO about the topic *Economy*. The main topic distribution of the cluster is represented in Figure 4.7b.



(a) *Black* cluster. (b) *Green2* cluster.

Figure 4.7: Main topics' distribution - Ward Linkage.

Table 4.7: Center of each cluster - Ward linkage solution.

	Coronavirus	*Energy& Oil*	*Family Children*	*Security (War)*	*Jobs work*	*Economy*	*External Relations*	*Tax*	*System*	*Human Rights*
Ward										
red	0.0000	0.1024	0.0015	0.4096	0.1062	0.1614	0.1106	0.0001	0.0277	0.0801
green	0.0094	0.0086	0.4122	0.0534	0.0444	0.1714	0.1243	0.0056	0.1146	0.0558
purple	0.0009	0.0088	0.1145	0.1891	0.1918	0.2592	0.0820	0.0030	0.0242	0.1258
blue	0.0112	0.0172	0.1135	0.4078	0.0548	0.0059	0.3384	0.0003	0.0027	0.0477
pink	0.0000	0.1024	0.0015	0.4096	0.1062	0.1614	0.1106	0.0001	0.0277	0.0801
yellow	0.0000	0.0000	0.0123	0.0001	0.4530	0.0137	0.0000	0.4992	0.0212	0.0000
grey	0.1516	0.0053	0.1097	0.0460	0.0961	0.2922	0.0988	0.0159	0.1307	0.0533
black	0.0115	0.1389	0.0724	0.2876	0.0001	0.2103	0.0084	0.0020	0.0388	0.2294
red2	0.0000	0.0011	0.0001	0.4847	0.0918	0.0799	0.2126	0.0020	0.0234	0.1040
green2	0.0363	0.0195	0.0638	0.2255	0.0084	0.4593	0.0765	0.0113	0.0092	0.0897

Table 4.8: Characteristics of each cluster - Ward linkage solution.

Cluster identifier	Dimension	Topics distribution	Author distribution
red	2	{Security(War), (50%); Economy, (50%)}	DT (50%); BO (50%)
green	8	{Family/Children, (62,5%); External Relations, (25%); System, (12,5%)}	DT (62,5%); BO (37,5%)
purple	7	{Jobs/work, (29%); External Relations, (14%); Security (War), (14%); Human Rights, (14%); Economy, (29%)}	DT (43%); BO (57%)
blue	6	{External Relations, (50%); Security (War), (50%)}	DT (100%)
pink	6	{Economy, (50%); Family/Children, (16,6%); Jobs/work, (16,6%); System, (16,6%)}	DT (67%); BO (33%)
yellow	2	{Tax, (50%); Jobs/work, (50%)}	DT (50%); BO (50%)
grey	20	Various topics	DT (45%); BO (55%)
black	11	Various topics	BO (100%)
red2	5	{Jobs/work, (20%); Security (War), (80%)}	DT (80%); BO (20%)
green2	16	Various topics	DT (25%); BO (75%)

The last clustering method applied was Complete Linkage. The optimal number of clusters was obtained through the Silhouette index. It was determined as represented in Figure E3 of *Appendix* that the optimal number of clusters is two, where the Silhouette index is maximum. When the number of clusters is ten the Silhouette and Calinski Harabaz indices take high values. However, since this solution is not optimal in any of the indices we chose not to detail it.

The partition obtained using Complete Linkage is represented in Figure 4.8.

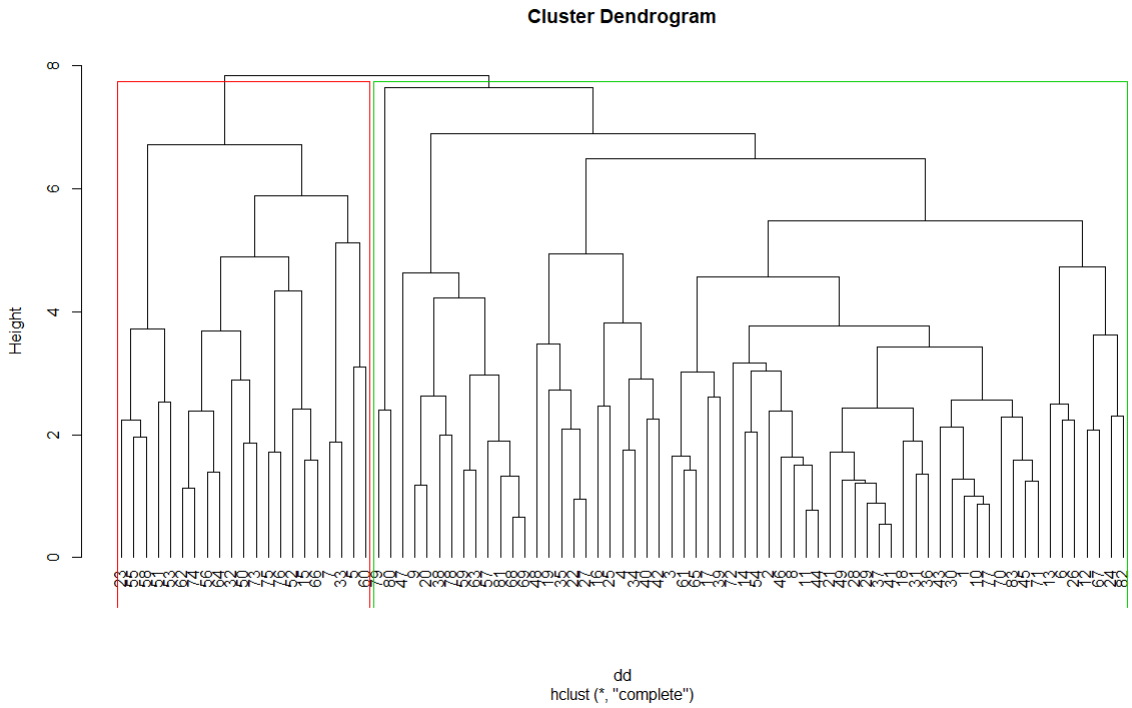


Figure 4.8: Cluster dendogram – Compositional data - Complete Linkage.

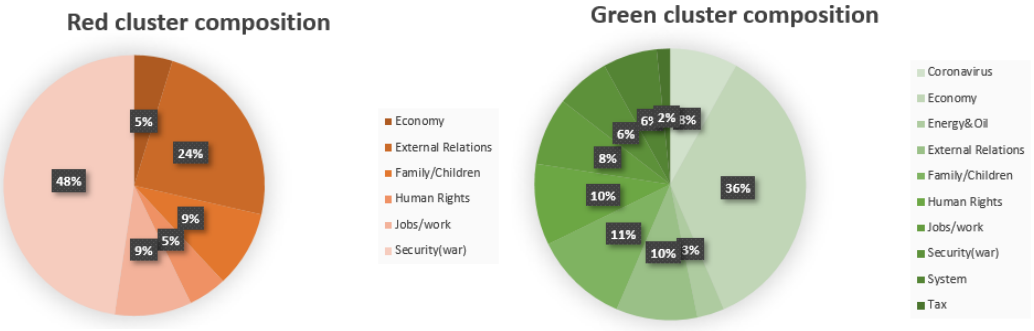
The clusters are organized as follows:

Red={'23', '55', '58', '51', '53', '62', '74', '56', '64', '32', '50', '73', '75', '76', '52', '15', '66', '7', '33', '5', '60' }

Green={'79', '80', '47', '9', '20', '38', '78','59', '63', '57', '61', '68', '69', '48', '19', '35', '22', '27', '16', '25', '4', '34', '40','42', '3', '61', '65', '17', '39', '72', '14', '54', '2', '46', '8', '11', '44', '21', '49', '28', '29', '37', '41', '18', '31', '36', '43', '30', '1', '10', '77', '70', '83', '45', '71', '13', '6', '26', '12', '67', '24', '82'}

The first group is mostly composed by speeches from Donald Trump about

the main topics *Security(War)* (48%), *External Relations* (24%) as represented in Figure 4.9a. The second cluster is composed by 62 speeches where 70% of them are from Barack Obama. The main topic distribution of the cluster is represented in Figure 4.9b. The main topics with greatest weight are *Economy*(36%) and *Family/Children*(11%). The center of each cluster is detailed in Table 4.9.



