



Common Legislative Procedure and Political Parties
A Text Mining Approach

by

Humberto José do Vale Ferreira

Master Dissertation in Modeling, Data Analysis and
Decision Support Systems

Supervised by

Professor Pavel Bernard Brazdil

2019

To my unborn son.

Biographical Note

I have always been deeply inspired by mathematics and statistics and I have a genuine interest in political matters. I should also mention that my fascination for lifelong learning comes from the fact that I like to be intellectually stimulated, and this was the ideal time to undertake this project.

Considering that my present job at the Institute of Employment and Vocational Training manages European and National funds for vocational training by public and private entities in the North of Portugal, I deal with a massive amount of data, not only to find possible irregularities, but also to capture relevant information for the correct allocation of financial resources by policy makers.

Finally, from an academic standpoint, since data analytics does not match the initial studies made in the field of Psychology, and given my postgraduate studies in Public Administration, I believe I now have the tools to be a more active citizen, politically speaking.

Acknowledgements

I would like to thank my advisor, Professor Pavel Bernard Brazdil, for all the support on this dissertation.

I would also like to thank all the developers for their spirit of mutual help and sharing of knowledge.

I thank free software, namely the R project for statistical computing.

I would also like to thank Alexandra Guimarães for helping me with the English revision of the text.

Finally, I would like to acknowledge my wife, Sandra Guimarães, and my parents, Humberto Ferreira and Teresa Ferreira.

.

Abstract

Our aim in this work was to analyze the legislative activity of politicians in the Portuguese Parliament and capture and characterize the differences between political parties. For this purpose, we have used and compared the results of six different methods for extracting keywords from plain texts.

The collection of documents selected for this task was the Member's bills produced in the 13th Portuguese Legislature (October 2015 – July 2019). In our comparison of keyword extraction methods, the best algorithm was RAKE. However, we decided also to combine the results of different techniques to capture the ideologies of the political parties represented in the Portuguese Parliament. The keywords showed that there are significant differences in the Member's bills of the different political parties. So, for instance, we have confirmed the importance of the term “*work*” in the legislature of left-wing parties (BE, PCP and PEV), “*municipality*” for centre parties (PS and PSD), “*inland*” for the right-wing party (CDS) and “*animal*” for the PAN party.

In the second part of our work we compared what the politicians delivered in the area of legislature and what they had promised. The applied procedure consisted in analyzing document similarity between the Member's bills and their respective electoral programs presented in the legislative elections held on October 2015. The results of cosine similarity suggested a higher degree of similarity between these texts in left wing parties (BE-0.446; PCP-0.449; PEV-0.526) then other parties (PS-0.351; PAN-0.301; PSD-0.390; CDS-0.337). Regarding textual proximity between parties, there was a degree of relatedness between the Member's bills produced by all political parties and the electoral program of PEV (between 0.436 and 0.540). The electoral program of PAN seems to be the most distant from all Member's bills that were produced (between 0.289 and 0.334).

Keywords: keyword extraction, document similarity, Member's bills, political parties.

Resumo

Neste trabalho, decidimos analisar a atividade legislativa de políticos no Parlamento Português e capturar e caracterizar as diferenças entre partidos políticos. Para esse objetivo utilizámos seis métodos diferentes de extração de palavras-chave e comparámos os resultados.

A coleção de documentos selecionada para esta tarefa foi o conjunto dos projetos de lei elaborados na 13ª Legislatura Portuguesa (de outubro de 2015 a julho de 2019). Na nossa comparação de métodos de extração de palavras-chave, o melhor algoritmo foi o RAKE. No entanto, decidimos combinar os resultados das técnicas aplicadas para melhor capturar a ideologia dos partidos políticos representados no Parlamento Português. As palavras-chave confirmaram a existência de diferenças nos projetos de lei dos partidos políticos revelando, por exemplo, a importância do termo “*trabalho*” nos partidos de esquerda (BE, PCP e PEV), “*freguesia*” nos partidos do centro (PS e PSD), “*interior*” no partido de direita (CDS) e “*animal*” no partido do PAN.

Na segunda parte do trabalho, decidimos comparar o que os políticos fizeram na área legislativa e o que prometeram. O procedimento envolveu a quantificação de similaridade de documentos relativos a projetos de lei e respetivos programas eleitorais apresentados para as eleições legislativas de outubro de 2015. Os resultados da similaridade cosseno sugeriram uma maior proximidade entre estes textos nos partidos de esquerda (BE-0.446; PCP-0.449; PEV-0.526) do que nos restantes (PS-0.351; PAN-0.301; PSD-0.390; CDS-0.337). No que diz respeito à proximidade textual entre partidos, houve um grau de semelhança entre os projetos de lei produzidos por todos os partidos políticos portugueses e o programas eleitoral do PEV (entre 0.436 e 0.540). O programa eleitoral do PAN parece ser o mais distante face aos diferentes projetos de lei apresentados (entre 0.289 e 0.334).

Palavras-chave: extração de palavras-chave, similaridade de documentos; projetos de lei, partidos políticos.

Contents

Biographical Note	v
Acknowledgements.....	vii
Abstract	ix
Resumo.....	xi
Contents	xiii
1. Introduction.....	1
1.1 The Problem to be Studied	2
1.2 Document's Structure	4
PART I – Relation to Other Work.....	5
2 Text Mining	6
2.1 Definition of Concepts.....	6
2.2 The Document Collection and the Document	7
2.3 General Architecture of Text Mining Systems.....	9
2.4 Text Mining Preprocessing Techniques	9
3 Keyword Extraction.....	11
3.1 TextRank	11
3.2 RAKE	12
3.3 SingleRank.....	13
3.4 YAKE!	14
3.5 Parts of Speech Tagging	15
3.6 Collocations and Co-occurrences	15
3.7 Parts of Speech Phrase Sequence Detection	16
3.8 Dependency Parsing	17
3.9 Baseline TF-IDF.....	17
4 Document Similarity.....	19
4.1 Jaccard Similarity	20
4.2 Cosine Similarity	20
4.3 Euclidean Distance	21

PART II – Our Project and Experimental Study	22
5 Objectives, Methodology and Framework	23
5.1 Objectives	23
5.2 Universe	23
5.3 Document Collection	24
5.4 Research Questions and Hypotheses	25
5.5 Methodology and Steps	26
5.6 Software	26
6 Creating a Corpus and Preprocessing	28
6.1 Creation of a Corpus	28
6.2 Preprocessing Procedures Used	28
7 Used Keyword Extraction Techniques	34
7.1 Method based on POS Tagging	34
7.2 Collocations and Co-occurrences	38
7.3 Keyword Extraction with TextRank	45
7.4 Keyword Extraction with RAKE	47
7.5 POS Phrase Sequence Detection	50
7.6 Dependency Parsing	53
7.7 Comparison of Keyword Extraction Methods	55
8 Similarity Between Member’s Bills and Electoral Programs	57
8.1 Applied Methodology	57
8.2 Cosine Similarity Results	59
9 Conclusion	60
Bibliography	64
Appendices	69
A. Automatic Keyword Extraction Results – BE	69
B. Automatic Keyword Extraction Results – CDS-PP	72
C. Automatic Keyword Extraction Results – PAN	74
D. Automatic Keyword Extraction Results – PCP	76
E. Automatic Keyword Extraction Results – PEV	79
F. Automatic Keyword Extraction Results – PS	82
G. Automatic Keyword Extraction Results – PSD	84

List of Figures

Figure 1 - The stages of analysis in natural language processing	8
Figure 2 - Relative frequency of Member's bills authorship in the 13th Legislature .	24
Figure 3 - Document collection and Member's bills by political party	25

List of Tables

Table 1 - Left vs. right differences from a global perspective (adaptation).....	1
Table 2 - Absolute and relative frequency of Member's bills by legislative session .	23
Table 3 - Keyword Extraction Comparative	55
Table 4 - Top 5 Keyword Extraction Techniques.....	56
Table 5 - Document Similarity.....	59

List of Equations

Equation 1 - Term Frequency (TF)	17
Equation 2 - Inverse Document Frequency (IDF)	18
Equation 3 - Term Frequency – Inverse Document Frequency (TF-IDF).....	18
Equation 4 - Jaccard Similarity	20
Equation 5 - Cosine Similarity	21
Equation 6 - Euclidean Distance	21

1. Introduction

Throughout our democratic life, we have witnessed a growing lack of interest by the Portuguese population regarding the activity of political parties, the politicians themselves and the elections. One of the reasons for this alienation is the popular conviction that politicians "are all the same", they "say one thing and do another", and that "deputies only defend the interests of certain economic groups", which results in the common belief that there is little to no difference in the political results, whether they come from the "left" or the "right".

On the other hand, and in the light of the knowledge acquired in the Master of Modeling, Data Analysis and Decision Support Systems, along with an individual interest in political science, we find it useful to apply data analysis techniques to political activity in order to create models for better understanding and decision support when we are called upon to choose between different ideologies (e.g., elections, referendums, party activity).

It is well known that the main concepts of the different ideologies are not at all similar and the values they advocate are manifestly unlike. We would like, at this time, to make a reference to the work developed by McCandless and Posavec (2018), regarding these differences, on the site "*information is beautiful*".

To facilitate understanding, Table 1 reflects some of the differences highlighted in this research.

Table 1 - Left vs. right differences from a global perspective (adaptation)

	LEFT	RIGHT
Principles	• Progressive nature • Looks to the future • Egalitarian • Idealism • Equality	• Conservative nature • Looks to the past • Survival of the fittest • Law of the jungle • Equity
Social Progress	= Evolution	= Status quo
Society and Social Lives	• Interfere with	• Don't interfere with
Trade	• Fair Trade	• Free trade
Support	• Workers	• Employers
Goal	• Personal freedom	• Economic freedom
Focus	• Society	• Individual
Economy	• Regulated	• De-regulated
Business & Industry	• Tax and spend	• Don't tax and spend

	LEFT	RIGHT
Society & Culture	• Community based on ethics; • "The world can be improved" (Utopianism) • Inclusive, Multicultural & Evolving	• Community based on morals; • "The world is fine as it is" (Preservation) • Exclusive, Nationalistic & Conservative
Rights Criminals Society	• Others should observe • Social and economic victims • "One for all and all for one" (protect minorities)	• Others must not interfere • Choose to be criminals • "Survival of the fittest" (immigration control)
Support	• Abortion right (66%) • War (34%) • Tax Cuts (24%)	• Abortion right (43%) • War (91%) • Tax cuts (84%)
Votes for	• Fairness • Positive role models • Champions of downtrodden • Diplomacy • Pacifism	• Upholding order • Strong role models • Champions of opportunity • Aggression • Militancy

It should be noted that, if these differences are present in the right and left supporters, there are values, principles and beliefs that overcome ideological frontiers, not being exclusive to each group. In Portugal, and specifically in the Portuguese Parliament, we do not know whether the activity of the deputies and the different political parties reveal such ideological differences.

Let us elucidate the objective and motivation of this work: we intend, with this dissertation, to contribute to the ideological clarification of a specific Portuguese parliamentary activity - the bills of law (or draft legislation or Member's bills) presented by different political parties in the Assembly of the Portuguese Republic - using textual data analysis tools.

1.1 The Problem to be Studied

In this section we discuss the type of parliamentary activity targeted by our study and the textual data analysis techniques used.

The Assembly of the Republic of Portugal has many duties, among which we highlight the legislative initiatives, interventions in plenary, parliamentary inquiries, motions (e.g., censure, confidence and rejection) and reports (e.g., for monitoring the activity of the Government, the appointment of members of external entities and subject to parliamentary consideration). The competence to initiate legislation lies with the Members of the Assembly of the Republic, the parliamentary groups and the groups of

registered electors – called *Member's bills* - and by the Government or Regional legislative assemblies - in this case, labeled *Government bills* (Assembleia da República 2015, 2019a).

The choice of draft legislation or Member's bills, within parliamentary activity, to establish the possible ideological differences of political parties, appears to be the right choice for our purpose: it is not dependent on its approval in law, nor is it contaminated with eventual "ideological" alterations of other parties or presidential vetoes.

The Member's bills that we focus on are those produced in the current legislative cycle - the 13th Legislature of the Assembly of the Republic of Portugal of the 3rd Portuguese Republic (resulting from the last legislative elections held on October 4, 2015) – which began on October 23, 2015 (date of first plenary session), representing in total, until July 19, 2019 (last legislative session), 1248 documents (of which 36, less than 3%, were drawn by more than one party).

Regarding the specific domain of our thesis, it is our purpose to use text mining tools in the Member's bills of the different Portuguese political parties in order to withdraw information and establish previously unknown and unsuspected relations among the features of these text's databases (Albrecht and Merkl 1998).

To clarify, our research question is – what are the major textual differences in the Member's bills produced in the 13th Legislature of Portugal by the existing political parties?

Related to the main research question, the first hypothesis of this study is the following: textual differences can be identified to a great extent by examining keywords.

As the methodology required to extract textual differences among the above-mentioned documents relies on Automatic Keyword Extraction, we introduce, immediately, our second hypothesis, intrinsically connected to our research question: there are significant differences regarding keywords in the Member's bills produced by the political parties in the 13th Legislature of Portugal.

Proximity between objects can be quantified through a similarity function, being very appropriate in the text domain where the objects can be of different granularities (Aggarwal and Zhai 2012). Once we gathered the information of the authors of the different Member's bills, as well as the different electoral programs presented by the Portuguese political parties in the last legislative elections, we decided to analyze and

measure the relatedness between these two "realities" via document similarity functions. For this additional dimension we formulated our third and last hypothesis: Member's bills produced by the political parties in the 13th Legislature of Portugal are related to the electoral programs presented in the legislative elections held on October 2015.

1.2 Document's Structure

Our dissertation is divided in two major parts: in the first part we describe the "state of art" related to text mining and in the second part we present our case study.

In the first part we cover the definition of text mining, the different domains of knowledge related to this subject and the areas of interest in this field that are closely linked to our empirical work: preprocessing techniques (Chapter 2), keyword extraction methodologies (Chapter 3) and document similarity (Chapter 4). We present the available techniques and some theoretical basis for each area of interest.

Our second part is composed by our empirical work. In Chapter 5 we introduce our objectives, the target document collection, the research question and hypotheses of this study, and also the methodologies and software that we used (in a succinct way), and their chronological sequence. Chapter 6 presents the program code in R for creating document collection and the preprocessing procedures that were used, as well as their graphical facilities. The following chapters deal with the techniques and comparison of automatic keyword extraction used in this work (Chapter 7), as well as the document's similarity methods that were applied between Member's bills and electoral programs and their results (Chapter 8). Finally, we present the conclusion of our dissertation in Chapter 9.

PART I – Relation to Other Work

2 Text Mining

In this chapter we cover the definition of text mining, the different domains of knowledge involved in getting high-quality information from text and the main areas of interest in this field, namely the preprocessing techniques.

2.1 Definition of Concepts

Text mining or extracting information from text can be defined as an intensive knowledge process in which a “user interacts with a collection of documents over time by using a suite of analysis tools” (Feldman and Sanger 2007, p. 1). The goal of this area of research is to try to solve the problem of information overload that we are witnessing today. As for the “suite of analysis tools”, we can mention the “techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR), and knowledge management” (idem, p. x).

To avoid any confusion of concepts, data mining – a territory of work and research belonging to Artificial Intelligence¹ – is “the implicit extraction of previously unknown and potentially useful knowledge from data”² (Cordeiro 2003, p. 15).

Machine learning, also a subarea of artificial intelligence, intends that the created systems have the ability to acquire new knowledge relating to their environment, that has not been previously provided (Cordeiro 2003). For Gama et al. (2017) the process of inducing a hypothesis from a past experience is called automatic learning or data extraction knowledge.

Usually, the Natural Language Processing (also a subarea of artificial intelligence), tends “to view the process of language analysis as being decomposable into a number of stages, mirroring the theoretical linguistic distinctions drawn between syntax, semantics, and pragmatics” (Dale 2010, p. 4), the last one related to the speaker’s intended meaning.

It should be noted that this triple distinction serves only as a starting point; a more refined decomposition of the process is needed in order to deal with real language data (we will address this question more thoroughly in the next section).

¹ Systems that think and act rationally (Russell and Norvig 1995)

² Translated text of the original, in portuguese: “a extração implícita de conhecimento, previamente desconhecido, e potencialmente útil, a partir de dados” (Cordeiro, 2003, p.15)

The meaning of Information Retrieval can be very wide. According to Manning, Raghavan and Schütze, is “finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)” (2009, p. 1). However, it should be noted that this definition is too restrictive, since the extraction of information is not exclusive to materials of unstructured nature, and can also occur in structured data (e.g., relational database) or be used to facilitate a search for semi-structured data (e.g., finding documents with a certain word in the title in a collection of documents).

To finish the question about the domains of text mining, “Knowledge Management is the process of creating, sharing, using and managing the knowledge and information” (Girard e Girard 2015, p. 14; Feldman et al. 1998).

2.2 The Document Collection and the Document

As we mentioned earlier, text mining could be defined as a process of knowledge in which a user deals with a collection of documents over time (Feldman and Sanger 2007).

In its simplest form, the document collection can be any set of text-based documents. However, in general, text mining solutions aim at discovering patterns across large collections of documents. According to Feldman and Sanger (2007), the first distinction we can make regarding document collections is whether they are “static” (i.e., the initial document repository remains unchanged) or “dynamic” (i.e., with the inclusion of new documents and/or change in the initial documents).

At this moment we would like to draw attention to the fact that text mining systems generally do not perform their knowledge discovery algorithms on collections of untreated documents: “considerable emphasis in text mining is devoted to what are commonly referred to as preprocessing operations” (idem, p. 2).

Another relevant element in the extraction of text information is the document *per se*, which “can be very informally defined as a unit of discrete textual data within a collection” (idem p. 3).

Although many characteristics can be used to “represent” documents, we will describe the most common ones (Feldman and Sanger 2007), namely:

a) Characters – the individual constituents (e.g., letters, numbers, special symbols and spaces); the building blocks of higher-level semantic elements described below.

b) Words – precise words selected directly from the original document; “can be described as the basic level of semantic richness” (p. 5).

c) Terms – single words and expressions with multiple words, selected straight from the document through term-extraction procedures, “for which they are meant to be generally representative” (p. 6).

d) Concepts – “features generated for a document by means of manual, statistical, rule-based, or hybrid categorization methodologies” (p. 6); they can be generated by hand, however they are usually extracted from documents through automatic preprocessing procedures.

This division features points of connection with the linguistic theoretical distinction referred above, concerning the natural language processing and its tripartite distinction, by level of granularity, between syntax, semantics and pragmatics analyses (Dale 2010). To deepen this topic, “a finer-grained decomposition” is required in order to “take into account the current state of the art in combination with the need to deal with real language data” (idem, p. 4), as shown in the picture below.

The purpose of Natural Language Processing (NLP) is to study the mechanisms of written and spoken language processing, with the intention of transforming it into a more formal representation. We identified the stage of tokenization and sentence segmentation, in which text preprocessing techniques have an important role, as the first and crucial step in this study. The morphological function tries “to identify isolated words or phrases in a sentence (...) based on delimiters such as punctuation and blanks” and classify them “according to their grammatical category” (Sites.google.com 2019).

Although the “lexical-morphological analyzer deals with the structure of words and their classification into different categories, the parser works with the grouping of words” (idem), examining the construction of sentences. The syntactic analysis (parsing) intends to capture the formal relations between the words of a phrase based on the grammatical rules of a language and the function of the word in the phrase (Dale 2010, Sites.google.com 2019, Hack et al. 2013).

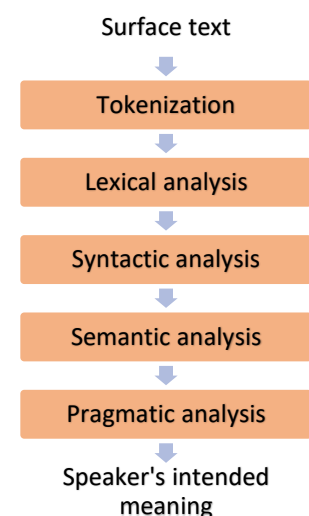


Figure 1 - The stages of analysis in natural language processing

The semantic analysis is linked to “the meaning, not only of each word, but also the expression” (Sites.google.com 2019). This process seeks "the mapping" of the sentences of a language, aiming at the representation of its meaning based on the constructions obtained in the morphological and syntactic analyses (Hack et al. 2013).

Finally, the pragmatic analysis is a process of studying the meaning of certain words of the text, based on the circumstance in which they happen (the meaning of a phrase often depends on its context and positioning in the document; Hack et al. 2013). This last stage “is often seen as being concerned with discourse (...) whereby the meaning of the utterance or text in context is determined” (Dale 2010, p. 4).

2.3 General Architecture of Text Mining Systems

In general, Text Mining Systems follow the model from classic data mining applications, and can be grouped into 4 main areas (Feldman and Sanger 2007):

- a) Preprocessing Tasks – includes all the routines, processes and methods that are necessary to prepare the data in order to apply knowledge discovery operations through the text mining systems;
- b) Core Mining Operations - are the basis of any text mining system and involve the discovery of patterns (e.g., distributions, proportions, associations) trend analysis, and incremental knowledge discovery algorithms.
- c) Presentation Layer Components – comprehend graphical user interface, pattern browsing service and query language approach.
- d) Refinement Techniques – in a simplified way, “include methods that filter redundant information and cluster closely related data (...) aimed at discovery optimization; these techniques have also been described as postprocessing” (p. 14).

Arriving at this point, it is essential to focus on the main techniques of preprocessing texts, because we need them in the present study.

2.4 Text Mining Preprocessing Techniques

The purpose of this phase is to try to recognize similarities, depending on the morphology or meaning of the terms in the textual documents (Hack et al. 2013), as well as contribute to the resolution of “problems relating to high feature dimensionality (...) typically of much greater magnitude in text mining systems than in classic data mining

systems” (Feldman and Sanger 2007, p. 4) with the aim of preparing raw unstructured textual data for text mining. We can even say that the utility of any data mining method relies on the quantity of noise present in the characteristics used for this intent (Aggarwal and Zhai 2012).

