UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Extractive document summarization using complex networks

Jorge Andoni Valverde Tohalino

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Jorge Andoni Valverde Tohalino

Extractive document summarization using complex networks

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Diego Raphael Amancio

USP – São Carlos
August 2018

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

V215e

Valverde Tohalino, Jorge Andoni
Extractive document summarization using complex networks / Jorge Andoni Valverde Tohalino; orientador Diego Raphael Amancio. -- São Carlos, 2018.
116 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2018.

1. Automatic summarization. 2. Complex networks. 3. Natural Language Processing. 4. Artificial intelligence. I. Amancio, Diego Raphael, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Jorge Andoni Valverde Tohalino

Sumarização extrativa de documentos usando redes complexas

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Diego Raphael Amancio

USP – São Carlos
Agosto de 2018

Este trabalho é dedicado
aos meus queridos pais
pela sua ajuda e amor incondicional

ACKNOWLEDGEMENTS

Agradeço primeiramente a Deus, pelo seu amor infinito, paciência e pelas diferentes oportunidades que ele me deu para continuar.

Agradeço a minha família, epecialmente a meus pais Jorge Antonio e Juana Irenia, pelo seu grande apoio e compreensão nestes dois anos que eu estive longe de casa. Todo esse trabalho é dedicado a vocês meus queridos pais. Eu amo muito vocês! Eu também dedico esse trabalho de mestrado para minha pequena sobrinha Maria José, que desde o seu nascimento foi uma razão de muitas alegrias para mim e minha família.

Aos meus amigos peruanos, que me apoiaram muito em suas mensagens de incentivo e motivação durante todo esse período que eu estive longe do Peru. Agradeço especialmente à minha segunda mãe, Glynis, por todas as suas palavras de orientação e encorajamento durante todo esse tempo.

Ao meu orientador Dr. Diego Raphael Amancio, pelo seu apoio, confiança e todos os ensinamentos ao longo deste projeto. Também aprecio sua paciência e sua disposição em me ajudar em qualquer momento.

Aos meus colegas e amigos do NILC, pelo apoio e aprendizado que tive com vocês.

Aos meus amigos da Igreja do Nazareno de São Carlos, por sua amizade, orações, conselhos e apoio incondicional durante minha estadia em São Carlos. Especialmente ao Pr. Rodrigo, Pr. Orivaldo, Pr. Victor, Pra. Lucineia, Maria Helena e Thiago. Muito obrigado família Nazarena.

Ao CNPq e FAPESP pelo apoio financeiro durante este trabalho de mestrado.

Ao ICMC e à USP por me darem a oportunidade de estudar e crescer como pessoa através deste trabalho de mestrado.

“Posso todas as coisas em Cristo que me fortalece. ”
(Filipenses 4:13)

RESUMO

VALVERDE, J. A. **Sumarização extrativa de documentos usando redes complexas**. 2018. 116 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Devido à grande quantidade de informações textuais disponíveis na Internet, a tarefa de sumarização automática de documentos ganhou importância significativa. A sumarização de documentos tornou-se importante porque seu foco é o desenvolvimento de técnicas destinadas a encontrar conteúdo relevante e conciso em grandes volumes de informação sem alterar seu significado original. O objetivo deste trabalho de Mestrado é usar os conceitos da teoria de grafos para o resumo extrativo de documentos para Sumarização mono-documento (SDS) e Sumarização multi-documento (MDS). Neste trabalho, os documentos são modelados como redes, onde as sentenças são representadas como nós com o objetivo de extrair as sentenças mais relevantes através do uso de algoritmos de ranqueamento. As arestas entre nós são estabelecidas de maneiras diferentes. A primeira abordagem para o cálculo de arestas é baseada no número de substantivos comuns entre duas sentenças (nós da rede). Outra abordagem para criar uma aresta é através da similaridade entre duas sentenças. Para calcular a similaridade de tais sentenças, foi usado o modelo de espaço vetorial baseado na ponderação Tf-Idf e word embeddings para a representação vetorial das sentenças. Além disso, fazemos uma distinção entre as arestas que vinculam sentenças de diferentes documentos (inter-camada) e aquelas que conectam sentenças do mesmo documento (intra-camada) usando modelos de redes multicamada para a tarefa de Sumarização multi-documento. Nesta abordagem, cada camada da rede representa um documento do conjunto de documentos que será resumido. Além das medições tipicamente usadas em redes complexas como grau dos nós, coeficiente de agrupamento, caminhos mais curtos, etc., a caracterização da rede também é guiada por medições dinâmicas de redes complexas, incluindo simetria, acessibilidade e tempo de absorção. Os resumos gerados foram avaliados usando diferentes corpus para Português e Inglês. A métrica ROUGE-1 foi usada para a validação dos resumos gerados. Os resultados sugerem que os modelos mais simples, como redes baseadas em Noun e Tf-Idf, obtiveram um melhor desempenho em comparação com os modelos baseados em word embeddings. Além disso, excelentes resultados foram obtidos usando a representação de redes multicamada de documentos para MDS. Finalmente, concluímos que várias medidas podem ser usadas para melhorar a caracterização de redes para a tarefa de sumarização.

Palavras-chave: Summarização Automática, Redes Complexas, Processamento de Linguagem Natural, Inteligência Artificial.

ABSTRACT

VALVERDE, J. A. **Extractive document summarization using complex networks**. 2018. 116 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Due to a large amount of textual information available on the Internet, the task of automatic document summarization has gained significant importance. Document summarization became important because its focus is the development of techniques aimed at finding relevant and concise content in large volumes of information without changing its original meaning. The purpose of this Master's work is to use network theory concepts for extractive document summarization for both Single Document Summarization (SDS) and Multi-Document Summarization (MDS). In this work, the documents are modeled as networks, where sentences are represented as nodes with the aim of extracting the most relevant sentences through the use of ranking algorithms. The edges between nodes are established in different ways. The first approach for edge calculation is based on the number of common nouns between two sentences (network nodes). Another approach to creating an edge is through the similarity between two sentences. In order to calculate the similarity of such sentences, we used the vector space model based on Tf-Idf weighting and word embeddings for the vector representation of the sentences. Also, we make a distinction between edges linking sentences from different documents (inter-layer) and those connecting sentences from the same document (intra-layer) by using multilayer network models for the Multi-Document Summarization task. In this approach, each network layer represents a document of the document set that will be summarized. In addition to the measurements typically used in complex networks such as node degree, clustering coefficient, shortest paths, etc., the network characterization also is guided by dynamical measurements of complex networks, including symmetry, accessibility and absorption time. The generated summaries were evaluated by using different corpus for both Portuguese and English language. The ROUGE-1 metric was used for the validation of generated summaries. The results suggest that simpler models like Noun and Tf-Idf based networks achieved a better performance in comparison to those models based on word embeddings. Also, excellent results were achieved by using the multilayered representation of documents for MDS. Finally, we concluded that several measurements could be used to improve the characterization of networks for the summarization task.

Keywords: Automatic Summarization, Complex Networks, Natural Language Processing, Artificial Intelligence.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

DT	Decision Tree
DUC	Document Understanding Conferences
Idf	Inverse document frequency
MDS	Multi Document Summarization
MLN	Multilayer Networks
NB	Naive Bayes
NILC	Núcleo Interinstitucional de Linguística Computacional
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
POST	Part Of Speech Tagging
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SDS	Single Document Summarization
SVM	Support Vector Machine
Tf	Term frequency
Tf-Idf	Term frequency - Inverse document frequency
VSM	Vector Space Models

CONTENTS

CHAPTER

1

INTRODUCTION

The amount of information available on the Internet has steadily increased over the last few years. A considerable part of this information is available as textual documents. Blog posts, emails, Facebook, Twitter, online news and scientific articles are some examples of text-oriented information. In one day, for instance, 2 million blog posts are written, 92 thousand newspapers are posted on the web and more than 400 million tweets are sent on Twitter (SPENCER, 2016). The manipulation and understanding of such a huge volume of textual data is a very difficult task for users because it demands a lot of time and resources. Due to this great increase of textual information, the need to use automatic methods to understand, index, classify and show the information in a clear and concise way arises, in this sense, it has become an important task in the natural language processing area (RIBALDO *et al.*, 2012).

Automatic document summarization techniques are one of the many solutions addressing the problem of managing large volumes of data (ANTIQUEIRA *et al.*, 2009; RIBALDO *et al.*, 2012). Document summarization is defined as the process of automatically creating a compressed version of one or more texts by extracting their most relevant or important content (FERREIRA *et al.*, 2013a). These summaries must present the information in a condensed way, maintaining consistency in the document and avoiding redundancy (LEE; BELKASIM; ZHANG, 2013a). When a summary of a single source document is generated, the process is called Single Document Summarization (SDS). On the other hand, the Multi Document Summarization (MDS) task is applied when a single summary is produced from a set of source documents that share the same topic (MANI, 2001). In recent years, document summarization has become important because it can be useful in a wide variety of tasks (NENKOVA; MCKEOWN, 2012). For example, automatic summarization is very effective for evaluating the relevance of a topic in news articles, because it was found that summaries with 17% length in relation to the full text accelerate the decision making based on topic relevance (NENKOVA; MCKEOWN, 2012). Query-based summarization is useful for determining the relevance or importance of retrieved documents. This type of summarization allows users to find relevant and reliable documents,

therefore, there is a minor need to consult the full text (TOMBROS; SANDERSON, 1998). For document indexation, summaries are useful for results displaying in a document search tool, because the summary is seen instead of accessing the entire document. Summaries also help in making purchasing decisions for books. For the MDS task, summaries are useful for a better understanding and organization of a set of news or articles sharing the same issue or subject. In this sense, MDS methods help to optimize retrieval time, especially when the user goal is to find as much information as possible about some topic (HIRSCHBERG *et al.*, 2005). For example, in Maña-López, Buenaga and Gómez-Hidalgo (2004), a task was given to a group of people for producing reports about some topic. Each people received several grouped news and the corresponding summary. Researchers concluded that people tend to write better reports when the summaries are provided.

Automatic summarization techniques are divided into two major groups: extractive summarization and abstractive summarization. Extractive summarization aims to generate summaries by joining the most relevant sentences, which are included without any modification. On the other hand, abstractive summarization is a more challenging task since it involves paraphrasing sections of the source document and may reuse clauses or phrases from such texts (NENKOVA; MASKEY; LIU, 2011). Systems that generate abstracts require more sophisticated resources and tools with the aim of inferring the meaning of source documents as well as language generators to compose the final summary (ANTIQUEIRA *et al.*, 2009). Extractive summarization systems do not require linguistic knowledge to select the most relevant information for the extract. Most of the works found in the literature adopt the extractive summarization task (MANI, 2001). Sentence extraction approaches for extractive document summarization are mainly divided into the following techniques (NENKOVA; MCKEOWN, 2012): word frequency based methods, sentence clustering methods, graph-based methods, and machine learning approaches. This research focuses on graph-based and complex network methods for extractive summarization.

In recent years, studies in complex networks have received more attention, since networks are an adequate tool to model several phenomena that occur in the real world. Some of these phenomena occur for example in social networks, citation networks, Internet topology networks, medicine and computer science (COSTA *et al.*, 2011). Complex networks are complex types of graphs. Like traditional graphs, complex networks contain nodes and edges connecting these nodes. Real-world complex networks have certain statistical and topological properties that do not occur in simple networks (PARDO *et al.*, 2006). Some of these properties are observed, for instance, in small-world networks, where the typical distance between two nodes is small. Social networks generally have the small-world property. In social networks, nodes are usually organized in groups (friends of a person are usually also friends with each other), in this sense, these networks have high local connectivity. Another important property is found in scale-free networks. These networks contain hubs, i.e. highly connected nodes. For example, in citation networks, a hub could be an article that is cited many times by other articles (COSTA *et al.*, 2011).

Natural Language Processing (NLP) has benefited from the use of complex network concepts. The use of these graphs has become a powerful technique for modeling and processing texts. For example, linguistic networks were included in the context of complex networks (COSTA *et al.*, 2011). The nodes of such networks could be syllables, words, sentences or paragraphs; and the interconnections between the nodes are determined in several ways. Language networks are divided into large groups: semantic networks and surface networks (COSTA *et al.*, 2011). The semantic networks are constructed from dictionaries of lexicons and they usually contain information about semantic relationships between words, such as synonyms or antonyms. The WordNet (MILLER, 1995), for example, groups words into sets of synonyms and it stores other semantic relationships such as antonyms, hypernyms, hyponyms, etc. Surface networks are based on the internal structure of words such as their morphological properties or their position in sentences or syntactic structures. For example, the set of syllables of a language can be used to construct another type of network, where the nodes are the syllables and the edges indicate if two syllables co-occur in the same word. Some NLP applications that used complex network concepts are the following: evaluation of the quality of machine translators and texts using word adjacency networks (AMANCIO *et al.*, 2008); authorship networks to identify the authors of a set of books (ANTIQUEIRA *et al.*, 2007); lexical networks related to the construction of spell checkers, where each network edge is linked with the orthographic distance between words (CHOUDHURY *et al.*, 2007); syntactic networks for the study of language acquisition (ANTIQUEIRA *et al.*, 2007); among others.

The use of complex network concepts is also ideal for extractive summarization. Such networks can capture the text structure in several ways: nodes can represent words, sentences or paragraphs of a document to be summarized. The links between the nodes can be established in different ways, for example, in Amancio *et al.* (2012) each distinct word is represented as a node and the edges are obtained by joining nodes whose corresponding words are immediately adjacent. In the work of Antiqueira *et al.* (2009), the nodes are the sentences and the edges are represented according to the number of common words between two sentences. According to some network measurement, relevance scores are assigned to each node (sentence) and a subset of those best-ranked nodes are selected to compose the final extract. The purpose of this Master's research is to address the extractive summarization task by using the concepts and metrics studied in complex networks.

1.1 Hypothesis and Objectives

Nowadays, with the great increase of textual information available on the web and little time to read and understand such information; manual analysis of documents becomes impossible. Due to this significant grow of information, it becomes clear that automatic summarization methods are fundamental. It is necessary to develop techniques which find the most relevant document content in a condensed way and without changing the original meaning. Although the

research in the document summarization area has been widely studied for many years, there are still few works that used complex network concepts for extractive summarization.

A complex network model emphasizes a structural organization model of words or sentences rather than their content. When documents are modeled as networks, we believe that it is possible to identify the most relevant or central sentences by analyzing their topological importance. Therefore, an extractive summary can be constructed by selecting the most central sentences in a text.

Although the extractive summarization problem has already been addressed using network science tools, there are still many issues that have not been considered. For example, dynamical measurements of complex networks have not been used yet for extracting the most important sentences. Also, we considered new methods for network creation. These methods include the use of new techniques for text representation such as word embeddings approaches and traditional Vector Space Models (VSM). Another approach considered here is the representation of multiple documents for Multi-Document Summarization by using a multilayered network representation. The main objectives of this Master's work are explained below:

This Master's research aims to investigate and employ concepts developed in the complex network area for extractive document summarization for Portuguese and English documents for both Single Document Summarization (SDS) and Multi-Document Summarization(MDS). By using complex network measurements, we recognize the most informative and relevant sentences to be incorporated into a final summary. We used the network models employed in previous works and we also proposed to evaluate the relevance of new methods for document representation by networks such as those models based on word embeddings and Multilayer Networks (MLN) for the extractive summarization task. In the same way, apart from the traditional measurements successfully used in previous works, we validated the performance of novel dynamical network measurements for sentence ranking. We also make a comparison of our systems with the results of other works found in the literature.

1.2 Organization

This manuscript is organized as follows: Chapter 2 presents the background; where the concepts related to complex networks and document summarization are briefly explained. In Chapter 3 we present some of the main words related to the use of complex network concepts for document summarization. In Chapter 4 we show the methodology we applied for achieving the objectives of this Master's work. Finally, Chapter 5 shows the results we obtained by evaluating the proposed systems while Chapter 6 presents the conclusion which discusses the contributions, limitations, and prospects for future works derived from this research.

CHAPTER

2

BACKGROUND

In this chapter is explained the main concepts related to the goals of this Master's research, which addresses the problem of automatic document summarization using complex network concepts. First, Section 2.1 shows some concepts that must be taken into consideration for automatic document summarization and a brief description of the main works found in the literature to address this problem. Section 2.2 presents some important definitions about complex network concepts such as node centrality and multilayer networks. Section 2.3 describes the methods employed in this research for text representation. Finally, Section 2.4 presents the final remarks.

2.1 Document summarization

Document summarization is the process of automatically creating a compressed version of one or more documents with the aim of building a summary that retains the most important ideas of the original document(s) (FERREIRA *et al.*, 2013a). Document summarization aims to automatically create a summary or abstract by finding the most informative and relevant sentences. The generated summaries should be concise and fluent. According to Nenkova and McKeown (2012), the main document summarization methods found in the literature are explained as follows:

- The summaries generated from an *extractive summarization method* are usually produced by joining several sentences, which are taken exactly as they appear in the original document. On the other hand, *abstractive summarization* is a more difficult task since it aims at paraphrasing sections of the source document. In addition, abstractive methods may reuse phrases or clauses of the source document (NENKOVA; MASKEY; LIU, 2011).

- *Single Document Summarization (SDS)* systems produce a summary of a single document. In *Multi Document Summarization (MDS)* case, systems produce a single summary from a

set of documents that share the same topic. Works on MDS became important when the amount of redundant information on the web grew enormously since MDS could provide a brief summary of many documents on the same topic or event.

- In some scenarios, it is key to produce small size summaries. For example, a *keyword summary* consists of a set of indicative words or phrases that are indicated in the input documents. In *headline summarization*, the input documents are summarized by a single sentence.

- *Generic summarization* makes few assumptions about the audience or some goal to generate the summary. In this case, the audience is general, that is, anyone can read the summary. In *query focused summarization*, the goal is to summarize only the input document information that is relevant to a specific user query.

- *Update summarization* is a special type of Multi Document Summarization method, where a summary is generated based on the assumption that the user has already read a given set of documents (AGGARWAL; SUMBALY; SINHA, 2009).