(a) *Red* cluster.

(b) *Green* cluster.

Figure 4.9: Main topics' distribution - Complete Linkage.

Table 4.9: Center of each cluster - Complete Linkage solution.

	Coronavirus	*Energy & Oil*	*Family Children*	*Security (War)*	*Jobs work*	*Economy*	*External Relations*	*Tax*	*System*	*Human Rights*
red	0.0617	0.0337	0.1132	0.1238	0.0731	0.3230	0.0685	0.0270	0.0859	0.0897
green	0.0041	0.0156	0.1126	0.3688	0.0841	0.1086	0.2109	0.0006	0.0103	0.0838

4.3.1 Variables' Cluster Analysis

As we have a relatively large number of variables we can also use cluster analysis to cluster variables instead of speeches. The variation matrix was used as dissimilarity measure for clustering the compositional parts, since it is symmetric and the diagonal elements are zero. Ward's method suggests a partition into four clusters, one composed by the topics *Energy & Oil* and *Tax*, the other composed by the topics *System*, *Coronavirus* and *Economy*, a third one composed by the topics *Jobs/Work* and *External Relations* and the last cluster composed by the topics *Security(War)*, *Family/Children* and *Human Rights*.

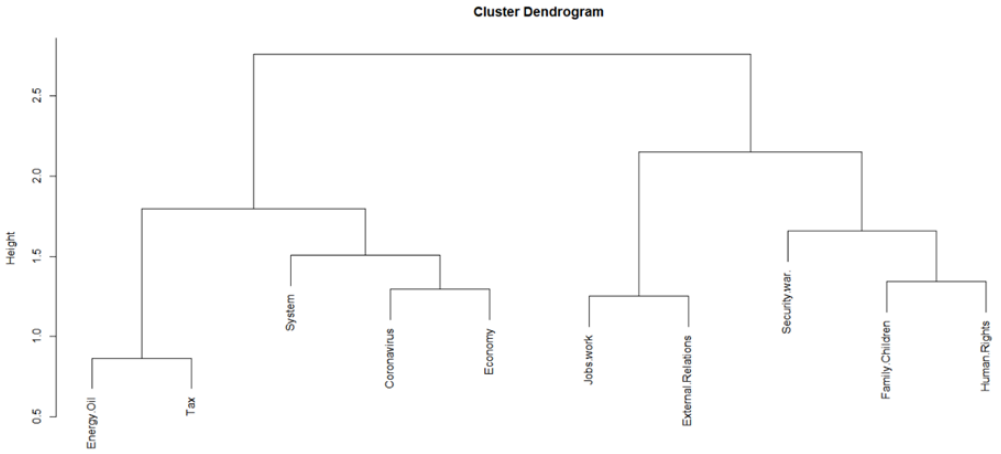


Figure 4.10: Cluster Dendogram - Variables.

4.4 Clustering techniques applied to symbolic data

In this section the solutions obtained with the symbolic approach will be analyzed. We have applied the method developed by Korenjak-Černe et al. (2011). The aggregation method applied using the package *Clamix* was the adapted Ward's method and the distance measure used was the weighted squared Euclidean distance. To define the optimal number of clusters we have used the within-cluster sum of squares that measures the variability of the observations within each cluster. It was determined that the optimal number of clusters is two, where the within-cluster sum of squares has a low value. However, we also decided to detail the solutions with three and ten clusters.

The partition with two clusters obtained from the solution is represented in Figure 4.11. The first group (represented in *red*) is mostly composed by speeches from Barack Obama about the main topics *Human Rights*(26%) - five speeches from BO - and *Security(War)*(74%) - seven speeches from BO and other seven speeches from DT. The second cluster (*Green* cluster) is composed by 64 speeches where 58% are from Barack Obama. The main topics with higher weight are *Economy*(36%) and *External Relations*(11%).

Regarding the solution with three clusters, the partition obtained is represented in Figure 4.12. The first cluster (*Red* cluster) is composed by the same speeches of the *Red* cluster described above in the solution with two clusters. The second cluster (*green*) is composed essentially by speeches whose probability of being related with theme *Economy* is higher. The last cluster is composed essentially by spechees from Barack Obama about the theme *Family/Children* and from Donald Trump about the themes *External relations* and *Jobs/Work*.

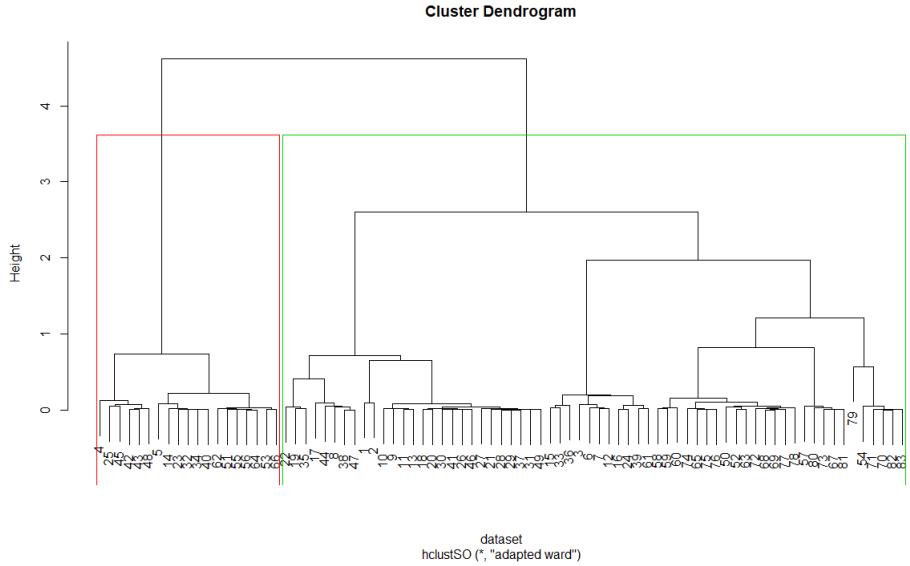


Figure 4.11: Cluster Dendogram - symbolic data analysis aggregation - Adapted Ward - 2 clusters.