In this subchapter we will present the most commonly used preprocessing tasks in text-based documents, in a synthetic way (Manning et al. 2008, Palmer 2010, Hack et al. 2013, Feldman and Sanger 2007, Kilian 2016):

a) Filtering – consists in deleting punctuation marks (generally these characters do not modify the meaning of the document, although can change, in some cases, the meaning of the expressions), removing numbers or converting all terms to lowercase, for instance.

b) Tokenization – is the “task of chopping it up into pieces, called tokens (...) a useful semantic unit for processing” (Manning et al. 2008, p. 22). These are often referred to as terms or words of the phrases.

c) Stopword removal – dropping common terms; “sometimes, some extremely common words which would appear to be of little value (...) are excluded” (idem, p. 27). A token that is not a noun, verb or adjective can be eliminated (Hack et al. 2013).

d) Stemming e lemmatization – after the words are disconnected and organized, they can be "reduced" to their most elementary form - the primitive word; while “stemming usually refers to a crude heuristic process that chops off the ends of words (...) and often includes the removal of derivational affixes, lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma” (Manning et al. 2008, p. 32)

e) Normalization (equivalence classing of terms) – “token normalization is the process of canonicalizing tokens so that matches occur despite superficial differences in the character sequences of the tokens” (idem, p. 29; typically we create implicit equivalence classes); on the other hand, and as distinct terms may have identical meaning, these can also be normalized (e.g., words with analogous meanings are used to normalize the texts in a well-standardized body).

3 Keyword Extraction

As we mentioned earlier, the techniques that we will use to analyze the Member's bills will be the automatic keyword extraction. Keywords, which Rose et al. defined as a “sequence of one or more words, provide a compact representation of a document's content” and “ideally (...) represent in condensed form the essential content of a document” (2010, p. 3).

Nowadays, with the huge explosion of data, manual document processing is a herculean task, if not nearly impossible. As a result, several automated solutions for extracting relevant elements or keywords of documents “have emerged over the last few years, some following a supervised approach (...), others following an unsupervised methodology” (Campos et al. 2018a, p. 684); inside the latter we find well-known techniques, such as *TextRank* (Mihalcea and Tarau 2004), RAKE (Rose et al. 2010) and SingleRank (Wan and Xiao 2008). In the following sections, we will present the techniques described above, as well as a lesser-known Portuguese technique called YAKE! (Campos et al. 2018a, Campos et al. 2018b). We will also demonstrate keyword extraction through Parts of Speech tagging (e.g., finding nouns, adjectives and verbs), Collocations ordering (using Pointwise Mutual Information) and Co-occurrences, by analyzing Parts of Speech phrase sequence detection (noun phrases / verb phrases), and lastly, built on results of dependency parsing (BNOSAC 2019, R-bloggers 2018).

3.1 TextRank

This is a graph-based ranking³ algorithm used successfully in natural language applications (Mihalcea and Tarau 2004). To apply these algorithms “we have to build a graph that represents the text, and interconnects words or other text entities with meaningful relations” (p. 1). According to its authors, this model offers pioneering unsupervised methods for keyword extraction.

This technique consists of the following critical phases (Mihalcea and Tarau 2004):

³ In a summarized manner, graph-based ranking algorithms “are essentially a way of deciding the importance of a vertex within a graph, based on global information recursively drawn from the entire graph” (Mihalcea and Tarau 2004, p. 1)

- “1. Identify text units that best define the task at hand, and add them as vertices in the graph.*
- 2. Identify relations that connect such text units, and use these relations to draw edges between vertices in the graph. Edges can be directed or undirected, weighted or unweighted.*
- 3. Iterate the graph-based ranking algorithm until convergence.*
- 4. Sort vertices based on their final score. Use the values attached to each vertex for ranking/selection decisions” (p. 2)*

What TextRank algorithm does is overwhelming simple: it discovers how similar each sentence is relative to every other sentence in the text; the most important being the one that is most similar to every other sentence. Using the word similarity function, we can also use this algorithm to extract keywords from the text: the word most similar to all others is the most important one (Argerich 2016). Evidently, filtering stopwords is a noteworthy procedure in this situation.

The TextRank is a word network ordered by Google Pagerank – “this is done by constructing a word network by looking if words are following one another; on top of that network the ‘Google Pagerank’ algorithm is applied to extract relevant words after which relevant words which are following one another are combined to get keywords” (R-bloggers 2018).

3.2 RAKE

The RAKE, an acronym for Rapid Automatic Keyword Extraction is built on the finding that keywords are often composed of several words, but rarely include punctuation or stopwords (e.g., "e", "o", "de"), or other words with negligible, poor, vocabulary sense. Words bearing meaning in a document are described as having substance and are usually designated as “content words” (Rose et al. 2010).

The input parameters in this algorithm include a list of stopwords, a set of phrase delimiters (e.g., “.”, “!”) and a set of word delimiters (e.g., spaces). The RAKE “uses stopwords and phrase delimiters to partition the document text into candidate keywords, which are sequences of content words as they occur in the text”; the “co-occurrence of words within these candidate keywords are meaningful”, allowing an “adaptive” and

detailed measurement “that will be used to score candidate keywords” (Rose et al. 2010, p. 6).

Thus, this algorithm searches for keywords by focusing on a contiguous sequence of words that do not contain irrelevant vocabulary, namely by:

- “1. calculating a score for each word which is part of any candidate keyword, this is done by*
 - among the words of the candidate keywords, the algorithm looks how many times each word is occurring and how many times it co-occurs with other words;*
 - each word gets a score which is the ratio of the word degree (how many times it co-occurs with other words) to the word frequency.*
- 2. a RAKE score for the full candidate keyword is calculated by summing up the scores of each of the words which define the candidate keyword”* (R-bloggers 2018, BNOSAC 2019).

3.3 SingleRank

The existing approaches for extracting keywords from a single document typically use only the information confined in that document. The SingleRank method proposes the use of a small number of nearest neighbor documents (and their information) to improve the extraction of keywords from individual documents (Wan and Xiao 2008).

Neighbor documents are those which are "spatially" close to a specific document and construct the neighborhood data setting for that particular document. We present below the outline of this algorithm:

- “1. Neighborhood Construction: Expand the specified document d_0 to a small document set $D=\{d_0, d_1, d_2, \dots, d_k\}$ by adding k neighbor documents (...);*
- 2. Keyphrase Extraction: Given document d_0 and the expanded document set D , perform the following steps to extract keyphrases for d_0 :*
 - a) Neighborhood-level Word Evaluation: Build a global affinity graph G based on all candidate words restricted by syntactic filters in all the documents of the expanded document set D , and employ the graph-based ranking algorithm to compute the global saliency score for each word.*

- b) *Document-level Keyphrase Extraction*: For the specified document d_0 , evaluate the candidate phrases in the document based on the scores of the words contained in the phrases, and finally choose a few phrases with highest scores as the keyphrases of the document” (idem, p. 856).

3.4 YAKE!

Campos et al. (2018a, 2018b) proposed an unsupervised method of extracting keywords (terms of a single word or multiple words) of a single document, without the need to rely on document sets, external sources or language tools (i.e., a “collection-independent automatic keyword extractor”, p. 806). This method is called YAKE!, an acronym for Yet Another Keyword Extractor, and has six main components:

1. *Text pre-processing* – “a pre-processing step which splits the text into individual terms whenever an empty space or a special character (e.g., line breaks, brackets, comma, period, etc.) delimiter is found” (p. 807).
2. *Feature extraction* – “a set of five features to capture the characteristics of each individual term” (although they can be applied to any language, they are particularly suited to western languages), namely: *Casing* (“reflects the casing aspect of a word”); *Word Positional* (“values more those words occurring at the beginning of a document”); *Word Frequency* (“scoring more those words that occur more often”); *Word Relatedness to Context* (“computes (...) the number of different terms that co-occur with the candidate word”); *Word DifSentence* (“quantifies how often a candidate word appears within different sentences”) (p. 807).
3. *Individual terms score* – “heuristically combine all these features into a single measure such that each term is assigned a score” (p. 807); the minor the value, the more significant the word would be.
4. *Candidate keywords list generation* – “the fact that a keyword may consist of more than one word, forces us to consider a further step where the final score of a keyword (be it one, two or n-terms) is determined” (Campos et al. 2018b, p. 687); additionally, it is important to mention that no condition is defined in relation to the minimum word or sentence frequency that a candidate keyword should have.

5. *Data Deduplication* – “eliminate similar candidates”, coming from the previous steps, through the use of “Levenshtein distance, which measures the similarity between two strings” (p. 688).
6. *Ranking* – “finally, the method will output a list of potential relevant keywords, formed by 1, 2 or 3-g, such that the lower the (...) score, the more important the keyword will be” (p.688).

3.5 Parts of Speech Tagging

According to Manning and Schütze (1999), the final goal of research on Natural Language Processing (NLP) is to deconstruct and comprehend language. Since that goal, in the beginning of the century was far from reaching, NLP has engaged on “tasks that make sense of some of the structure inherent in language without requiring complete understanding” (idem, 341): one such task is part-of-speech tagging (POS Tagging).

Tagging is the task of classifying (or tagging) each individual word in a sentence according to its appropriate part of speech (i.e., adjective, verb, noun and so on). In the majority of dialects, nouns are the most common categories of words, next to the verbs, and these are the most pertinent for analytical purposes; adjectives and proper nouns immediately follow (BNOSAC 2019, R-bloggers 2018). For an exhaustive list of all parts of speech tags we can look at <http://universaldependencies.org/u/pos/index.html>.

Parts of speech tags are a way to select words for frequency analysis; and for this task we do not really need all types of words (e.g., stopwords) – if we extract only nouns, verbs or adjectives, we will have the most relevant “tags”.

Finally, in this kind of frequency statistics we can be “trapped” in words which only seem right in combination with other words (i.e., we want to find keywords that are a combination of words, like in the previous procedures).

3.6 Collocations and Co-occurrences

It is also possible to get multi-word expression by observing either collocations (words following one another) or co-occurrences (within each sentence or which are close in a specific neighborhood). According to BNOSAC (2019) and R-bloggers (2018) if we associate these approaches with mining only the nouns and adjectives, the result is very satisfactory.

To identify keywords in a text we can resort to collocation ordering using pointwise mutual information. According to Manning and Schütze “mutual information between X and Y (...) is the reduction in uncertainty of one random variable due to knowing about another, or in other words, the amount of information one random variable contains about another” (1999, p. 68).

With co-occurrences we can see how words are used either in the same sentence or next to each other. A common procedure is observing how many times nouns and adjectives are used in the same sentence. On the other hand, we can try to find nouns and adjectives which follow one another, and this can be done by calculating word co-occurrences of specific parts of speech types that follow one another or we can postulate how distant away to look (BNOSAC 2019).

Keyword correlations show how words are combined in the same document/sentence; while co-occurrences pay attention to frequency, correlation measurements between two words can also be high even if two words occur only a minor number of times but always emerge together (idem).

3.7 Parts of Speech Phrase Sequence Detection

Other possibility is to extract phrases, and these are defined as an arrangement of parts of speech tags. Languages have constraints on word sequence (i.e., words do not occur in just any order, “instead, words are organized into phrases, groupings of words that are clumped as a unit” (Manning and Schütze, 1999, p. 93)). The central idea is that certain groupings of words behave as constituents of the phrase structure, and syntax “is the study of the regularities and constraints of word order and phrase structure” (idem).

The most common types of phrases are noun phrases and verb phrases, but there are also prepositional phrases and adjective phrases (Manning and Schütze 1999). According to BNOSAC “parts of speech tags are recoded” (e.g., A - adjective, C - coordinating conjunction, D - determiner, M - modifier of verb, N - noun or proper noun, P - preposition) and then we “can define a regular expression to indicate a sequence of parts of speech tags (...) to extract from the text” (2019).

Another and intuitive way to represent phrase structure is as a tree or, finally, via a labeled bracketing to show the referred constituency (Manning and Schütze 1999).

3.8 Dependency Parsing

According to NLP-progress, “dependency parsing is the task of extracting a dependency parse of a sentence that represents its grammatical structure and defines the relationships between “head” words and words, which modify those heads” (2019).

In dependency-based methods, the head-dependent connection is made explicit by directly connecting heads to the words that are proximately dependent on them, avoiding the necessity for constituent structures. Furthermore, dependency grammars allow us to classify grammatical relations (e.g., clausal argument relations, nominal modifier relations) and get familiar with notions such as subject, direct object and indirect object (Jurafsky and Martin 2018). Thirty-seven universal syntactic relations can be observed at <http://universaldependencies.org/u/dep/index.html>.

In this option, for instance, we can be looking for the words which have “nominal subject” as dependency relation and add to that the adjective which is changing the nominal subject: “in this way we can combine what are people talking about with the adjective they use when they talk about the subject” (R-bloggers 2018).

3.9 Baseline TF-IDF

We began this chapter with the assertion that there are three eminent unsupervised approaches (RAKE, TextRank and SingleRank) for keyword extraction; however, there is also a baseline called “Term Frequency – Inverse Document Frequency” (TF-IDF), according to Campos et al. (2018a), which demands the presence of additional documents and also the use of a Parts of Speech Tagger.

According to Manning et al. (2008) we can “assign to each term in a document a weight for that term that depends on the number of occurrences of the term in the document” (...) denoted $tf_{t,d}$ ” (p. 107):

$$tf_{t,d} = n^o \text{ occurrences of term } t \text{ in a document } d$$

Equation 1 - Term Frequency (TF)

Raw term frequency as defined above suffers from a critical problem: “all terms are considered equally important when it comes to assessing relevancy on a query” (p. 108) and “instead, it is more commonplace to use for this purpose the document frequency df_t , defined to be the number of documents in the collection that contain a term t .” (p. 108).

Being N the total number of documents in a collection, we can define the inverse document frequency (idf) of a term t as follows:

$$idf_t = \log \frac{N}{df_t}$$

Equation 2 - Inverse Document Frequency (IDF)

We can associate the definitions of term frequency and inverse document frequency “to produce a composite weight for each term in each document: the tf-idf weighting scheme assigns to term t a weight in document d given by:

$$tf - idf_{t,d} = tf_{t,d} \times idf_t$$

Equation 3 - Term Frequency – Inverse Document Frequency (TF-IDF)

“In other words, $tf-idf_{t,d}$ assigns to term t a weight in document d that is

- 1. highest when t occurs many times within a small number of documents (thus lending high discriminating power to those documents);*
- 2. lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);*
- 3. lowest when the term occurs in virtually all documents.” (p. 109)*

Essentially, and according to Ventura, “a word is more important in a certain document the more it occurs in that document, but if that word occurs in other documents, its importance decreases” (2014, p. 18). Thus, we can say that TF-IDF can be used to extract important keywords from a document to get a sense of what typifies a document, to extract words that are exclusive to a particular document, to be used as a very unpretentious summary of a document (Ganesan 2019).

4 Document Similarity

According to Aggarwal and Zhai, proximity between objects can be “measured with the use of a similarity function”, being very suitable to the text domain “where the objects (...) can be of different granularities such as documents, paragraphs, sentences or terms.” (2012, p. 78). As in keyword extraction the value of any data mining method depends on the amount of noise present in the characteristics used for this purpose (idem). One of the major problems associated with document’s similarity is the high dimensionality associated with the huge numbers of terms in the documents (Alelyani et al. 2014, Aggarwal and Zhai 2012).

Likewise in text clustering, we can use several methods of feature selection to prepare the analysis of document-document similarity, namely:

- a) Document Frequency-based Selection - to allow an effective text similarity, the occurrences of words need to be normalized in terms of their relative frequency in a specific document and throughout the document collection using the representation TF-IDF based on vector space;
- b) Term Strength - is fundamentally used to measure how revealing a word is for detecting two associated documents. The central aspect of this technique relies on how we decide to evaluate this connection: whether manually or automatically, with Wilbur and Sirotkin (1992) demonstrating that it was possible to identify stopwords in a collection (in their analysis) through automated statistical testing based on the “cosine coefficient of document-document similarity” (1992, p. 45);
- c) Entropy-based Ranking - the quality of the term is valued by the entropy reduction when it is removed (Dash e Liu 1997 cit in Aggarwal and Zhai 2012). It should be noted that this technique becomes unrealistic for large collections of documents containing many terms, being more efficient when used with sampling procedures (Aggarwal and Zhai 2012);
- d) Term Contribution – the contribution of a term can be defined “as its contribution to document similarity” (Aggarwal and Zhai 2012, p. 83). In the dot-product based similarity, for instance, the “likeness” between two documents can be understood as the dot product of their normalized frequencies, and the contribution of the term (to the similarity of two

documents) as the product of their normalized frequencies in the two documents. This method can be impractical for large document collections.

If in feature selection we try to explicitly choose objects from the original document, in feature transformation the new components are defined as a functional representation of the original elements (Aggarwal and Zhai 2012, Sieg 2018), namely the method:

- e) Latent Semantic Indexing (LSI) – a dimensionality reduction technique of document vectors, in which documents are transformed into a new space, whose components are a linear combination of the elements of the original data. This methodology is closely related to Principal Component Analysis (PCA) and only “keeps the directions in our vector space that contain the most variance” (Sieg 2018).

Related to feature selection and transformation, the next sections present three different techniques to measure document similarity (Selivanov 2017, Sieg 2018).

4.1 Jaccard Similarity

Jaccard similarity is a ration of similarity between two sets, defined as the size of the intersection divided by the size of the union (i.e., $|A \cap B|/|A \cup B|$). For documents “we measure it as proportion of number of common words to number of unique words in both documents” (Selivanov 2017); the value “0” means the documents are totally unlike (i.e., no overlap), “1” means that they are undistinguishable (i.e., perfect overlap), and values between 0 and 1 represent a degree of similarity, or Jaccard coefficient (Manning and Schütze 1999, Manning et al. 2008).

$$J(doc_1, doc_2) = \frac{doc_1 \cap doc_2}{doc_1 \cup doc_2}$$

Equation 4 - Jaccard Similarity

We can also define Jaccard distance or Jaccard dissimilarity as $1 - \text{similarity}$.

4.2 Cosine Similarity

To compensate for the effect of document dimension, the way to quantify the similarity between two documents (doc_1 and doc_2) is to compute the cosine similarity of their vector representations (Manning et al. 2008); for this we have to represent documents

as bag-of-words, each document being a sparse vector, “and define measure of overlap as angle between vectors” (Selivanov 2017):

$$\text{similarity}(doc_1, doc_2) = \cos(\theta) = \frac{doc_1 * doc_2}{|doc_1| * |doc_2|} = \frac{\vec{V}(doc_1) * \vec{V}(doc_2)}{|\vec{V}(doc_1)| * |\vec{V}(doc_2)|}$$

Equation 5 - Cosine Similarity

Where the numerator is the dot product of the vectors $\vec{V}(doc_1)$ and $\vec{V}(doc_2)$, and the denominator represents the product of their Euclidean lengths (Manning et al. 2008).

The cosine similarity measures the cosine of the angle between two vectors. It ranges from “1” for vectors pointing in the same direction to “- 1” in opposite directions (Manning and Schütze 1999).

It can be useful to measure similarity with a bag-of-words matrix transformed with Term Frequency – Inverse Document Frequency and applying Latent Semantic Indexing in order to reduce noise and attain better quality similarities, as we said before (LSI “applies principal component analysis on our vector space and only keeps the directions in our vector space that contain the most variance”, Sieg 2018).

4.3 Euclidean Distance

The Euclidean distance between two vectors measures how far apart they are in the vector space:

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^M (x_i - y_i)^2}$$

Equation 6 - Euclidean Distance

In other words, Euclidean distance is the square root of the sum of squared differences between corresponding elements of the two vectors. Score means the distance between two objects, and if it is 0, it means that both objects are identical.

According to Manning and Schütze, “Euclidean space is not a well-motivated choice if the vectors we are dealing with are vectors of probabilities or counts” which is what document similarity is based on. To the same authors, “the cosine is the only vector space measure that accommodates quantitative information” (p. 303, 1999).

PART II – Our Project and Experimental Study

5 Objectives, Methodology and Framework

In this chapter we describe the objectives, the document collection and the methodology used in this empirical work, in a briefly manner (the subsequent chapters present them separately and extensively). We also introduce the chronological sequence of the applied procedures, as well as the software that was used.

5.1 Objectives

We intend, with this dissertation, to contribute to the ideological clarification of a specific Portuguese parliamentary activity - the Member's bills presented by different political parties in the Assembly of the Portuguese Republic -, using textual data analysis tools, namely keyword extraction and document similarity.

In other words, our goal is to demonstrate the variances between the Portuguese political parties by observing a specific legislative procedure, the draft legislation produced by each party, and then compare them to the electoral programs.

5.2 Universe

Specifically, the universe of documents that we studied was the Member's bills presented by the different Portuguese political parties produced in an actual legislative cycle (i.e., the 13th Legislature of the Assembly of the Republic of Portugal of the 3rd Portuguese Republic) ranging from October 23, 2015 (date of the first plenary session⁴) to July 19, 2019 (last plenary legislative session), with a total of 1248 documents.

A legislature comprises four legislative sessions, each lasting one year - normally from September 15 to Jun 15⁵; extraordinarily, the 4th legislative session of the 13th Legislature ended on July 19, 2019, with the last plenary session of the Assembly of the Republic of Portugal (Assembleia da República 2019b).

Table 2 - Absolute and relative frequency of Member's bills by legislative session

Session	Frequency	Percentage
1	297	23,80%
2	308	24,68%
3	387	31,01%
4	256	20,51%
Total	1248	100,00%

⁴ Resulting from the last legislative elections held on October 2015.

⁵ "Without prejudice to suspensions which the Assembly of the Republic decides by a two-thirds majority of the Members present" - Article 49, Rules of Procedure of the Assembly of the Republic (Assembleia da República 2015, p. 49).

Of the 1248 Member's bills produced in the 13th Legislature, we present their distribution by legislative session (Table 2), as well as the relative frequency of the political parties (or groups of registered electors or the non-attached member of the Socialist Party, since 6 December 2018) responsible for their elaboration (Figure 3).