A good automatic summarization method should reflect the various topics of the document while avoiding redundancies. Document summarization algorithms should look for headings and other subtopics markers in order to identify the key points of a document (GUPTA; LEHAL, 2010). The development of methods for MDS is a more complex task than SDS since this difficulty arises from thematic diversity within a large set of documents. The problem involves multiple sources of information that overlap and complement each other, which are sometimes contradictorily. Therefore, it is not only fundamental to identify and deal with redundancy, but also recognize the novelty and ensure that the final summary is coherent and complete, thus avoiding pending references to something that is not cited or approached in the summary (DAS; MARTINS, 2007).

2.1.1 Extractive document summarization: Sentence extraction

This section explains the main methods for the extraction of the most important sentences for the extractive document summarization task (NENKOVA; MCKEOWN, 2012).

- **Methods based on word frequency:** This method is based on the fact that the more frequently a word occurs in the text, the higher is its score. The sentence weight is calculated by counting the number of most frequent words in the sentence. The sentences which contain the most frequent words in a document have a higher chance of being selected for the final summary (FERREIRA *et al.*, 2013a).

  - *Word probability:* The frequency-based methods are fairly simple and effective, but they are heavily influenced by the document length. For example, a word that

appears twice in a 10-word document might be important, however, this word will not probably be relevant in a 1000-word document. This method makes an adjustment for document length. The word probability is calculated as the number of word occurrences, $c(w)$, divided by the number of all words in the input. For each sentence, it is assigned a weight equal to the average probability $p(w_i)$ of the words that the sentence contains.

- *Vector space model based on Tf-Idf weighting:* The Term frequency (Tf) is the frequency of a term within a document and the Inverse document frequency (Idf) is a well-known measure which calculates the relevance of a word within a document set. The Tf-Idf weights (Tf*Idf) of each word are a good indicator of importance. This method is easy and fast to compute, Section 2.3 describes the calculation details for Tf-Idf scores. The sentence weight is computed by averaging the Tf-Idf values of its content words.
- *Log-likelihood ratio for topic signatures:* These approaches not only use a background corpus, but also allow the definition of topic signature words that are unique words in input documents. These words are useful to gauge sentence importance, while other words are completely ignored in the computation of sentence relevance. Topic signatures are words that frequently occur in the input, but are unusual in other texts, similarly to words with high Tf-Idf weights. Unlike Tf-Idf weighting, the log-likelihood ratio provides a way to establish a threshold which divides all input words as descriptive or non-descriptive. The sentence relevance is calculated as the number of topic signatures it contains.

- **Sentence clustering:** For MDS case, the input consists of several articles that possibly come from different sources, therefore, it is very likely that different articles contain sentences with similar information. The information that appears in most of the input documents is considered relevant and should be included in a summary. Similar sentences can be grouped together. Clusters including several sentences could represent relevant topics in the source documents. An extractive summary could be generated by selecting one important sentence from each cluster. The greater number of sentences in a cluster, the more relevant information in the cluster is considered.

- **Graph-based methods for sentence ranking:** These methods represent the input documents as a highly connected graph. The nodes could be words, sentences or paragraphs and weights are assigned to edges between the nodes. The edge weight is usually established by computing a similarity measurement between the corresponding words, sentences or paragraphs. The cosine similarity is a measurement commonly used in these approaches. For the identification of most central or relevant nodes, graph-based algorithms such as Page Rank are generally used.

- **Machine learning approaches for document summarization:** Due to a variety of indicators to determine the sentence importance found in the literature, it is key to find and use new methods that could combine such indicators of sentence relevance. Machine learning methods have been introduced to address the summarization problem. Statistical methods are usually used to compute the features for machine learning. The most common features are grammatical or positional features, words contained in the title, functional information transmitted by the source text, proper names, among others. The goal of these approaches is to classify sentences as "Present" or "Not present" in the summary.

2.1.2 Automatic summary validation methods

The automatic summary validation is an important task because it allows us to compare the performance of the generated summaries from this work with the results of other works. Most automatic validation methods help to determine the summary quality by comparing system summaries with other summaries created by humans (ideal summaries). The summary validation is a complex task because of its high level of subjectivity. Generally, it uses handwork which takes a lot of time and availability of workforce (ANTIQUEIRA *et al.*, 2009). Because it is a difficult task, many automatic validation methods have been used in the literature. The most common methods for summary validation are the precision, recall and f-measure metrics and the ROUGE-1 metric. These methods are most frequently used because they allow comparing the different automatic summarization techniques proposed in several works. The precision and recall metrics consider a sentence as the basic unit, therefore, it is required to have an extract of ideal sentences to be compared with the generated summaries (ANTIQUEIRA, 2007). The precision measures the proportion of common sentences between two extracts in relation to the number of sentences of the automatic extract, while the recall measures the ratio between two extracts in relation to the number of sentences of the reference extract (ANTIQUEIRA, 2007). The f-measure is the harmonic mean of the precision and recall metrics.

Another relevant metric for summary validation is the ROUGE measurement (LIN, 2004). The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric is one of the most used tools for automatic validation of summarization systems. It is a widely used metric because it correlates strongly with human validation (RIBALDO *et al.*, 2012). ROUGE is as good as humans in ranking summaries according to their informativeness, being a good indicator of the quality of the summarization methods. This metric counts the number of overlapping units such as n-grams, word sequences, and word pairs between the system summaries and the human summaries (RIBALDO *et al.*, 2012). The ROUGE calculates the recall, precision, and f-measure using distinct attributes, which leads to the following variations: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. These metrics are briefly explained as follows (LIN, 2004): ROUGE-N is the overlap of N-grams between the system and reference summaries; ROUGE-L refers to Longest Common Subsequence based statistics; ROUGE-W is the Weighted

LCS-based statistics; and ROUGE-S is the Skip-bigrams based on co-occurrence statistics.

The systems proposed in this research were evaluated by using the ROUGE-N metric (where N was set to 1-gram), which is a metric based on N-gram statistics related to the number of common N-grams conveyed by the summaries. While the precision and recall are metrics for quantifying the number of correct sentences in a system summary, the ROUGE-N is a metric which counts the number of N-grams between generated and reference summaries. In this sense, ROUGE-N is ideal to evaluate summaries which are not of extractive type (ANTIQUEIRA, 2007). Some of the corpora we used do not have reference extracts, however, all corpus includes abstractive summaries. For this reason, we only used the ROUGE metric for the validation of our systems. Also, ROUGE-1 was used because this metric has been shown to be enough for comparing summary informativeness (RIBALDO *et al.*, 2012). The ROUGE-N is computed as follows:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}, \tag{2.1}$$

where n stands for the length of the n-gram, $gram_n$ and $Count_{match}(gram_n)$ represent the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries.

2.2 Complex networks

With the aim of understanding the growth of the complex network area, it is important to go back to the end of the 50's. At that time, the Erdos-Renyi model (ERDÖS; RÉNYI, 1959), consisted on the creation of a random graph. In this graph, for each edge was established a probability p, that is, each pair of nodes got the same probability p of being connected. Differently from the traditional research focused on random networks, scientists began to study and represent networks that model real and complex systems, and they come to the conclusion that the properties of these new network models were very far from the ones observed in random graphs. In this way, these type of graphs took the name of complex networks (WATTS; STROGATZ, 1998; STROGATZ, 2001). These new network models have particular properties and they are very different from random networks. The structure of these networks is strongly related to the behavior of the system they represent and, therefore, we could extract valuable information from such networks if they are analyzed efficiently. The properties of a complex network could refer to degree distributions that follow power laws, hierarchical structures, community structures, high local cohesiveness, etc. As a result of these researches, complex network concepts have been very useful to model real-world phenomena. Phenomena which is related to our social interactions, the environment we live or our own biological behavior (COSTA *et al.*, 2007).

A network or graph $G = \{V, E\}$ is defined by a set $V = V(G)$ of elements called nodes or vertices, and another set $E = E(G) \subset VxV$ of elements called links or edges that join the network nodes. An adjacency matrix A is used to represent the connectivity patterns in the network. The elements of this matrix take 0 or 1 values. In the adjacency matrix, $A_{ij} = 1$ if the

node i is connected to the node j, otherwise it has $A_{ij} = 0$. Another network-related matrix is the weighted matrix W, which contains the values or weights assigned to each network edge (in the case it had weights). These weights are related to the intensity of the connections between two nodes (COSTA *et al.*, 2007). Therefore, a weighted network $G = \{V, E, W\}$ could be defined. Networks can be directed or undirected. For undirected networks, the matrix A is symmetric.

Complex networks can be divided into four categories: social networks, information networks, technological networks, and biological networks (NEWMAN, 2010). In social networks, nodes are represented by people and the edges could represent a friendly relationship between people. For information networks, we highlight the citation networks, where nodes are the authors of some article, which interlace through the references that are given to the articles of other authors. Examples of technological networks are seen in electrical and Internet networks. Finally, biological networks represent information patterns between different biological elements. For example, neurons can be modeled as nodes, and their relationships with other neurons are determined by chemical reactions between cells.

In order to understand the properties and structure of complex networks, several mathematical models were constructed. These models provided a topology with statistical properties similar to the real world networks, thus allowing the application of several mathematical methods to analyze general behaviors of similar networks. The three main complex network models are described as follows (COSTA *et al.*, 2007; NEWMAN, 2010):

- **Erdos-Rényi Random Graph model:** It is the most basic network model. In this model, random networks consisting of N nodes and L edges are generated. Starting with N disconnected nodes, the networks are constructed through the addition of L random edges, avoiding multiple and independent connections (ERDÖS; RÉNYI, 1959). The maximum value of L is $\frac{N(N-1)}{2}$ possible edges.

- **Watts-Strogatz Small-World model:** In this model, the most of the nodes can be reached from the others through a small number of edges. Most nodes are neighbors of each other. This model also has the property that considers the presence of a large number of size three cycles. That is, if a vertex i is connected to the vertices j and k, then there is a high probability that their vertices j and k are connected (WATTS; STROGATZ, 1998).

- **Barabási-Albert Scale-Free model:** Networks have a degree distribution characterized by an uneven distribution and absence of a characteristic degree. That is, some vertices are highly connected, while other have hardly any connections, or in the worst case they do not have connections. This model is also characterized by the existence of hubs, which are vertices that are connected to an important fraction of the total edges of the network (ALBERT; BARABÁSI, 2002).

2.2.1 Centrality network measurements

The centrality measurements allow to identify the most relevant nodes in a network. The goal of these measurements is to rank the nodes according to their topological importance. The centrality is a structural attribute, that is, the value assigned to a node depends strictly on its location in the network, thus allowing to determine what is the contribution of this node to the network (BORGATTI, 2005). The detection of most central nodes in a network is a fundamental task, for example, in a social network, it allows to analyze the influence of a person, or it helps to determine how good is a path in a transport network, or how important a web page can be as well as being able to determine what are the essential components in a computer network (COSTA *et al.*, 2007). For the purposes of this research, the centrality measurements will allow to identify the most relevant document sentences which could belong to the final extract.

Below it is explained some of the most common measurements which are used to determine the network centrality. These measurements are node degree, strength, shortest paths, betweenness, Page Rank and clustering coefficient (COSTA *et al.*, 2007). Also, some slightly more advanced measurements such as concentric metrics, accessibility, symmetry and absorption time are explained as follows:

- **Node degree (k):** The degree of a node i is the number of edges connected to that node. The node degree is calculated as follows:

$$k_i = \sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n} a_{ji}, \tag{2.2}$$

  where a_{ij} is an element from the adjacency matrix A.

- **Strength (s):** For weighted networks, this measurement is calculated as the sum of the weights of all the corresponding edges of a reference node i. In the case of unweighted networks, the strength represents the node degree. The strength is calculated as follows:

$$s_i = \sum_{j=1}^{n} w_{ij} = \sum_{j=1}^{n} w_{ji}, \tag{2.3}$$

  where w_{ij} is an element from the weighted matrix W.

- **Clustering coefficient (cc):** It is a metric which characterize the presence of loops of order three in a network. This measurement computes the probability that two neighbors of a node are connected. The clustering coefficient is a typical measurement used in social networks: if two people (network nodes) have a friend in common, then, it is very likely that these people also have a friendly relationship. This measurement is calculated as follows:

$$cc_i = \frac{2e_i}{k_i(k_i - 1)}, \tag{2.4}$$

  where e_i represents the number of edges among the neighbors of node i and k_i is the degree of node i.

- **Shortest paths (l):** A shortest path is defined as a path that links two nodes i and j with a minimum length. The length of such a path is denoted as d_{ij}. The average shortest path length is defined as follows:

$$l_i = \sum_{j \neq i} \frac{d_{ij}}{(N-1)}, \tag{2.5}$$

where N is the total number of nodes in the network.

- **Betweenness (b):** This metric is computed as the ratio of shortest paths between any pair of nodes which pass over a special node. This measurement quantifies the node influence in the information diffusion from the network. In Equation 2.6 we show the estimation of this measurement (NEWMAN, 2003).

$$b_i = \frac{\sum_{s<t} g_i^{(st)}/n_{st}}{(1/2)N(N-1)}, \tag{2.6}$$

where $g_i^{(st)}$ is the number of shortest paths from vertex s to vertex t passing through i, and n_{st} is the total number of geodesic paths from s to t.

- **Page Rank (π):** According to this metric, a node i is relevant if it is connected to other relevant nodes. This measurement is calculated in a recursive way as follows:

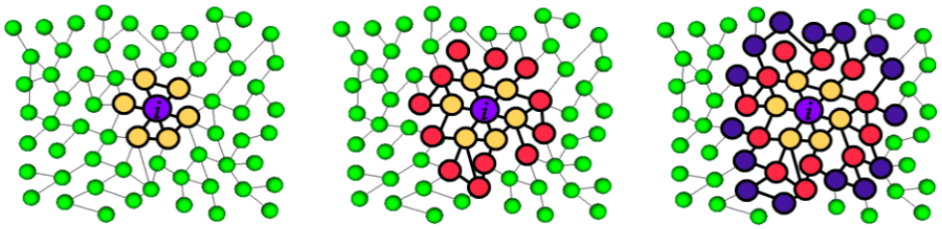
$$\pi_i = \gamma \sum_j a_{ij} \frac{\pi_j}{k_j} + \beta, \tag{2.7}$$

where γ and β are damping factors which take values from 0 to 1.

- **Concentric measurements:** The previously studied measurements such as node degree, clustering coefficient, etc. emphasize relevant network properties. However, even more valuable topological information can be retrieved using such measurements along hierarchical levels of networks. The concentric measurements are a set of eight metrics that are able to extract this type of information (COSTA; SILVA, 2006). A hierarchical level let to make a powerful extension of basic measurements with the aim of obtaining a complete network characterization. In Figure 1 we show an example of the first three hierarchical levels of a reference node i in a network.

  In order to compute any concentric measurement, it is important to identify the ring $R_d(i)$, which is the set of nodes that are d hops away from i. Next, we explain some metrics proposed in the work of Costa and Silva (2006).

  1. **Concentric number of nodes:** It is the number of nodes pertaining to the ring $R_d(i)$.
  2. **Concentric number of edges:** It represents the number of edges linking nodes inside the ring $R_d(i)$.
  3. **Concentric node degree:** Number of edges extending from the ring $R_d(i)$ to $R_{d+1}(i)$.



(a) Hierarchical level h=1 (b) Hierarchical level h=2 (c) Hierarchical level h=3

Figure 1 – Example of hierarchical levels of the reference node i. (a) First hierarchical level. (b) Second hierarchical level. (c) Third hierarchical level.

Source: Costa and Silva (2006).

4. **Concentric clustering coefficient:** It is the ratio between the number of existing edges in the ring $R_d(i)$ and the total number of possible edges in this ring.
5. **Convergence ratio:** Division of the concentric node degree by the number of nodes which are at the next concentric ring.
6. **Intra-ring node degree:** This metric represents the average degree of the nodes at the ring $R_d(i)$ which only include the edges located in the ring $R_d(i)$.
7. **Inter-ring node degree:** It is the proportion between the node degree and the number of nodes which are in the same ring.
8. **Concentric common degree:** This metric is the average degree which considers all the connections of nodes at a specific ring.

- **Accessibility (α):** The accessibility is a measurement which quantifies the number of accessible nodes from an initial node through the use of a self-avoiding random walk of length h (TRAVENçOLO; COSTA, 2008). A maximum accessibility value will imply the minimum scanning time. This metric takes into consideration not only the number of nodes at a given distance but also the probabilities of transition between the source and these nodes (TRAVENçOLO; COSTA, 2008). This metric considers different levels of hierarchy, which can be established by specifying the length h of the random walks. With the aim of computing this metric, let $p^{(h)}_{(i,j)}$ to represent the probability of reaching a node j from a starting node i through a self-avoiding random walk of length h. The accessibility measurement is then defined as the exponential of the true diversity of $p^{(h)}_{(i,j)}$, and it is calculated as follows:

$$a_i^{(h)} = \exp\left(-\sum_j p^{(h)}(i,j) \ln p^{(h)}(i,j) \right) \tag{2.8}$$

Figure 2 shows an example of the computation of the accessibility for two different graphs. In this example, we calculate the accessibility of a node A at the second hierarchical level

for both graphs. We can see how the accessibility varies according to the graph connections. These illustrations show that the accessibility can be interpreted as the effective number of accessed nodes.