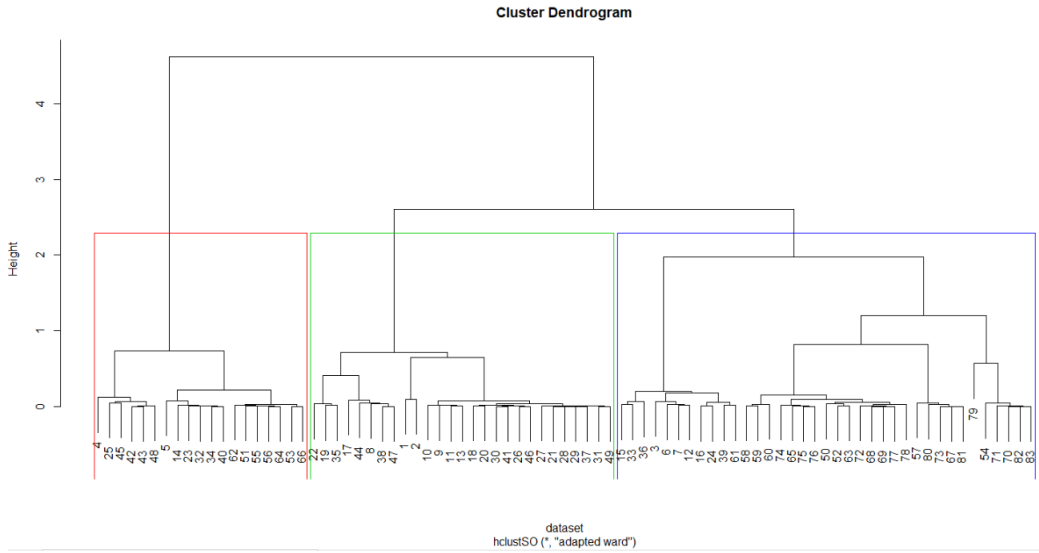


Figure 4.12: Cluster Dendogram - symbolic data analysis aggregation - Adapted Ward - 3 clusters.

The solution with ten clusters is represented below in Figure 4.13.

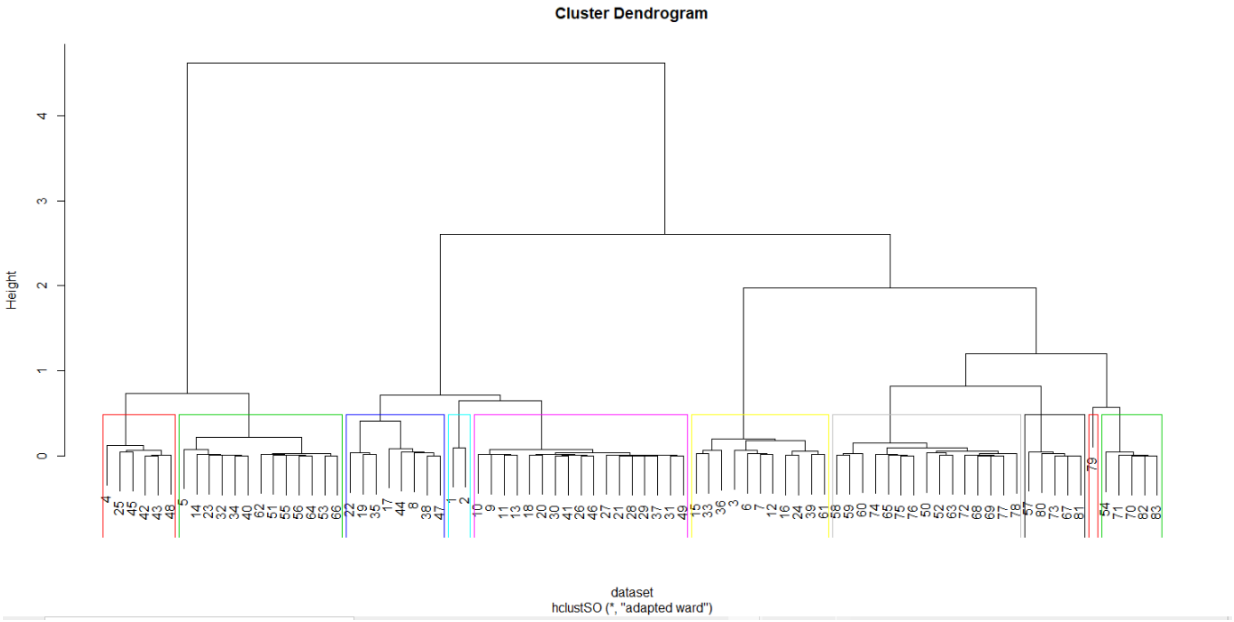


Figure 4.13: Cluster Dendogram - symbolic data analysis aggregation - Adapted Ward - 10 clusters.

The clusters are organized as follows:

Red={'4', '25', '45', '42', '43', '48'}
Green={'5', '14', '23', '32', '34', '40', '62','51','55', '56', '64', '53', '66'}
Purple={'22', '19', '35', '17', '44', '8', '38', '47'}
Blue={'1', '2'}
Pink={'10', '9', '11', '13', '18','20', '30', '41', '26', '46', '27', '21', '28', '29', '37', '31', '49'}
Yellow={'15', '33', '36', '3', '6', '7', '12', '16', '24', '39', '61'}
Grey={'58', '59', '60', '74', '65', '75', '76', '50','52', '63', '72', '68', '69', '77','78'}
Black={'57', '80', '73', '67', '81' }
Red2={'79'}
Green2={'54', '71', '70', '82', '83'}

Note that the first six clusters include mostly speeches from BO and the last four clusters include mostly speeches from DT. The *Red* cluster is composed by speeches whose topics *Human Rights* and *Security (War)* appear simultaneously with a probability higher than 40% and 15%, respectively. The *Green* cluster is composed by speeches whose main theme is *Security (War)*, except the ones included in the previous cluster. The *Black* cluster includes the speeches whose probability of the topic *Jobs/Work* is higher than 45%. The remaing speeches whose main topic is *Jobs/Work* belong to *Grey* cluster. This cluster also includes all the speeches whose main theme is *External Relations*. Using this method the *Green2* cluster is only composed by speeches whose main topic is *Coronavirus*. The topic composition of the remaining clusters is presented in Figure 4.10, where we have computed the average topic probability in each cluster.

Table 4.10: Center of each topic - Symbolic approach -10 clusters.

	Coronavirus	*Energy & Oil*	*Family Children*	*Security (War)*	*Jobs Work*	*Economy*	*External Relations*	*Tax*	*System*	*Human Rights*
red	0.0000	0.0036	0.0396	0.3251	0.0001	0.0869	0.0002	0.0005	0.0087	0.5348
green	0.0079	0.0012	0.0238	0.6787	0.0109	0.0648	0.1262	0.0007	0.0143	0.0709
purple	0.0139	0.1658	0.0623	0.0292	0.0127	0.3365	0.0110	0.0013	0.3289	0.0379
blue	0.0000	0.0006	0.2194	0.0751	0.0146	0.5719	0.0615	0.0018	0.0440	0.0106
pink	0.0017	0.0333	0.0372	0.0895	0.0131	0.7079	0.0092	0.0141	0.0540	0.0395
yellow	0.0037	0.0087	0.5300	0.0596	0.0263	0.1558	0.0093	0.0001	0.0573	0.1488
grey	0.0109	0.0210	0.0825	0.1096	0.1278	0.1658	0.4068	0.0087	0.0306	0.0359
black	0.0220	0.0000	0.0387	0.0524	0.6778	0.0832	0.0786	0.0069	0.0384	0.0015
red2	0.0000	0.0000	0.0003	0.0001	0.0000	0.0007	0.0000	0.9984	0.0000	0.0000
green2	0.6710	0.0150	0.0008	0.0357	0.0417	0.0907	0.0145	0.0515	0.0768	0.0016

4.5 Similarity between clustering solutions

In order to measure the similarity between the solutions obtained with the different hierarchical clustering methods we used the Adjusted Rand Index (ARI). Given a set S of n elements, and two clusterings of these elements, namely X=X_1,X_2,...,X_r and $Y = Y_1, Y_2, ..., Y_m$, the overlap between X and Y can be summarized in a contingency table $[n_{ij}]$ where each n_{ij} denotes the number of objects in common between X_i and Y_j;

Table 4.11: Contingency Table (retrieved from Gates and Ahn (2017)).