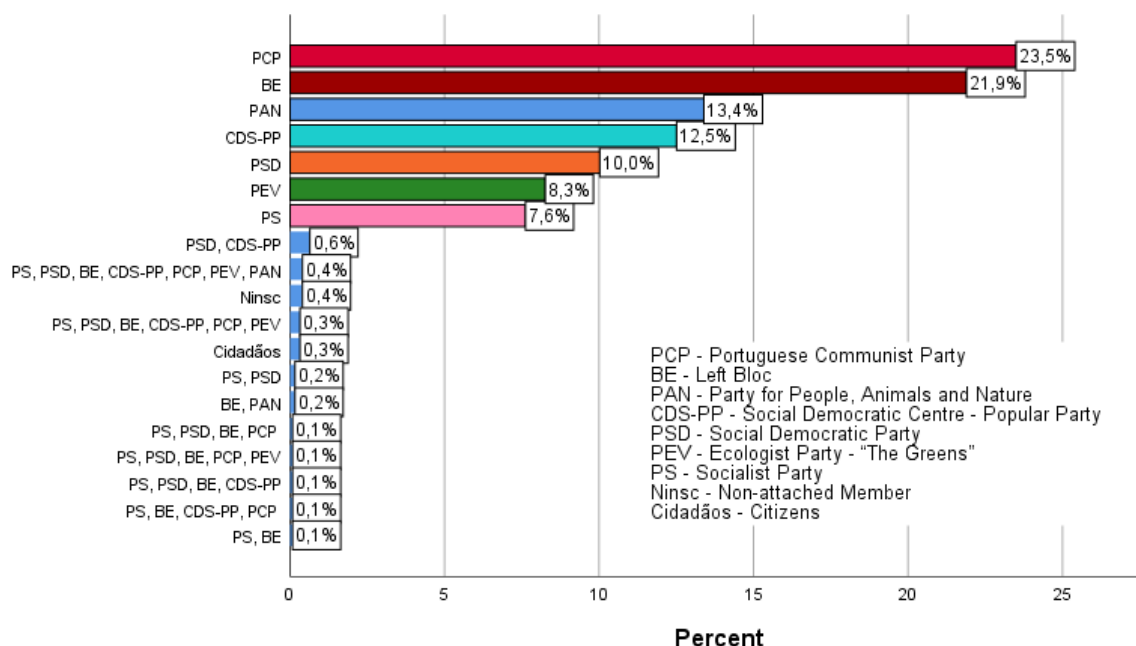


Figure 2 - Relative frequency of Member's bills authorship in the 13th Legislature

As we perceive in figure 2, there are residual Member's bills made by more than one political party, signed by citizens ("Cidadãos") and by the non-attached member ("Ninsc") that we don't consider for this dissertation⁶.

5.3 Document Collection

From a total of 1248 Member's bills produced in the 13th Legislature, we focused our attention on 1212 documents. We present in Figure 3 the absolute distribution, by political party, of the Member's bills, which constitutes the document collection of this study.

⁶ If we recognize that more than one political party contributed to the elaboration of a Member's bill, all parties are equally and ideologically responsible for the contents of the same. On the other hand, if we intend to automatically extract the ideological differences in the draft legislation, we have to restrict the document collection to the Member's bills of a single political party, by removing, also, the legislation signed by different groups, at least 20,000 registered electors, under the terms of Law n.º 17/2003, of 4 June (with disparate ideologies) and by the the non-attached member, since 12 December 2018, which "slammed the door of the Socialist Parliamentary Group" (Observer 2018).

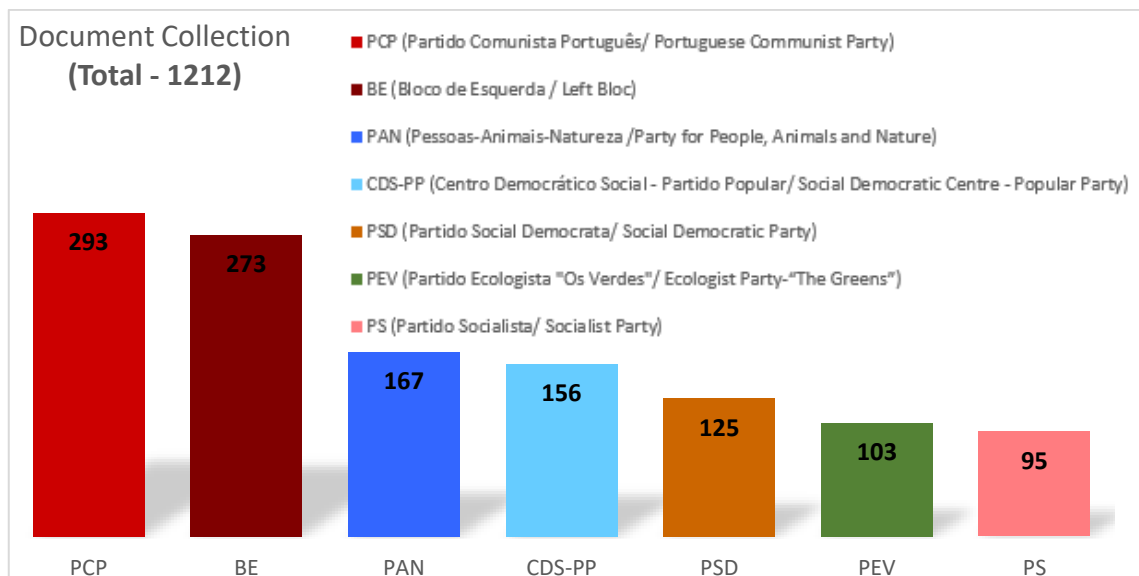


Figure 3 - Document collection and Member's bills by political party

Interestingly, in light of the number of deputies that make up the present legislature, the "production" of Member's bills by deputy is very disparate with respect to the different political parties: it varies between 1.3 Member's bills by deputy in the Socialist Party (PS) and 174 bills produced by deputy in the Party for People, Animals and Nature (PAN). The Portuguese Parliament, known as the Assembly of the Republic, has 230 deputies, in total.

5.4 Research Questions and Hypotheses

Let us recall our research question and the three formulated hypotheses of our dissertation:

Research Question – What are the major textual differences in the Member's bills produced by the political parties in the 13th Legislature of Portugal?

Hypothesis 1 – Textual differences can be identified to a great extent by examining keywords.

Hypothesis 2 – There are significant differences regarding keywords in the Member's bills produced by the political parties in the 13th Legislature of Portugal.

Hypothesis 3 – The Member's bills produced by the political parties in the 13th Legislature of Portugal are related to the electoral programs presented in the legislative elections held on October 2015.

5.5 Methodology and Steps

In this dissertation we use several texts mining methodologies in the document collection (i.e., the Member's bills), including preprocessing techniques and keyword extraction, and the measurement of document similarity between the Member's bills and the electoral programs of each political party.

Chapter 6 describes the corpus and the preprocessing techniques that were used. Chapter 7 discusses various keyword extraction techniques and shows the output obtained with each one. Chapter 8 discusses the similarity between Member's bills and electoral programs, shedding some light on Hypothesis 3.

All methodologies were applied to the Member's bills and electoral programs of each Portuguese political party. For convenience, we use the initials BE for Bloco de Esquerda, a left wing Portuguese political party.

As we do not have a golden standard, namely a set of correct keywords selected manually, it is difficult to estimate the effectiveness of the different keyword extraction techniques. For this reason, we relied on our own judgement and marked every keyword that seemed useful using the color **green**, while the ones that seemed useless are highlighted in **red**. The keywords that we considered neutral are in **black**.

5.6 Software

For the treatment of data, the use of automatic keyword extraction techniques and the employment of document similarity functions (the methodologies used in this work), we turned mainly to R Project - a free software environment for statistical computing and graphic - and occasionally to Microsoft Excel and IBM SPSS.

Amongst the R packages we used:

- *tm*, *tm.plugin.mail* - for corpus creation and preprocessing (e.g., DTM);
- *textstem* – for preprocessing (e.g., lemmatization of corpus);
- *readxl* – for the creation of a dictionary of words and lemmas from Excel
- *udpipe* – simple access in order to tokenize, tag, lemmatize or perform dependency parsing on raw text; with this package it was possible to prepare the raw text and/or perform all keyword extraction techniques used in this project (Chapter 7).
- *textrank* – for keyword extraction with TextRank

- *text2vec* – for document similarity
- *igraph* – for network analysis
- *wordcloud*, *wordcloud2* – for word cloud plots
- *ggraph*, *ggplot2* – for graph plots and data visualization
- *lattice* – for bivariate plots (e.g., bar charts).

In the next chapter we discuss the preprocessing procedures used in the Member's bills collection.

6 Creating a Corpus and Preprocessing

Our collection of documents was taken from a public site (Assembleia da República 2019c) and the corpus for each political party was created from there. The methodology used involved first some pre-processing operations, including, for instance, tokenization, filtering and stopword removal. After that, we performed lemmatization and document-term matrix creation, based on inverse document frequency. The details about each of these operations are given in the following sections.

6.1 Creation of a Corpus

Our first task was to create a corpus, i.e., a collection of text documents, and carry out the preprocessing tasks in order to solve the problem relating to high feature dimensionality in textual data.

The Package *tm* of R allows the effective processing of text documents by creating a corpus. Thus, we used the following command in R to create a corpus for each political party (e.g., BE - the Left Bloc):

```
library(tm)
library(tm.plugin.mail)
corpus_BE<- Corpus(DirSource("BE", encoding ="UTF-8"),
                  readerControl = list(reader = readMail,
                                     language = "pt"))
```

6.2 Preprocessing Procedures Used

In this subchapter we will present the preprocessing tasks used in the Member's bills, our document collection.

Tokenization, Filtering and Stop Word Removal

The following command provides an auxiliary function for preprocessing – changes a particular character to whitespace:

```
makewhitespace<- content_transformer(function(corpus, pattern)
                                     gsub(pattern, " ", corpus))
```

Next we introduce the functions that carry out various preprocessing operations, namely:

- a) *Content_transformer* - allows custom transformation of the corpus (e.g., convert texts to lower case);
- b) *Makewhitespace* - auxiliary function that changes particular characters to whitespace (e.g., /, @, °);
- c) *Removewords* - remove Portuguese stopwords and particular words without interest (e.g., “assembleia”, “artigo”, “república”, “deputado”, “deputada”);
- d) *RemovePunctuation* - remove punctuation marks while keeping hyphens between words, for example;
- e) *RemoveNumbers* - eliminate numbers in the corpus;
- f) *StripWhitespace* - remove extra whitespace.

The format of this preprocessing function is *tm_map*:

```
preprocess<- function(corpus) {
  corpus <-tm_map(corpus, content_transformer(tolower))
  corpus <-tm_map(corpus, makewhitespace, "/")
  corpus <-tm_map(corpus, makewhitespace, "@")
  corpus <-tm_map(corpus, makewhitespace, "°")
  corpus <-tm_map(corpus, removeWords, stopwords("portuguese"))
  corpus <-tm_map(corpus, removePunctuation,
                  preserve_intra_word_dashes=TRUE, ucp = TRUE)
  corpus <-tm_map(corpus, removeNumbers)
  corpus <-tm_map(corpus, removeWords, c("assembleia", "artigo",
    "república", "deputado", "deputada"))
  corpus <-tm_map(corpus, stripWhitespace)
  return(corpus)
}
```

In this example we perform these first preprocessing tasks in the “corpus_BE” originating “corpus_prep_BE”:

```
corpus_prep_BE<- preprocess(corpus_BE)
```

Lemmatization vs Stemming

As the Porter stemmer is not really suitable for Portuguese, we resorted to lemmatization using an adaptation of the linguistic toolkit for Portuguese language elaborated by Gamallo et al. (2018). We have performed an adaptation of this toolkit, with Microsoft Excel using two sheets for this process (lexico1 and lexico2) because this toolkit has more tokens/rows (1,115,611) than those allowed in Excel (1,048,576):

```
library(readxl)
lexico1 <- read_excel("lexico1.xlsx")
lexico2 <- read_excel("lexico2.xlsx")
```

```
lexico <- merge(lexicol, lexico2, all = T)
colnames(lexico) <- c("word", "lemma", "pos.tag")
```

Another problem found in Gamallo's toolkit (2018) was in the lemmatization itself. For some words, such as “esquerda”, “direita”, “estado”, which are nouns as “the left” in English, the first “lemma/pos.tag” suggested by this toolkit was “esquerdo”, “direito”, “estada”. We note that these words are identify as adjectives, i.e. incorrect syntactic category here.

As we cannot use the function/format *tm_map* for this particular lemmatization, we perform the following command from *textstem* package:

```
library(textstem)
for (i in 1:length(corpus_prep_BE)) {corpus_prep_BE[[i]][[1]]<-
  lemmatize_strings(corpus_prep_BE[[i]][[1]], dictionary =
    lexico)}
```

Due to the fact that our document collection consists of Member's bills, there are a lot of Roman numerals in the legislation that were converted to Arabic numerals by our lemmatization. So, we removed them as follows (in addition to other words and extra whitespace created):

```
library(tm)
corpus_prep_BE <-tm_map(corpus_prep_BE, removeNumbers)
corpus_prep_BE <-tm_map(corpus_prep_BE, removeWords, c("http",
  "email", "diploma", "palácio", "bloco", "aplicável"))
corpus_prep_BE <-tm_map(corpus_prep_BE, stripWhitespace)
```

To illustrate the diminution of characters in the documents (including spaces), after the preprocessing tasks, we can use, for example, the first Member's bill of BE, with a reduction of 43% in the number of characters: from the original 6848 to 3917.

Creation of Document-Term Matrix (DTM)

After doing the above preprocessing operations, we created a DTM: each row contains one case (in this case a Member's bill) and each column represents a specific feature (in this case a word). The function *DocumentTermMatrix* can be used to create this table. The main settings selected for this function in our document collection are:

- a) *Weighting* – by default is *term frequency weighting*, but we chose *term frequency - inverse document frequency weighting* (TF-IDF);

- b) *WordLengths* – it defines limits for term length; in our case we only considered words with more than 2 characters and less than 18.
- c) *Bounds* – the global bounds refer to the number of documents in which a word needs to appear in order to be considered (we will contemplate words that appear at least in 3 documents); the local bounds indicate the number of times a word needs to appear in one document to be registered (we will consider words that appear at least twice in the same document).

Thus, we perform the following command:

```
library(tm)
dtm_corpus_prep_BE<- DocumentTermMatrix(corpus_prep_BE, control =
  list(weighting = weightTfIdf, wordLengths =
    c(2,18), bounds = list(global=c(3,Inf),
    local=c(2,Inf))))
```

Let us compare the DTM of the original document collection (i.e., `corpus_BE`) and the document collection after de preprocessing functions and, finally, the new DTM with TF-IDF weighting and special bounds:

```
> DocumentTermMatrix(corpus_BE)
<<DocumentTermMatrix (documents: 273, terms: 32929)>>
Non-/sparse entries: 159470/8830147
Sparsity           : 98%
Maximal term length: 108
Weighting           : term frequency (tf)
> DocumentTermMatrix(corpus_prep_BE)
<<DocumentTermMatrix (documents: 273, terms: 9592)>>
Non-/sparse entries: 94541/2524075
Sparsity           : 96%
Maximal term length: 45
Weighting           : term frequency (tf)
> DocumentTermMatrix(corpus_prep_BE, control = list(weighting =
  weightTfIdf, wordLengths = c(2,18), bounds =
  list(global=c(3,Inf), local=c(2,Inf))))
<<DocumentTermMatrix (documents: 273, terms: 2086)>>
Non-/sparse entries: 31569/537909
Sparsity           : 94%
Maximal term length: 18
Weighting           : term frequency - inverse document frequency
                      (normalized) (tf-idf)
```

With the global bound, the words that appear at least in 3 of the 273 documents remain, while all the others are eliminated. Also, the terms that have fewer values in the columns are dropped (i.e., it required almost 1% were non-empty). As the number of Member's bills varies from one political party to another, we use the *removeSparseTerms*

function, where the level of sparseness is controlled by parameter *S*. If we consider the option of *S* = 0.95 for all political parties, it is imperative that 5% of the values/rows are non-empty (i.e., the words that appear at least in 13,65 of the 273 Member's bills of BE):

```
library(tm)
dtm95_corpus_prep_BE<-removeSparseTerms(dtm_corpus_prep_BE, 0.95)
```

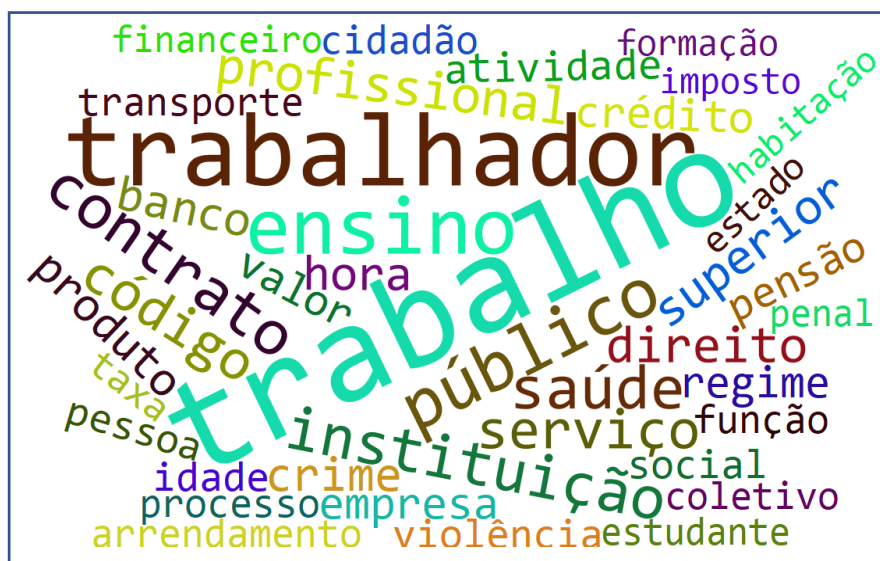
If we compare the dimensions of the original “*dtm_corpus_prep_be*” and this new “*dtm95_corpus_prep_be*”, we observe that only 616 terms were kept in the 273 documents:

```
> dim(dtm_corpus_prep_BE)
[1] 273 2086
> dim(dtm95_corpus_prep_BE)
[1] 273 616
```

To visualize the final terms weighted by TF-IDF we use the package *wordcloud2*, after converting the DTM into a Data Frame (“*df_dtm_corpus_prep_BE*”):

```
library(wordcloud2)
df_dtm95_corpus_prep_BE = as.data.frame(as.matrix(
  dtm95_corpus_prep_BE))
sort.df_dtm95_corpus_prep_BE = sort(colSums(
  df_dtm95_corpus_prep_BE), decreasing = TRUE)
cloud = data.frame(word = names(sort.df_dtm95_corpus_prep_BE),
  freq = sort.df_dtm95_corpus_prep_BE)
wordcloud2(head(cloud, 40), color = 'random-dark', fontWeight =
  'normal', fontFamily = 'Consolas')
```

The result of the visualization is the following:



We can use *head ()* to find the highest 40 TF-IDF of BE:

```
> head (data.frame(freq = sort.df_dtm95_corpus_prep_BE), 40)
```

	freq		freq		freq
trabalho	5.981086	crime	2.099222	cidadão	1.772554
trabalhador	4.452375	banco	2.073832	transporte	1.768085
ensino	3.502236	produto	2.052729	processo	1.746107
público	3.095374	empresa	1.988363	taxa	1.730695
contrato	2.791213	crédito	1.979079	arrendamento	1.726695
instituição	2.703266	regime	1.976681	função	1.722535
saúde	2.696011	valor	1.945551	coletivo	1.705410
serviço	2.473061	pensão	1.921382	penal	1.684694
código	2.469430	atividade	1.849407	estado	1.671155
profissional	2.334341	violência	1.812656	estudante	1.649569
direito	2.221428	social	1.810278	financeiro	1.628216
superior	2.206700	idade	1.795517	habitação	1.610461
hora	2.137397	pessoa	1.780748	formação	1.605860
				imposto	1.604000

Base on DTM with TF-IDF weighting and a level of sparseness of 95%, it is possible to associate some words with BE. Apparently, the Member's bills of this particular political party focus primarily on certain subjects, namely:

- Labour issues – “trabalho”, “trabalhador” and “profissional” are in the top 10 of the TF-IDF values, providing a strong indication about the importance of this subject to BE;
- Education issues – “ensino” and “estudante” reveal this concern;
- Social issues – the words “pensão”, “social”, “pessoa”, “cidadão”, “coletivo”;
- Health issues – the word “saúde”;

There are also other topics that appear to be significant:

- Economic issues – words such as “banco”, “produto”, “empresa”, “crédito”, “financeiro”, “taxa” and “imposto”;
- Housing issues – words like “arrendamento” and “habitação”;
- Security issues – words like “crime”, “violência” and “penal”;
- Mobility issues – the word “transporte”.

For the words “público”, “contrato”, “superior”, “hora”, “valor”, “atividade”, “idade”, “função”, “estado” and “formação” our position is neutral (their scope is only clear when accompanied by other words).

There are some words apparently useless, such as “instituição”, “serviço”, “código”, “direito”, “regime” and “processo”. In this case, they were added to the domain of stopwords and removed from the corpus subjected to automatic keyword extraction.

7 Used Keyword Extraction Techniques

Similarly, we will use BE Member's bills to exemplify the following keyword extraction techniques.

After converting the information to a canonical format, we applied various types of feature extraction methods, namely keyword extraction. There are many keyword extraction techniques. In this study we decided to explore the following ones:

- Method based on POS tagging;
- Method based on collocations and co-occurrences;
- Keyword extraction with TextRank;
- Keyword extraction with RAKE;
- Method based on POS phrase sequence detection;
- Method based on results of dependency parsing.

In order to carry out a comparison between these different techniques we decided to restrict our study to the first 30 keywords generated by each method.