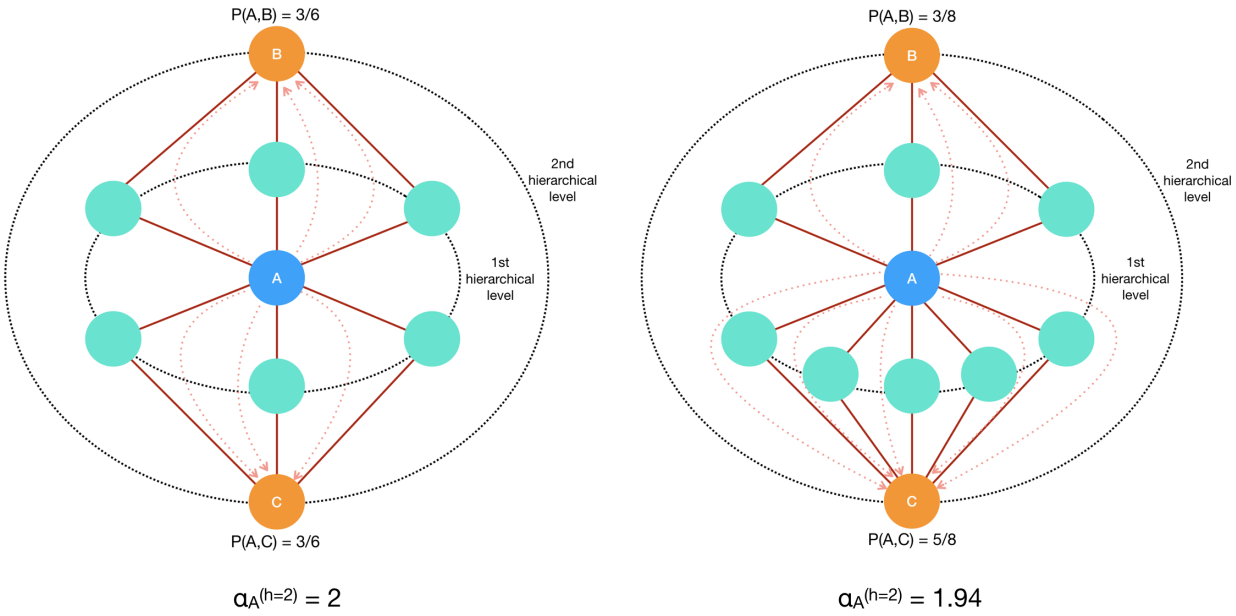


Figure 2 – Example of computation of accessibility for two networks at the second hierarchical level. In the first graph, nodes B and C are equally accessed from node A, because both nodes have the same probability ($p = 3/6$). In this sense, the accessibility reaches its maximum value $a_A^{(h=2)} = 2$. In the second example, the access to the nodes becomes uneven, because, from the initial node A, it is much more likely that a random walk ends at node C rather than node B. Therefore, the accessibility value decrease to $a_A^{(h=2)} = 1.94$.

Source: Elaborated by the author.

- **Generalized accessibility (ã):** Because the accessibility measurement depends on the parameter h for its calculation, Arruda *et al.* (2014) proposed an adaptation of this metric, called generalized accessibility. This novel metric is an improvement of the accessibility because it can be computed without any previous choice of the parameter h. This metric considers walks of all lengths between any pair of nodes. For the calculation of the generalized accessibility metric, first let P be a stochastic matrix whose element $p(i, j)$ is the probability of a random walker to go from an initial node i to final node j in the next step of the random walk. The transition probability which considers all lengths is calculated as follows:

$$P^{(\infty)} = \frac{1}{e} \sum_{j=0}^{\infty} \frac{1}{j!} P^j \tag{2.9}$$

Finally, we computed the generalized accessibility by using the true diversity of the transition probabilities of $P^{(\infty)}$ as follows:

$$\tilde{a}_i = \exp\left(-\sum_j p^{(\infty)}(i,j) \ln p^{(\infty)}(i,j) \right), \tag{2.10}$$

where $p^{(\infty)}$ is an element of $P^{(\infty)}$.

- **Symmetry (S):** The symmetry measurement is a normalized adaptation of the accessibility, where the number of accessible nodes is used as normalization factor (AMANCIO; SILVA; COSTA, 2015). This measurement evaluates the hierarchical levels at h hops of a reference node i. The goal of the symmetry is to quantify how diverse is the exploration of a neighborhood, in this sense, this metric considers that at each step, the random walker can access the next concentric level. In order to compute the probability transitions, all edges connecting nodes in the same concentric level are ignored. The symmetry metric is computed as follows:

$$S_i^{(h)} = |\xi_i^{(h)}|^{-1} \exp\left(-\sum_j p^{(h)}(i,j) \ln p^{(h)}(i,j) \right), \tag{2.11}$$

where $\xi_i^{(h)}$ is the set of accessible nodes that are at a distance h from the node i.

- **Absorption time (τ):** The absorption time is a metric which is defined as the time it takes for a particle in an internal node to reach an output node (absorbent node) by using a random walk. This measurement calculate how fast a randomly-walking particle is absorbed by the output nodes, considering that such particle starts the random walk at the initial node (AMANCIO; JR.; COSTA, 2011).

  In order to compute the absorption time, first let P be a stochastic transition matrix. Then, we used the matrix $\Psi = (I - \Theta)^{-1}$, where I represents the identity matrix while Θ is a submatrix of P which stores the transitions between transient nodes or non absorbent nodes. The time spent in transient nodes can be calculated as follows:

$$t_i = \sum_j \Psi(i,j) \tag{2.12}$$

  Finally, in order to compute Ψ, we could define i as a initial node and j be an absorbent node for each pair (i,j). Therefore, the absorption time τ_i of a node i could be defined as:

$$\tau_i = < t_i > = \frac{1}{n-1} \sum_{k \neq i} t_k \tag{2.13}$$

2.2.2 Multilayer networks

The study of single or monolayer networks allowed us to better understand different complex phenomena of the real world. However, despite the advances in the last years, we need to consider and understand more complex phenomena which can not be represented in single networks. This is because many real-world systems do not operate in isolation. They are interconnected and what happens at a single level of interaction affects the structure and function at another interconnected layer (SYSTEMS; LAB, 2018). For example, in a social system, the nodes represent people that interact with each other. In this system, we could see different types of interactions between the same people: a single person has personal, social, professional, etc., circles in the off-line world, but also could have multiple accounts in online social systems (Facebook, Twitter, etc.) (SYSTEMS; LAB, 2018). Due to these complex phenomena with a diversity of connections, the study of multilayer network arises.

A multilayer network is a type of network that explicitly incorporates multiple channels of connectivity and constitutes the natural environment for describing interconnected systems across different categories of connections: each channel is represented by a layer and the same node or entity may have different kinds of interactions (BOCCALETTI *et al.*, 2014). Each layer represents a given operation mode, social circle, or temporal instance. In social networks, for example, we can consider several types of different relationships between actors: friendship, vicinity, kinship, membership of the same cultural society, partnership or coworker-ship, etc.

In Figure 3, we show two examples of multilayer networks: an air transportation network (CARDILLO *et al.*, 2013), where each layer contains the flights from a different airline; and a bank-wiring room network (ROETHLISBERGER; DICKSON, 2003), where each individual uses a different node color and each layer contains the ties from a different type of relationship.

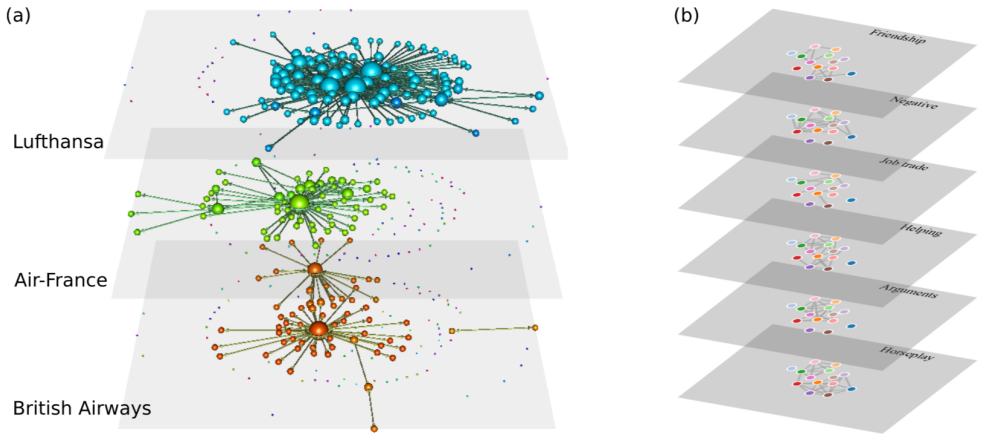


Figure 3 – Examples of visualization of two multilayer networks: (a) Air transportation network, and (b) Bank-wiring room network.

Source: Kivelä *et al.* (2014).

2.3 Models for text representation

In most NLP applications, the document set is required to be converted and modeled in a more appropriate way. The modeling of the document includes the task of finding a mathematical form that best represents the document structure and content from such document. There are several approaches for the mathematical representation of documents. One of the most simple approaches is the so called vector space model where texts are represented as vectors. The values of each element of these vectors depend on the approach that is being used. For example, the boolean model (LANCASTER; GALLUP, 1973) is based on the presence or absence of words in a document. Another widely used approaches are the word frequency model and the Tf-Idf model (SALTON; WONG; YANG, 1975). Other models which require a more advanced statistical analysis are the LSA (DUMAIS, 2004) and LDA (BLEI; NG; JORDAN, 2003) models. Finally, models based on word embeddings have been recently studied. These methods use neural networks, dimensionality reduction or probabilistic models to represent words or phrases to vectors of real numbers (MIKOLOV *et al.*, 2013b).

2.3.1 Vector Space Model based of Tf-Idf weighting

The Vector Space Model (VSM) is one of the most used models for the mathematical document representation (LEE; CHUANG; SEAMONS, 1997). In this model, texts are represented as vectors of identifiers in a multidimensional linear space. More formally, the vector $d_j = (w_{1j}, w_{2j}, ..., w_{mj})$ represents a sentence, paragraph or a set of paragraphs, where m is the total number of unique terms that appear in documents and $0 \leq w_{ij} \leq 1$ represents the contribution of the term t_i for the semantic representation of the document d_j. The weight of each vector element could be represented by considering the presence or absence of a term in a document (boolean model) or the frequency of occurrence of that term (frequency based model). Because of their simplicity, these models are not suitable to represent the semantic content of documents. Hence, the Tf-Idf weighting of terms deals with the errors or uncertainties that could occur in the simple models for document representation (boolean or frequency based model).

Term frequency - Inverse document frequency (Tf-Idf) is a measurement that indicates how relevant is a word in a document collection. The Tf-Idf value increases proportionally to the number of times a word appears in the document, but it is compensated by the word frequency in the document collection. While Tf estimates how often a term occurs in a document, Idf estimates whether the term is common or rare throughout all documents. The term frequency (Tf) of a term t in a document d is calculated as follows:

$$Tf(t,d) = \frac{f(t,d)}{|d|} \tag{2.14}$$

where $f(t,d)$ is the number of times the term t appears in the document d and $|d|$ is the total numbers of terms of d. The inverse document frequency (Idf) of term t in the document collection

D is computed as follows:

$$Idf(t,D) = log\left(\frac{|D|}{DF}\right) + 1 \tag{2.15}$$

where D represents the total number of documents and DF is the number of document in which the term t appears. Finally, the computation of the Tf-Idf value of a term t is shown in Equation 2.16.

$$\text{Tf-Idf(t,d,D)} = Tf(t,d) * Idf(t,D) \tag{2.16}$$

The representative vector of each document d of a collection D is then calculated according to the Tf-Idf value of each of its corresponding terms.

2.3.2 Word embeddings

The previous model, despite of its popularity, has several disadvantages. For example, the vector space model ignores the semantics of words. For instance, words like "powerful", "strong", and "Paris" are equally distant despite the fact that semantically "powerful" should be closer to "strong" than "Paris" (MIKOLOV *et al.*, 2013b). Another disadvantage of vector space model is the high dimensionality of data. For example, the vector size of traditional vector space models depends on the number of words the vocabulary contains. On the other hand, the size of word embedding models is fixed.

Word embeddings are models for word vector representation that allows words with similar meaning to have similar representations. The goal of semantic vector space models is to represent each word with a real-valued vector. These representations are based on the usage of words. This allows words which are used in similar ways to have similar vector representations, and naturally captures their meaning. This is an advantage that overcomes the vector space model, where different words have different representations, regardless of how they are used. Word embedding models were used with success in several NLP applications such as information retrieval (HUANG *et al.*, 2012), document classification (SEBASTIANI, 2002), question answering (TELLEX *et al.*, 2003), name entity recognition (TURIAN; RATINOV; BENGIO, 2010), etc. In this Master's work, we studied three word embedding models: Word2Vec, GloVe, and FastText.

- **Word2Vec:** The Word2Vec model (MIKOLOV *et al.*, 2013b; MIKOLOV *et al.*, 2013a), is a three layer neural network with one input, one hidden and one output layer. The idea of this model is to learn word representations that can predict a word given its surrounding words. The input layer corresponds to signals for context (surrounding words) and output layer corresponds to signals for the predicted target word. In Figure 4 we show an example of this neural network. Let's suppose we have the input sentence "The cat sat on the mat". The neural network tries to learn features (weights W and W') which look at words in

a window, for example, "the cat sat" and it tries to predict the next word "on". Hence, with the input words "the", "cat", and "sat"; the training process adjusts the weight of the network, so the probability of output "on" is maximized; as compared to other words which are in the vocabulary. As the training procedure repeats this process over a large number of sentences, the weights stabilize. These weights are then used as the vectorized representation of words.

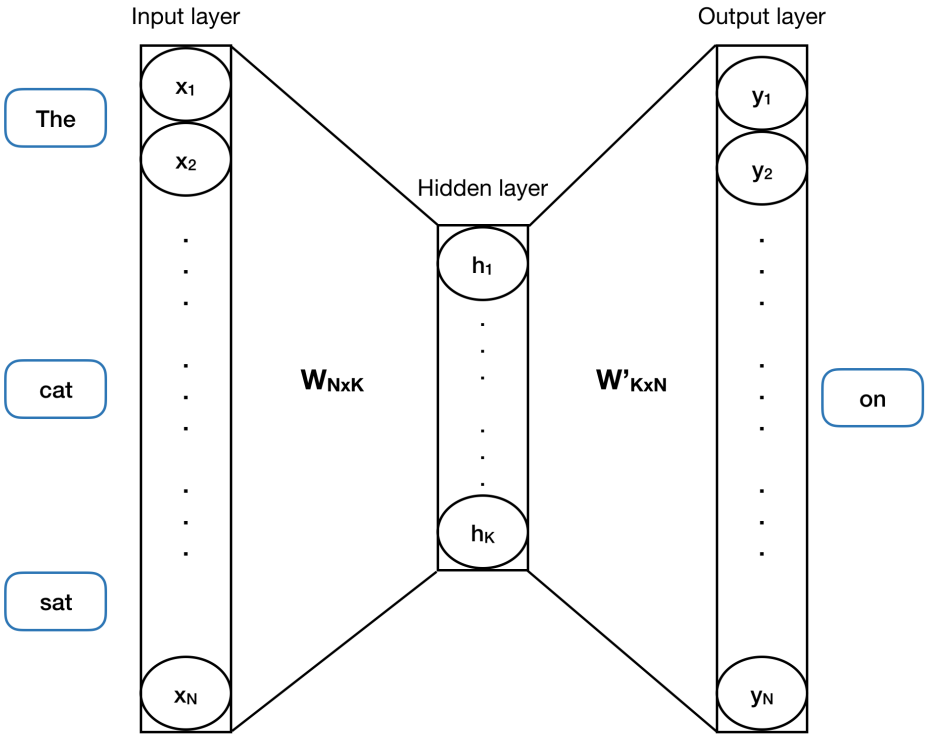


Figure 4 – Example of Word2Vec neural network.

Source: Mikolov *et al.* (2013b).

More specifically all words are mapped into single vectors, which are represented by a column matrix W. This column is indexed by the position of a word in the vocabulary. For the prediction of the next word in a sentence, then the concatenation or sum of vectors is used. Given a set of training words $w_1, w_2, ..., w_t$ the goal of the model is to maximize the average log probability, which is shown in the following equation:

$$\frac{1}{T}\sum_{t=k}^{T-k} \log p(w_t|w_{t-k},\ldots,w_{t+k}) \tag{2.17}$$

Then the softmax multiclass classifier is commonly used for the prediction task. In Equations 2.18 and 2.19 we show the required calculations for this task.

$$p(w_t|w_{t-k},\ldots,w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}, \tag{2.18}$$

$$y = b + Uh(w_{t-k}, \ldots, w_{t+k}; W), \tag{2.19}$$

where y_i represents the normalized log-probability for each output word i, U and b are softmax parameters while h is composed by joining the average of the word vectors that are extracted from W. After the training stage converges, words with similar meanings are mapped in a similar position in the space vector.

- **GloVe:** Models for learning word representation are mainly divided into two groups: global matrix factorization methods and local context window methods (PENNINGTON; SOCHER; MANNING, 2014). The first models learn their vectors by doing dimensionality reduction on the co-occurrence counts matrix (for example LSA model (DEERWESTER *et al.*, 1990)), while the latter are predictive models which learn their vectors in order to improve their predictive ability of loss, i.e., loss of predicting the target words from the context words given the vector representations (for example Word2Vec model (MIKOLOV *et al.*, 2013b)).

  Both models have significant disadvantages. For example, while LSA related methods efficiently leverage statistical information, they have many problems with the word analogy task. On the other hand, some context window methods do better on the analogy task, but they hardly use the corpus statistics (PENNINGTON; SOCHER; MANNING, 2014). The Global Vectors for Word Representation, or GloVe (PENNINGTON; SOCHER; MANNING, 2014) is a method which combines the advantages of both main models explained before. It is called global vector method because the global corpus statistics are captured by this model. Rather than using a window to define local context, GloVe builds a word-co-occurrence matrix using statistics across the whole text corpus. GloVe efficiently leverages statistical information by training only on the nonzero elements in a word-word co-occurrence matrix, rather than on the entire sparse matrix.

  Pennington, Socher and Manning (2014) proposed a weighted least squares model which trains on global word-word co-occurrence counts, in this way, this model makes a more efficient use of statistics. GloVe creators believe that the ratio of the co-occurrence probabilities of two words is what contains information, therefore, it is crucial to encode this information as vector differences. In order to accomplish this information, Pennington, Socher and Manning (2014) proposed the weighted least squares objective J that aims to reduce the difference between the dot product of the vectors of two words and the logarithm of their number of co-occurrences:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2, \tag{2.20}$$

where V is the vocabulary size, w_i and b_i are the word vector and bias of word i respectively, $\tilde{w}_j$ and b_j are the context word vector and bias of word j respectively, X denotes the word

co-occurrence matrix, X_{ij} represents the number of times word i occurs in the context of word j, and f is a weighting function that assigns relatively lower weight to rare and frequent co-occurrences. This function f helps to prevent common word pairs and it is calculated as follows:

$$f(x) = \begin{cases} \left(\frac{x}{x_{max}}\right)^{\alpha} & \text{if } x < x_{max} \\ 1 & \text{otherwise} \end{cases} \tag{2.21}$$

In the case that extremely common word pairs are found (where $X_{ij} > x_{max}$), the function f will cut off its normal output and will return 1. For other word pairs, this function will return some weight in the range (0,1). The weight distribution in this range is decided by the parameter α.