XY	Y_1	Y_2	...	Y_m	$sums$
X_1	n_{11}	n_{12}	...	n_{1m}	a_1
X_2	n_{21}	n_{22}	...	n_{2m}	a_2
...	...	...	...	...	...
X_r	n_{n1}	n_{n2}	...	n_{nm}	a_r
$sums$	b_1	b_2	...	b_m	

The Adjusted Rand Index is defined as follows:

$$\frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]}{\binom{n}{2}}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - \frac{[\sum_i \binom{a_i}{2}\binom{b_j}{2}]}{\binom{n}{2}}} \tag{4.2}$$

where n_{ij}, a_i and b_j are values from the contingency table (as represented in Table 4.11).

The summary of the ARI values when comparing the solutions obtained in the previous sections is represented in Table 4.12. Considering the clustering methods applied to compositional data, the value of ARI when comparing with the symbolic solutions is below 0.3, which means that the solutions obtained with the different approaches are quite different.

The most similar solutions are those corresponding to the Average and Complete Linkage, both obtained by the compositional approach, with an ARI value of 0.3885.

Table 4.12: Adjusted Rand Index - application.

ARI	Compositional - average	Compositional - ward	Compositional - complete
Compositional - average	1	0.2345	0.3885
Compositional - ward	0.2345	1	0.1118
Compositional - complete	0.3885	0.1118	1
Symbolic - 2 clusters	0.2814	0.0534	0.2063
Symbolic - 3 clusters	0.1301	0.0555	0.0862
Symbolic - 10 clusters	0.0721	0.0729	0.0536

There are also some qualitative considerations that are relevant when comparing the two approaches. First, it seems that the concept of "main topic" is more taken into account in the symbolic approach. For instance, the speeches '79' and

'80' in the compositional approach are isolated in one cluster and considered in two different clusters in the symbolic approach. Although these speeches have similar distributions (whose main topic has a weight of more than 90%), the two speeches relate to very different topics. This difference is taken into account by the symbolic approach. In the symbolic analysis, speeches with the same topic appear more frequently in the same cluster than in the compositional approach.

4.6 Discriminant Analysis

We also applied Discriminant Analysis to investigate the possible relation between the categorical dependent variable - the author of the speech - and a set of quantitative independent variables - the various topics.

We have ten numerical predictor variables and we aim at predicting one categorical variable that has two levels, consisting in the two authors of the speech, Barack Obama and Donald Trump.

We have obtained the prior probabilities of two groups, 59,04% of the speeches are from BO and 40,96% of the speeches are from DT.

We applied classic linear discriminant analysis adapted to compositional data and the group means obtained for each transformed variable by author are represented in Table 4.13.

Table 4.13: Group means for each transformed variable by author - classic linear discriminant analysis.

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9
BO	-4.6939	-2.8745	0.6610	1.9307	-1.3722	3.7214	-1.8292	-3.7096	-0.5787
DT	-2.4059	-4.1299	-0.6198	2.1690	1.0942	2.3298	4.2853	-1.3979	-0.6413

The transformed variables Z=$\{Z_1; Z_2; Z_3; Z_4; Z_5; Z_6; Z_7; Z_8; Z_9\}$ correspond to the isometric logratio transformation, a transformation from the simplex to the D-1 compositional space.

The linear coeficients of the *Fisher* discriminant function are represented in

the Table 4.14.

Table 4.14: Linear coeficients.

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9
BO	-0.6897	-0.7480	0.3517	0.7927	-1.2479	4.0013	-0.8840	-0.7188	1.0022
DT	-0.1248	-0.5840	-0.2638	0.7438	-0.3485	2.2508	1.0229	-0.1385	0.4193

Using this rule in the training dataset, the 49 speeches from Barack Obama are actually correctly classified, as represented in Table 4.15. This rule also correctly classifies 32 out of 34 speeches from Donald Trump. There are two cases of misclassification, speeches from Donald Trump are predicted as being from Barack Obama.

Table 4.15: Confusion matrix.

Actual group/Predicted Group	BO	DT
BO	49	0
DT	2	32

The apparent error rate is 0.0241. It is called "apparent" due to the fact that the data were not split into training and test set, so that the estimate of the error rate tends to be optimistic.

We also applied classic quadratic discriminant analysis adapted to compositional data and have obtained the group means by author as represented in Table 4.16.

Table 4.16: Group means for each transformed variable by author - classic quadratic discriminant analysis.

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9
BO	-3.5771	-2.1432	0.5196	1.9214	-1.3864	3.7352	-1.6265	-3.2916	-0.1514
DT	-1.8009	-3.4030	-1.0073	1.4834	1.1808	2.0485	2.8706	-0.8986	-0.0742

The apparent error rate is null. This rule correctly classifies all speeches from Donald Trump and from Barack Obama.

Finally, we applied robust linear discriminant analysis and robust quadratic discriminant analysis, however the prediction results aren't more accurate than

any of the classic analysis prediction results. In particular, the apparent error rate is 0.0723 for the solution of robust linear discriminant analysis and 0.0843 for the solution of robust quadratic discriminant analysis.

Chapter 5

Conclusions

5.1 Main Conclusions

Natural Language Processing is one of the main challenges in Text Mining, as one word may have multiple meanings depending on the context on which it is used, and multiple words can have the same meaning. Additionally, today we face an exponential growth of textual data available. Therefore, new techniques of summarization have to be applied such as Latent Dirichlet Allocation, Clustering etc.

There is few literature available where text data are represented as distributions on topics and then considered as symbolic and compositional data.

In this work, after the pre-processing phase we applied Latent Dirichlet Allocation to identify the ten relevant topics *Coronavirus*, *Energy & Oil*, *Family/Children*, *Security (War)*, *Jobs/Work*, *Economy*, *External Relations*, *Tax*, *System* and *Human Rights* and then represent the textual data as distributional data, on those topics.

Although the two methods are hardly comparable since in the Compositional approach data are points in a specific space - Simplex Space - where the elements are ratios which need logaritmic transformations, we can establish some

conclusions or considerations based on our application.

In the case of compositional clusterig analysis various aggregation methods were applied, namely Average, Ward and Complete Linkage. For the symbolic clustering analysis we applied the approach developed by Korenjak-Černe et al. (2011), where the aggregation method is the adapated Ward.

We have defined the concept of main topic as the topic with highest weight on the distribution of the speech. In the symbolic approach the speeches whose main topics are the same tend to be aggregated in the same cluster. We have observed that six out of ten clusters are composed by speeches whose main topic is the same. On the other hand, in the compositional approach we have observed that speeches with the same type of distribution regardless of the main topic tend to appear together in the same cluster.