7.1 Method based on POS Tagging

Finding keywords through Parts of Speech (POS) tagging in order to identify nouns (and also verbs and adjectives, if applicable) is an acceptable approach. This is feasible because each word has a "Parts of Speech tag" that can be identified with *udpipe* package, as follows:

```
library(udpipe)
ud_model <- udpipes_download_model(language = "portuguese")
ud_model <- udpipes_load_model(ud_model$file_model)
df_corpus_BE <- data.frame(text = sapply(corpus_BE, paste,
                                         collapse = " "), stringsAsFactors = FALSE)
BE<- udpipes_annotate(ud_model, x = df_corpus_BE$text)
BE <- as.data.frame(BE)
```

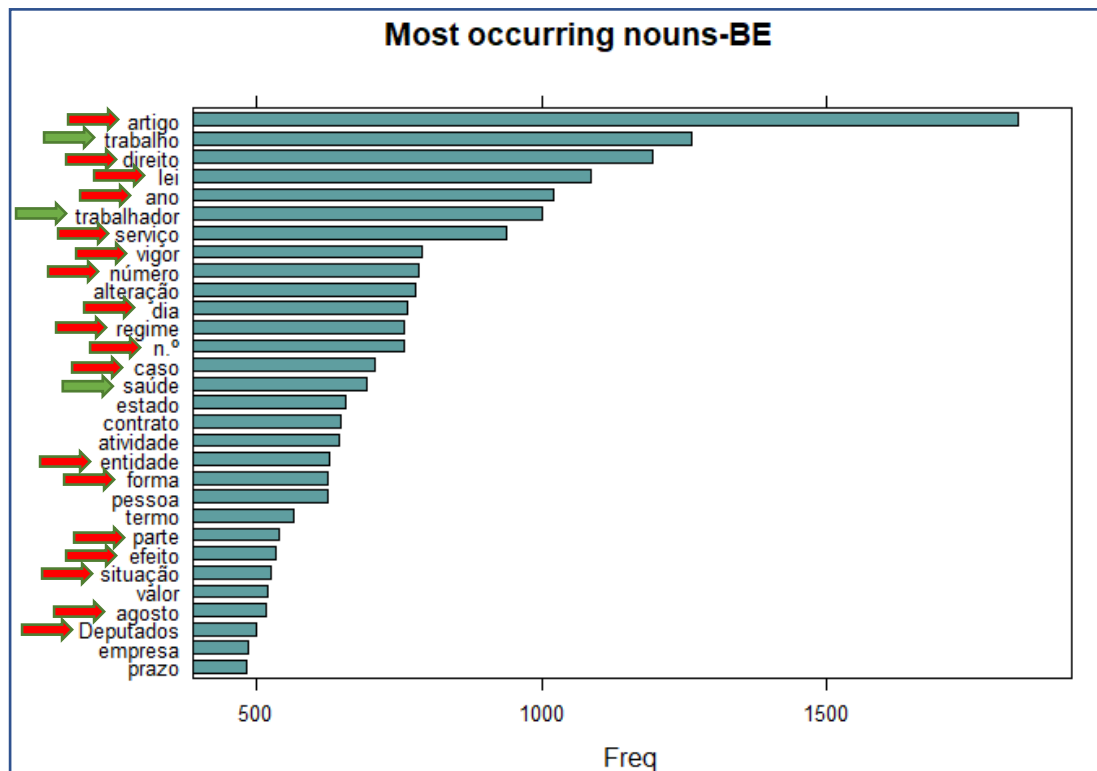
Parts of speech tags enable us to select the words of interest for analysis, because we do not need all type of words. If we extract only nouns, verbs or adjectives, we have the most relevant "tags". The *udpipe* package identifies each word in the phrase as a noun (NOUN), a verb (VERB), an adjective (ADJ) and a proper noun (PROPN), the four most relevant "types" for analytical purposes. The resulting data frame has a field called "upos"

tag (i.e., Universal Parts of Speech) and a field called “lemma”, which is the root form of each token in the text.

In this keyword extraction technique, as in the others presented further on in this chapter, we consider, for analytical purposes, two types of corpus for the “udpipe” package: the original corpus (“BE”, in this exemplifying case) and, if applicable, and for more reliable results, the corpus after the preprocessing tasks presented in the previous chapter (“BE_prep” in this demonstrating case). So, for “nouns” we use the following command:

```
library(lattice)
BE_nouns <- subset(BE, upos %in% "NOUN")
BE_nouns <- txt_freq(BE_nouns$lemma)
BE_nouns$key <- factor(BE_nouns$key, levels = rev(BE_nouns$key))
barchart(key ~ freq, data = head(BE_nouns, 30), col = "cadetblue",
          main = "Most occurring nouns-BE", xlab = "Freq")
```

with this result:



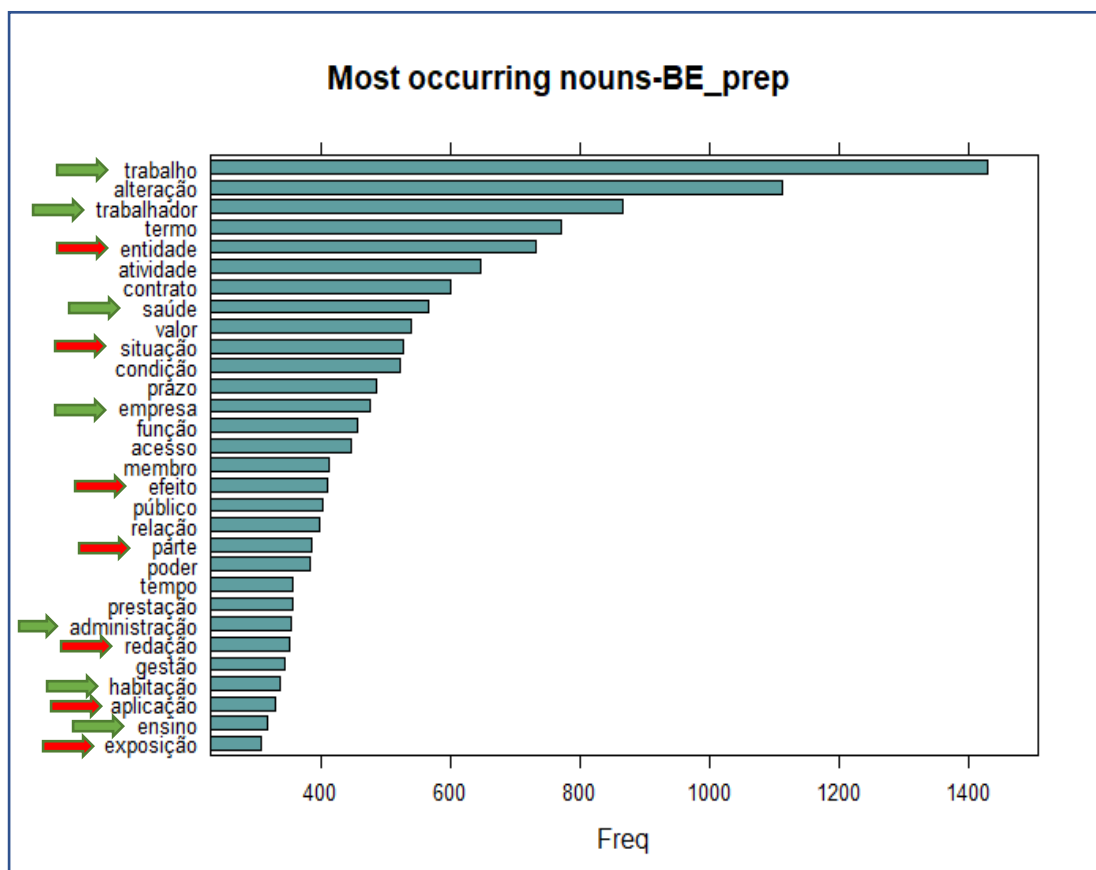
We note that, in most cases, the keywords are marked using green (rarely) and red arrows, depending on whether they are considered relevant or not, as outlined in Chapter 5. Analyzing these results, we decided to use preferably the preprocessed corpus, eliminating all non-relevant words from the set of top 40 TF-IDF candidates:

```

library(tm)
library(udpipe)
library(lattice)
corpus_prep_BE <-tm_map(corpus_prep_BE, removeWords,
                        c("instituição", "serviço", "código", "direito",
                          "superior", "hora", "regime", "processo"))
corpus_prep_BE <-tm_map(corpus_prep_BE, stripWhitespace)
df_corpus_prep_BE <- data.frame(text = sapply(corpus_prep_BE,
                                              paste, collapse = " "), stringsAsFactors =
                                              FALSE)
BE_prep<- udpipe_annotate(ud_model, x = df_corpus_prep_BE$text)
BE_prep <- as.data.frame(BE_prep)
BE_prep_nouns <- subset(BE_prep, upos %in% "NOUN")
BE_prep_nouns <- txt_freq(BE_prep_nouns$lemma)
BE_prep_nouns$key <- factor(BE_prep_nouns$key, levels =
                           rev(BE_prep_nouns$key))
barchart(key ~ freq, data = head(BE_prep_nouns, 30), col =
         "cadetblue", main = "Most occurring nouns-BE_prep", xlab
         = "Freq")

```

with this result:



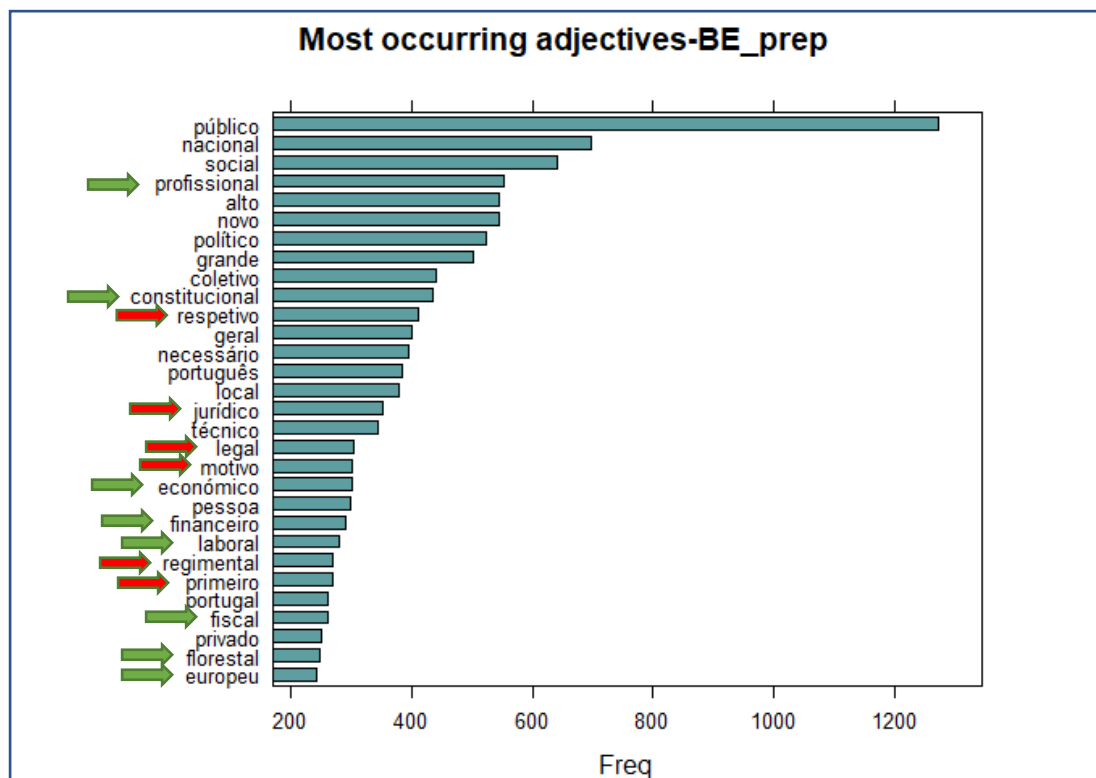
Analyzing the results of this technique, which involves identifying keywords by detecting nouns with POS tagging, we observe that BE focuses on certain topics, namely:

- Labour issues – “trabalho” and “trabalhador” are in the top 3 of noun term frequency;

- Health issues – “saúde” is in the top 8;
- Economic issues – “empresa” and “administração” is in the top 25;
- Housing issues – “habitação”;
- Education issues – “ensino”;

From our point of view, some nouns that were extracted were rather non-informative and, hence, useless (e.g., “entidade”, “situação”, “efeito”, “parte”, “redação”, “aplicação” and “exposição”); regarding others words/nouns our position is neutral.

Concerning the adjectives’ analysis, the results are:



Analyzing the results of this technique we can say that BE apparently focuses on some subjects, namely:

- Labour issues – “profissional” and “laboral”;
- Ideological issues – “constitucional”;
- Economic issues – “económico”, “financeiro” and “fiscal”
- European issues – “europeu”;
- Environmental issues – “florestal”.

There are some nouns apparently useless for keyword extraction, such as “respetivo”, “jurídico”, “legal”, “motivo” (i.e., the cause or origin of something),

“regimental” and “primeiro”; regarding others words/adjectives our position is neutral (we need further details).

In respect to proper nouns, and as we convert texts to lower case during preprocessing, we have to consider the original BE Corpus “udpipe” data frame. None the less, this analysis did not bring any added value, with many words without interest (in red), one with real value (in green) and others in which our position is neutral (in black):

```
library(udpipe)
library(lattice)
BE_proper nouns <- subset(BE, upos %in% "PROPN")
BE_proper nouns <- txt_freq(BE_proper nouns$lemma)
BE_proper nouns$key <- factor(BE_proper nouns$key, levels =
                             rev(BE_proper nouns$key))
```

with this result:

```
> head(BE_proper nouns, 30)
```

key	freq	freq_pct	key	freq	freq_pct	key	freq	freq_pct
artigo	1964	4.2015189	.º	537	1.1487860	Trabalho	321	0.6867045
n.º	1878	4.0175420	DE	535	1.1445074	Nacional	320	0.6845652
lei	1183	2.5307520	Deputadas	513	1.0974436	2	317	0.6781474
Lei	1165	2.4922452	S.	495	1.0589368	Projeto	312	0.6674511
República	1020	2.1820516	Palácio	475	1.0161515	alteração	286	0.6118301
Assembleia	922	1.9724035	Bento	475	1.0161515	1.º	274	0.5861589
Bloco	894	1.9125040	telefone	475	1.0161515	entrada	271	0.5797411
Esquerda	864	1.8483260	fax	475	1.0161515	objeto	266	0.5690448
decreto-lei	837	1.7905658	Portugal	382	0.8171997	grupo	258	0.5519307
Lisboa	564	1.2065462	N.º	379	0.8107819			
			Governo	361	0.7722751			

Analyzing the results of POS tagging for proper nouns, we can say that BE apparently focuses on labour.

7.2 Collocations and Co-occurrences

Even though POS tagging is a good start, our aim is also to identify multi-word expressions. One way of doing this is by identifying collocations.

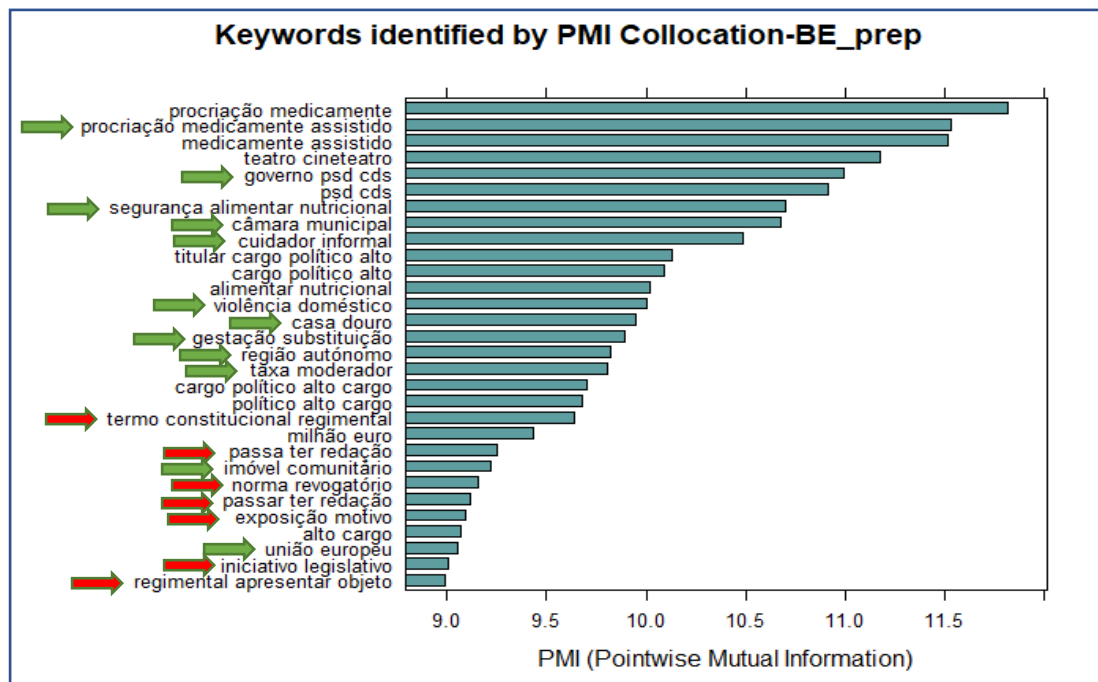
Collocations are an arrangement of words or terms that co-occur more often than would be expected by chance. The function *keywords_collocation*, from *udpipe* package, extracts relevant collocations and computes the following statistics: pointwise mutual information (i.e., PMI), mutual dependency (i.e., MD) and log-frequency biased mutual dependency (i.e., LFMD).

```
library(udpipe)
BE_prep_colloc <- keywords_collocation(x = BE_prep, term =
                                       "token", group = c("doc_id", "paragraph_id",
                                                         "sentence_id"), ngram_max = 4)
```

As we have 273 BE Member's bills, we are going to analyze the frequency greater than 54 (20% * Member's bills), and using only PMI:

```
library(udpipe)
library(lattice)
BE_prep_colloc$key <- factor(BE_prep_colloc$keyword, levels =
                             rev(BE_prep_colloc$keyword))
barchart(key ~ pmi, data = head(subset(BE_prep_colloc, freq > 54),
                                  20), col = "cadetblue", main = "Keywords identified by
                                  PMI Collocation-BE_prep", xlab = "PMI (Pointwise Mutual
                                  Information)")
```

with this result:



Many terms seem useful because they add insights about issues that matter to BE. Regarding Health issues, for instances, you can say BE tried to legislate the following subjects:

- Medically assisted procreation (“procriação medicamento assistido”);
- Food and nutritional security (“segurança alimentar nutricional”);
- Substitute pregnancy (“gestação substituição”);
- Health care costs (“taxa moderador”; this arrangement of words better classifies the term “taxa”, initially placed in Economic issues based on DTM with TF-IDF weighting and 95% sparseness in the previous chapter).

Related to the Local issues:

- “casa douro” gives us information about the local subject that matters for BE – the Douro Wine Company;
- The keywords “câmara municipal” and “região autónoma” show us other local topics important for BE: local and regional authorities.

About Social issues:

- The topics informal caregiver (“cuidador informal”) and domestic violence (“violência doméstica”) seem important for BE.

Housing issues:

- Communal property – “imóvel comunitário”.

Ideological issues:

- With regard to the keyword “governo psd cds”⁷, it appears that BE tried to legislate matters related to the last Portuguese right-wing coalition government.

European issues:

- Keyword european union (“união europeia”) seems important to BE.

The reason for keywords marked with red arrows is the existence of arrangements of words linked to the creation of the draft legislation itself, and thus do not add any relevant data.

Regarding co-occurrences, they allow us to see how words are used either in the same sentence or next to each other, and if we combine this with selecting only nouns and adjectives, this becomes very useful for our analysis. There are 3 types of co-occurrences:

1. Observing which nouns and adjectives are in the same document, sentence or paragraph (i.e., “BE_prep_cooc1”):

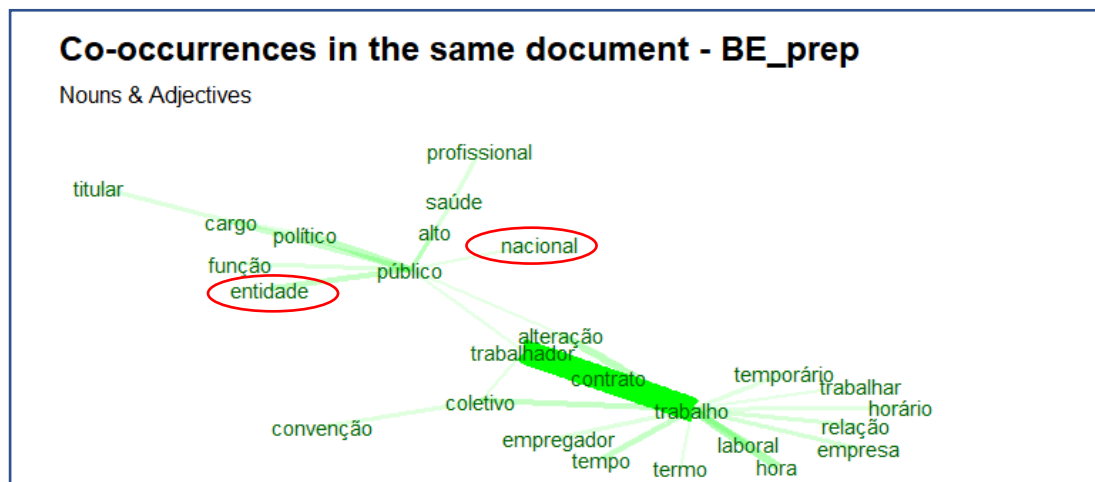
```
library(udpipe)
library(igraph)
library(ggraph)
library(ggplot2)
BE_prep_cooc1 <- cooccurrence(x = subset(BE_prep8, upos %in%
                                         c("NOUN", "ADJ")), term = "lemma", group =
                                         c("doc_id", "paragraph_id", "sentence_id"))
wordnetwork <- head(BE_prep_cooc1, 30)
wordnetwork <- graph_from_data_frame(wordnetwork)
```

⁷ During preprocessing all capital letters were substituted by lower case letters. This is why the abbreviations of the parties appear as “psd” and “cds”, and not “PSD” and “CDS”, which is the normal standard.

⁸ Although in the preprocessed corpus we reduce each document to a level in which they only have one paragraph and one sentence, the use of the original documents does not bring any advantage, on the contrary.

```
gggraph (wordnetwork, layout = "fr") + geom_edge_link(aes(width =
  cooc, edge_alpha = cooc), edge_colour = "green") +
  geom_node_text(aes(label = name), col = "darkgreen", size
  = 4) + theme_graph(base_family = "Arial Narrow") +
  theme(legend.position = "none") + labs(title = "Co-
  occurrences in the same document - BE_prep", subtitle =
  "Nouns & Adjectives")
```

with this result:



This keyword technique allows us to go further in our analysis: the noun “trabalho” (labour) co-occurs in the same document, sentence or paragraph, with other nouns or adjectives related to:

- Working hours (“tempo”, “horário”, “hora”);
- Collective contract of employment (“convenção”, “coletivo”, “contrato”, “relação”);
- Relations between employer (“empregador”) and employee (“trabalhador”)
- Questions about occasional employment contract (“contrato”, “trabalho”, “termo” and “temporário”);

We also observe a strong relation between the term labour (“trabalho”) and the term change/ amendment (“alteração”). In this co-occurrence analysis we also observed a link between the term public (“público”) and:

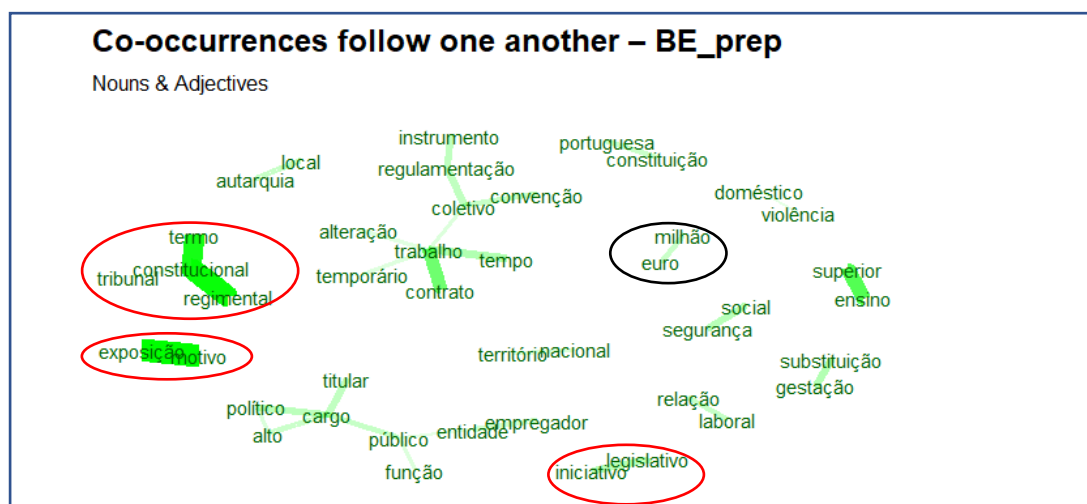
- Civil service (“função”)
- Health (“saúde”) and health workers (“profissional”)
- Political appointees (“cargo”, “titular”, “alto” and “político”).