- **FastText:** The most popular models for semantic vector representation ignore the morphology of words, by assigning a distinct vector to each word (BOJANOWSKI *et al.*, 2016). These models ignore the internal structure of words, therefore, this limitation could be a problem for those morphologically rich languages with large vocabularies and many rare words. In order to overcome these difficulties, Bojanowski *et al.* (2016) proposed the FastText method. In this model, which is an extension to Word2Vec, each word is represented as a bag of character n-grams.

  The process of FastText is explained as follows: instead of training individual words in the neural network, FastText divides corpus words into several N-grams (sub-words). Each word w is represented as a bag of characters N-gram. For example, to learn the vector representation for a word "control", it will be represented by its character n-grams: <co, on, nt, tr, ro, ol, con, ont, ntr, tro, rol, cont, ontr, ntro, trol, contr, ontro, ntrol, contro, ontrol> and the word <control>. Bojanowski *et al.* (2016) extracted all the n-grams for n greater or equal to 3 and smaller or equal to 6. After the training phase, all the n-grams given the training set will have their respective word embedding representation. For each character n-gram, a vector representation is associated, thus, words are represented as the sum of these representations. In the example, the word embedding which represents the word "control" will be the sum of all its n-grams. This model has the following advantages:

  - It trains models on large corpora quickly.
  - FastText computes word representations for words that did not exist in the training data, since these words can be divided into character n-grams.
  - It is helpful to find the vector representation for rare words. Such words could be represented adequately because it is highly likely that some of their n-grams are present in other words.
  - Character n-grams embeddings generally have a better performance in comparison to Word2Vec and GloVe on smaller datasets.

This Master's research focused on two mathematical models for document representation: the traditional vector space model based on Tf-Idf weighting and models based on word embeddings (Word2Vec, GloVe, and FastText). Table 1 shows a brief comparison between the previously studied models.

Table 1 – Comparison between both traditional methods and word embeddings for word vector representation

Traditional methods	**Word embeddings**
Generally use one hot encoding	Each word is represented by a fixed number of dimensions, generally of size 300, 600, or 1000
In most cases, each word in the vocabulary is represented by one bit position in a huge vector	Dimensions are basically projections along different axes, more of a mathematical concept
Context information is not utilized	Context information is one of the key features for vector representation

2.4 Final remarks

In order to understand the proposal of this Master's research, in this chapter, we studied the main concepts related to the methodology of this work. We explained the main approaches for extractive summarization and complex network measurements. Also, we described the main methods for text representation we used in this research. Chapter 3 presents the main related works that addressed the extractive document summarization task with complex network concepts.

CHAPTER

3

RELATED WORKS

The task of selecting the most prominent sentences in a document is key for extractive document summarization. According to Nenkova and McKeown (2012) the main methods for determining sentence importance are divided into four groups: methods based on word frequency, methods based on sentence clustering, graph-based methods for sentence ranking and methods based on machine learning. Since graph-based methods have become a powerful tool for document representation and thus network metrics are good indicators of sentence importance, such methods were investigated and implemented in this Master's research. The main works that used complex network concepts and graph theory for extractive document summarization are presented and described in this chapter.

In Section 3.1, we present some main works that used the concepts and measurements of complex networks for extractive document summarization for Portuguese and English language. In Section 3.2 we show a brief summary of the previously studied works. In the works we presented, summaries generally have a compression rate of 70%. The compression rate defines the summary size in relation to the original source text. Several authors defined the summary size according to the number of words, while others defined the size based on the number of sentences. For summary validation, the ROUGE metric (LIN, 2004) was used by the majority of works found in the literature. This metric is widely used because it is highly correlated with the evaluation made by humans. It is also important to highlight that most of the Portuguese-based approaches employed the news corpus called TeMário (PARDO; RINO, 2003). For the validation of English summaries, the corpus belonging to the Document Understanding Conferences (DUC) (OVER; LIGGETT, 2002) were mainly used.

3.1 Main graph based approaches for the extractive summarization task

There are a large number of works based on graph theory that addresses the summarization task. However, few works have used complex network concepts for the extractive summary of documents. The reason is simple, complex networks is a relatively new area that is just beginning to grow in the scientific world. Next, we make a brief description of some of the main works found in the literature which addresses the extractive document summarization task by using complex networks and graph-based concepts.

- **Work of Antiqueira *et al.* (2009):** In the work of Antiqueira *et al.* (2009), the authors proposed the creation of a network, where nodes represent sentences while there is an edge between two nodes if the corresponding sentences share significant nouns. Then, the method selects a subset of sentences (nodes) that were previously ranked according to some network measurement.

  The work is divided into four stages. In the first stage, the documents are pre-processed. The pre-processing stages include the segmentation of each document into sentences, then, the recognized nouns in each sentence are lemmatized. In the second stage, the resulting texts are mapped as a network. Each sentence is represented by a node and there is an edge between two nodes if their corresponding sentences have at least one noun in common. The edge weight is defined as the number of shared words between two sentences. In the third stage, in order to give a numerical or ranking value to each node, some complex network measurement is used. Finally, the n first nodes (sentences) of the ranking made in the previous stage, are selected to compose the final extract.

  A total of 14 different summarizers from 7 different complex network measurements were proposed: node degree, minimum paths, locality index, dilatation strategies, k-cores, w-cuts and community-based measurements. In addition to these summarizers, the authors proposed a summarizer based on a voting system, which combines the results of the 14 summarizers by selecting the best-ranked sentences by most of the proposed systems.

  Complex network metrics such as degree and strength were considered under the concept that those sentences that have a large number of links to other sentences probably convey relevant information that complements many sentences. Shortest paths measurements are based on the idea that the smaller the distances between nodes, the greater the interrelationships between sentences. The locality index was used to point out those central nodes or isolated groups that could represent important sentences that summarize the meaning of their neighbors. Measurements based on graph dilatation were used in order to select sentences that complement the central idea of a text. The k-cores and w-cuts metrics were used to represent those groups of sentences that are strongly related. Network communities could be useful to represent important topics in texts. Finally, the voting system joins

all previous strategies in an integrated approach, giving priority to the sentences that constantly appear at the top of the sentence rankings defined in each strategy.

The summary evaluation was carried out using the TeMário corpus (PARDO; RINO, 2003). The authors carried out two experiments. In the first experiment, the proposed summarizers were evaluated with the reference extracts of TeMário, which were generated by the GEI tool (PARDO; RINO, 2004). In the second experiment, the systems were evaluated using the reference abstracts of TeMário. According to the experiments, some of the proposed systems had similar results to the top summarizers for Portuguese. The summarizer based on voting systems achieved the best result. The systems based on degree and strength also obtained the best results, thus concluding that degree metric plays an important role for summarization purposes. According to the results of k-cores and w-cuts, it was possible to see that groups of highly interconnected nodes could be relevant for summarization. Measurements based on communities and locality index did not achieve the best results.

- **Work of Ribaldo *et al.* (2012):** The work of Ribaldo *et al.* (2012) addresses the problem of Multi-Document Summarization for texts in Brazilian Portuguese. In this approach, all sentences in the document set are recognized and treated as if they came from a single document. As the approach proposed by Antiqueira *et al.* (2009), the pre-processed sentences are represented as nodes. The links between nodes indicate how similar the sentences are. The cosine metric was used to calculate the similarity between the sentences. The degree, clustering coefficient and shortest path measurements were used to give a score to the sentences, allowing to select the best n ranked sentences.

  The summary is made with the most prominent sentences of all texts. Due to the large redundancy that could exist between sentences, the authors proposed to use an anti-redundancy method. Ribaldo *et al.* (2012) established a redundancy limit which a new sentence could have in relation to the previously selected sentences. This redundancy limit was defined as the sum of the highest value and the lowest value of the graph edges divided by 2. In the process of sentence selection, if the redundancy limit is reached, the new sentence is considered redundant and it is ignored, and the summarization process continues with the following candidate sentence; otherwise, the sentence is included in the final summary. In order to determine the summary size, authors established a compression rate of 70%, which indicates that the summaries should be 30% long in relation to the length of the largest text in the document group.

  The evaluation of the proposed methods was performed using the corpus CSTNews (CARDOSO *et al.*, 2011), which is a set of news in Brazilian Portuguese for Multi-Document Summarization. The authors compared their systems with the results of other works which used the CSTNews corpus: GistSumm (PARDO; RINO; NUNES, 2003), CSTSumm (JORGE; PARDO, 2010), MEAD (RADEV, 2001), and a baseline method that randomly selects the sentences. The work of Ribaldo *et al.* (2012) shows that graph-based

methods for MDS for Portuguese provide very good results close to the best systems available for this language. Measurements based on degree and shortest paths achieved good results. In contrast, the clustering coefficient obtained fairly low results. In the works of Agostini and Pardo (2011) and Leite and Rino (2008), however, the clustering coefficient measurement is used as a feature for a system based on machine learning, obtaining good results. The authors believe that this measurement probably is better used as complementary information than being used as an indicator of sentence importance.

- **Work of Amancio *et al.* (2012):** In the work presented by Amancio *et al.* (2012), new complex network measurements are used for the summarization task. Betweenness, vulnerability and diversity measurements were used to analyze texts written in Brazilian Portuguese. The authors extended the study carried out by Antiqueira *et al.* (2009) proposing new metrics and new network models.

  The network proposed in Amancio *et al.* (2012) is created as follows: texts are lemmatized and stopwords are removed. Then, each distinct word is as a single node. The edges are obtained by joining nodes whose corresponding words are immediately adjacent. The edge weight is determined by the number of repeated associations between two words. In order to improve the information of the summaries, the authors proposed an additional network which is identical to the previous network, adding new edges related to the syntactic dependence between words.

  The authors evaluated the following measurements: diversity centrality, vulnerability, betweenness, shortest paths and strength metrics. The diversity measurement quantifies how close is a vertex to the network edges, thus, the central vertex has high relevance and the marginal vertices have low relevance. Another relevant measurement is the global efficiency, GE, which is a geodesic measurement. This measurement is key for the definition of the vulnerability measurement. GE is related to the speed of information exchange between two nodes, that is, a small distance contributes more significantly than a longer distance. In order to quantify the node relevance by using vulnerability, changes in the network structure are determined when the target node is removed. Therefore, the GE measurement is used as a performance measurement to quantify these variations in the network. The betweenness measurement was defined in Chapter 2.

  For sentence selection, first, some network measurements (betweenness, vulnerability, etc.) are calculated for each node or word. Then a weight is assigned to each sentence according to the weight of each word. This weight is defined as the average weight of the words that compose the sentence. Thus, top-sentences with high average are included in the final summary because they probably provide a good level of information.

  TeMário corpus is used again for the evaluation of the proposed systems. The authors found that diversity-based methods have the highest score, indicating that the diversity measurement is not only effective for detecting borders in complex networks. Diversity-

based methods achieved results higher than the best technique proposed in the work of Antiqueira *et al.* (2009). In contrast, methods that employ vulnerability and betweenness metric yielded the poorest results.

The authors concluded that diversity metrics could be effective for detecting keywords with high precision, which would be favorable for producing good summaries. They also showed that the incorporation of linguistic knowledge improves the performance of the summarizers, but this improvement is minimal.

- **Work of Leite and Rino (2008):** Leite and Rino (2008) explore multiple features for extractive summarization using machine learning. The authors considered 11 features previously used in the work of Leite and Rino (2006) (SuPor-v2), which is a supervised summarizer for Brazilian Portuguese, and 26 different features based on complex network metrics. The authors used feature sets and four classifiers to define automatic extractive summarization models.

  The automatic summarization features used in Leite and Rino (2006) were the following: lexical chaining, sentence length, proper nouns, sentence location, word frequency, relationship maps, and topic importance (BARZILAY; ELHADAD, 1997; KUPIEC; PEDERSEN; CHEN, 1995a; EDMUNDSON, 1969; LUHN, 1958; SALTON *et al.*, 1997; NETO *et al.*, 2000).

  Regarding the complex network features, the authors used the network and metrics proposed in Antiqueira *et al.* (2009). Leite and Rino (2008) used 26 features in total, which are based on the following measurements: degree, clustering coefficient, minimal paths, locality index, dilatation, hubs, k-cores, w-cuts, and communities.

  A complex problem in machine learning is the feature selection stage, which consists of finding an optimal subset of features that maximizes the system's ability to classify correct instances. In this work, the authors explored the Correlation Feature Selection (CFS) method (HALL, 2000). This method was used because it does not need a previous definition of the number of selected features, resulting in a subset of recommended features.

  In machine learning approaches for document summarization, each sentence is usually represented in the training set as a tuple of binary features. Such tuples are labeled belonging to the class "Present" or "Not Present" in the summary. Then, the classifiers are trained to calculate the probability that any sentence belongs to one of these classes. The authors used four classifiers: Flexible-Bayes, C4.5, SVM and Logistic Regression (JOHN; LANGLEY, 1995; QUINLAN, 1993; VAPNIK, 1995; WITTEN; FRANK; HALL, 2011).

  For the evaluation of the systems proposed in Leite and Rino (2008), the authors used the TeMário corpus. Different feature combinations and different classifiers were used. Such feature combinations included, for example, the SuPor-v2-based features, features based on complex network measurements, and an approach based on the combination

of SuPor-v2 features and complex network metrics. In addition, the authors performed additional experiments in order to determine the influence of using the CFS method for feature selection. All combinations of features and classifiers originated a total of 24 automatic summaries. The evaluation results showed that the Logistic Regression and Flexible Bayes classifiers outperformed most of the C4.5 and SVM based methods. This fact demonstrates that probabilistic classifiers probably are the best for extractive summarization using machine learning. Regarding the use of CFS method, results showed that the efficiency of the methods depends largely on the set of source features. Even so, the CFS method improved the average recall for all but one system. This could mean that using machine learning techniques to find the best feature subset might be useful for some models. The authors concluded that automatic summarization can be improved by providing the adequate classifiers and features.

- **Work of Salton *et al.* (1997):** In the work of Salton *et al.* (1997), paragraphs are represented as nodes and there is an edge between two nodes according to a similarity measurement based on the number of shared words. Every paragraph is represented as a term vector and the similarity between these vectors is calculated using the scalar product. The $1.5N$ highest similarity values were selected to represent the network edges, where N is the number of paragraphs or nodes. The following routing algorithms were used to select the most important paragraphs:

  - *Global bushy path:* This algorithm selects all bushy nodes (high degree nodes) to compose the extract. These nodes are traversed in the order they appear in the text.
  - *Depth first path:* First, a bushy node is selected. Then, the node with the highest edge weight that joins the current node is visited, verifying that it is in a later position in the text. Finally, the previous step is repeated by sequentially selecting the most similar nodes until they reach the previously established size limit.
  - *Segmented bushy path:* The bushy paths are constructed for each segment and then they are concatenated maintaining the original order.

  The algorithms proposed in Salton *et al.* (1997) were evaluated using a corpus of 50 English documents. The best algorithm, Global bushy path, selected 45.6% of the paragraphs which were chosen by human summarizers. The other algorithms achieved a performance slightly better than the baseline system.

- **Work of Mani and Bloedorn (1999):** In the work proposed in Mani and Bloedorn (1999), each term or word is a node and the edge between two nodes exists according to cohesion relations. The relations of cohesion can be proximity, repetition, synonymy, and co-reference. The algorithm receives as input a topic and the extract is produced to satisfy such topic. Then, the terms that are present in the given topic are selected in the network. Then the algorithm called spreading activation traverses the other nodes related to the

topic in order to give a relevance weight to each node. Finally, the nodes with the highest relevance weights are selected to compose the final extract.

Experiments were performed using a set of five English texts. Results showed that this algorithm has a better performance in comparison to the Tf-Idf metric (ANTIQUEIRA *et al.*, 2009; SALTON; MCGILL, 1986) and other works that used node degree as the main feature to characterize networks. Also, the authors concluded that their system correlated with human summaries.

- **Work of Mihalcea (2005):** Mihalcea (2005) defines a network of sentences, which are connected according to the number of terms they share. In this work, the author uses web page recommendation algorithms for the selection of the most relevant sentences (best ranked) that will belong to final extract. The algorithms for page recommendation were Google's Page Rank and HITS algorithms (PAGE *et al.*, 1998; KLEINBERG, 1999). The author defined two network types: undirected, forward (edges follow the reading flow of the text) and backward (edges follow the flow contrary to reading the text).

  The proposed systems were evaluated with the corpus for English documents DUC-2002 and the TeMário corpus for Portuguese texts. The systems were validated by using the ROUGE-1 metric. The results showed that HITS algorithms were superior to the best DUC-2002 system when forward and backward networks were used. However, the Page Rank algorithm achieved slightly lower results than these systems when the backward network was used. For Portuguese summarization, the backward network and the Page Rank algorithm achieved the best performance.

- **Work of Erkan and Radev (2004b):** In the work of Erkan and Radev (2004b), the sentences are represented as nodes. The authors used the bag-of-words model to represent each sentence (vector elements were the Tf-Idf value of each word). In order to represent the connectivity between sentences, the cosine angle was calculated between the representative vectors of each sentence. Then, there is an edge between two nodes if the respective angle was greater than a previously established threshold. Edges were constructed without considering their weight. In order to give a relevance weight to each node network for MDS, the authors used the following measurements: degree centrality, LexRank (Page Rank algorithm applied to the network), and continuous LexRank(LexRank which uses the network with weights given by the cosine similarity).

  The proposed systems were evaluated for the corpus of English documents DUC-2003 and DUC-2004. Experiments showed that the work of Erkan and Radev (2004b) produced one of the best systems for the respective DUC conferences (2003 and 2004). LexRank was the second best method for DUC-2003, while at least one of the proposed measurements achieved the best score for DUC-2004.