Applying the discriminant analysis where we tried to find what distinguishes the speeches of BO and DT, and we concluded that speeches with higher probability of occurence of the topics *External Relations* and *Coronavirus* tend to be classified as from DT. The rule created also allowed predicting to which president corresponds the speech. The conclusion obtained in this approach is corroborative with the quaternary diagram in the descriptive statistics section, where it was also concluded that the topics *Family/Children* and *Economy* are more likely to be included in speeches from Barack Obama.

5.2 Final considerations and Future Work

Natural Language Processing can lead to some challenges. These challenges emphasize the importance of pre-processing for the correct identification of topics. For instance, the main topic of speech '80' is *Jobs/Work*. However, after listening to the speech we noted that these terms appear in a different context. In this speech Donald Trump makes some remarks in the day after the US Senate voted to acquit him of the impeachment charges, so the word "Job" or "Work" are uttered in a context of congratulating the work of some specific persons. In order to mitigate that we could have explored a correspondence analysis in which topics and subtopics were defined.

We did not consider the fact that the dataset is not balanced in terms of the authors of the speeches, as there are more speeches from Barack Obama than Donald Trump.

Finally, time series analysis of the themes of the speeches could be made, understanding how the themes/topics of the speeches have changed along the time, extending our database to speeches made by other presidents in other periods.

Bibliography

Aggarwal, C. C. and Zhai, C. (2012). An introduction to text mining. In *Mining Text Data*, pages 1–10. Springer.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, 70(1):57–65.

Aitchison, J. (1985). A general class of distributions on the simplex. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(1):136–146.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. London: Champman & Hall (Reprinted in 2003 with additional material by Blackburn Press).

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609.

Atchison, J. and Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272.

Baker, F. B. and Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70(349):31–38.

Beale, E. (1969). *Euclidean Cluster Analysis*. Scientific Control Systems Limited.

Benzécri, J. (1981). Pratique de l'Analyse des Données,Tome iii. *Linguistique & Lexicologie. Dunod, Paris.*

Benzécri, J.-P. (1992). *Correspondence Analysis Hanbook*. CRC Press LLC.

Benzécri, J.-P. et al. (1973). *L'Analyse des Données, Vol. 2*. Paris: Dunod.

Beranger, B., Lin, H., and Sisson, S. A. (2018). New models for symbolic data analysis. *arXiv preprint arXiv:1809.03659*.

Blasco-Duatis, M. and Coenders, G. (2020). Sentiment analysis of the agenda of the spanish political parties on twitter during the 2018 motion of no confidence. a compositional data approach. *Revista Mediterránea de Comunicacíon/ Mediterranean Journal of Communication*, 11.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Brito, P. (1991). Analyse de données symboliques: Pyramides d'héritage. *Thèse, Mathematiques de la Decision. University Paris-IX Dauphine.*

Brito, P. (1994). Use of pyramids in symbolic data analysis. In *New Approaches in Classification and Data Analysis*, pages 378–386. Springer.

Brito, P. (1998). Symbolic clustering of probabilistic data. In *Advances in Data Science and Classification*, pages 385–390. Springer.

Brito, P. (2014). Symbolic data analysis: another look at the interaction of data mining and statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(4):281–295.

Brito, P. and Diday, E. (1990). Pyramidal representation of symbolic objects. In *Knowledge, Data and Computer-Assisted Decisions*, pages 3–16. Springer.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Carson, V., Tremblay, M. S., Chaput, J.-P., and Chastin, S. F. (2016). Associations between sleep duration, sedentary time, physical activity, and health indicators among canadian children and youth using compositional analyses. *Applied Physiology, Nutrition, and Metabolism*, 41(6):S294–S302.

Celeux, G., Diday, E., and Govaert, G. (1989). *Classification Automatique des Données*. Dunod informatique.

Dahiya, R. and Sachar, D. (2021). Discriminant analysis application to understand the usage of digital channels while buying a car. *South Asian Journal of Marketing*.

Dias, C. S. V. (2015). *Análise de Dados Textuais: Análise de Correspondências e Classificação*. Dissertação de Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão, Faculdade de Economia, Univ. Porto, Portugal.

Diday, E. (1972). Optimisation en classification automatique et reconnaissance des formes. *Revue Française d'Automatique, Informatique, Recherche Opérationnelle. Recherche Opérationnelle*, 6(V3):61–95.

Diday, E. (1988). The symbolic approach in clustering and related methods of data analysis. *Proceedings of IFCS, Classification and Related Methods of Data Analysis*, pages 673–384.

Duda, R. O., Hart, P. E., et al. (1973). *Pattern Classification and Scene Analysis*, volume 3. Wiley New York.

El-Sonbaty, Y. and Ismail, M. A. (1998). Fuzzy clustering for symbolic data. *IEEE Transactions on Fuzzy Systems*, 6(2):195–204.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, number 34, pages 226–231. AAAI Press.

Etemad, K. and Chellappa, R. (1997). Discriminant analysis for recognition of human face images. *Josa a*, 14(8):1724–1733.

Fan, W., Wallace, L., Rich, S., and Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*, 49(9):76–82.

Filzmoser, P., Hron, K., and Templ, M. (2012). Discriminant analysis for compositional data and robust parameter estimation. *Computational Statistics*, 27(4):585–604.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Gates, A. J. and Ahn, Y.-Y. (2017). The impact of random models on clustering similarity. *arXiv preprint arXiv:1701.06508*.

Godichon-Baggioni, A., Maugis-Rabusseau, C., and Rau, A. (2019). Clustering transformed compositional data using k-means, with applications in gene expression and bicycle sharing system data. *Journal of Applied Statistics*, 46(1):47–65.

Gowda, K. C. and Krishna, G. (1978). Disaggregative clustering using the concept of mutual nearest neighborhood. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 8(12):888–895.

Hardy, A. and Lallemand, P. (2004). Clustering of symbolic objects described by multi-valued and modal variables. In *Classification, Clustering, and Data Mining Applications*, pages 325–332. Springer.

Hirschfeld, H. (1935). A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31, pages 520–524. Cambridge University Press.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57.

Hubert, L. J. and Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 83(6):1072.

Hughes, W. L., Kalbfleisch, J. M., Brandt, E. N., and Costiloe, J. P. (1963). Myocardial infarction prognosis by discriminant analysis. *Archives of Internal Medicine*, 111(3):338–345.

Kamalja, K. K. and Khangar, N. V. (2017). Multiple correspondence analysis and its applications. *Electronic Journal of Applied Statistical Analysis*, 10(2):432–462.

Kejžar, N., Korenjak-Černe, S., and Batagelj, V. (2021). Clustering of modal-valued symbolic data. *Advances in Data Analysis and Classification*, 15(2):513–541.

Kim, S.-H., Lee, N., and King, P. E. (2020). Dimensions of religion and spirituality: A longitudinal topic modeling approach. *Journal for the Scientific Study of Religion*, 59(1):62–83.

Kitchens, J. T. and Powell, J. L. (1975). Discriminant analysis as an instrument for political analysis. *Southern Journal of Communication*, 40(3):313–320.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., and Saarela, A. (2000). Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585.

Korenjak-Černe, S. and Batagelj, V. (2002). Symbolic data analysis approach to clustering large datasets. In *Classification, Clustering, and Data Analysis*, pages 319–327. Springer.

Korenjak-Černe, S., Batagelj, V., and Japelj Pavešić, B. (2011). Clustering large data sets described with discrete distributions and its application on timss data set. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(2):199–215.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Oakland, CA, USA.