For the term entity (“entidade”) and national (“nacional”) we think their link is a redundancy, so we marked it in red.

2. Looking at which words are followed by another word (i.e., “BE_prep_cooc2”):

```
library(udpipe)
library(igraph)
library(gggraph)
library(ggplot2)
BE_prep_cooc2 <- cooccurrence(x = BE_prep$token9, relevant =
                             BE_prep$upos %in% c("NOUN", "ADJ"))
wordnetwork <- head(BE_prep_cooc2, 30)
wordnetwork <- graph_from_data_frame(wordnetwork)
gggraph(wordnetwork, layout = "fr") + geom_edge_link(aes(width =
  cooc, edge_alpha = cooc), edge_colour = "green") +
  geom_node_text(aes(label = name), col = "darkgreen", size =
  4) + theme_graph(base_family = "Arial Narrow") +
  theme(legend.position = "none") + labs(title = "Co-
  occurrences follow one another - BE_prep", subtitle =
  "Nouns & Adjectives")
```

with this result:



This keyword technique gives us new subject areas in BE Member’s bills, others observed in previous methodologies, some that are useless - marked in red -, and finally the subject areas in which our position is neutral – marked in black. Regarding the new (or more accurate) information about BE Member’s bills we can point out:

- Education issues - the higher education sector (co-occurrence “ensino” and “superior”);
- Social issues – the social security (co-occurrence “segurança” and “social”);
- Ideological issues – reference to the Portuguese constitution (“constituição” “portuguesa”) traditionally more connoted with the values of the left.
- Local issues – local municipalities (“autarquia local”)

⁹ “lemma” in the original *udpipe* package command.

Let's look at some of the useless keywords: the co-occurrence of words formally linked to the creation of the bill itself do not add any relevant data (e.g., deputies have the right to draft the Member's bill in the applicable constitutional and regimental terms ("termo" "constitucional" "regimental").

This keyword technique alerts us for the imperfect result of lemma "alto" for the token "superior" in the *udpipe* package; because of that, we chose the lemmatization of Gamallo et al. (2018) and substituted the original codification "lemma" for "token" which was, in fact, already lemmatized in the preprocessing stage.



- Identifying which words are in the neighborhood of the word while considering skips of n (e.g. 2) words (*skipgram*; in other words, how often do we identify words with skips up to 2 words in between; i.e., "BE_prep_cooc3"):

```
library(udpipe)
BE_prep_cooc3 <- cooccurrence(x = BE_prep$token10, relevant =
                             BE_prep$upos %in% c("NOUN", "ADJ"), skipgram = 2)
```

This specific co-occurrence within 3 words distance analysis did not add new information comparing to the prior methodology:

> head (BE_prep_cooc2, 30)				> head (BE_prep_cooc3, 30)			
	term1	term2	cooc		term1	term2	cooc
1	exposição	motivo	273	1	exposição	motivo	273
2	constitucional	regimental	266	2	constitucional	regimental	266
3	termo	constitucional	254	3	termo	constitucional	265
4	ensino	superior	200	4	termo	regimental	259
5	contrato	trabalho	175	5	contrato	trabalho	226
6	tempo	trabalho	116	6	ensino	superior	207
7	gestação	substituição	115	7	político	público	138
8	iniciativo	legislativo	110	8	alteração	trabalho	136
9	segurança	social	110	9	trabalho	trabalho	136
10	cargo	político	107	10	tempo	trabalho	132
11	instrumento	regulamentação	104	11	titular	cargo	117
12	regulamentação	coletivo	103	12	gestação	substituição	116
13	titular	cargo	102	13	iniciativo	legislativo	110
14	relação	laboral	98	14	segurança	social	110
15	autarquia	local	98	15	cargo	político	109
16	cargo	público	98	16	instrumento	coletivo	108
17	convenção	coletivo	97	17	regulamentação	coletivo	105
18	milhão	euro	95	18	instrumento	regulamentação	104
19	constituição	portuguesa	94	19	convenção	coletivo	103
20	alto	cargo	94	20	cargo	público	102
21	entidade	empregador	87	21	milhão	euro	98
22	território	nacional	81	22	relação	laboral	98
23	político	alto	80	23	autarquia	local	98
24	função	público	78	24	titular	político	97
25	coletivo	trabalho	78	25	constituição	portuguesa	95
26	trabalho	temporário	77	26	alto	público	95
27	alteração	trabalho	74	27	alto	cargo	94
28	violência	doméstico	71	28	entidade	empregador	88
29	tribunal	constitucional	70	29	político	cargo	86
30	entidade	público	70	30	titular	alto	85

¹⁰ "lemma" in the original *udpipe* package command

While co-occurrences focus on frequency, keyword correlations merely indicate how terms are together in a same document/sentence (i.e., correlation measures between 2 terms can also be high even if these occur only a small number of times but always together). Thus, and for this correlation (i.e., “BE_prep_corr”) we use the following commands:

```
library(udpipe)
BE_prep$id <- unique_identifier(BE_prep, fields = c("sentence_id",
  "doc_id"))11
dtm_BE_prep <- subset(BE_prep, upos %in% c("NOUN", "ADJ"))
dtm_BE_prep <- document_term_frequencies(dtm_BE_prep, document =
  "id", term = "token")12
dtm_BE_prep <- document_term_matrix(dtm_BE_prep)
dtm_BE_prep <- dtm_remove_lowfreq(dtm_BE_prep, minfreq = 7)13
termcorrelations <- dtm_cor(dtm_BE_prep)

BE_prep_corr <- as_cooccurrence(termcorrelations)
BE_prep_corr <- subset(BE_prep_corr, term1 < term2 & abs(cooc) >
  0.99)
BE_prep_corr <- BE_prep_corr[order(abs(BE_prep_corr$cooc),
  decreasing = TRUE), ]
```

This high keyword correlation offers a different information about the terms (they appear in a few BE Member’s bills, but always together):

```
> head (BE_prep_corr, 30)
```

term1	term2	cooc	term1	term2	cooc
besa	capitalismo	1	besa	polos	1
besa	complacência	1	capitalismo	polos	1
capitalismo	complacência	1	complacência	polos	1
besa	financeirização	1	financeirização	polos	1
capitalismo	financeirização	1	montar	polos	1
complacência	financeirização	1	permissibilidade	polos	1
besa	montar	1	besa	regulatórios	1
capitalismo	montar	1	capitalismo	regulatórios	1
complacência	montar	1	complacência	regulatórios	1
financeirização	montar	1	financeirização	regulatórios	1
besa	permissibilidade	1	montar	regulatórios	1
capitalismo	permissibilidade	1	permissibilidade	regulatórios	1
complacência	permissibilidade	1	polos	regulatórios	1
financeirização	permissibilidade	1	besa	terrorismo	1
montar	permissibilidade	1	capitalismo	terrorismo	1

For visualization we use the code:

```
library(igraph)
library(ggraph)
library(ggplot2)
wordnetwork <- head(BE_prep_corr, 30)
wordnetwork <- graph_from_data_frame(wordnetwork)
```

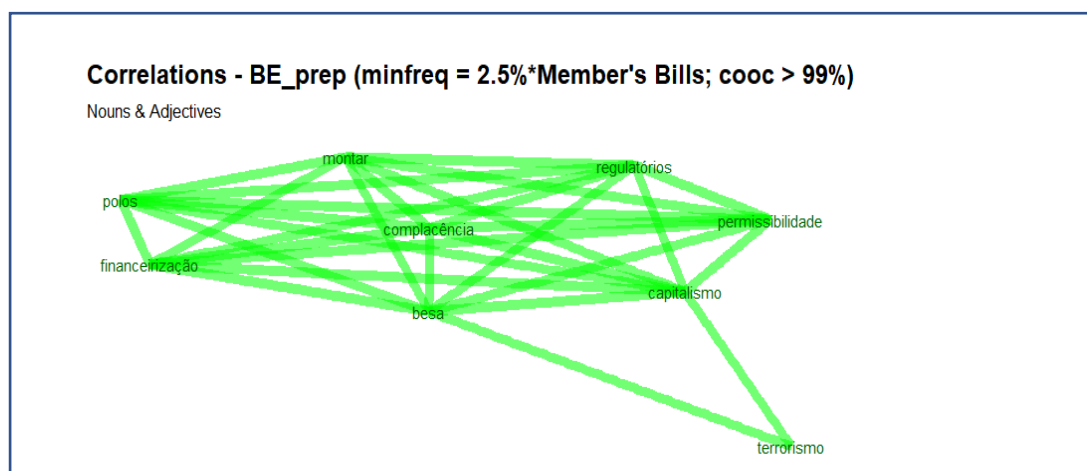
¹¹ In the preprocessed corpus we reduce each document to a level in which they have only one paragraph and one sentence (the use of the original documents or sentence division provides no advantage).

¹² “lemma” in the original *udpipe* package command.

¹³ We used as reference a frequency higher than 2.5% for all Member’s bills produced by each political party.

```
gggraph(wordnetwork, layout = "fr") + geom_edge_link(aes(width =
  cooc, edge_alpha = cooc), edge_colour = "green") +
  geom_node_text(aes(label = name), col = "darkgreen", size =
  4) + theme_graph(base_family = "Arial Narrow") +
  theme(legend.position = "none") + labs(title =
  "Correlations - BE_prep (minfreq = 2.5%*Member's Bills;
  cooc > 99%)", subtitle = "Nouns & Adjectives")
```

with this result:



This keyword technique allows us to go further in our analysis about BE specific subjects addressed in a few Member’s bills related to Economic/Ideological issues:

- The problems of complacency (“complacência”) and permissibility (“permissibilidade”) associated with capitalism (“capitalismo”) and financialization (“financiarização”),
- The particular problem of BESA’s (“besa”- Banco Espírito Santo Angola) bankruptcy and the topics above; there are also a heavy correlation with the term terrorism (“terrorismo”)

7.3 Keyword Extraction with TextRank

Another approach for keyword detection is TextRank, a word network that follows the same principles as Google PageRank. This method allows us to summarize text and extract keywords through a word network by observing if the words, either nouns or adjectives, are following one another, with a frequency higher than 27¹⁴ (i.e., “BE_prep_textrank”):

¹⁴ We used as reference a frequency higher than 10% for all Member’s bills produced by each political party.

```
library(udpipe)
library(textrank)
BE_prep_textrank <- textrank_keywords(BE_prep$token15, relevant =
  BE_prep$upos %in% c("NOUN", "ADJ"), ngram_max = 8,
  sep = " ")
BE_prep_textrank <- subset(BE_prep_textrank$keywords, ngram > 1 &
  freq >= 27)
```

However, and before presenting the plots, let us introduce the top 30 keywords produced by this technique:

```
> head (BE_prep_textrank, 30)
```

keyword	ngram	freq	keyword	ngram	freq
constitucional regimental	2	266	segurança social	2	50
termo constitucional regimental	3	256	violência doméstico	2	48
ensino superior	2	82	cargo político alto cargo público	5	48
exposição motivo	2	80	político alto cargo público	4	48
norma revogatório	2	79	tempo trabalho	2	47
casa dourado	2	79	trabalho temporário	2	45
contrato trabalho	2	69	titular cargo político alto cargo público	6	45
psd cds	2	66	autarquia local	2	44
iniciativo legislativo	2	66	governo psd cds	3	44
constituição portuguesa	2	66	posterior alteração	2	44
gestação substituição	2	64	convenção coletivo	2	43
cargo público	2	57	função público	2	42
alto cargo público	3	55	união europeu	2	40
entidade empregador	2	54	cuidado saúde	2	40
relação laboral	2	50	alimentar nutricional	2	40

As usual, we chose the color green to highlight the keywords that seemed to characterize BE reasonably well. This technique introduced various keywords, including a new topic related to health care (“cuidado saúde”).

When we plot the results, with *wordcloud2* package, some multi-words with high frequency do not appear:

```
library(wordcloud2)
cloud2 = data.frame(keyword = BE_prep_textrank$keyword, freq =
  BE_prep_textrank$freq)
wordcloud2(head(cloud2, 50), color = 'random-dark', fontWeight =
  'normal', fontFamily = 'Consolas')
```

with this result:



¹⁵ “lemma” in the original *udpipe/textrank* package command.

7.4 Keyword Extraction with RAKE

The next algorithm is called RAKE (Rapid Automatic Keyword Extraction) and looks for keywords by observing a contiguous sequence of words with no irrelevant words:

```
library(udpipe)
BE_rake <- keywords_rake(x = BE, term = "token", group =
  c("doc_id", "paragraph_id", "sentence_id"), relevant =
  BE$upos %in% c("NOUN", "ADJ"), ngram_max = 4)
```

Observing the highest 30 RAKE values at the original corpus (i.e., BE) we gather a different information regarding keywords with low frequency and (key)words without interest.

```
> head (BE_rake, 30)
```

	keyword	ngram	freq	rake
1	regra juridica explícita impositiva	4	4	7.785481
2	regra jurídica implícita permissiva	4	4	7.585481
3	outra/s pessoa/s beneficiária/s	3	2	6.666667
4	atividades económicas secundárias compatíveis	4	2	5.302538
5	períodos contratuais mínimos excessivos	4	2	5.127140
6	falsos recibos verdes	3	10	4.976190
7	subsídio social parental inicial	4	2	4.698571
8	longuíssimas carreiras contributivas	3	3	4.660473
9	práticas concorrenciais agressivas	3	7	4.519262
10	medidas restritivas adicionais respeitantes	4	2	4.483324
11	iniciativa legislativa cidadã	3	2	4.362436
12	comprovada carência económica	3	8	4.300144
13	Decretos-Leis números 261/2001	3	5	4.281929
14	altos cargos públicos	3	56	4.274637
15	retribuição mínima mensal garantida	4	2	4.256878
16	tecido empresarial nacional	3	2	4.175012
17	medicina tradicional chinesa	3	5	4.039773
18	longas carreiras contributivas	3	7	3.993806
19	único escândalo bancário	3	7	3.969841
20	procedimentos inspetivos pendentes	3	4	3.962963
21	Decretos-Leis n.ºs 272/2001	3	5	3.954695
22	profundas crises económicas	3	3	3.940429
23	sucessivas crises bancárias	3	7	3.920996
24	sistema financeiro português frágil	4	3	3.891471
25	normas imperativas absolutas	3	4	3.857907
26	entidade patronal interessada	3	4	3.842952
27	domínio público hidrico	3	12	3.808143
28	ativos fixos tangíveis	3	2	3.800000
29	licença parental inicial	3	8	3.784138
30	nova iniciativa legislativa	3	2	3.776615

If we consider our preprocessed corpus (i.e., BE_prep), as an alternative, we get higher RAKE values, but an even lower frequency of keywords:

```
> head (BE_prep_rake, 30)
```

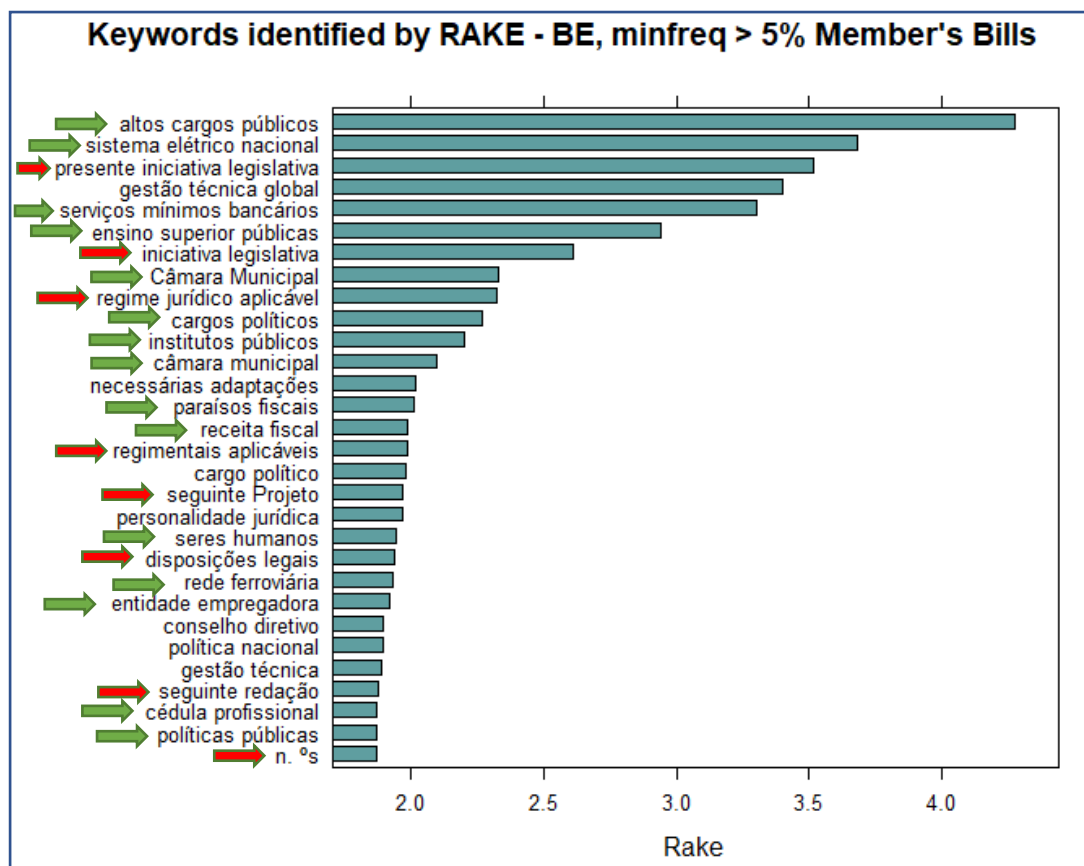
	keyword	ngram	freq	rake
1	exposição campo elétrico magnético	4	3	24.51221
2	sistema nacional gás natural	4	2	21.67210
3	gestão técnico global sen	4	2	20.76291
4	humano alimentação nutrição adequado	4	2	19.81412
5	nacional segurança alimentar nutricional	4	2	19.48080
6	titular cargo político	3	5	18.44828
7	matéria segurança alimentar nutricional	4	2	17.90952
8	ensino superior público regulamentação	4	2	17.66002
9	hora semanal médio trabalho	4	2	17.56437
10	comunicação eletrónicas acessível público	4	2	17.35474
11	aproveitamento racional recurso natural	4	2	17.27585
12	juridico ensino superior rjies	4	3	17.13807

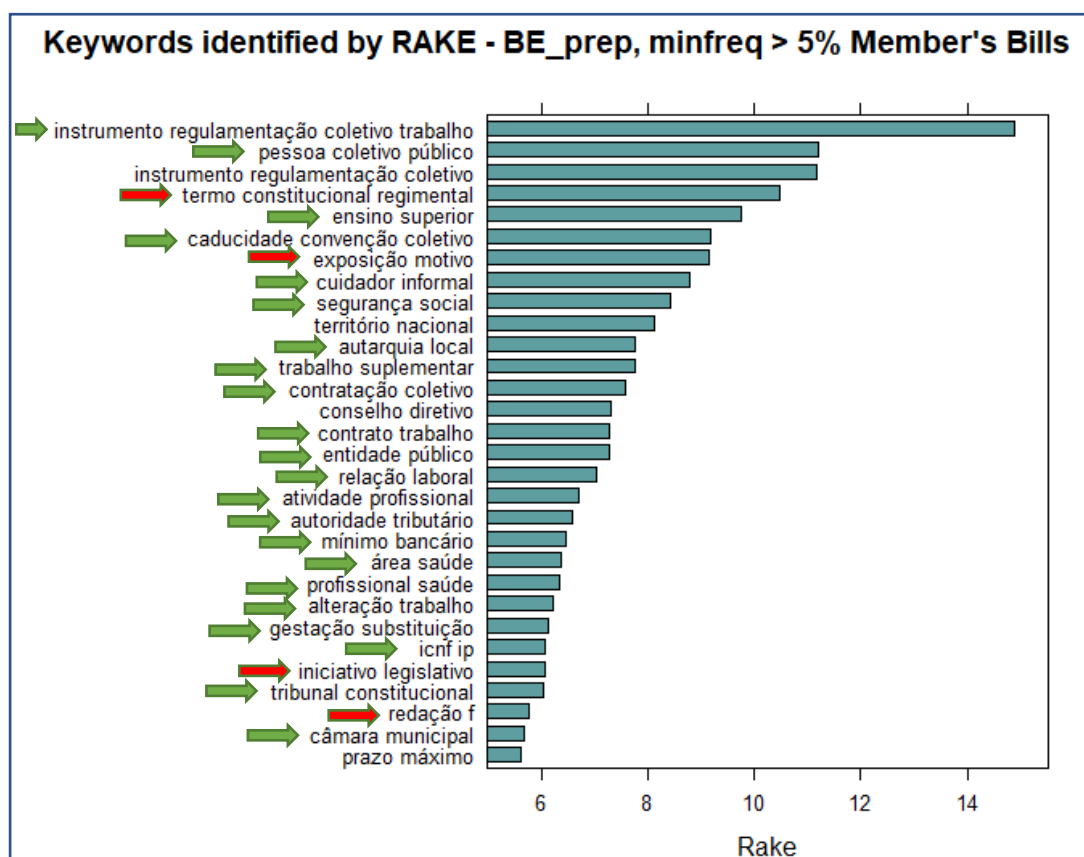
	keyword	ngram	freq	rake
13	especial acesso ensino superior	4	2	17.09674
14	órgão gestão hospitalar respetivo	4	2	17.09566
15	contrato trabalho termo resolutivo	4	2	17.07721
16	trabalho novo ciclo político	4	2	17.06037
17	científico estabelecimento ensino superior	4	2	17.03292
18	fixado órgão ensino superior	4	2	16.98936
19	domínio público hídrico	3	2	16.98219
20	ensino superior público estudante	4	2	16.92635
21	cultivo importação comercialização organismo	4	3	16.80211
22	ensino superior público situação	4	2	16.63564
23	finanças público crise bancário	4	3	16.59620
24	ciclo estudo conducente licenciatura	4	2	16.55546
25	sistema ecológico produção alimento	4	2	16.52178
26	gestão meio produção comunitário	4	2	16.51287
27	nível nacional regional local	4	2	16.43540
28	ação representativo capital social	4	2	16.39796
29	público financiamento partido político	4	2	16.26929
30	condição frequência ensino superior frequência	4	3	16.22001

We were interested, also, in keywords that occur more often, e.g. with a frequency greater than 13 (representing 5% of the total 273 BE Member's bills). The following bar charts present the results of algorithm RAKE for these higher frequencies.

```
library(lattice)
BE_rake$key <- factor(BE_rake$keyword, levels =
  rev(BE_rake$keyword))
barchart(key ~ rake, data = head(subset(BE_rake, freq >= 13), 30),
  col = "cadetblue", main = "Keywords identified by RAKE -
  BE, minfreq > 5% Member's Bills", xlab = "Rake")
```

with these results (BE, with original sentence structure, and BE_prep):





In this keyword methodology we conducted “2 x 2” experiments: we used two kinds of corpora, the original one (i.e., BE) and the preprocessed one (i.e., BE_prep). For each of them, we used the highest 30 RAKE values, first with no term frequency limit or when the frequency of keywords was greater than a given threshold ($5\% \times \text{number of Member's bills of political party, respectively}$). We observed various informative keywords of different types in each corpus:

- Economic issues – national electrical system (“sistema elétrico nacional”); minimum banking services (“serviços mínimos bancários”, “mínimo bancário”¹⁶); fiscal paradises (“paraísos fiscais”); tax revenue (“receita fiscal”); railway network (“rede ferroviária”); tax authority (“autoridade tributário”).
- Labour issues - professional license (“cédula profissional”), expiration of a collective agreement (“caducidade convenção/ contratação coletivo”).
- Public Administration issues - public institutes (“institutos públicos”), public policies (“políticas públicas”); legal system related (decree law “272/2001”¹⁷);

¹⁶ The elimination of the word “serviço” in our preprocessing stage was perhaps hastily unadvised.