- **Work of Balinsky, Balinsky and Simske (2011):** Generally, in the majority of the works, edges are created by applying some similarity measurement between the sentences or paragraphs. By using these approaches, there might not exist a control on the resulting graph type. This could generate graphs with a lot of noise and some credible ranking function could not be applied. The range of classification functions could be very narrow or only a small number of nodes would have small values of the classification functions. In this sense, in the work of Balinsky, Balinsky and Simske (2011), documents are modeled as a small-world network. Nodes are sentences or paragraphs, and the edges are defined by the Helmholtz principle (DADACHEV *et al.*, 2012). The authors tried to build a network with the same properties of a social network. Graphs are built as affiliation networks (LATTANZI; SIVAKUMAR, 2009), which are based on Helmholtz principle. This principle is used to find topics and detect unusual behavior in textual data.

- **Work of Khushboo, Dharaskar and Chandak (2010):** Extractive methods generally produce summaries that are unattractive to read. This happens because there is a lack of flow in the text, since the extracted pieces of texts are formed from different parts of the original text. In Khushboo, Dharaskar and Chandak (2010), it is proposed to extract sentences that compose a trajectory, where each sentence is similar to the previous one, allowing thus that final extracts display a better flow. First, texts are divided into sentences and words. Sentences become the graph nodes. Sentences that are similar to each other have an edge between them. The similarity between sentences was defined according to the number of common words between them. The authors also defined that all sentences have an edge to the next sentence. The edge weight was defined by a cost function. The more similar the sentences, the lower the edge cost. On the contrary, the more dissimilar the sentences, the edge cost will be higher. The final summary was constructed taking into consideration the shortest path that begins with the first sentence of source text and ends with the last sentence. This approach achieved good results because it is not only based on the local sentence context (node), but also considers the information recursively extracted from the entire text (graph).

- **Work of Ferreira *et al.* (2013b):** In Ferreira *et al.* (2013b), the authors defined a four dimension graph model where nodes represent the document sentences and the edges were defined in several ways. The authors combined different strategies for edge establishment. The strategies were the following: traditional sentence similarity, semantic similarity, co-reference resolution and discourse relations. The authors focused on this four dimension approach because it combines more dimensions than other works, which generally apply only one or two techniques to graph creation. In addition, they used co-reference resolution, which is a feature not found in graph-based related work.

  The approach presented by Ferreira *et al.* (2013b) performs a syntactic analysis of the text discourse considering semantic and linguistic aspects. In addition, the authors find

relationships between sentences by using co-references which were not used in previous works. As said previously, sentences from the document are the network nodes, and four types of edges are defined:

- *Traditional similarity:* These methods measure the overlapping content between a pair of sentences. Four different similarity metrics were used: centrality, cosine similarity, entropy measure and word co-occurrence.
- *Semantic similarity:* The traditional similarity methods are not able to find semantic relations between sentences. Such semantic relations include the meaning of words, such as synonyms, hyponym, and hypernym. The following semantic similarity measurements were used: Resnik measure, Wu and Palmer metric, Path metric, among others (RESNIK, 1995; WU; PALMER, 1994; WUBBEN; BOSCH, 2009).
- *Co-reference resolution:* This method consists of identifying the noun that is referring to the same entity. It could be useful to link sentences that are about the same subject. Three types of co-reference were used: named, nominal, and pronominal.
- *Discourse relations:* In order to consider the relationships between sentences, the authors used discourse relations. Louis, Joshi and Nenkova (2010) defined different discourse relations: cause, comparison, condition, contrast, attribution, background, elaboration, among others.

In order to rank the sentences, the authors used the TextRank algorithm (MIHALCEA; TARAU, 2004), which measures the importance of a sentence by counting the number of links to it. The authors used the CNN corpus. This corpus is a dataset of 400 texts extracted from the news of CNN website. Ferreira *et al.* (2013b) made different experiments by using only one, two, three or all of the proposed dimensions. The experiments showed that the TextRank algorithm using the four dimensions achieved the best results. The results also showed that only using semantic similarity did not achieve good results. Finally, the co-reference resolution also improved the performance in all cases it was used.

- **Work of Samei *et al.* (2014):** In Samei *et al.* (2014), the authors proposed a directed weighted graph for multi-document summarization, where the nodes represent sentences and the edges are based on distortion measures. They constructed a graph by applying an iterative ranking algorithm based on theoretical distortion measures. The sentences were represented in a directed graph. Every two sentences were examined by a distortion measure, which represents the semantic relation between such sentences. Edges between two sentences were set if the distortion was below a predefined threshold. The threshold was set according to the average of sentence distortion. Samei *et al.* (2014) used the square error as distortion measure, which is a statistical way of quantifying the difference between values.

With the aim of calculating the distortion between two sentences, each sentence is considered as a bag of words. Then, the authors assigned a score to each word based on its frequency and the position of the sentence in the whole text. The distortion of two sentences is computed based on the score of their words. After the network is created, the Page Rank algorithm is used to select the most important sentences. The sentences are ranked based on its coverage of the whole content. The authors used the DUC-2002 corpus for English documents. The results showed that the proposed system by Samei *et al.* (2014) was one of the best systems in comparison to works of DUC conferences (2002).

3.2 Final remarks

In the present chapter, we briefly reviewed some works related to document summarization by using complex network concepts and graph theory. We studied many ways of representing the documents as networks. Also, we studied a myriad of measurements to characterize the network nodes. Several works found that the modeling of documents as networks is an efficient and simple way to represent documents. Also, we found that network measurements are an excellent tool to give an importance value for the network nodes for the sentence selection stage. For the purposes of this Master's research, the works presented by Antiqueira *et al.* (2009), Ribaldo *et al.* (2012), and Leite and Rino (2008) were very important. Some of the proposed network models of this work follow the methodology of Antiqueira *et al.* (2009) and Ribaldo *et al.* (2012), where the relationship between sentences was established based on the number of common nouns or cosine similarity between such sentences. Although the construction of these models is simple, they achieved very good results, for that reason they were considered for this Master's work. Also, the methodology proposed by Ribaldo *et al.* (2012) was followed for the multi document summarization task, where all documents from the document set were processed in a single network. Finally, the proposal of this research was also inspired by the work of Leite and Rino (2008), which combined traditional summarization features with complex network measurements for a machine learning approach. Finally, in Table 2, we show a comparative table with a brief review of the studied works in the previous section.

Table 2 – Description of the main works found in literature about graph based methods for document summarization. The measurements used in these works are the following: A:node degree, B:shortest paths, C:locality index, D:dilatation strategies, E:k-cores, F:w-cuts, G:communities, H:clustering coefficient, I:diversity centrality, J:vulnerability, K:betweenness, L:global bushy path, M:depth first path, N:segmented bushy path, O:spreading activation, P:page rank, Q:HITS, R:transitivity, S:text rank

Work	Nodes	Edges	Meas.	Methodology	Language	Corpus
Antiqueira *et al.* (2009)	Sentences	Common nouns	A, B, C, D, E, F, G	Selection of best ranked sentences by any measurement	Portuguese	TeMario
Ribaldo *et al.* (2012)	Sentences	Cosine similarity	A, B, H	Selection of best ranked sentences by any measurement	Portuguese	CSTNews
Amancio *et al.* (2012)	Words	Co-occurrence	B, I, J, K	Each sentence is weighted based on the weight of its content words, and then the sentence is ranked	Portuguese	TeMario
Leite and Rino (2008)	Sentences	Common nouns	A, B, C, D, E, F, G	Machine learning to classify sentences as present in summary. Measurements are used as features	Portuguese	TeMario
Salton *et al.* (1997)	Paragraphs	Common words	L, M, N	Routing algorithms are used to construct the summaries	English	50 document corpus
Mani and Bloedorn (1999)	Words	Cohesion relations	O	Each sentence is weighted based on the weight of its content words, and then the sentence is ranked	English	5 document corpus
Mihalcea (2005)	Sentences	Common words	P, Q	Selection of best ranked sentences by any measurement	Portuguese English	TeMario DUC-2002
Erkan and Radev (2004b)	Sentences	Cosine angle	P	Selection of best ranked sentences by any measurement	English	DUC-2002 DUC-2003
Balinsky, Balinsky and Simske (2011)	Sentences Paragraphs	Helmholtz principle	H, R	Selection of best ranked sentences by any measurement	English	3 big document corpus
Khushboo, Dharaskar and Chandak (2010)	Sentences	Common words Co-occurrence	B	Sentences that compose the shortest path between first and last sentence are selected	English	DUC-2002
Ferreira *et al.* (2013b)	Sentences	Similarity, semantic similarity, co-reference, discourse relations	S	Selection of best ranked sentences by any measurement	English	CNN corpus
Samei *et al.* (2014)	Sentences	Distortion measures	P	Selection of best ranked sentences by any measurement	English	DUC-2002

CHAPTER

4

SENTENCE EXTRACTION METHOD FOR DOCUMENT SUMMARIZATION

Chapter 3 presented several works that used complex network concepts to address the automatic document summarization task. These works used different methodologies for representing documents as networks. However, in most works, only the most traditional complex network measurements for sentence selection were used. Therefore, we believe that it is important to evaluate new complex network measurements and new methodologies for network creation for automatic document summarization purposes.

In this chapter, we explain the methodology of this Master's research. First, Section 4.1 explains the corpus we used for summarizing documents for both Portuguese and English language. The methodology, which is divided into four stages, is described in the following sections: The first stage, which is explained in Section 4.2, the documents are conveniently pre-processed with the aim of model each document as a network. Section 4.3 describes the second stage, where a network is created for each pre-processed documents. We proposed the creation of different network models, where each node represents a sentence and the connection between two sentences is established in several ways. In Section 4.4, we explain the third stage, where it is employed a set of complex network measurements with the aim of giving a value of importance (relevance weight) to each network node; and in this way, the nodes could be ranked. The last stage, which is explained in Section 4.5, the best-ranked sentences (nodes) are selected to belong to the final summary. Finally, Section 4.6 the final remarks of this chapter are reviewed. In Figure 5, we show the architecture of this Master's research.

4.1 Datasets

For summary validation, it is important to have reference summaries with the aim of comparing the generated summaries by our proposed methods and the summaries made by

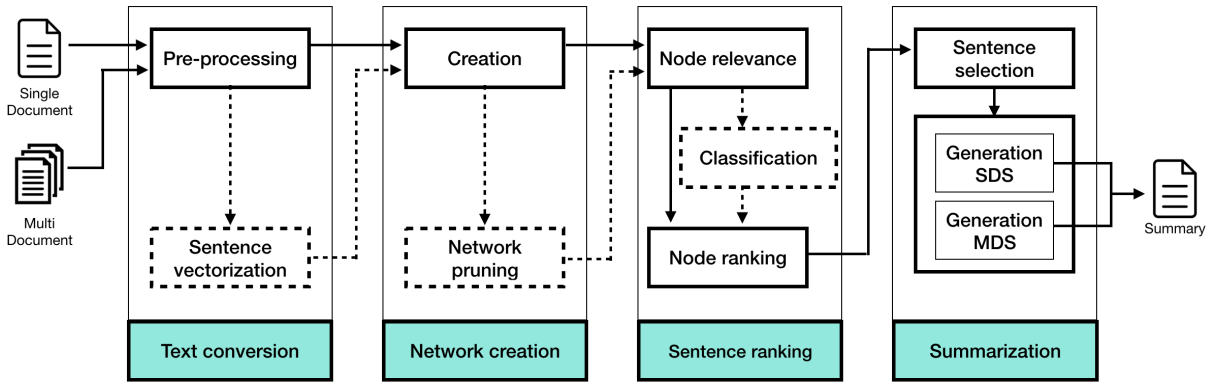


Figure 5 – Architecture of our system for this Master's research.

Source: Elaborated by the author.

humans. In order to validate the methods proposed in this research, we selected a set of corpus for document summarization for both Portuguese and English language. For Portuguese we selected the TéMario and CSTNews corpus (PARDO; RINO, 2003; CARDOSO *et al.*, 2011) for SDS and MDS respectively. In the case of English, the DUC-2002 corpus (OVER; LIGGETT, 2002) was used for both SDS and MDS.

We chose these corpus for the following reasons (ANTIQUEIRA, 2007): (i) availability of reference summaries which help us to validate all the proposed summarizers. (ii) Other works have already used this corpus, in this sense, we can compare our results with the scores of such works. (iii) We evaluated corpus for Brazilian Portuguese for being the local language; and we validated our summaries for English, because this language is the most studied for document summarization, and therefore, we can compare our systems with several works based on document summarization. (iv) Also, the evaluation of English documents allows us to use several tools available for this language. The description of each selected corpus is explained as follows:

- *Temário for Portuguese SDS:* This corpus includes a set of 100 news texts with their respective summaries. The documents were extracted from the newspapers *Folha de São Paulo* and *Jornal do Brasil*. Summaries were constructed by professional human summarizers, considering that summaries should be 25 to 30% of the size of original texts.

- *CSTNews for Portuguese MDS:* This corpus presents a set of documents extracted from the following online Brazilian news agencies: *Folha de São Paulo, Estadão, O Globo, Gazeta de Povo*, and *Jornal do Brazil*. CSTNews corpus consists of 140 news documents, which are grouped into 50 clusters. Each cluster includes 2 or 3 documents which share the same topic. This corpus contains two reference manual multi-document summaries for each group of documents. Each summary was created with a 70% compression rate.

- *DUC-2002 for English SDS and MDS:* This corpus includes a collection of 567 documents divided into 59 clusters. The documents were extracted from the following online news journals: Financial Times, AP Newswire, San Jose Mercury News, LA Times, FBIS, and Wall Street Journal. Each document set contains two 100-word and 200-word reference summaries.

In Table 3 we show the main features of the corpus we used for extractive document summarization.

Table 3 – Main features of the corpus we used for Document Summarization.

Description	TeMário	CSTNews	DUC 2002
Goal	Portuguese SDS	Portuguese MDS	English SDS-MDS
Documents	100	140	567
Clusters	-	50	59
Docs. per cluster	-	2 to 3	6 to 10
Avg. sentences per doc.	29.37	16.8	27.92
Summary size	70% compression rate	70% compression rate	100-200 words

4.2 Stage 1: Text conversion

The documents to be summarized are represented as networks, where each extracted sentence from each document represents a node in the network. In order to model sentences as network nodes, we must apply a set of changes to the source texts. Such changes include, for instance, the elimination of irrelevant words and the transformation of words into their canonical form. The proposed method also requires representing sentences in numerical vectors to be used in embedding models. This stage comprises the document pre-processing and sentence vectorization sub-stages.

4.2.1 Document pre-processing

In this phase, the following pre-processing steps are applied: text segmentation, removal of unnecessary words, morphosyntatic labeling, and lemmatization. For text segmentation, the documents are divided into sentences. A sentence is defined as any text segment which is separated by a period, exclamation or question mark. We decided to divide documents into sentences because the sentences are the basic unit for extractive summarization. Punctuation marks and stopwords were also removed. Finally, the remaining words are lemmatized with the aim of obtaining their canonical forms. Next, we explain how we implemented the pre-processing steps commented before.

- **Text segmentation:** Because the nodes of the proposed network models are sentences, at this stage, it is required to recognize each document sentence. We used the Python Natural

Language Toolkit (NLTK) (BIRD, 2006) for text segmentation. Some of the corpus we used already had the documents segmented into sentences.

- **Elimination of stopwords and punctuation marks:** Stopwords are words that usually are filtered out in the pre-processing stage in any NLP task. These words are removed because they do not contribute semantic content. Although there is no a single universal list of stopwords, this list commonly includes prepositions, adverbs, and articles. The inclusion or exclusion of this type of words depends on the NLP task that is being considered. For the purposes of this research, we employed a stopword list which contains prepositions, adverbs, articles, and punctuation marks. In Appendix A, the stopword lists for both Portuguese and English are displayed.

- **Morphosyntatic labeling:** The Part Of Speech Tagging (POST) consists of associating each word with its morphosyntatic identification, in this sense, the word is classified as preposition, verb or noun. POS Tagging is important for word lemmatization and for the identification of all nouns composing a sentence. In this phase, we used the MXPost Tagger (RATNAPARKHI *et al.*, 1996) for Portuguese and Python NLTK library (BIRD, 2006) for English.

- **Lemmatization:** This phase consists of transforming the words into their respective canonical form. These canonical forms are called lemmas. This stage allows us to process in a unique way the different variations of a word. In this way, plural nouns are transformed to their singular version while conjugated verbs are converted to their infinitive forms. We used the WordNet Lematizer from NLTK library for English text lemmatization. For the lemmatization of Portuguese documents we used the Portuguese Lemmatizer developed by the researchers of the Núcleo Interinstitucional de Linguística Computacional (NILC) of the Universidade de São Paulo.

In order to illustrate the document pre-processing stage, we provide in Table 4 a small piece of text including the previously mentioned pre-processing steps.

4.2.2 Sentence vectorization

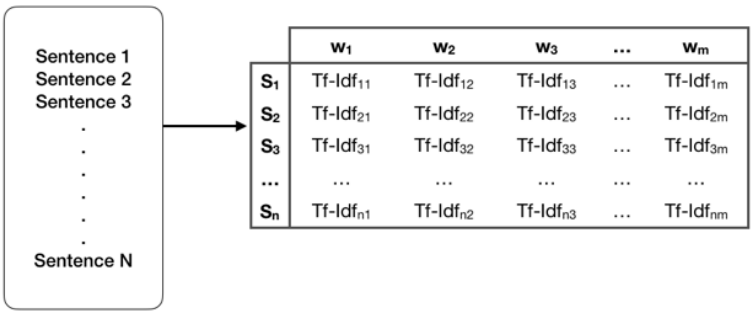
In the next stage, referred as network creation, we constructed different network models such as Noun, Tf-Idf or Embedding-based network. For the creation of some of these network models, our method requires the sentences to be transformed into their vector form. By using this new representation, we generated new relationships between sentences. These relationships are established based on the lexical similarity between them. This similarity is computed according to the distance between the representative vectors of the sentences. As explained in Chapter 2, we used the vector space model based on the Tf-Idf weighting and models based on word embeddings. Below we describe how these vector representation models are used for the purposes of this Master's work.

Table 4 – Example of the pre-processing steps extracted from Tohalino and Amancio (2017). In this example, we applied all the pre-processing phases for a small piece of text extracted from Wikipedia. First, we show the original text divided into six sentences. Then, we present the corresponding pre-processed sentences. In addition, we highlight the shared nouns between sentences.