Martín-Fernández, J., Barceló-Vidal, C., Pawlowsky-Glahn, V., Buccianti, A., Nardi, G., and Potenza, R. (1998). Measures of difference for compositional data and hierarchical clustering methods. In *Proceedings of IAMG*, volume 98, pages 526–531.

Mcauliffe, J. and Blei, D. (2007). Supervised topic models. *Advances in Neural Information Processing Systems*, 20:121–128.

Morin, A. (2004). Intensive use of correspondence analysis for information retrieval. In *26th International Conference on Information Technology Interfaces, 2004*, pages 255–258. IEEE.

Palarea-Albaladejo, J., Martín-Fernández, J. A., and Soto, J. A. (2012). Dealing with distances and transformations for fuzzy c-means clustering of compositional data. *Journal of Classification*, 29(2):144–169.

Petrović, S., Dalbelo Bašić, B., Morin, A., Zupan, B., and Chauchat, J.-H. (2009). Textual features for corpus visualization using correspondence analysis. *Intelligent Data Analysis*, 13(5):795–813.

Ralambondrainy, H. (1995). A conceptual version of the k-means algorithm. *Pattern Recognition Letters*, 16(11):1147–1157.

Reinert, M. (1993). Les «mondes lexicaux» et leur «logique» à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et Société (Maison des Sciences de l'Homme)*, 66:5–39.

Richardson, M. and Kuder, G. (1933). Making a rating scale that measures. *Personnel Journal*, 12:36–40.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Sadika R., Summa M. S., B. S. (2014). Joint analysis of closed and open-ended questions in a survey about the tunisian revolution. In *COMPSTAT 2014 21st International Conference on Computational Statistics*, page 685. Citeseer.

Salton, G. and McGill, M. (1983). Introduction to Modern Information Retrieval. *McGraw Hill Book Company, New York*.

Široki, T., Koska, S., Čorak, N., Futo, M., Domazet-Lošo, T., and Domazet-Lošo, M. (2019). Correspondence analysis applied to large scale evo-devo data. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 262–265. IEEE.

Summa, M. G. and Brito, P. (2018). Processing symbolic modal valued textual data through compositional approaches. In *Symbolic Data Analysis Workshop SDA 2018 Programme and Abstracts*, page 75.

Tahmasebi, P., Hezarkhani, A., and Mortazavi, M. (2010). Application of discriminant analysis for alteration separation; sungun copper deposit, East Azerbaijan, Iran. *Australian Journal of Basic and Applied Sciences*, 6(4):564–576.

Talib, R., Hanif, M. K., Ayesha, S., and Fatima, F. (2016). Text mining: techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, 7(11):414–418.

Tufféry, S. (2005). *Data Mining et Statistique Décisionnelle: l'Intelligence dans les Bases de Données*. Editions Technip.

Verde, R., de Carvalho, F. d. A., and Lechevallier, Y. (2000). A dynamical clustering algorithm for multi-nominal data. In *Data Analysis, Classification, and Related Methods*, pages 387–393. Springer.

Wang, F., Zhang, C., and Li, T. (2007). Regularized clustering for documents. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 95–102.

Wang, W., Yang, J., Muntz, R., et al. (1997). *STING: A Statistical Information Grid Approach to Spatial Data Mining*, volume 97. UCLA Computer Science Department.

Appendix A

Knime overview

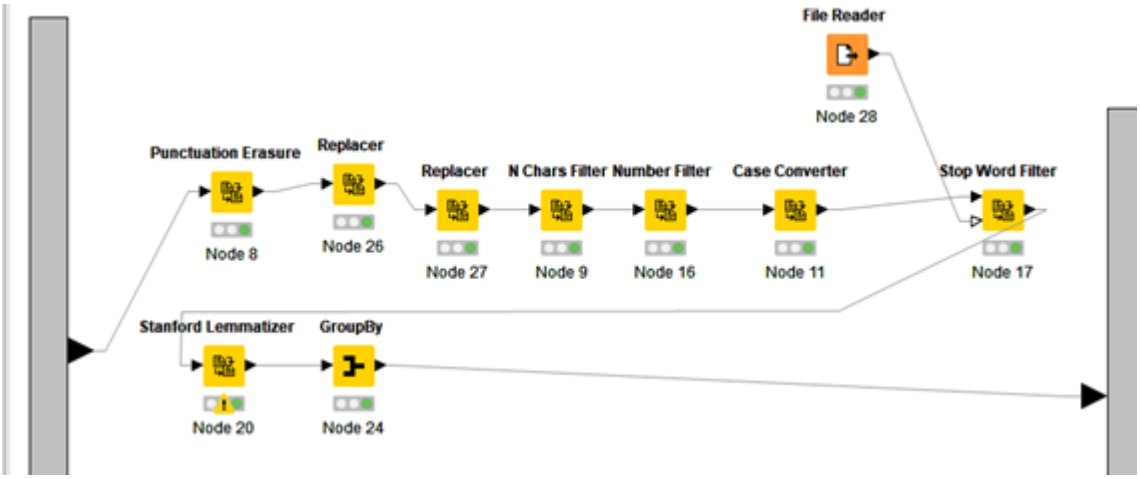


Figure A1: General View KNIME and Pre-processing Nodes.

Appendix B

Details about the speeches

Table B1: Details about the speeches - Author, Date and Main Topic.

Speech number	Author	Date	Main topic
1	BO	28/08/2008	Economy
2	BO	20/11/2014	Economy
3	BO	16/12/2012	Family/Children
4	BO	04/06/2009	Human Rights
5	BO	22/06/2011	Security(war)
6	BO	08/04/2013	Family/Children
7	BO	04/11/2008	Family/Children
8	BO	03/11/2010	System
9	BO	15/03/2010	Economy
10	BO	20/01/2015	Economy
11	BO	09/09/2009	Economy
12	BO	12/01/2011	Family/Children
13	BO	08/09/2011	Economy
14	BO	21/10/2011	Security(war)
15	BO	21/01/2013	Human Rights
16	BO	07/03/2015	Family/Children
17	BO	19/07/2013	System
18	BO	29/01/2009	Economy
19	BO	15/04/2010	Energy&Oil
20	BO	29/01/2013	Economy

Speech number	Author	Date	Main topic
21	BO	24/01/2012	Economy
22	BO	28/04/2010	Economy
23	BO	01/12/2009	Security(war)
24	BO	26/06/2015	Family/Children
25	BO	10/12/2009	Human Rights
26	BO	24/07/2013	Economy
27	BO	07/02/2009	Economy
28	BO	25/01/2011	Economy
29	BO	13/02/2013	Economy
30	BO	06/09/2012	Economy
31	BO	12/01/2016	Economy
32	BO	01/05/2011	Security(war)
33	BO	06/11/2012	Family/Children
34	BO	10/09/2013	Security(war)
35	BO	15/06/2010	Energy&Oil
36	BO	22/03/2016	Human Rights
37	BO	24/02/2009	Economy
38	BO	01/03/2013	System
39	BO	26/05/2009	Family/Children
40	BO	31/08/2010	Security(war)
41	BO	27/01/2010	Economy
42	BO	23/09/2010	Human Rights
43	BO	19/05/2011	Human Rights
44	BO	15/05/2016	Economy
45	BO	21/03/2013	Human Rights
46	BO	04/12/2013	Economy
47	BO	09/02/2010	System
48	BO	25/05/2011	Security(war)
49	BO	28/01/2014	Economy
50	DT	20/01/2017	External Relations
51	DT	19/09/2017	Security(war)
52	DT	28/02/2017	External Relations
53	DT	24/09/2019	Security(war)
54	DT	11/03/2020	Coronavirus