¹⁷ As we consider the original corpus (i.e., BE) the keywords generated incorporate numbers that, in turn, indicate subjects dealt with.

- Environmental issues - Institute for Nature Conservation and Forests (“icnf”);
- Local issues - Porto metro light rail related (decree law “261/2001”¹⁸).

7.5 POS Phrase Sequence Detection

Another possibility is to extract phrases or, in others words, a sequence of terms that follow each other based on a sequence of POS tags (i.e., A: adjective, C: coordinating conjunction, D: determiner, M: modifier of verb, N: noun or proper noun, P: preposition). We studied two common type phrases (defining a sequence of parts of speech tags to extract from the text), presented below:

Simple Noun Phrase

The goal is to find a phrase with this pattern: "(A|N)*N(P+D*(A|N)*N)*" (Wijffels 2019, p. 32), which essentially starts with an adjective or noun, another noun, composed a preposition and determiner followed by an adjective or noun and then a noun again.

For this keyword technique we used the *udpipe* package, and the functions *as_phrasemachine*, which convert POS tags to the one-letter tags mentioned above, and then *keywords_phrases* to extract a sequence of terms which follow each other (i.e., phrases).

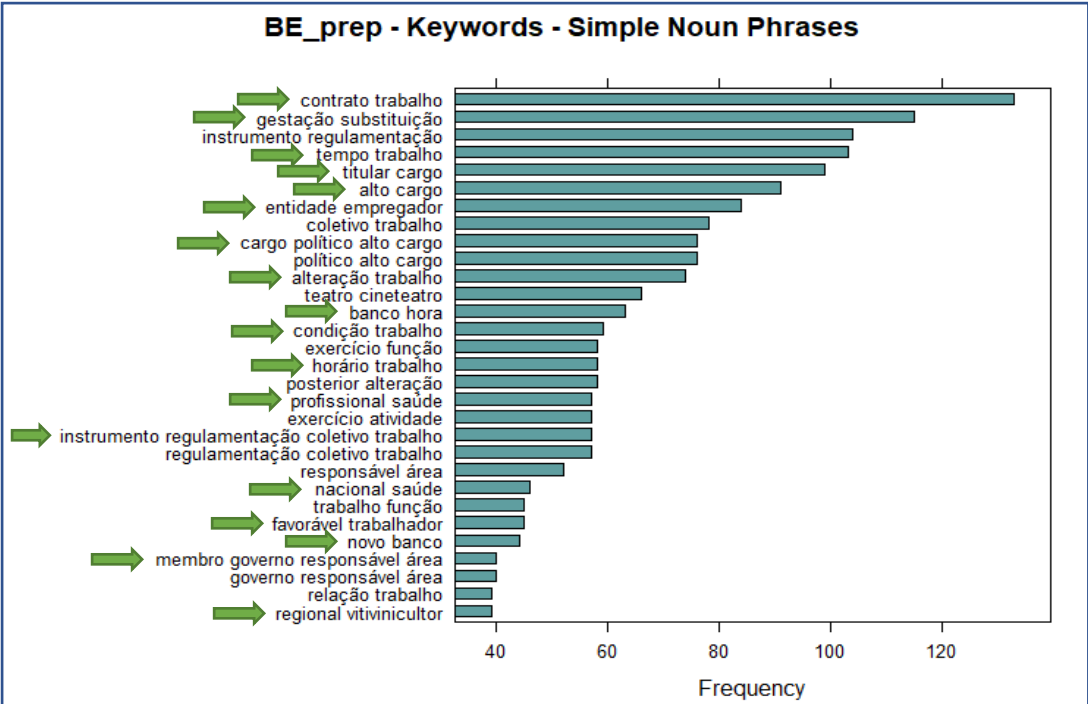
```
library(udpipe)
BE_prep$phrase_tag <- as_phrasemachine(BE_prep$upos, type =
  "upos")
BE_prep_nphrases <- keywords_phrases(x = BE_prep$phrase_tag, term
  = BE_prep$token, pattern =
  "(A|N)+N(P+D*(A|N)*N)*", is_regex = TRUE,
  ngram_max = 4, detailed = FALSE)
```

As we are interested in higher frequencies, for instance, bigger than 27 (i.e., 10%* BE Member’s bills), we present a bar chart presentation with this data:

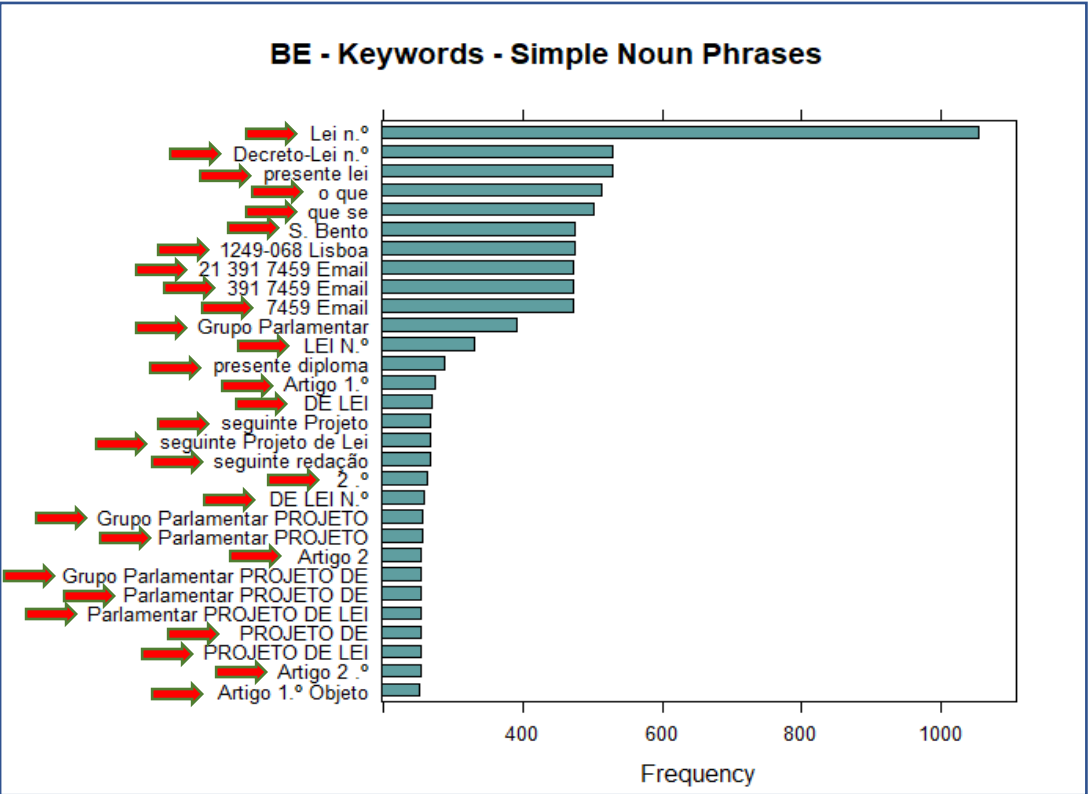
```
library(lattice)
BE_prep_nphrases <- subset(BE_prep_nphrases, ngram > 1 & freq >=
  27)
BE_prep_nphrases$key <- factor(BE_prep_nphrases$keyword, levels =
  rev(BE_prep_nphrases$keyword))
barchart(key ~ freq, data = head(BE_prep_nphrases, 30), col =
  "cadetblue", main = "BE_prep - Keywords - Simple Noun
  Phrases", xlab = "Frequency")
```

¹⁸ See footnote 17.

with this result:



Unlike the last technique used, the choice of the original corpus (i.e., BE) is counterproductive, as reflected in the next bar chart results:



In this noun phrase extraction, we observe new dimensions in BE Member's bills topics, not seen in the previous methodologies.

- Labour issues - keyword “banco hora”, a bank of hours, job related.
- Economic issues – name of a Portuguese bank (“novo banco”) in which the Portuguese state had to intervene; regional winemakers related (“regional vitivinicultor”)

Simple Verb Phrase

According to Wijffels, the purpose is to find a (verb) phrase with a pattern:

"((A|N)*N(P+D*(A|N)*N)*P*(M|V)*V(M|V)*|
 (M|V)*V(M|V)*D*(A|N)*N(P+D*(A|N)*N)*|
 (M|V)*V(M|V)*(P+D*(A|N)*N)+|
 (A|N)*N(P+D*(A|N)*N)*P*((M|V)*V(M|V)*D*(A|N)*N(P+D*(A|N)*N)*|
 (M|V)*V(M|V)*(P+D*(A|N)*N)+)" (2019, p. 32).

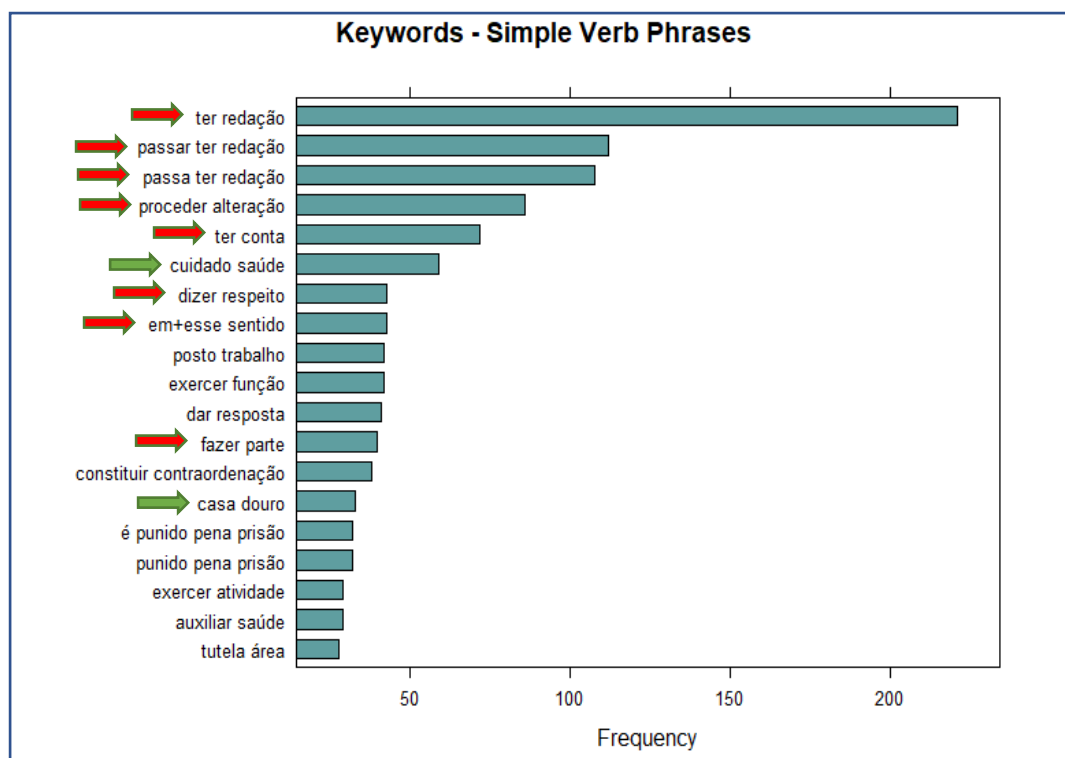
The noun phrase discussed earlier can be seen in this grammar description underline. Likewise, the first, we used the functions *as_phrasemachine* and *keywords_phrases* of *udpipe* package:

```
library(udpipe)
BE_prep$phrase_tag <- as_phrasemachine(BE_prep$upos, type =
  "upos")
BE_prep_vphrases <- keywords_phrases(x = BE_prep$phrase_tag, term
  = BE_prep$token, pattern =
  "( (A|N)*N(P+D*(A|N)*N)*P*(M|V)*V(M|V)*| (M|V)*V
  (M|V)*D*(A|N)*N(P+D*(A|N)*N)*| (M|V)*V(M|V)*(P+
  D*(A|N)*N)+| (A|N)*N(P+D*(A|N)*N)*P*((M|V)*V(M|
  V)*D*(A|N)*N(P+D*(A|N)*N)*| (M|V)*V(M|V)*(P+D*(
  A|N)*N)+)", is_regex = TRUE, ngram_max = 4,
  detailed = FALSE)
```

And bar chart representation from *lattice* package:

```
library(lattice)
BE_prep_vphrases <- subset(BE_prep_vphrases, ngram > 1 & freq >=
  27)
BE_prep_vphrases$key <- factor(BE_prep_vphrases$keyword, levels =
  rev(BE_prep_vphrases$keyword))
barchart(key ~ freq, data = head(BE_prep_vphrases, 30), col =
  "cadetblue", main = "Keywords - Simple Verb Phrases",
  xlab = "Frequency")
```

We used a minimum frequency equal to 27, corresponding a 10% Member's bills.



In this case the result would be the same if we did not incorporate this frequency restriction (i.e., `head (subset (BE_prep_vphrases, ngram > 1), 30))`).

The results of the verb phrase extraction were very weak, not adding new topics to BE Member's bills issues (only a known local wine company subject – “casa douro”, and a phrase related to health care - “cuidado saúde”).

7.6 Dependency Parsing

In this section we use dependency parsing output to extract the nominal subject and the associated adjective. In this analysis we are searching for the words which have a dependency relation between the nominal subject and an adjective that changes the nominal subject (i.e., BE_deparsing):

```
library(udpipe)
BE_deparsing <- merge(BE, BE, by.x = c("doc_id", "paragraph_id",
    "sentence_id", "head_token_id"), by.y = c("doc_id",
    "paragraph_id", "sentence_id", "token_id"), all.x =
    TRUE, all.y = FALSE, suffixes = c("", "_parent"),
    sort = FALSE)
BE_deparsing <- subset(BE_deparsing, dep_rel %in% "nsubj" & upos
    %in% c("NOUN") & upos_parent %in% c("ADJ"))
BE_deparsing$term <- paste(BE_deparsing$lemma_parent,
    BE_deparsing$lemma, sep = " ")
BE_deparsing <- txt_freq(BE_deparsing$term)
```

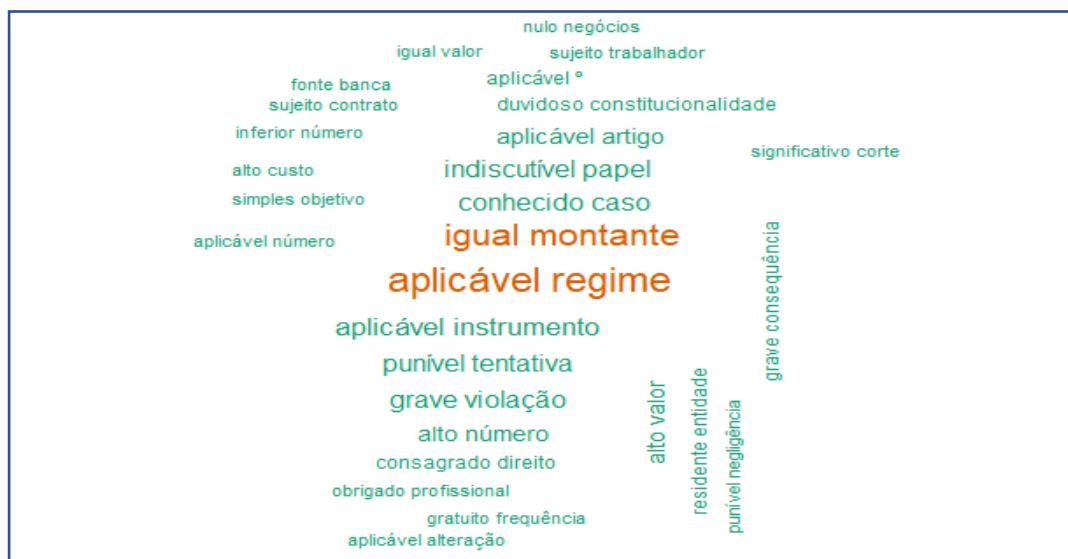
In a canonical way of presentation (and compared with preprocessed BE):

> head(subset(BE_deparsing), 30)				> head(subset(BE_prep_deparsing), 30)			
	key	freq	freq_pct		key	freq	freq_pct
1	aplicável disposto	50	4.6948357	1	preciso fator	7	0.8641975
2	aplicável regime	13	1.2206573	2	conhecido presidente	7	0.8641975
3	igual montante	12	1.1267606	3	indiscutível banif	7	0.8641975
4	aplicável instrumento	8	0.7511737	4	vital trabalho	4	0.4938272
5	punível tentativa	7	0.6572770	5	salvo parte	4	0.4938272
6	conhecido caso	7	0.6572770	6	necessário alteração	4	0.4938272
7	indiscutível papel	7	0.6572770	7	grande possibilidade	4	0.4938272
8	grave violação	7	0.6572770	8	necessário adaptação	4	0.4938272
9	aplicável artigo	6	0.5633803	9	necessário utilização	4	0.4938272
10	alto valor	5	0.4694836	10	mil coisa	3	0.3703704
11	alto número	5	0.4694836	11	necessário definição	3	0.3703704
12	grave consequência	4	0.3755869	12	possível método	3	0.3703704
13	duvidoso constitucionalidade	4	0.3755869	13	igual montante	3	0.3703704
14	consagrado direito	4	0.3755869	14	alto contrato	3	0.3703704
15	residente entidade	4	0.3755869	15	vantajoso salvo	3	0.3703704
16	aplicável °	4	0.3755869	16	incentivo criação	3	0.3703704
17	obrigado profissional	3	0.2816901	17	possível substituição	2	0.2469136
18	simples objetivo	3	0.2816901	18	necessário caso	2	0.2469136
19	aplicável número	3	0.2816901	19	punível tentativo	2	0.2469136
20	sujeito trabalhador	3	0.2816901	20	terceiro ónus	2	0.2469136
21	gratuito frequência	3	0.2816901	21	inequívoco extinção	2	0.2469136
22	nulo ato	3	0.2816901	22	alto carência	2	0.2469136
23	necessário política	3	0.2816901	23	possível aumento	2	0.2469136
24	sujeito contrato	3	0.2816901	24	possível reequilíbrio	2	0.2469136
25	inferior número	3	0.2816901	25	alargado sociedade	2	0.2469136
26	nulo negócios	3	0.2816901	26	imenso utilização	2	0.2469136
27	punível negligência	3	0.2816901	27	obrigatório órgão	2	0.2469136
28	alto custo	3	0.2816901	28	respetivo conselho	2	0.2469136
29	aplicável alteração	3	0.2816901	29	pessom coletivo	2	0.2469136
30	igual valor	3	0.2816901	30	incompatível função	2	0.2469136

The key “pessom coletivo” is apparently a bug in lemmatization of some “pessoa” tokens on *udpipe* package.

Let us visualize the terms using the original BE corpus, with some restrictions on the layout:

```
library(wordcloud)
wordcloud(words = BE_deparsing$key, freq = BE_deparsing$freq,
  min.freq = 3, max.words = 30, random.order = FALSE,
  colors = c("#1B9E77", "#D95F02", "#7570B3", "#E7298A",
    "#66A61E", "#E6AB02"))
```



This keyword extraction methodology was not particularly valuable to our analysis, adding only a new dimension to *Economic issues* in preprocessed BE (“banif” – another bank in which the Portuguese state had to intervene) and a repetition of a topic related to *Ideological issues* regarding doubts of a matter (unknown) and its concordance with the Portuguese Constitution (“duvidoso constitucionalidade”) in the original BE corpus.

With this last methodology we finished our presentation about automatic keyword extraction in BE Member’s bills.

7.7 Comparison of Keyword Extraction Methods

To compare the effectiveness of each keyword extraction technique applied, and as we don’t have the correct keywords prepared by an expert (i.e., a golden rule mentioned in Chapter 5), we have to rely on our judgement¹⁹, captured by the colors green, red and black that were used to highlight certain keywords. We compare the different keyword extraction methods by analyzing the top 30 keywords extracted in each case.

The result was calculated by the following criterion: green keyword – 3 points; black keyword – 1 point; red keyword – 0 points; and issues – 3 points.

The issues punctuation is related to the discriminative power of a technique, i.e. the ability of addressing different topics in its results, which we considered positively in this assessment.

When we applied different approaches to a given methodology (e.g., the original BE corpus or our preprocessed corpus, or even different frequencies or n-grams) we obtained the best score. We summarize bellow the results of each methodology and approach used on the 273 BE Member’s bills (MB)²⁰.