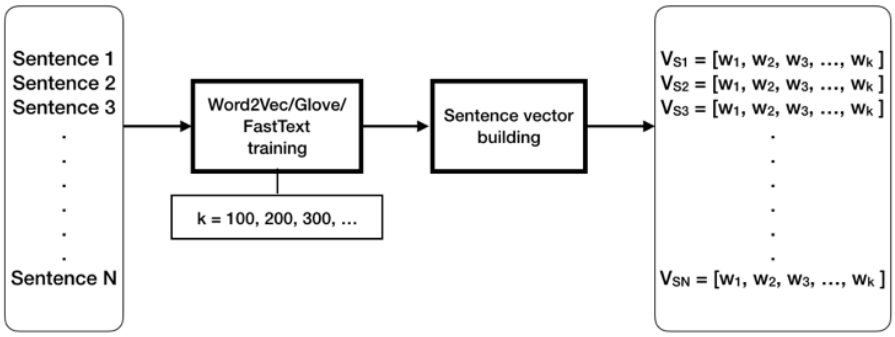
Original Sentences
1. Brazil is the largest country in South America
2. It is the world's fifth-largest country by both area and population
3. It is the largest country to have Portuguese as an official language and the only one in America
4. Bounded by the Atlantic Ocean on the east, Brazil has a coastline of 7,491 kilometers
5. It borders all other South American countries except Ecuador and Chile
6. Brazil's economy is the world's ninth-largest by nominal GDP of 2015
Pre-proccesed Sentences
1. **brazil** be large **country south america**
2. be **world** five large **country** area population
3. be large **country** have portuguese official language **america**
4. bound atlantic ocean east **brazil** have coastline kilometer
5. border **south america country** ecuador chile
6. **brazil** economy be **world** nine large nominal gdp

- **Sentences as Tf-Idf vector:** Since the Tf-Idf weighting has been commonly used with good results for many NLP tasks (ROBERTSON, 2004), we applied this metric for the vector representation of sentences. To calculate the representative vector of each sentence, we compute the Tf-Idf value of each of its content words, where Tf is the term frequency and Idf is the inverse document frequency. The calculation of the Tf-Idf weighting was explained in Chapter 2. In Figure 6 we show an example of the sentence vector computation based on Tf-Idf model.

- **Sentences as embedding vector:** With the aim of improving the document representation and addressing the disadvantages of the Tf-Idf weighting model, we used other models based on word embeddings. We used three different word embedding models for sentence representation: Word2Vec, GloVe, and FastText embedding. Several issues we should take into consideration for the training stage of these models. First, it is needed to determine the ideal size of the representative vectors of the sentences. The size could depend on the corpus length and other factors. For this reason, we made several tests in order to find the ideal number of elements of the sentence vectors. Second, we found several pre-trained models for these embeddings for both the Portuguese and the English languages. These pre-trained models were trained over big corpus by other researchers. We made a myriad of tests by using the found pre-trained versions and the embeddings trained with our corpus. Finally, the models we used are based on word embeddings, i.e. we have the vector representation of each corpus word. Therefore, we need to build the sentence vectors from these word vector representations. For the sentence vector construction, first we get the embedding representation of each of the content words of a sentence, and then, we simply



(a) Sentence representation based on Tf-Idf model



(b) Sentence representation based on word embeddings

Figure 6 – Example of the sentence vectorization stage. (a) Example of the process for sentence vectorization using the Tf-Idf model. The vector size for this model depends on the number of words the vocabulary has. (b) Example of the process for sentence vectorization based on word embeddings. The vector size for this model is fixed.

Source: Elaborated by the author.

calculate the average of each element from word vectors. In Equation 4.1 we show how the construction of the vector of a sentence *Sn* is made.

$$V_{Sn} = [avg(v_1), avg(v_2), avg(v_3), ..., avg(v_k)], \tag{4.1}$$

where v_m represents a vector of the elements at m position in each word vector representation and *avg* is the average of such vector. We show in Figure 6 an example of the process of sentence vector creation based on word embeddings.

4.3 Stage 2: Network creation

After the document pre-processing step, the next goal is the construction of the network representing a single document (SDS) or multiple documents(MDS). In all proposed networks, nodes represent the sentences. For MDS, all document sentences from the document set are modeled in a single network, except for the multilayer network approach. In this section, we explain the different network models we used.

4.3.1 Noun-based network

Following the work of Antiqueira *et al.* (2009), each node is a sentence which is represented by its lemmatized nouns. An edge between two sentences occurs when there is at least one noun in common between such sentences. The edge weight is determined by the number of word repetitions between both sentences. As Antiqueira *et al.* (2009) suggest, we used only nouns with the aim of avoiding a large number of edges in the network. If all sentence words were included, the discrimination between sentences would be hampered because we are creating many additional relationships between words that are not necessarily indicators of main concepts in the sentences. In preliminary experiments, we found that a better performance is obtained when only nouns are considered in the sentences. Figure 7 shows an example of a noun based network which was generated from the example shown in Table 4.

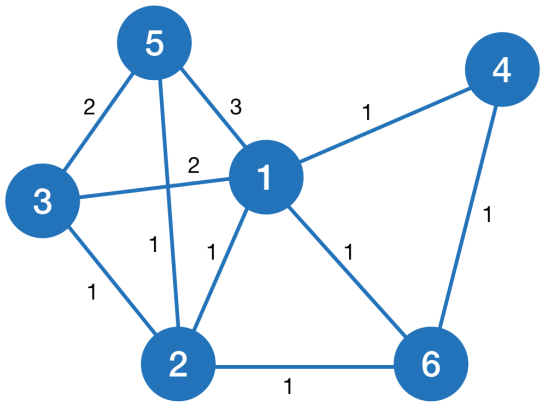


Figure 7 – Example of a Noun based network generated from the piece of text of Table 4. Each node represents the document sentences and the edges represent the number of common nouns between two sentences.

Source: Tohalino and Amancio (2017).

4.3.2 Tf-Idf based network

We create this network following the work of Ribaldo *et al.* (2012). In order to build this network, we first need to calculate the Tf-Idf vector representation of each document sentence. Then, each node network is represented by a sentence. We used the similarity between the Tf-Idf vectors of two sentences for edge establishment. The similarity is computed as the cosine similarity obtained from the Tf-Idf vectors. In Equation 4.2 we show the calculation of cosine similarity.

$$cosSim(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \tag{4.2}$$

where A and B are the representative vectors of two sentences. In Figure 8 we show an example of the Tf-Idf based network generated from the example shown in Table 4.

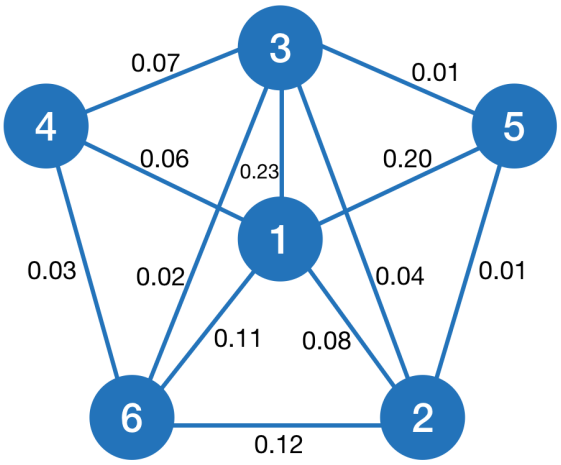


Figure 8 – Example of a Tf-idf based network generated from the piece of text of Table 4. Each node represents the vectorized value of the sentences and the edges are the cosine similarity between such sentences.

Source: Tohalino and Amancio (2017).

4.3.3 Embedding-based network

For the creation of this network, we first follow the creation process based on Tf-Idf model, with the difference that in this model we use the vectors based on word embeddings instead of Tf-Idf vectors. As said before, we generate three different network models based on word embeddings: Word2Vec, GloVe, and FastText. In this approach, an important issue arises. The generated graphs have a large number of edges. It may even be the case that completed graphs are built, where all the network nodes are connected. This is a big problem because it would affect the performance of most complex network measurements. In Figure 9 we show an example of a complete graph. In the case we apply some network measurement to this graph, for example, the degree metric, all nodes will have the same value (in the example of Figure 9 the node degree is 5), in this way, we could not make a distinction between the network nodes. However, if we use the strength measurement, the edge weights are considered; thus, the strength of nodes A and B from the graph of Figure 9 will be 11 and 7 respectively. So, we can see that node A has a greater relevance weight in comparison to node B. However, the consideration of edge weight is not the best solution since some network measurements adopted in this work do not use the edge weights. For this reason, the best solution is to remove a subset of edges from the network.

Hence, we applied a method to remove some redundant edges in this network. Redundant edges mean those edges with the lowest weight values. For edge removal purposes, we removed a fraction r of the weakest edges. Figure 10 shows an example of a complete graph which was generated from the sentences of Table 4 based on sentence embeddings. In the same illustration, we additionally show two graphs generated from the original network; which were applied the

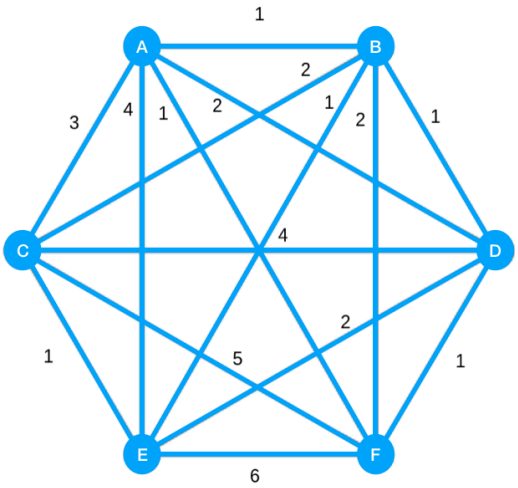


Figure 9 – Complete graph where all nodes have degree 5

Source: Elaborated by the author.

edge redundant removal stage.

4.3.4 Multilayer network for MDS

In this model, differently from previous studies in MDS, we make a distinction between edges linking sentences from different documents (inter-layer edges) and those which connect sentences from the same document (intra-layer edges). In order to create a multilayer based representation of a set of documents, the following steps are required:

- **Tf-idf based network creation:** The network is created following the same steps required for the Tf-Idf based network previously studied in this research.
- **Edge type identification:** This step requires to identify two edge types. The first edge type is called intra-layer edge, which connects sentences from the same document. The second edge type, inter-layer edge, connects sentences from different documents. We need to differentiate these edge types in order to give a more relevance to the sentences according to the edge types which are established.
- **Type-based edge weighting:** For MDS tasks the proposed method needs to give priority to multi-document relationships (PADMANABHAN *et al.*, 2005; WEI *et al.*, 2010). In this sense, we believe it would be relevant to increase or decrease the weight for inter-layer edges. Such an increase or decrease is made by using the following function:

$$w_{i,j}^{(inter)} = \alpha w_{i,j}^{(inter)} \tag{4.3}$$

where α is a factor that increases or decreases the weight value of the inter-layer connections (if $\alpha > 1$ or $\alpha < 1$ respectively), and $w_{i,j}^{(inter)}$ represents the original inter-layer edge

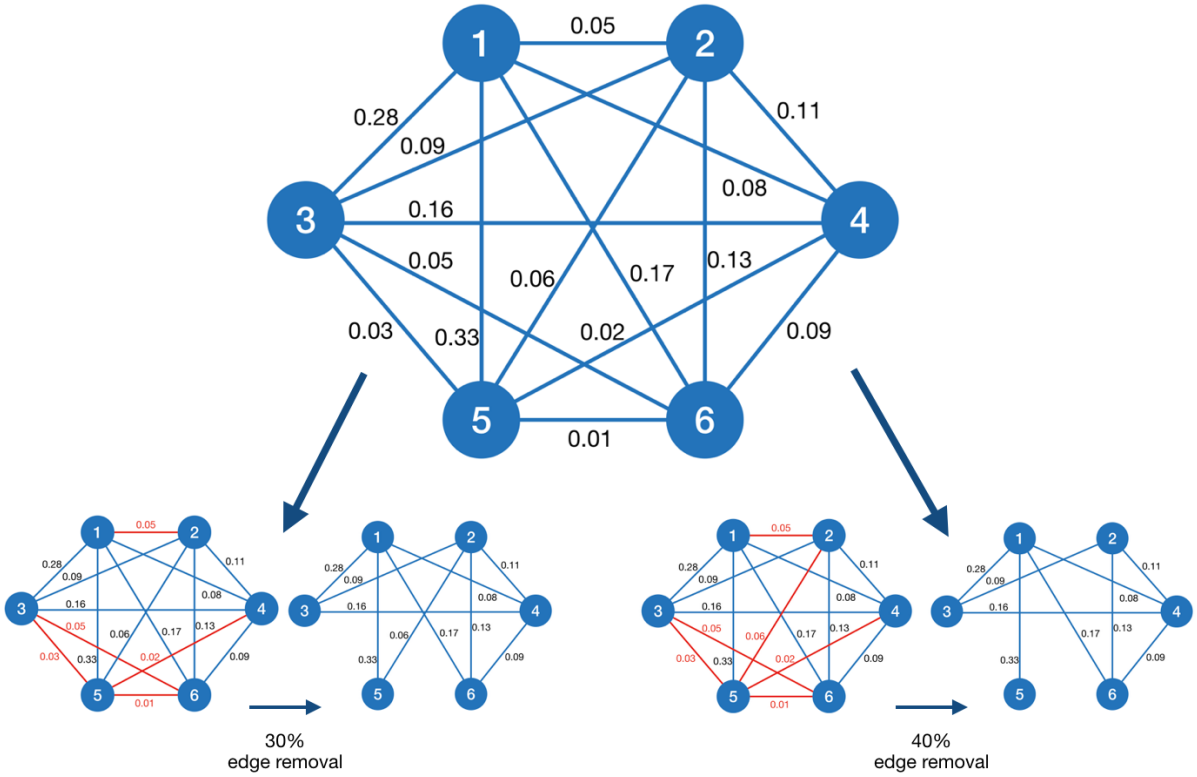


Figure 10 – Example of an embedding-based network generated from the piece of text of Table 4. Each node represents the embedding vector of a sentence and the edges are based on the cosine similarity of such sentences. We see above that a complete graph was generated. The graphs below represent the same graph with the elimination of 30 and 40% of its weakest links.

Source: Elaborated by the author.

weight which connects nodes i and j. Such a reinforcement of inter-layer edges could also be relevant for the representation of the set of documents. In our experiments, we considered $\alpha < 1$ and $\alpha > 1$ with the aim of simulating these effects.

- **Edge removal for non-weighted measurements:** Several measurements used in this research are not defined for weighted networks, therefore, it is required to remove a portion r of the weakest links when we consider such metrics.

Figure 11 represents an example of a multilayer network which was generated from a document set containing three small texts.

4.4 Stage 3: Sentence ranking

In this stage, we use a set of network measurements with the aim of giving a value of importance (relevance) to each node. This relevance weight allows us to rank the nodes so that the best-ranked sentences (nodes) compose the final summary. This stage is divided into two sub-stages: node relevance and node ranking.

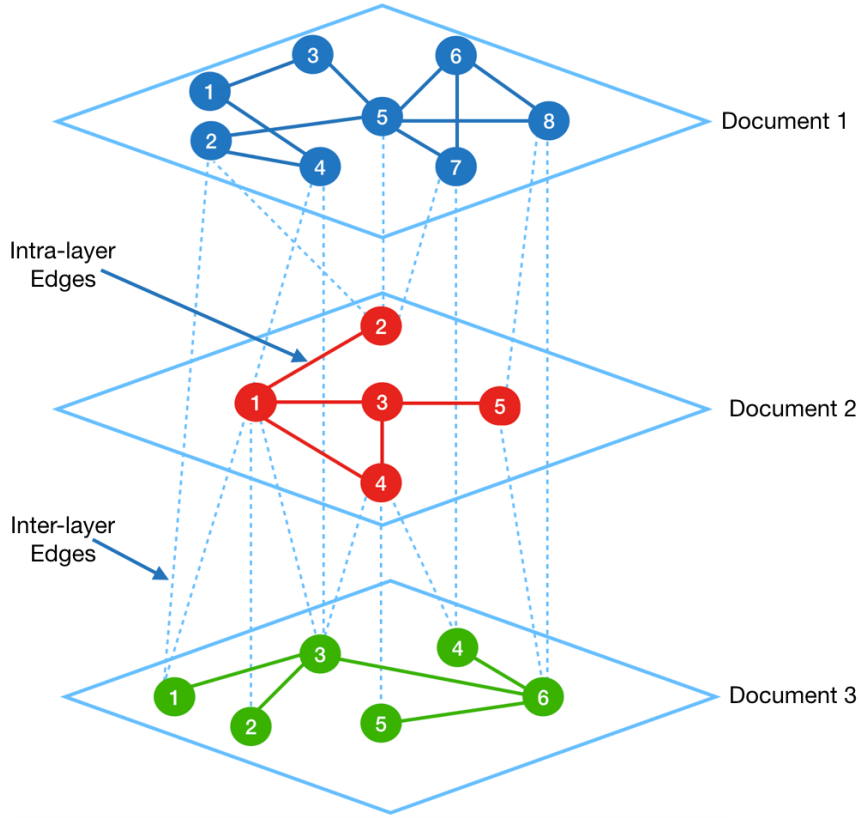


Figure 11 – Multilayer network representation of a document set of three documents. Each layer represents a document. Continuous lines are the edges which connect sentences from the same document (intra-layer edges), while dashed lines connect sentences from different documents(inter-layer edges).

Source: Tohalino and Amancio (2018).

4.4.1 Node relevance

In the summarization context, the goal of a centrality network measurement is to rank the nodes according to their relevance. The importance assigned by network measurements allows us to determine which are the best-ranked sentences that could compose the final summary. Therefore, here we use a set of network measurements to rank each node of the network. In this Master's research, we used not only traditional network measurements such as degree, strength, etc.; but also additional measurements to take into account both topological structure of the networks and their dynamical behavior. This behavior can be achieved by recognizing the dynamical processes which occur on the top of the networks. In order to analyze such dynamical behavior, we considered the variation of random walks (MASUDA; PORTER; LAMBIOTTE, 2017). Such a dynamics originated the following network measurements: concentric metrics, accessibility, symmetry and absorption time. In Chapter 2, we described the definition and computation of these measurements, then we explain how these network measurements are used in this research.