Speech number	Author	Date	Main topic
55	DT	18/12/2017	Security(war)
56	DT	08/01/2020	Security(war)
57	DT	25/09/2019	Jobs/work
58	DT	24/07/2018	Jobs/work
59	DT	01/02/2018	Jobs/work
60	DT	29/06/2017	Economy
61	DT	15/02/2018	Family/Children
62	DT	27/10/2019	Security(war)
63	DT	26/01/2018	Economy
64	DT	03/01/2020	Security(war)
65	DT	24/01/2020	External Relations
66	DT	25/09/2018	Security(war)
67	DT	24/07/2017	Jobs/work
68	DT	30/01/2018	External Relations
69	DT	04/02/2020	External Relations
70	DT	23/04/2020	Coronavirus
71	DT	13/03/2020	Coronavirus
72	DT	19/01/2019	External Relations
73	DT	23/02/2018	Jobs/work
74	DT	01/06/2020	External Relations
75	DT	04/07/2020	External Relations
76	DT	03/06/2020	External Relations
77	DT	05/02/2019	External Relations
78	DT	19/03/2018	External Relations
79	DT	08/08/2020	Tax
80	DT	06/02/2020	Jobs/work
81	DT	20/06/2020	Jobs/work
82	DT	13/04/2020	Coronavirus
83	DT	15/04/2020	Coronavirus

Appendix C

List of stopwords

Table C1: List of stopwords.

America	before	not	his	other	your
American	same	that	want	hes	say
lot	she	this	very	like	really
were	our	to	just	ever	theyre
Americans	have	are	been	never	doing
nothing	and	but	can	today	time
ago	whats	they	them	some	weve
next	youve	its	people	point	many
biden	let	what	must	think	had
happening	applause	who	own	because	him
thought	that	going	her	some	see
Applause	everything	has	vey	there	good
always	country	was	ahead	get	over
actually	nation	with	please	when	much
might	beautiful	did	from	these	way
move	wont	can	know	said	need
obama	thats	make	yeah	would	including
ones	for	new	also	those	laughter
thank	fact	years	last	when	great
called	saying	now	one	than	got
saw	agree	year	dont	each	back
tough	more	all	okay	should	only
you	here	united	about	still	youre
their	the	states	well	every	didnt
under	will	her	how	those	even

could	come	making	talking	forwards	bad
where	sir	issues	little	everybody	around
down	theres	between	theyve	forward	went
few	which	cant	Donald	through	end
then	being	made	Trump	both	some
two	sure	any	first	better	together
things	ive	give	call	president	change
right	something	why	guy	too	question
done	out	look	after	anything	kind
thing	tell	litle	into	additional	take
important	ways	went	big	cuban	most
anybody	part	coming	nobody	may	help
along	came	again	remember	mean	Barack
tremendous	happen	cuba	able	tonight	

Appendix D

Top words by topic

Table D1: List of top words by topic.

Human Rights	*Energy & Oil*	*External Relations*	*Coronavirus*	*Security (War)*
drug	energy	nations	country	security
immigration	space	world	virus	war
border	industry	countries	testing	peace
congress	oil	trade	governors	world
heroes	financial	military	working	iraq
administration	clean	country	cases	rights
drugs	wall	iran	secretary	israel
justice	nasa	security	question	nations
enforcement	program	always	tests	afghanistan
family	crisis	administration	health	human

Economy	*Family/Children*	*System*	*Jobs/Work*	*Tax*
economy	country	grace	country	democrats
jobs	world	gun	job	money
health	nation	cuban	incredible	tax
care	future	cuba	love	relief
work	work	law	beautiful	bill
congress	home	court	money	additional
tax	together	sometimes	bad	jobs
businesses	believe	violence	went	economy
insurance	women	rights	man	actually
plan	tonight	african	actually	vets

Appendix E

Determination of the optimal number of clusters

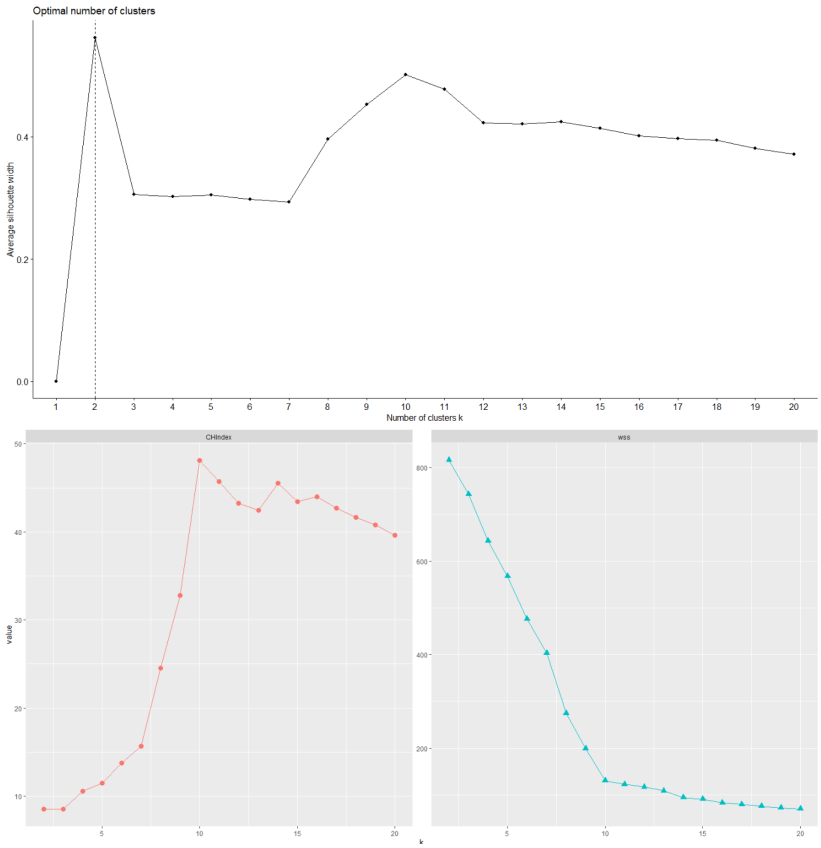


Figure E1: Optimal number of clusters determination according to the Silhouette index (top graph) and to the Calinski Harabaz index (bottom graph) - Average Linkage.