Table 3 - Keyword Extraction Comparative

	Green	Black	Red	Issues	TOTAL
1. Method based on POS Tagging					55
1.1. Most occurring nouns – BE	3	9	18	2	24
1.2. Most occurring nouns – BE_prep	7	16	7	5	52
1.3. Most occurring adjectives - BE_prep	8	16	6	5	55
1.4. Most occurring proper nouns – BE	1	3	26	1	9

¹⁹ In the absence of a known ontology of relevant words in political science.

²⁰ Consultation of the detailed results of keywords extraction is available, by political party, in the appendices A to G.

	Green	Black	Red	Issues	TOTAL
2. Collocations and Co-occurrences					94
2.1. PMI collocation – BE_prep	12	11	7	6	65
2.2. Co-occurrences in the same document – BE_prep	28	0	2	3	93
2.3. Co-occurrences follow one another – BE_prep	25	1	4	6	94
2.4. Co-occurrences within 3 words distance – BE_prep	23	2	5	6	89
2.5. Correlations – BE_prep, freq $\geq 2.5\% \cdot MB$	19	11	0	3	77
3. Keyword Extraction with TextRank					85
3.1. TextRank – BE_prep, freq $\geq 10\% \cdot MB$	18	7	5	8	85
4. Keyword Extraction with RAKE					110
4.1. RAKE – BE	22	1	7	7	88
4.2. RAKE – BE_prep	28	2	0	8	110
4.3. RAKE – BE, freq $\geq 5\% \cdot MB$	15	7	8	6	70
4.4. RAKE – BE_prep, freq $\geq 5\% \cdot MB$	22	4	4	9	97
5. POS Phrase Sequence Detection					84
5.1. Simple Noun Phrase – BE_prep	18	12	0	6	84
5.2. Simple Noun Phrase – BE	0	0	30	0	0
5.3. Simple Verb Phrases – BE_prep	2	9	8	2	21
6. Dependency Parsing					31
6.1. Dependency Parsing – BE	1	23	6	1	29
6.2. Dependency Parsing – BE_prep	1	25	4	1	31

For readability we present the top 5 of this comparative analysis:

Table 4 - Top 5 Keyword Extraction Techniques

Keyword Extraction Technique	TOTAL
RAKE – BE_prep	110
RAKE – BE_prep, freq $\geq 5\% \cdot MB$	97
Co-occurrences follow one another – BE_prep	94
Co-occurrences in the same document – BE_prep	93
Co-occurrences within 3 words distance – BE_prep	89

Comparing the different methods, two conclusions can be formulated: in each methodology, the preprocessed procedure had better results and, apparently, the best keyword extraction technique for BE Member's bills is RAKE (the top two were accomplished with this procedure).

8 Similarity Between Member's Bills and Electoral Programs

Our aim is to determine document similarity between Member's bills and electoral programs using the following procedure for each political party: we chose the RAKE algorithm to extract keywords and compare these documents because we obtained good results with this methodology in the study described in the previous chapter.

8.1 Applied Methodology

The main concern was the identification of an appropriate set of keywords, obtained with the RAKE algorithm, in order to establish the similarities of these two types of documents. If, on the one hand, regarding Member's bills, the higher RAKE values would constitute a simple criterion, on the other hand, adding a minimum frequency would allow a clearer image of the obtained set of keywords. As each party, or coalition, only has one electoral program, we chose to increase the set of keywords to ensure a better representativeness, given the more generalist character of these documents.

Therefore, and due to the low frequency of the keywords corresponding to the 30 higher RAKE values in Member's bills (i.e., RAKE – BE_prep) we established a threshold and added the keywords that appeared more often than the number of times defined by that threshold (i.e., RAKE – BE_prep, $\text{freq} \geq 5\% \cdot \text{MB}$). Then, we compared them with the 60 highest RAKE values of the electoral programs. We used the same preprocessing procedures and codes as in the previous chapters:

```
library(udpipe)
BE_prep_rake <- keywords_rake(x = BE_prep, term = "token", group =
  c("doc_id", "paragraph_id", "sentence_id"),
  relevant = BE_prep$upos %in% c("NOUN", "ADJ"),
  ngram_max = 4)
BE_prep_rake_freq <- subset(BE_prep_rake, freq >= 13)
BEp_prep_rake <- keywords_rake(x = BEp_prep, term = "token", group =
  c("doc_id", "paragraph_id", "sentence_id"),
  relevant = BEp_prep$upos %in% c("NOUN", "ADJ"),
  ngram_max = 4)
```

To measure document-document similarity we used the “text2vec” package. We defined two sets of documents, as previously stated, and evaluated their similarity:

```
library(text2vec)
BE1_1 = itoken(c(head((BE_prep_rake$keyword), 30),
  head((BE_prep_rake_freq$keyword), 30)),
  progressbar = FALSE)
BEp60 = itoken(head((BEp_prep_rake$keyword), 60), progressbar =
  FALSE)
```

To compare documents in a vector space we need to define a common space and project documents into it (this package uses vocabulary-based vectorization):

```
BE_prep_rake_BE1_1_BEp60<- c(head((BE_prep_rake$keyword), 30),
                              head((BE_prep_rake_freq$keyword), 30),
                              head((BEp_prep_rake$keyword), 60))
BE1_1_BEp60 = itoken(BE_prep_rake_BE1_1_BEp60, progressbar =
                     FALSE)
vBE1_1_BEp60 = create_vocabulary(BE1_1_BEp60)
vectorizerBE1_1_BEp60 = vocab_vectorizer(vBE1_1_BEp60)
```

To measure the similarity between the two sets of documents we used the function *sim2* (*x*, *y*, *method*), which calculates the “similarity between each row of matrix *x* and each row of matrix *y* using a given method” (Selivanov 2017). But first we had to create DTM matrices in the same vector space:

```
library(text2vec)
dtmBE1_1_BEp60 = create_dtm(BE1_1, vectorizerBE1_1_BEp60)
dtmBEp60_BE1_1 = create_dtm(BEp60, vectorizerBE1_1_BEp60)
```

The two methods that were used to measure document similarity between the Member’s bills and the electoral programs were the Jaccard Index and the Cosine function.

Jaccard Similarity:

```
library(text2vec)
BE1_1_BEp60_jac_sim = sim2(dtmBE1_1_BEp60, dtmBEp60_BE1_1, method
                           = "jaccard", norm = "none")
Jaccard3 <- BE1_1_BEp60_jac_sim
```

Cosine Similarity:

```
library(text2vec)
BE1_1_BEp60_cos_sim = sim2(dtmBE1_1_BEp60, dtmBEp60_BE1_1, method
                           = "cosine", norm = "l2")
Cosine3 <- BE1_1_BEp60_cos_sim
```

With the results:

```
> mean (Jaccard3@x)           > mean (Cosine3@x)
[1] 0.274                     [1] 0.446
```

The two sets of RAKE keywords chosen to measure document similarity are suitable because they faithfully represent the issues brought to the assembly of the republic by the Portuguese political parties and their respective electoral programs.

9 Conclusion

In view of the purpose of our dissertation – which is to contribute to the ideological clarification of a specific Portuguese parliamentary activity – we used several automatic keyword extraction methodologies in order to answer our research question: what are the major textual differences in the Member’s bills produced by the political parties in the 13th Legislature of Portugal?

According to the results of our study, we could established two main observations, answering positively to the first hypotheses formulated, as follows: *there are significant differences in the Member’s bills produced by the political parties in the 13th Legislature of Portugal, and those differences can be identified to a great extent by examining keywords in such documents.*

Grounded on baseline TF-IDF weighting with a level of sparseness of 95%, and the applied preprocessing techniques, we observed differences between the left-wing parties (BE, PCP and PEV) the centre parties (PS, PAN, PSD) and the right-wing party (CDS): while in all left-wing parties the words with the highest frequency were *work* (“trabalho”) and *employee* (“trabalhador”), with the third word varying between *education* (“educação”) in BE, *health* (“saúde”) in PCP and *animal* (“animal”) in PEV; as for the larger Portuguese parties - the centre parties PS and PSD - the term with the highest frequency was *municipality* (“freguesia”), alongside *administrative* (“administrativo”) and *union* (“união”) in PSD, *village* (“vila”) and *financial* (“financeiro”) in PS; as for the right-wing party CDS, issues related to *inland* (“interior”), *son* (“filho”) and *tax* (“imposto”) appeared to be the most important. For the PAN party, the word *animal* (“animal”) was the main theme, following by the words *work* (“trabalho”) and *child* (“criança”).

All these words also appeared in the keyword extraction technique based on POS tagging. Although this methodology was a good start, our aim was to identify multiword expressions in order to have more information about the issues that mattered most to the different Portuguese political parties.

As for the comparison between keyword extraction methods, according to our assessment, the RAKE algorithm and the methodology based on Co-occurrences delivered the best results.

In this work we presented the main differences regarding the first 30 keywords generated by each keyword extraction method in the Member's bills produced by the political parties during the last legislature (2015-2019). For instance, regarding labour issues, although the most frequent words in the left-wing parties' texts were "*work*" and "*employee*", there were some differences that should be noted: the keywords in BE Member's bills revealed concerns about "*working hours*", "*occasional contract of employment*", "*expiration of a collective agreement*", "*professional license*", "*long contributory careers*" and a strong association with the word "*amendment*"; PCP Member's bills exposed similar concerns about "*collective agreement*", "*working hours*" and "*long contributory careers*", but also focused on "*wage guarantee*" and "*labour accidents*". As for the PEV party, there were some common points of interest with PCP and BE ("*working hours*", "*expiration*", "*collective agreement*", "*labour accidents*" and "*amendment*") and some differences with the introduction of new topics such as "*national minimum wage*", "*termination of contracts of employment*" and "*workers' dismissal*", as well as the notion of "*huge setback at work conditions*" and the need of "*favourable treatment of the employee*".

As for the centre parties PS and PSD, and probably due to their significant involvement in municipal issues, local matters appeared to be the most important for both parties. However, we identified more differences than commonalities: the word "*municipality*" for PS is used regarding "*territorial and administrative boundary*", while PSD associates the term with "*territorial and administrative union*" and "*specific municipal problems of Alfena and Campo Sobrado*". The "*Douro Wine Company*", "*regional related*" and "*rural areas*" are terms that got some attention on PS's documents (as well as BE's and PCP's) while for PSD the "*autonomous region of Madeira*" and "*city council*" were more important. The only common position between the centre parties was the importance of "*local authority*". To the right-wing party CDS, the importance of "*inland*" is associated with "*municipal taxes*", "*municipal veterinarians*" and the "*greatest wealth in mainland coastal strip*".

Finally, regarding the newest party PAN, the Member's bills related to "*animal*" disclosed their thematic concern for social issues ("*animal protection*" and "*keep track*"), health issues ("*veterinary doctor*", "*death/sick animal*", "*animal suffering and slaughter*" and "*animal public service*"). As for the PEV party, the word "*animal*" is also important,

but it often appears related to “circus”. A detailed observation of the results obtained through keyword extraction is available (per political party and per topic) in the appendices.

Once we gathered the above information, as well as the different electoral programs presented in the last legislative elections, we tried to answer our last hypothesis: *Member’s bills produced by the political parties in the 13th Legislature of Portugal are related to the electoral programs presented in the legislative elections held on October 2015.* We applied the cosine similarity to the sets of RAKE keywords, which obtained good results in the comparisons between keyword extraction techniques used on the Member’s bills. Our aim was to compare the promises of political parties and what they actually accomplished through a specific Common Legislative Procedure - the Member’s bills.

The results of the cosine similarity method showed that the left-wing Portuguese parties had a higher degree of similarity between the Member’s bills and their electoral program (BE-0.446; PCP-0.449 and PEV-0.526) then other parties (PS-0.351 and PAN-0.301). Regarding the electoral program of the centre-right coalition, each party had different similarity results to the post-split Member’s bills (PSD-0.390 and CDS-0.337), indicating different commitment or problems in balance power over this political solution, specially designed for the elections held on October 2015.

Regarding possible points of proximity in the texts of the different parties, the results obtained through the document similarity methodology suggest, interestingly, a certain degree of similarity between the Member’s bills produced by the Portuguese political parties and the electoral program of PEV (between 0.436 and 0.540). The PAN electoral program seems to be the furthest from all the Member’s bills, perhaps emphasizing its non-ideological origin and the fact that it is a new party in the Portuguese political panorama.

Limitations and Future Work

We concede that there were some limitations in the methodology, namely in the document similarity analysis, in the use of beta version software package *text2vec* and also in the fact that we resorted to our own judgement to evaluate the extracted keywords. In our view, it would be beneficial, also, to use various keyword extraction methods and then combine the results. This could be done sequentially, using a kind of *cascade*

generalization (Gama and Brazdil 2000), or in parallel, by pinpointing the most frequent keywords identified by the different methods.

Coincidentally, a newspaper (Público 2019) recently published an article analyzing the frequency and word associations of the electoral programs presented in the upcoming legislative elections, to be held on October 2019. We hope that our work can help people make a more informed decision when called upon to choose a Government, and decrease abstention in Portugal.

We believe that further studies should be performed to increase the reliability of document similarity and keyword extraction analyses in the field of political science so that in the future we can assess - perhaps even automatically - the ideology of a document, newspaper or website.

Bibliography

- Aggarwal, C. and C. Zhai (2012), “A Survey of Text Clustering Algorithms”, in *Mining Text Data*, C. Aggarwal C. and C. Reddy (editors), pp. 29-60, Boston: Springer.
- Albrecht, R. and D. Merkl (1998), “Knowledge discovery in literature data bases”, *Library and information services in astronomy III*, ASP Conference Series, vol. 153.
- Alelyani S., J. Tang and H. Liu (2014), “Feature Selection for Clustering: A Review”, in *Data Clustering: Algorithms and Applications*, C. Aggarwal C. e C. Zhai (editors), pp. 77-128, Boca Raton: CRC Press.
- Argerich, L. (2016), “A Non-mathematical Approach to TextRank”, <https://www.quora.com/What-is-a-simple-but-detailed-explanation-of-TextRank>, last accessed 2019-01-13.
- Assembleia da República (2015), *Regimento da Assembleia da República*, Lisboa: Assembleia da República – Divisão de Edições.
- Assembleia da República (2019a), “Sobre as iniciativas legislativas”, <http://www.parlamento.pt/ActividadeParlamentar/Paginas/Iniciativas-sobre.aspx>, last accessed 2019-01-08.
- Assembleia da República (2019b), “Balanço da Atividade Parlamentar: Nota à Comunicação Social”, https://www.parlamento.pt/Documents/2019/julho/Balanco_Atividade_Parlamentar_CS_XIII_4_1.pdf, last accessed 2019-07-25.
- Assembleia da República (2019c), “Iniciativas”, <https://www.parlamento.pt/ActividadeParlamentar/Paginas/IniciativasLegislativas.aspx>, last accessed 2019-09-01.
- BNOSAC (2019), “Support in text mining”, <http://www.bnosac.be>, last accessed 2019-03-10.
- Campos, R., V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes e A. Jatowt (2018a), “A Text Feature Based Automatic Keyword Extraction Method for Single Documents”, in Pasi G., B. Piwowarski, L. Azzopardi e A. Hanbury A. (editors), *Lecture Notes in Computer Science - Advances in Information Retrieval - 40th*

European Conference on Information Retrieval (ECIR'18). Grenoble, France. March 26 – 29. (Vol. 10772 (2018), pp. 684 - 691).

- Campos, R., V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes e A. Jatowt (2018b), “YAKE! Collection-independent Automatic Keyword Extractor”, in Pasi G., B. Piwowarski, L. Azzopardi e A. Hanbury A. (editors), *Lecture Notes in Computer Science - Advances in Information Retrieval - 40th European Conference on Information Retrieval (ECIR'18)*. Grenoble, France. March 26 – 29. (Vol. 10772 (2018), pp. 806 - 810).
- Cordeiro, J. (2003), “Extracção de Elementos Relevantes em Texto/Páginas da World Wide Web”, Tese submetida à Faculdade de Ciências da Universidade do Porto para a obtenção do grau de Mestre em Inteligência Artificial e Computação, Faculdade de Ciências da Universidade do Porto.
- Dale, R. (2010), “Classical Approaches to Natural Language Processing”, in *Handbook of Natural Language Processing*, N. Indurkha e F. Damerau (editors), 2nd edition, pp. 3-7, Boca Raton: CRC Press.
- Feldman, R. and J. Sanger (2007), *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, New York: Cambridge University Press.
- Feldman, R., M. Fresko, H. Hirsh, Y. Aumann, O. Liphstat, Y. Schler and M. Rajman (1998), “Knowledge Management: A Text Mining Approach”, *Proceedings of the 2nd International Conference on Practical Aspects of Knowledge Management (PAKM98)*, Basel, Switzerland.
- Gama, J., A. Carvalho, K. Faceli, A. Lorena and M. Oliveira (2017), *Extração de Conhecimento de Dados: Data Mining*, 3ª edição, Lisboa: Edições Sílabo.
- Gama, J. and P. Brazdil (2000), “Cascade Generalization”, *Machine Learning*, N° 41, pp. 315-343.
- Gamallo, P., M. Garcia, C. Piñeiro, R. Martínez-Castaño and J. (2018), “LinguaKit: a Big Data-based multilingual tool for linguistic analysis and information extraction” In *Proceedings of The Second International Workshop on Advances in Natural Language Processing (ANLP 2018)* co-located at SNAMS-2018, pp. 239-244, <https://github.com/citiususc/Linguakit>, last accessed 2019-03-01.

- Ganesan, K. (2019), “How to extract keywords from text with TF-IDF and Python’s Scikit-Learn”, <https://www.freecodecamp.org/news/how-to-extract-keywords-from-text-with-tf-idf-and-pythons-scikit-learn-b2a0f3d7e667/>, last accessed 2019-07-12.
- Girard, J. P. and J. L. Girard (2015), "Defining Knowledge Management: Toward an Applied Compendium", *Online Journal of Applied Knowledge Management*. 3 (1): 14.
- Hack, A., L. Nunes, M. Silva and T. Lima (2013), “Text Mining: Data Mining-INE5644”, Universidade Federal de Santa Catarina.
- Jurafsky, D. and J. Martin (2018), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd Edition draft, <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>, last accessed 2019-07-07.
- Kilian, T. (2016), “Introduction to the KNIME Text Processing Extension”, in *Text Mining and Visualization: Case Studies Using Open-Source Tools*, M. Hofmann and A. Chisholm (editors), pp. 63-80, Boca Raton: CRC Press.
- Manning, C. and H. Schütze (1999), *Foundations of Statistical Natural Language Processing*, Cambridge: The MIT Press.
- Manning, C., P. Raghavan e H. Schütze (2008), *An Introduction to Information Retrieval*, online edition: Cambridge University Press.
- McCandless, D. and S. Posavec (2018), “Left vs Right (World)”, <https://informationisbeautiful.net/visualizations/left-vs-right-world/>, last accessed 2018-09-11.
- Mihalcea, R. and P. Tarau (2004), “TextRank: Bringing Order into Texts”, <https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf>, last accessed 2019-01-13.
- NLP-progress (2019), “Dependency parsing”, <http://nlpprogress.com>, last accessed 2019-07-07.
- Observador (2018), “Como vão ser os dias do mais recente deputado não inscrito. E como ficou o grupo que deixa para trás”, Observador, 14 de fevereiro, <https://observador.pt/2018/12/14/como-vaio-ser-os-dias-do-mais-recente->

deputado-nao-inscrito-e-como-ficou-o-grupo-que-deixa-para-tras/, last accessed 2019-07-13.

- Palmer, D. (2010), “Text Preprocessing”, in *Handbook of Natural Language Processing*, N. Indurkha e F. Damerau (editors), 2nd edition, pp. 9-30, Boca Raton: CRC Press.
- Polk, J., J. Rovny, R. Bakker, E. Edwards, L. Hooghe, S. Jolly, J. Koedam, F. Kostelka, G. Marks, G. Schumacher, M. Steenbergen, M. Vachudova and M. Zilovic (2017), "Explaining the salience of anti-elitism and reducing political corruption for political parties in Europe with the 2014 Chapel Hill Expert Survey data", *Research & Politics* (January-March): 1-9.
- Público (2019), “Não são só palavras. O que eles mais dizem no programa é aquilo em que mais pensam?”, <https://www.publico.pt/2019/09/06/politica/noticia/nao-sao-so-palavras-programa-pensam-1885647>, last accessed 2019-09-06.
- R-bloggers (2018), “An Overview of Keyword Extraction Techniques”, <https://www.r-bloggers.com/an-overview-of-keyword-extraction-techniques/>, last accessed 2019-09-01.
- Rose, S., D. Engel, N. Cramer and W. Cowley (2010), “Automatic keyword extraction from individual documents”, in *Text Mining: Applications and Theory*, M. Berry and J. Kogan (editors), pp. 3-20, United Kingdom: John Wiley & Sons, Ltd.
- Russel, S and P. Norvig (1995), *Artificial Intelligence: A Modern Approach*, New Jersey: Prentice Hall.
- Selivanov, D. (2017), “Documents Similarity”, <http://text2vec.org/similarity.html>, last accessed 2019-09-05.
- Sieg, A. (2018), “Text Similarities : Estimate the degree of similarity between two texts”, <https://medium.com/@adriensieg/text-similarities-da019229c894>, last accessed 2019-09-05.
- Sites.google.com (2019), “Processamento de Linguagem Natural (PLN)”, <https://sites.google.com/site/inteligenciaartificialist/home>, last accessed 2019-01-09.

- Ventura, J. (2014), “Automatic Extraction of Concepts from Texts and Applications”, Dissertação para obtenção do Grau de Doutor em Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.
- Wan, X. and J. Xiao (2008), “Single document keyphrase extraction using neighborhood knowledge”, AAAI Conference on Artificial Intelligence 2008, 13–17 July, pp. 855–860.
- Wijffels, J. (2019), Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit”, version 0.8.3, <https://cran.r-project.org/web/packages/udpipe/udpipe.pdf#page=31&zoom=100,0,89>, last accessed 2019-08-04.
- Wilbur, J. e K. Sirotkin (1992), “The Automatic Identification of Stopwords”, J. Inf. Sci., 18, pp. 45–55.