- **Traditional network measurements:** It includes the following measurements: degree, strength, clustering coefficient, Page Rank, shortest paths, betweenness and their weighted versions. The edge weights are determined according to the number of common words of the sentences or it is based on the cosine similarity between such sentences. We used these metrics for several reasons, for example, sentences (nodes) with high degree values suggest that such sentences are related to several others in the document. Another useful example we can see for shortest path metric. By analyzing such measurement, we can consider a sentence as relevant if its average distance to any other sentence takes low values. In other NLP works, the shortest path metric has been used with success to identify textual concepts (AMANCIO *et al.*, 2011). We used the complex network Igraph software package (CSARDI; NEPUSZ, 2006) for the network creation and the application of network measurements.

- **Concentric measurements:** With the aim of capturing important topological information from the network, we used a set of eight concentric measurements to rank the network sentences. These measures are the following: concentric number of nodes (conc_1), concentric number of edges (conc_2), concentric node degree (conc_3), concentric clustering coefficient (conc_4), convergence ratio (conc_5), intra-ring node degree (conc_6), inter-ring node degree (conc_7), and concentric common degree (conc_8). In addition, these measurements were evaluated at hierarchical levels $h = 2$ and $h = 3$. These concentric measurements were provided by Costa and Silva (2006) and Silva and Costa (2013).

- **Accessibility based measurements:** By using these measurements, we intend to select the most accessible nodes (sentences) in order to be included in the summary. The accessibility was evaluated at hierarchical levels $h = 2$ and $h = 3$; and also we used the generalized accessibility measurement, which was successfully used in other text classification tasks (AMANCIO; SILVA; COSTA, 2015). These measurements were developed by Viana, Batista and Costa (2012) and Arruda *et al.* (2014).

- **Symmetry based measurements:** The goal of using these measurements is to determine if nodes with a higher degree or lower degree of symmetry are good indicators of sentence relevance. We made different experiments by selecting the nodes with the highest and lowest symmetry values. Symmetry measurements were evaluated at hierarchical levels $h = 2$ and $h = 3$. These measurements were developed by Silva *et al.* (2016b).

- **Measurement based on absorption time:** According to this measurement, we believe that sentences with low values of absorption time are more semantically related to the other sentences in the document, therefore, they are ideal to belong to the final summary. This measurement was implemented as explained in the work of Amancio, Jr. and Costa (2011).

4.4.2 Node ranking

After the relevance of each node is weighted, such relevance weights will allow us to rank the sentences in two different ways: sentences could be ranked directly by every network measurement or we could rank the sentences by using a hybrid approach that combines complex networks and machine learning concepts. Next, we explain both methods.

- **Ranking by network measurements:** In this step, when network nodes are evaluated, the sentences corresponding to these nodes are re-ordered according to the relevance weight they received. Therefore, the top-ranked sentences have a high probability to be included in the final summary. Every network measurement is used in an individual way, therefore, there is one different summary for each measurement. The selection strategy used for each network measurement is summarized in Table 5.

Table 5 – Adopted network measurements. We considered the weighted version of the networks only with the most traditional measurements

Selection strategy	Measurement	Abbr.
Highest values	Degree	dg
	Strength	stg
	Betweenness	btw/btw_w
	Page Rank	pr/pr_w
	Clustering Coefficient	cc/cc_w
	Concentric	conc_{1,2,3,...,8}
	Accessibility	access_h2/access_h3
	Gen. Accessibility	gAccess
	Symmetry	sym_h2/sym_h3
Lowest values	Shortest Paths	sp/sp_w
	Symmetry	sym_h2/sym_h3
	Absorption Time	absT

- **Ranking by machine learning:** In the first approach, we observe that every network measurement will produce a different summary. We apply this new approach with the aim of producing a single summary which uses all or some group of the proposed network measurements together. Here, we combine complex network concepts with traditional summarization methods by using a machine learning system. We intend to classify sentences as "Present in Summary" or "Not present in Summary" by using a set of supervised classifiers. We use the network measurements proposed in this work as features for a trainable classifier. In addition, following the work of (LEITE; RINO, 2008), we combined network features with several features used in traditional summarization works.

  In order to rank the sentences, we first build our training set, which comprises the document sentences and their labels (sentences belonging to summary or not). We used the reference extracts from corpus to determine the sentence labels. Second, we used 10

fold cross-validation for training and testing the samples (KUPIEC; PEDERSEN; CHEN, 1995b). As Leite and Rino (2008) suggests, our focus is not the predicted class, but also the probability of a sentence to belong to the desired class. In this way, sentences are ranked according to the likelihood they have to belong to "Present in Summary" class, i.e. sentences with the highest probabilities of pertaining to "Present in Summary" class are the best for being present in the final summary. The classifiers we used in this work are detailed as follows (KUPIEC; PEDERSEN; CHEN, 1995b):

- *Support Vector Machine (SVM):* It is a binary classifier. Given the training examples (labeled sentences), it constructs a hyperplane as a decision boundary with the property of the margin of separation between two classes being maximum (HAYKIN *et al.*, 2009). The likelihood of a sentence belonging to the "Present in Summary" class is calculated based on the distance between the hyperplane and the extract (LEITE; RINO, 2008).
- *Naive Bayes (NB):* Naive Bayes classifiers are a set of supervised learning algorithms based on applying Bayes theorem with strong (naive) independence assumptions between the features (ZHANG, 2004). By using this classifier, we computed the probability of a sentence to belong to the "Present in Summary" class.
- *Decision Tree (DT):* The goal of using decision trees is to create a model which predicts the value of a target variable by learning simple decision rules which are inferred from the features (QUINLAN, 2014). We calculate the sentence probability through the relative frequency of the "Present in Summary" class in the decision leaf (LEITE; RINO, 2008).

As we commented before, in this work we used a set of features which are commonly used for automatic summarization with machine learning approaches. The features we used are described as follows (NETO; FREITAS; KAESTNER, 2002; LEITE; RINO, 2008):

- *Word frequency (F1):* The sentence weight is calculated by counting the number of most frequent words which such sentence has.
- *Word Tf-Idf (F2):* This feature is calculated as the mean value of the Tf-Idf values of all the words that contain a sentence.
- *Indicator of main concepts (F3):* The main concepts are assumed to be found in the most relevant nouns. This is a binary feature which indicates the presence or absence in a sentence of words classified as main concepts.
- *Occurrence of proper names (F4):* The occurrence of proper names such as people or places are clues that a sentence is relevant for the summary. This is a binary feature which indicates if a sentence contains at least one proper name (True) or not (False).

- *Occurrence of non-essential information (F5):* Neto, Freitas and Kaestner (2002) consider the following words as indicators of non-essential information: "because", "furthermore", and "additionally". This feature takes the value True if the sentence contains at least one of these discourse markers, and False otherwise.
- *Topic modeling (F6):* Topic words are a set of words that best describe a set of documents. Each extracted topic word from documents is used to determine the sentence score (LEE; BELKASIM; ZHANG, 2013b). The LDA model (LEE; BELKASIM; ZHANG, 2013b) generates a set of topic words for corpus documents. The sentence weight is computed as the ratio of the number of topic words the sentence contains over the sentence size.
- *Sentence length (F7):* It is the ratio of the number of words occurring in the sentence over the number of words occurring in the longest sentence.
- *Sentence to sentence cohesion (F8):* For each sentence, we compute the similarity between this sentence and the other sentences and then we add up those similarity values, obtaining the value of this feature. Next, we calculate the normalized version of this feature by dividing the raw value between the largest raw value among all sentences.
- *Sentence to centroid cohesion (F9):* In order to compute this feature, we first calculate the vector which represents the document centroid, and then we compute the similarity between the centroid and each sentence. The normalized value of this feature is obtained by computing the ratio of the raw feature value over the largest raw feature value among all sentences.
- *Sentence position (F10):* This feature gives a higher score to the sentences which appear at the beginning of the document.

In Figure 12 we show a scheme that summarizes the sentence ranking stage.

4.5 Stage 4: Summarization

In this stage, the best-ranked sentences are selected to compose the summary. In the first place, this stage comprises the process of the selection of best sentences, i.e. the most informative and relevant. Also, in the case of MDS, it is important to avoid redundancy in the selected sentences. It is also important to mention that generated summaries must respect a previously established size.

To generate the summaries, it is key to determine the compression rate of the summaries in relation to the original document. The summary size should be adapted according to the size of the reference summaries we have. Generally, summaries have a compression rate of 70% of the original text (RIBALDO *et al.*, 2012). The size can be based on the number of words or

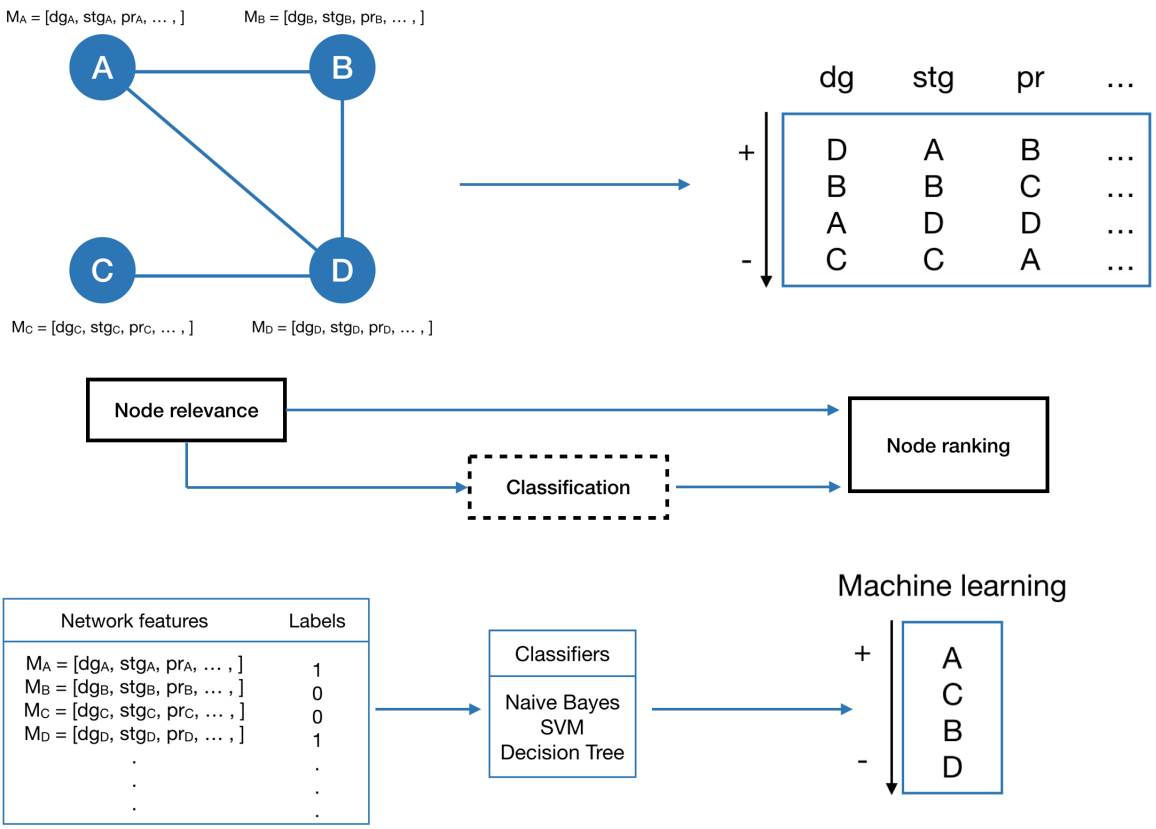


Figure 12 – Example of the process of the sentence ranking stage. This stage comprises the node weighting and node ranking steps. Above we could see the process of ranking by network measurements, which produces *n* different possible sentence rankings. Below we see the process of ranking by machine learning, which produces only one ranking.

Source: Elaborated by the author.

number of sentences which are present in relation to the original document. The summary size can also be established as a fixed number, for example, the summaries for DUC-2002 corpus must have a fixed size of 100 to 200 words.

After setting the summary size, the next step is the summary construction. For SDS the process is simple because it consists of selecting the best-ranked sentences up to the limit of the established size. However, the MDS approach requires an additional step in the sentence selection stage. This additional stage is the so-called redundancy treatment. In the context of automatic summarization, redundancy arises when identical or similar sentences composes the final summary. In network terms, two sentences in the final summary are considered redundant if they are linked by strong connections (RIBALDO *et al.*, 2012). In this Master's research we used two anti-redundancy detection methods (RIBALDO *et al.*, 2012; GAIZAUSKAS; SAGGION, 2004). In both methods, a similarity threshold is established to compare sentences. At each step, if the current best-ranked sentence is similar to any of the previously selected sentences in the summary, then it is considered redundant. In this sense, the redundant piece of text is not included in the final extract. In this case, the summarization process continues and the next candidate

sentence is evaluated. In the first anti-redundancy detection method (AR1), the threshold value (L1) is computed as follows:

$$L_1 = \frac{max(\sigma(i,j)) - min(\sigma(i,j))}{2}, \tag{4.4}$$

where $\sigma(i,j)$ represents the cosine similarity between two sentences i and j. The second anti-redundancy detection method (AR2) is based on the following similarity measurement:

$$\tilde{\sigma}(i,j,n) = \sum_{k=1}^{n} \gamma k. \frac{|g(i,k) \cap g(j,k)|}{|g(i,k) \cup g(j,k)|} \tag{4.5}$$

where n represents the number of n-grams to be considered, $g(i,k)$ is a group of k-grams of the sentence i, and γk is a weight which is associated with the k-gram similarity of two sets. According to Gaizauskas and Saggion (2004), we chose as threshold for $\tilde{\sigma}$ the value $L_2 = 0.1$.

4.6 Final remarks

In this chapter, we described the methodology of this Master's research. First, we explained the datasets we used as well as their importance for the objectives of this work. With the objective of efficiently managing each process of this research and a better understanding of such processes, the proposed methodology was divided into four stages. The text conversion, network creation, sentence ranking, and summarization stages were detailed in the present chapter. In Chapter 6, we briefly mentioned the contributions, limitations and prospects for future works for each of these stages.

CHAPTER

5

EXPERIMENTS AND RESULTS

In this chapter, we show and discuss the main results obtained from the evaluation of our systems produced for the summarization of Portuguese and English documents. Our goal is to make a comparative analysis between different techniques to generate a set of ideal extracts. After the analysis is made, we chose the best summarizers in order to meet the objectives of this Master's research. Additionally, the proposed methods are compared with the results of other works that address the extractive summarization task and which used the corpus and evaluation methods employed in this work. The proposed systems for document summarization were applied to three different corpus of news documents, which were transformed into sentence networks. The corpus we used were TeMário and CSTNews for Portuguese and DUC-2002 for English documents. The quality (informativity) of the generated summaries was validated by using the ROUGE-1 recall metric.

For comparison purposes, we show in Table 6 a set of other works that achieved the best results for the corpus we used. The methods based on the algorithms $HITS_AB$ and PageRank B. (MIHALCEA, 2005), defined in Chapter 3, were used for both Portuguese and English SDS. The methods SuPor2-LogistRegr and SuPor-2 (LEITE; RINO, 2008; LEITE; RINO, 2006), which were explained in Chapter 3, were used for Portuguese SDS. For Portuguese MDS the following systems were found in the literature: GistSumm (PARDO; RINO; NUNES, 2003) which was the first MDS system produced for Portuguese language, which selects sentences according to the frequency of their words; CSTSumm (JORGE; PARDO, 2010), which follows a CST-based method (Cross-document Structure Theory); and Bushy Path and Depth-first Path systems (RIBALDO *et al.*, 2012), which adapt the Relationship Map approach for MDS. In addition to the works proposed in (MIHALCEA, 2005), for English SDS the ntt.duc02 and ULeth131m systems (HIRAO *et al.*, 2002; BRUNN; CHALI; DUFOUR, 2002) were considered, which are the systems that achieved the best scores for the DUC-2002 Conference (OVER; LIGGETT, 2002). Finally, for English MDS the following systems were considered: DUC-best, which is the system with highest ROUGE scores for DUC conferences, BSTM (WANG *et al.*,

2009), which employs a Bayesian sentence-based topic model for document summarization, FGB (WANG *et al.*, 2008), which proposes a new language method to group and summarize the documents; and LexPR (ERKAN; RADEV, 2004a), which creates a sentence graph based on cosine similarity, which selects relevant sentences according to its eigenvector centrality. All systems were compared according to their ROUGE-1 recall value (RG-1).

The performance of our methods was also compared with baselines summaries. In this work we used two baselines summaries: in order to construct the extract, the first baseline, called Top baseline, selects the first n sentences of the source document, while the Random baseline makes a random selection of sentences from the source document (ANTIQUEIRA, 2007). For the Random baseline system, we make a generation of n different random summaries (where n was set to 100), and then, we selected the more repeated sentences to be included in a unique random summary. Both systems (top and random baselines) are very simple, and they are a baseline for the validation of the proposed systems. For example, if the performance of some system is close to the baselines performance, this system is considered irrelevant or critical, because simpler systems like baselines achieve better results. However, the Top baseline systems display a good performance when they are generated for journalistic documents (ANTIQUEIRA, 2007). Table 6 shows the scores of Top and Random baselines for document summarization.

Table 6 – ROUGE-1 scores (RG-1) for other works that achieved the best performance for Portuguese and English Document Summarization. Also we show the results obtained from the Top (Top B.) and Random (Random B.) baselines for each language and summarization method.