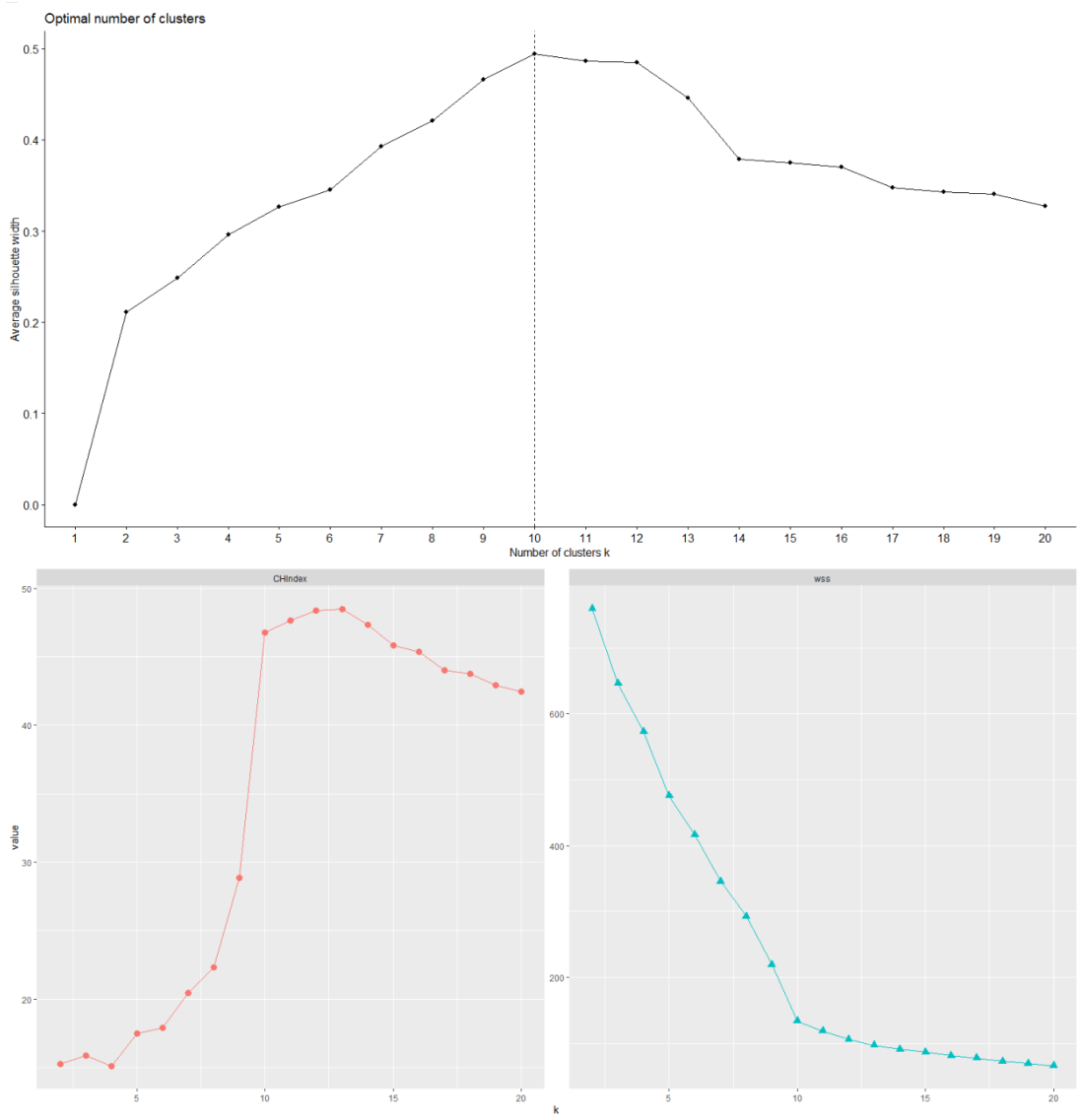


Figure E2: Optimal number of clusters determination according to the Silhouette index (top graph) and to the Calinski Harabaz index (bottom graph) - Ward Linkage.

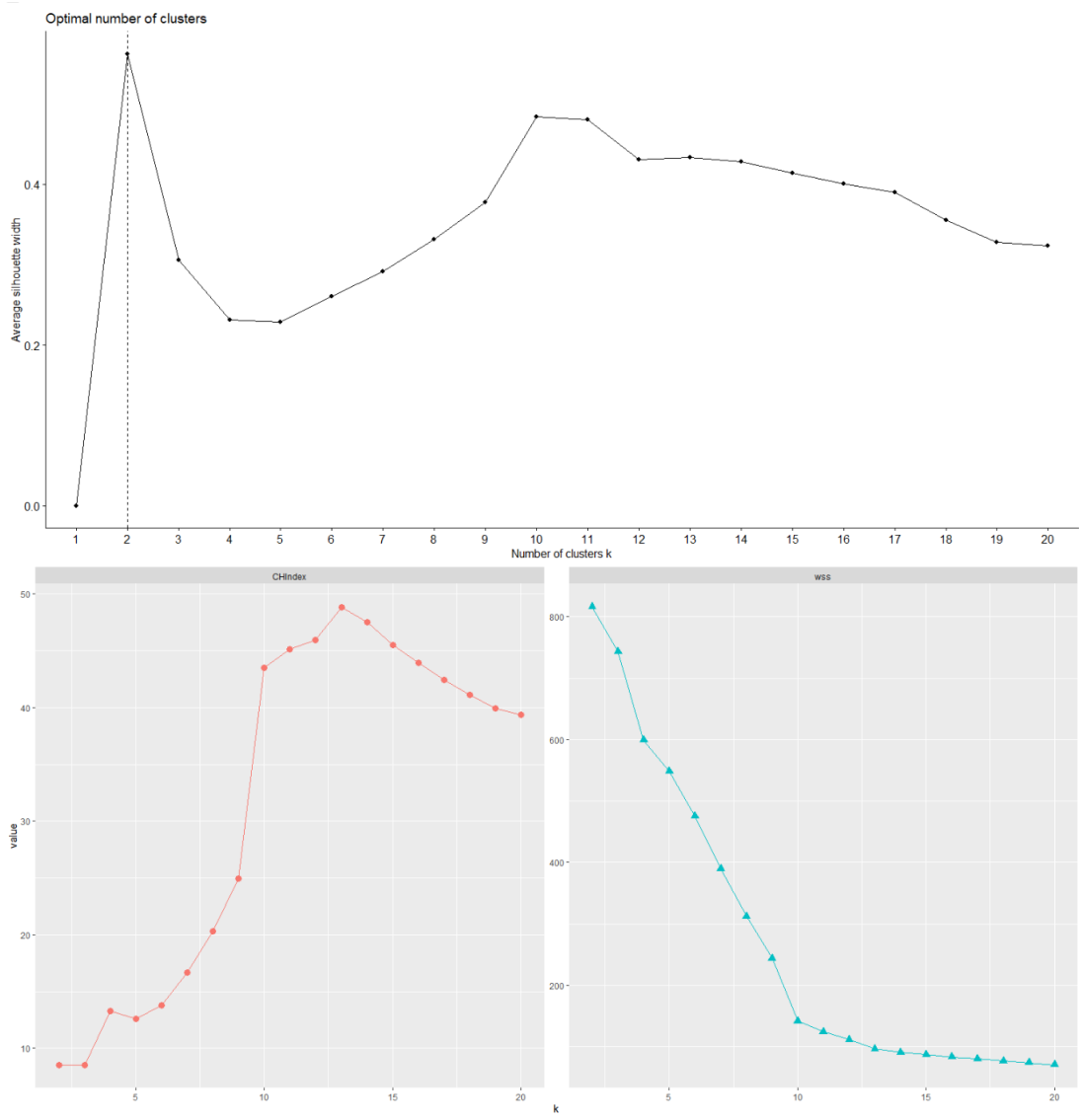


Figure E3: Optimal number of clusters determination according to the Silhouette index (top graph) and to the Calinski Harabaz index (bottom graph) - Complete Linkage.