Appendices

A. Automatic Keyword Extraction Results – BE

BE – Left Bloc (“Bloco de Esquerda”)							
ISSUES	KEYWORD TECHNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
ECONOMIC							
• económico		x					
• financeiro		x					
• empresa		x					
• administração		x					
• fiscal		x					
• capitalismo financeirização			x				
• atividades económicas secundárias compatíveis					x		
• práticas concorrenciais agressivas					x		
• tecido empresarial nacional					x		
• profundas crises económicas					x		
• sucessivas crises bancárias					x		
• sistema financeiro português frágil					x		
• domínio público hídrico					x		
• ativos fixos tangíveis					x		
• sistema nacional gás natural					x		
• gestão técnico global sen/sist. elétrico nacional					x		
• aproveitamento racional recurso natural					x		
• gestão meio produção comunitário					x		
• serviços mínimos bancários					x		
• paraísos fiscais					x		
• receitas fiscais					x		
• rede ferroviária					x		
• novo banco						x	
• indiscutível banif							x
LABOUR							
• trabalho		x					
• trabalhador		x					
• profissional		x					
• laboral		x					
• trabalho contrato trabalhador			x	x		x	
• trabalho temporário / termo / tempo			x	x	x	x	
• trabalho horário / hora /banco			x		x	x	
• trabalho empresa / empregador			x	x	x	x	
• trabalho coletivo convenção / regulamentação/ caducidade/ contratação			x	x	x	x	
• relação laboral			x	x	x		
• períodos contratuais mínimos excessivos					x		
• falsos recibos verdes					x		
• longuíssimas carreiras contributivas					x		
• cédula profissional					x		
• trabalho suplementar					x		

BE – Left Bloc (“Bloco de Esquerda”)							
ISSUES	KEYWORD TECNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
SOCIAL							
• cuidador informal			X		X		
• violência doméstico			X	X			
• segurança social			X	X	X		
• subsídio social parental inicial					X		
• comprovada carência económica					X		
• retribuição mínima mensal garantida					X		
• comunicação eletrónicas acessível público					X		
• seres humanos					X		
PUBLIC ADMINISTRATION							
• público função			X	X			
• público político cargo titular altos			X		X	X	
• público entidade empregador			X				
• transferências de competências judiciais					X		
• finanças público crise bancário					X		
• institutos públicos					X		
• políticas públicas					X		
• pessoa coletivo público					X		
• autoridade tributário					X		
HEALTH							
• saúde / área / nacional		X			X	X	
• procriação medicamente assistido			X				
• segurança alimentar nutricional adequado			X	X	X		
• gestação substituição			X	X		X	
• taxa moderador			X				
• público / saúde profissional			X		X	X	
• cuidado saúde				X		X	
• medicina tradicional chinesa					X		
• exposição campo elétrico magnético					X		
• órgão gestão hospitalar					X		
• cultivo importação comercialização organismo					X		
IDEOLOGICAL							
• constitucional		X					
• governo psd cds			X	X			
• constituição portuguesa			X	X			
• complacência permissibilidade capitalismo			X				
• capitalismo terrorismo			X				
• único escândalo bancário					X		
• trabalho novo ciclo político					X		
• público financiamento partido político					X		
• tribunal constitucional					X		
• favorável trabalhador						X	
• duvidosa constitucionalidade							X
EDUCATION							
• ensino		X					
• ensino superior			X	X	X		
• ensino superior público regulamentação/ acesso					X		

BE – Left Bloc (“Bloco de Esquerda”)							
ISSUES	KEYWORD TECNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
ENVIRONMENTAL							
• florestal		x					
• sistema ecológico produção alimento					x		
• icnf ip (Inst. Conservação Natureza e Florestas)					x		
EUROPEAN							
• europeu		x					
• união europeu			x	x			
LOCAL							
• câmara municipal			x		x		
• casa douro			x	x		x	
• região autónomo			x				
• autarquia local			x		x		
• Metro do Porto					x		
• regional vitivinicultor						x	
HOUSING							
• habitação		x					
• imóvel comunitário			x				
SECURITY							
•							

B. Automatic Keyword Extraction Results – CDS-PP

CDS-PP - Social Democratic Centre - Popular Party ("Centro Democrático Social - Partido Popular")							
ISSUES	KEYWORD TECNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
ECONOMIC							
• rendimento		x					
• fiscal		x					
• financeiro		x					
• económico		x					
• bpp funchal			x				
• confederação associação empresarial sector					x		
• produto petrolífero energético continente					x		
• crédito fiscal extraordinário investimento					x		
• carga fiscal					x		
• imposto sobre rendimento						x	
• fosso prosperidade							x
LABOUR							
• trabalho		x					
• situação incompatibilidade exercício profissão					x	x	
SOCIAL							
• filho		x					
• natalidade		x					
• social		x					
• idoso		x					
• jovem		x					
• parental		x					
• filho natalidade			x				
• idade população idoso índice envelhecimento			x	x	x	x	x
• idade ativo / renovação população			x		x	x	
• segurança social / serviço nacional			x	x			
• licença parental inicial / subsídio			x		x		
• cuidador informal			x				
• primeiro filho / idade tardio			x	x	x	x	
• assistente social				x			
• adiamento maternidade/ decréscimo				x	x		
• indexante apoio social					x		
• retribuição mínima mensal garantida					x		
• saldos migratórios negativos					x		
• particularidade risco fragilidade idoso					x		
• índice sintético fecundidade						x	
PUBLIC ADMINISTRATION							
• tributário		x					
• INE		x			x		
• tribunal administrative tributário			x				
• proteção jurídico			x	x			
• apoio judiciário			x				
• orçamento estado				x		x	
• défice crónico					x		

CDS-PP - Social Democratic Centre - Popular Party ("Centro Democrático Social - Partido Popular")							
ISSUES	KEYWORD TECNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
HEALTH							
• saúde / área		x		x			
• saúde serviço social profissional			x			x	
• saúde nacional publico			x	x		x	
• cuidado saúde				x		x	
• majoração subsídio doença					x		
• serviço nacional saúde						x	
IDEOLOGICAL							
• governo		x					
• político		x					
• português		x					
• constitucional		x					
• filho natalidade político			x				
• cor étnico			x				
• fertilidade / dilema / listagem			x				
• nação unido						x	
• obrigatório pai							x
• justo sociedade							x
EDUCATION							
• estatuto ensino particular cooperativo					x		
• sistema nacional educação sne					x		
• ensino básico secundário					x		
• ensino profissional privado					x		
ENVIRONMENTAL							
•							
EUROPEAN							
• agenda política europeia					x		
LOCAL							
• imposto municipal				x			
• faixa litoral continente riqueza					x		
• médicos veterinários municipais					x		
HOUSING							
• habitação própria permanente					x		
SECURITY							
• segurança anpc entidade					x		
• vítima responsabilidade							x

C. Automatic Keyword Extraction Results – PAN

PAN - Party for People, Animals and Nature (“Pessoas-Animais-Natureza”)							
ISSUES	KEYWORD TECNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
ECONOMIC							
• económico		X					
• concorrencial gpl			X				
• azeitona olival			X				
• zona caça					X	X	
• grandes superfícies comerciais					X		
LABOUR							
• trabalho		X					
• trabalho contrato alteração			X		X	X	
• entidade empregador					X	X	
• salário mínimo nacional mensal					X		
SOCIAL							
• animal		X					
• violência		X					
• criança / jovem		X		X			
• consumidor		X					
• cuidador informal			X	X			
• matriz braille / boletim voto			X	X	X	X	
• pessoa idoso			X	X	X		
• tauomático actividade			X		X		
• violência doméstico			X	X	X		
• licença parental responsabilidade/ inicial			X	X	X		
• ovário / nascer/ mama			X				
• instituição particular solidariedade social					X		
• comandante corpo bombeiro					X		
• envolvimento parental igualitário					X		
• protecção criança						X	
• tema occisão							X
• censurável cão							X
• animal monitorização							X
• protegido animal							X
PUBLIC ADMINISTRATION							
• estado		X					
• Governo		X					

PAN - Party for People, Animals and Nature (“Pessoas-Animais-Natureza”)							
ISSUES	KEYWORD TECNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
HEALTH							
• saúde / profissional		x			x	x	
• veterinário / médico		x			x	x	
• morte medicamente assistido			x			x	
• animal veterinário medico doente morte			x				
• animal público serviço			x				
• animal /alimentação /sofrimento/matadouro /companhia			x	x	x x	x x	x
• saúde /publico /privado			x	x			
• hiperactividade / metilfenida			x				
• cédula acss			x				
• produto fitofarmacêuticos				x			
• autoridade sanitário veterinário nacional					x		
• animal gás							x
• humano carcinogénio							x
• imunológico protecção							x
• letal abandono							x
IDEOLOGICAL							
• constitucional		x					
• politico partido			x				
• eclesiástico /papa							
• constituição portuguesa				x	x		
• uniforme país							x
EDUCATION							
• ensino superior			x	x			
• instituição ensino						x	
ENVIRONMENTAL							
• resíduo urbano			x	x	x		
• organismo geneticamente modificar			x				
• desperdício alimentar					x		
• extrema riqueza genética vegetal							
• Autoridade Florestal Nacional					x		
• calendário venatório (caça)					x		
EUROPEAN							
• europeu		x					
• união europeu			x	x	x		
• directivo parlamento europeu					x		
LOCAL							
• medico veterinário municipal			x	x			
• câmara municipal				x			
HOUSING							
•							
SECURITY							
• crime		x					x
• mau trato			x		x	x	
• é punido pena prisão			x	x		x	
• autoridade autuante					x		
• polícia criminal					x		
• coima sanção						x	
• ineficiente cadeia							x

D. Automatic Keyword Extraction Results – PCP

PCP - Portuguese Communist Party ("Partido Comunista Português")							
ISSUES	KEYWORD TECHNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
ECONOMIC							
• empresa		x					
• financeiro		x					
• acesso mobilidade intermodalidade			x				
• imposto produto petrolífero			x				
• off - shore			x				
• determinante estrutura empresarial nacional					x		
• plano regional ordenamento território					x		
• micro pequeno médio empresa					x		
• serviço mínimo bancário					x		
• indústrias biotecnológicas agroalimentares					x		
• grandes grupos económicos nacionais					x		
• paraísos fiscais					x		
• gratuito taxa							x
LABOUR							
• trabalho		x					
• trabalhador		x					
• garantia salarial					x		
• horário trabalho /tempo					x	x	
• longas carreiras contributivas					x		
• trabalho / contrato /condição /função/ coletivo						x	
• trabalho / acidente						x	
SOCIAL							
• social / nacional		x		x			
• população perante encarecimento			x				
• passe social intermodal			x				
• maternidade paternidade			x				
• trabalho /segurança/ contrato /alteração			x				
• trabalho trabalhador			x				
• serviço segurança social ação			x				
• telespectador ouvinte			x				
• arquivo musical			x				
• segurança social				x	x		
• retribuição mínima mensal garantida					x		
• subsídio parental inicial exclusivo					x		
PUBLIC ADMINISTRATION							
• estado		x					
• governo membro		x	x				
• função público			x				
• orçamento estado				x		x	
• serviço estrangeiro fronteiro					x		
• guarda nacional republicano					x		
• estado capital estatutário rtp					x		
• institutos públicos					x		

PCP - Portuguese Communist Party ("Partido Comunista Português")							
ISSUES	KEYWORD TECNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
HEALTH							
• saúde		x					
• hospital		x					
• serviço alimentação veterinário dsav			x				
• saúde serviço público trabalhador			x				
• internato médico			x	x		x	
• bactéria legionella			x				
• serviço nacional saúde				x			
• medicamento antipsicóticos serviço nacional					x		
• médico trabalho					x		
• autoridade sanitária veterinária nacional					x		
• saúde segurança trabalho						x	
• prestação cuidado saúde						x	
IDEOLOGICAL							
• público / serviço		x		x			
• político		x					
• privado		x					
• coletivo		x					
• revogar		x					
• governo psd cds				x			
• interesse público					x		
• grandes gigantes bancários					x		
• entidades patronais pressões diretas					x		
• serviço público postal					x		
• público contrato							x
• único escolha							x
• grave despedimento							x
EDUCATION							
• ensino / instituição		x				x	
• escolar		x					
• agrupamento escola			x		x		
• superior ensino instituição			x	x	x		
• manual escolar				x			
• pedagógico ensino básico ensino secundário					x		
• físico desportivo escolar popular					x		
• vencimento anual médio docente					x		
• ensino superior particular cooperativo					x		
• instituto politécnico					x		
• limitação ação social escolar					x		
• âmbito ação social escolar					x		
• necessidades educativas especiais					x		
• formação inicial					x		
ENVIRONMENTAL							
• construção perfuração sondagem			x				
• agricultura convencional e/ou biológica					x		
• produção agroalimentar sadia					x		
EUROPEAN							
•							

PCP - Portuguese Communist Party ("Partido Comunista Português")							
ISSUES	KEYWORD TECHNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
LOCAL							
• município		x					
• casa dourado			x	x			
• município competência territorial / área			x	x	x	x	
• local autarquia			x		x		
• união freguesia poceirão marateca					x		
• autoridade sanitária veterinária concelhia					x		
• passe social intermodal/Andante					x		
• Metro do Porto					x		
• médicos veterinários municipais					x		
• câmaras municipais					x		
HOUSING							
•							
SECURITY							
• elemento policial equipado			x	x			
• criminal sede				x			
• apoio judiciário					x		

E. Automatic Keyword Extraction Results – PEV

PEV - Ecologist Party-“The Greens” (“Partido Ecologista "Os Verdes"")							
ISSUES	KEYWORD TECNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
ECONOMIC							
• empresa		X					
• consumidor		X					
• económico		X					
• época balnear			X		X		
• submarino estaleiro naval			X				
• grave prejuízo sector táxi					X		
LABOUR							
• trabalhador / trabalho		X					
• trabalho enorme recuo			X				
• trabalho trabalhador / período /hora / alteração			X	X	X	X	
• trabalhador tratamento favorável			X		X	X	
• relação laboral			X	X			
• caducidade diuturnidade			X				
• período normal trabalho referência					X		
• instrumento regulamentação coletivo trabalho					X		
• cessação contrato trabalho despedimento					X		
• salário mínimo nacional					X		
• acidente trabalho						X	
SOCIAL							
• animal		X					
• social		X					
• grande tradição carnavalesco			X				
• atleta alto competição			X				
• carnaval feriado			X	X	X	X	
• animal circo			X				
• pessoa mobilidade / reduzido				X		X	
• idade normal acesso pensão					X		
• especial acesso pensão invalidez					X		
• pensão velhice beneficiário geral					X		
• longas carreiras contributivas					X		
• licença parental inicial					X		
PUBLIC ADMINISTRATION							
• governo		X					
• público		X					
• função público trabalho contrato / alteração			X	X	X		
• administração público				X			
• orçamento estado				X		X	
• serviço público				X	X		
• ordenamento jurídico					X		
• governo orçamento							X

PEV - Ecologist Party-“The Greens” (“Partido Ecologista "Os Verdes"")							
ISSUES	KEYWORD TECNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
HEALTH							
• saúde		x					
• doente		x					
• taxa moderador			x	x	x		
• médico doente comissão/ verificação			x				
• insalubridade penosidade risco			x	x	x	x	
• observação sintoma			x				
• subfinanciamento serviço nacional saúde					x		
• doença mental psíquico incapacitar					x		
• doença profissional					x		
• estabelecimento saúde						x	
• cuidado saúde						x	
• doente médico							x
IDEOLOGICAL							
• constitucional		x					
• correio ctt encerrado			x				
• psd cds			x				
• constituição portuguesa			x	x			
• governo psd				x			
• trabalho poder				x		x	
• concorrência desleal					x		
• parado país							x
EDUCATION							
• escolar		x					
• ensino superior			x				
• politécnico universitário			x				
• necessidades educativas especiais					x		
• ensino obrigatório					x		
• manuais escolares					x		
ENVIRONMENTAL							
• ambiental		x					
• ecologista		x					
• gás efeito estufa			x				
• organismo geneticamente modificar			x			x	
• recurso florestal			x		x		
• campo golfe / ambiental gestão			x	x	x		
• sustentabilidade fator			x			x	
• raposa correção caça rabo saca vulpes			x				
• resíduo embalagem /produto				x			x
• melhor desempenho ambiental sociedade					x		
• ração transgénico união europeu					x		
• águas residuais urbanas					x		
• olival super intensivo					x		
• bebida vegetal alternativa					x		
• recursos naturais					x		
• impacte ambiental					x		
• espécie protegido						x	

PEV - Ecologist Party-“The Greens” (“Partido Ecologista "Os Verdes"”)							
ISSUES	KEYWORD TECHNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
EUROPEAN							
• europeu		x					
• união europeu			x	x	x		
LOCAL							
• autarquia local respetiva regulamentação					x		
HOUSING							
•							
SECURITY							
• grave violação							x

F. Automatic Keyword Extraction Results – PS

PS - Socialist Party ("Partido Socialista")							
ISSUES	KEYWORD TECNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
ECONOMIC							
• financeiro		x					
• gás petróleo liquefazer			x				
• bancário setor /comissão inquérito			x	x	x		
• mobiliário valor			x	x			
• intermediário financeiro instrumento			x	x	x		
• cae diversão itinerante				x			
• estabelecimento entidade fim lucrativo					x		
• instituição crédito						x	
LABOUR							
• trabalho / contrato		x				x	
• profissão / exercício		x		x		x	
• profissional / atividade		x	x				
• desportivo contrato trabalho			x	x	x		
• profissional profissão /exercício ordem			x	x			
• profissional membro direção órgão conselho			x	x			
• corporação ator				x			
• conselho geral vitivinicultor					x	x	
• corporações profissionais públicas					x		
• salário mínimo nacional					x		
• reduções remuneratórias					x		
• entidade empregador						x	
SOCIAL							
• proteção		x					
• social		x					
• desportivo		x					
• antecipação morte			x				
• praticante desportivo formação			x	x			
• assistente social			x	x	x		
• circo cativo cicense				x			
• acucar teor				x			
• atividade profissional território nacional					x		
• exercício parentalidade					x		
• conselho económico social					x		
• liberdade autodeterminação sexual menor					x		
• património azulejar português					x		
• comunicação social audiovisual					x		
• urgente filho							x
PUBLIC ADMINISTRATION							
• estado		x					
• público		x					
• eleitoral		x					
• Governo		x					
• ordenamento território			x				
• entidade público politico cargo titular			x		x	x	
• geral trabalho função público					x		
• transferências de competências judiciais					x		
• orçamento estado						x	

PS - Socialist Party (“Partido Socialista”)							
ISSUES	KEYWORD TECHNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
HEALTH							
• saúde /profissional		x				x	
• terapeuta fala			x				
• anatomia patológico citológico			x				
• bem - estar			x				
• prótese dentário			x				
• estratégia político nacional saúde					x		
• procriação medicamente assistido						x	
IDEOLOGICAL							
• poder		x					
• político		x					
• constitucional / tribunal / anonimato		x		x			x
• privado		x					
• racial étnico			x				
EDUCATION							
• Informação / dever		x				x	
• ensino exercício profissão organização					x		
• língua oficial portuguesa					x		
ENVIRONMENTAL							
• gás efeito estufa			x				
EUROPEAN							
• europeu		x					
• europeu união			x				
• diretivo ce parlamento europeu					x		
• europeu espaço económico					x		
• carta social europeu revisto					x		
• processo legislativo europeu					x		
•							
LOCAL							
• local		x					
• regional		x					
• Douro casa		x		x			
• estado região autónomo autarquia			x	x		x	
• sub-região			x				
• autarquia local			x	x			
• regeneração bairro informal					x		
• territorial limite administrativo freguesia					x		
• respetivos espaços rurais					x		
HOUSING							
• habitação / nacional		x		x	x	x	
• habitacional		x					
• público habitação/ politico/ local			x			x	
• parque habitacional apoio público					x		
• património imobiliário público					x		
SECURITY							
• pena prisão			x			x	
• infração disciplinar				x			x
• público crime							x
• punível eutanásia							x

G. Automatic Keyword Extraction Results – PSD

PSD - Social Democratic Party (“Partido Social Democrata”)							
ISSUES	KEYWORD TECHNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
ECONOMIC							
• Financeiro / sistema		X			X		
• económico		X					
• falsificação contrafação			X				
• banco portugal / função consultivo			X	X	X		
• concorrência autoridade / infração			X			X	
• compra aeroporto			X				
• adc anteprojeto / autoridade concorrência			X	X			
• cae diversão			X				
• instituição financeiro					X		
• veículos automóveis ligeiros					X		
• tecido empresarial nacional					X		
• transportadoras aéreas					X		
LABOUR							
• trabalho / contrato		X	X			X	
• profissional		X					
• praticante desportivo contrato trabalho			X		X		
• arquitecto directiva cee			X				
• estimular emprego científico tecnológico					X		
• salário mínimo nacional					X		
• entidade empregadora					X		
SOCIAL							
• proteção / dado		X				X	
• social		X					
• desportivo / competição / época		X			X		
• civil		X					
• morrer mulher assassinado			X				
• violência racismo xenofobia			X				
• cuidador informal			X		X		
• subsídio social mobilidade			X		X	X	
• violência doméstico / crime			X	X			
• praticante desportivo			X		X		
• remuneração mínimo nacional anual					X		
• segurança social					X		
• isento voluntário							X
PUBLIC ADMINISTRATION							
• estado		X					
• Governo/ membro		X	X				
• imposto valor acrescentado			X				
• perda mandato demissão			X				
• função pública / serviço			X				
• titular cargo público / alto			X	X	X	X	
• ministério público							
• ip / instituto público				X		X	
• novo regime jurídico					X		
• divisão administrativa					X	X	
• orçamento estado						X	

PSD - Social Democratic Party (“Partido Social Democrata”)							
ISSUES	KEYWORD TECNIQUE	POS Tags	Colloc. Co-occ.	Text Rank	RAKE	POS Phrase	Depend. Parsing
HEALTH							
• saúde / cuidado / área		x		x	x	x	
• bem-estar			x				
• inmlcf ip			x				
• prestação saúde serviço nacional			x	x			
• adn perfil base			x	x		x	
• genético biológico							x
IDEOLOGICAL							
• público		x					
• autoridade		x					
• constitucional / tribunal		x			x		
• político		x					
• estrangeiro/ residente		x	x	x			
• nacionalidade portuguesa			x				
• bioética cnevc			x				
• constituição portuguesa				x			
• território português filho estrangeiro					x		
• caderno eleitoral existente					x		
EDUCATION							
• informação		x					
• estatuto carreiro docente universitário					x		
• necessidades educativas especiais					x		
• social educação							x
ENVIRONMENTAL							
•							
EUROPEAN							
• europeu / união / membro		x	x	x	x		
LOCAL							
• local / autarquia		x	x	x			
• valongo freguesia alfena			x				
• freguesia campo sobrado			x				
• concelho santo tirso			x				
• região autónomo madeira			x				
• união freguesia / territorial				x		x	
• nível nacional regional local					x		
• câmara municipal					x		
HOUSING							
• enorme alojamento							x
• igual arrendamento							x
SECURITY							
• problema investigação criminal					x		
• ministério público					x		
• autoridade judiciária competente					x		
• pena prisão / alta / inferior						x	x
• punível tentativa							x