Portuguese Summarization				**English Summarization**			
SDS		**MDS**		**SDS**		**MDS**	
System	**RG-1**	**System**	**RG-1**	**System**	**RG-1**	**System**	**RG-1**
Supor2-LogistRegr	0.5316	GistSumm	0.6643	*HITS$_A$B.*	0.5023	DUC-best	0.4986
Supor-2	0.5227	*Top B.*	*0.5497*	ntt.duc02	0.5013	BSTM	0.4881
PageRank B.	0.5121	Bushy Path	0.5397	PageRank B.	0.5008	FGB	0.4850
HITS$_A$B.	0.5002	Depth-first Path	0.5340	ULeth131m	0.4911	LexPR	0.4796
Top B.	*0.4757*	CSTSumm	0.5065	*Top B.*	*0.4860*	*Top B.*	*0.3928*
Random B.	*0.4565*	*Random B.*	*0.4622*	*Random B.*	*0.4258*	*Random B.*	*0.3623*

The remaining of this chapter is organized as follows: First, in Section 5.1, we make a comparison between the noun and Tf-Idf based networks. Second, Section 5.2 shows the performance of the different word embedding based methods we used for document summarization. In third place, in Section 5.3, we show the evaluation of the multilayer approach for Multi-Document Summarization. In Section 5.4, we also show the evaluation of a hybrid approach which combines the complex network concepts with machine learning methods for sentence classification for document summarization. Finally, in Section 5.5, we discuss and show a comparative analysis of all proposed methods in this work. Also, the results of other works that achieved the best scores for each language and summarization method are shown.

5.1 Noun and Tf-Idf based Network

Here we compare the performance of two network models: Noun and Tf-Idf based networks for SDS and MDS. We compared these models together because both models displayed a similar performance. We believe they have a similar performance because their construction is based on the number of shared words between each document sentences. While the noun-based network considers the number of common nouns between two sentences, the Tf-Idf based network uses a similarity measurement (cosine similarity) to quantify the number of shared words between such sentences. Table 7 shows the average ROUGE-1 Recall scores (RG-1) for Portuguese documents while Table 8 presents the results for the evaluation of English documents. Note that for MDS we used two anti-redundancy detection methods (ARD): cosine similarity based method (AR1) and n-gram similarity method (AR2). In Table 7 and Table 8 we only show the results of the ARD methods that achieved the best scores. The network measurements we used were grouped as follows: traditional network measurements (degree, shortest paths, Page Rank, betweenness, and clustering coefficient), concentric metrics (set of eight measurements), accessibility based measurements (accessibility and generalized accessibility), symmetry, and absorption time.

The results showed that generally both noun and Tf-Idf based networks achieved good scores for some network measurements, close to the best summarizers for Portuguese and English. We can also see that our systems achieved a better performance for Portuguese documents. All of our best summarizers for Portuguese outperformed the top baseline performance and they were very close to the best systems found in the literature for both SDS and MDS. Also, for MDS we achieved excellent results which were superior to such best systems. In the case of English document summarization, our best systems did not outperform the top baseline performance for SDS, however, they were close to the results of other works. For English MDS, our best systems outperformed the top baseline, nonetheless, they achieved a poor performance because their scores were inferior to the results of other works.

For MDS we made different types of experiments for sentence selection. In a first approach, we make a simple selection of best-ranked sentences without using some anti-redundancy detection method (ARD). In a second approach, we used two ARD methods to remove redundant sentences. The results we obtained show that applying anti-redundancy detection methods (ARD) does not have a big impact on the summary quality because the ARD methods had a slightly better performance than the simple selection method. We could conclude there is not great relevance in applying the adopted ARD methods for the corpus we analyzed. The ARD methods proposed in this work (AR1 and AR2) achieved a similar performance, in some cases the AR1 method outperformed the AR2 methods, however, in other cases the AR2 method achieved the best performance. More specifically, for Portuguese, the AR2 method was better than AR1 method for both noun and Tf-Idf based networks, while the two ARD methods achieved a very similar performance for English documents.

Table 7 – RG-1 results for Portuguese SDS-MDS. The first two columns show the performance of the noun and Tf-Idf based network for SDS. In last columns, the performance of noun and Tf-Idf based network for MDS are shown. Also, we show the results of the anti-redundancy detection method (ARD) that achieved the best scores. Results in blue represent the six best systems, while those in orange represent the six worst systems. Additionally, results framed in green represent the best score for each category, while results framed in red represent the worst score.

		SDS		**MDS**			
		Noun	**Tf-Idf**	**Noun**		**Tf-Idf**	
	Meas.	**RG-1**	**RG-1**	**ARD**	**RG-1**	**ARD**	**RG-1**
1	dg	0.4826	0.4796	AR2	0.5499	AR2	0.5555
2	stg	0.4822	0.4793	AR1	0.5430	AR1	0.5554
3	sp	0.4786	0.4774	AR2	0.5502	AR2	0.5581
4	sp_w	0.4820	0.4750	AR2	0.5518	AR2	0.5567
5	pr	0.4812	0.4770	AR2	0.5438	AR2	0.5542
6	pr_w	0.4842	0.4807	AR2	0.5474	AR2	0.5553
7	btw	0.4812	0.4748	AR2	0.5462	AR1	0.5376
8	btw_w	0.4670	0.4571	AR2	0.5037	AR1	0.4893
9	cc	0.4439	0.4488	AR2	0.4266	AR2	0.4435
10	cc_w	0.4465	0.4507	AR1	0.4335	AR1	0.4523
11	conc_1	0.4479	0.4480	AR2	0.4145	AR2	0.4149
12	conc_2	0.4407	0.4453	AR2	0.4117	AR2	0.4179
13	conc_3	0.4472	0.4424	AR1	0.4265	AR1	0.4300
14	conc_4	0.4375	0.4388	AR1	0.4180	AR2	0.4152
15	conc_5	0.4543	0.4580	AR2	0.4343	AR2	0.4668
16	conc_6	0.4490	0.4565	AR1	0.4319	AR2	0.4569
17	conc_7	0.4402	0.4361	AR1	0.4055	AR2	0.4143
18	conc_8	0.4528	0.4568	AR1	0.4451	AR2	0.4577
19	access_h2	0.4685	0.4737	AR1	0.4968	AR2	0.5185
20	access_h3	0.4584	0.4546	AR2	0.4606	AR2	0.4791
21	gAccess	0.4821	0.4792	AR2	0.5523	AR2	0.5541
22	sym_h2	0.4768	0.4779	AR2	0.5348	AR1	0.5387
23	sym_h3	0.4683	0.4779	AR1	0.5128	AR1	0.5357
24	absT	0.4796	0.4763	AR2	0.5466	AR1	0.5612

Table 8 – RG-1 results for English SDS-MDS. The first two columns show the performance of the noun and Tf-Idf based network for SDS. In last columns, the performance of noun and Tf-Idf based network for MDS are shown. Also, we show the results of the anti-redundancy detection method (ARD) that achieved the best scores. Results in blue represent the six best systems, while those in orange represent the six worst systems. Additionally, results marked in green represent the best score for each category, while results framed in red represent the worst score.

		SDS		**MDS**			
		Noun	**Tf-Idf**	**Noun**		**Tf-Idf**	
	Meas.	**RG-1**	**RG-1**	**ARD**	**RG-1**	**ARD**	**RG-1**
1	dg	0.4712	0.4688	AR2	0.4023	AR2	0.3884
2	stg	0.4663	0.4699	AR2	0.4015	AR1	0.4104
3	sp	0.4677	0.4680	AR2	0.4033	AR2	0.3883
4	sp_w	0.4620	0.4652	AR2	0.4004	AR2	0.4007
5	pr	0.4692	0.4623	AR1	0.4024	AR1	0.3918
6	pr_w	0.4671	0.4723	AR1	0.4021	AR1	0.4115
7	btw	0.4650	0.4606	AR1	0.3955	AR2	0.3830
8	btw_w	0.4569	0.4203	AR2	0.3853	AR1	0.3286
9	cc	0.4029	0.3928	AR1	0.3248	AR2	0.3250
10	cc_w	0.4021	0.4016	AR1	0.3268	AR1	0.3433
11	conc_1	0.4177	0.3920	AR2	0.3146	AR1	0.2945
12	conc_2	0.4058	0.3889	AR1	0.3021	AR1	0.2947
13	conc_3	0.3951	0.3972	AR2	0.2942	AR2	0.3019
14	conc_4	0.3984	0.3863	AR1	0.2968	AR1	0.2934
15	conc_5	0.3955	0.4185	AR1	0.2884	AR1	0.3205
16	conc_6	0.3987	0.4175	AR1	0.2914	AR1	0.3199
17	conc_7	0.3877	0.3835	AR1	0.2928	AR1	0.2976
18	conc_8	0.4123	0.4213	AR2	0.3031	AR1	0.3199
19	access_h2	0.4546	0.4381	AR2	0.3821	AR2	0.3608
20	access_h3	0.4351	0.4179	AR2	0.3338	AR2	0.3102
21	gAccess	0.4689	0.4660	AR2	0.4006	AR2	0.3902
22	sym_h2	0.4298	0.4690	AR2	0.3489	AR1	0.3866
23	sym_h3	0.4464	0.4536	AR2	0.3808	AR1	0.3729
24	absT	0.4649	0.4734	AR2	0.3996	AR1	0.4106

In Figure 13 we show a performance comparison of the two network models explained in this section. We compared the behavior of these networks in order to determine which network model achieved the best performance. The first two graphics compare these networks for Portuguese SDS and MDS, while the last two graphics show the behavior of such networks for English SDS and MDS. We observe that both networks displayed very similar results. Only for Portuguese SDS the noun-based network outperformed the Tf-Idf approach, while in the other cases, the Tf-Idf based network had a slightly better performance than the network which is based on nouns. As we said before, both network models obtained a similar performance because they are based on the number of common words between two sentences. Although the construction of these models is relatively simple and it was not necessary to use complex NLP resources, they achieved prominent results.

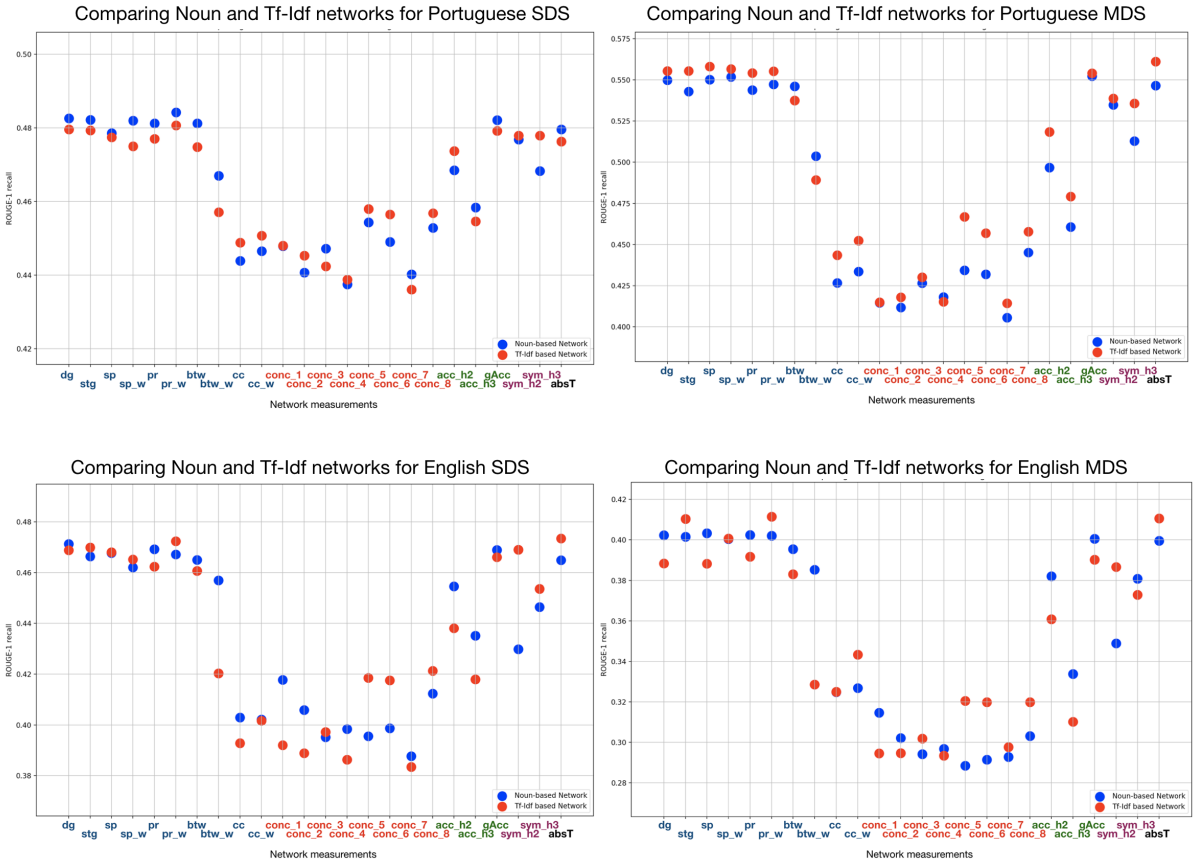


Figure 13 – Comparison analysis of the performance of noun and Tf-Idf based network.

Source: Elaborated by the author.

According to Tables 7 and 8, traditional network measurements like degree, shortest paths, Page Rank, and some of their weighted versions yielded the best scores. The measurements based on the dynamical behavior of the networks, such as absorption time and generalized accessibility also displayed excellent performances. In the case of symmetry measurements, we considered two approaches for sentence selection. The first approach selected the highest weighted nodes, while the second approach considered the nodes with lowest symmetry values.

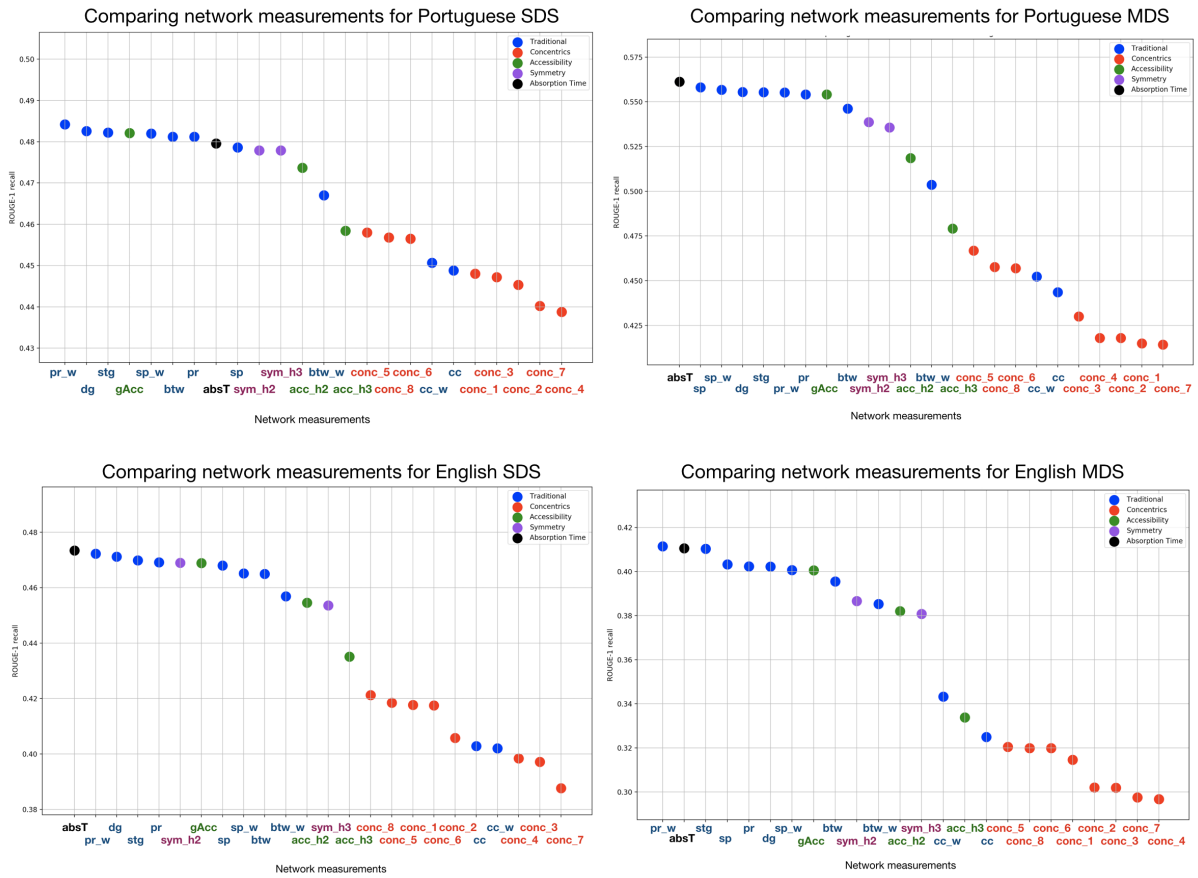


Figure 14 – Comparison analysis of the performance of the network measurements for both noun and Tf-Idf based networks. For each subplot, the measurements were ranked according to their ROUGE-1 score.

Source: Elaborated by the author.

The symmetry ($h = 2$) achieved a good performance when the least symmetric nodes were taken into consideration; however, in other cases, symmetry measurements yielded low ROUGE scores. In all cases, the accessibility measurement was outperformed by the top baseline scores, when it was evaluated at the second hierarchical level. Such a performance decreased when further hierarchical levels were taken into account. Finally, the systems based on concentric and clustering coefficient measurements yielded the lowest results, which were lower than the random baseline performance. The observations previously explained could be seen in a more clear way in Figure 14, which shows a performance comparison between all proposed network measurements for both noun and Tf-Idf based networks.

As we said previously, these first two models did not use sophisticated NLP tools for its construction like those mechanisms which try to find deeper linguistic information. In this sense, our language-independent methods rely on the document structure rather than its meaning or semantic information. The following network models intended to use more complex methods like word embeddings and machine learning with the aim of using not only the complex network concepts but also the summarization approaches found in other works. We used these combined

approaches in order to generate most informative summaries.

5.2 Embedding-based network

In this approach, we generated several network models based on word embeddings. We constructed the networks based on the cosine similarity between the word embedding vector representation of the sentences. In order to obtain the vector representation of a sentence, we used different word and sentence embeddings for both Portuguese and English. For both languages, at the beginning, we used the sentence embeddings Sent2Vec (PAGLIARDINI; GUPTA; JAGGI, 2017) and Doc2Vec (LE; MIKOLOV, 2014) for document representation. However, these models did not achieve good results. These models were trained by using the corpus employed in this work. The success of these models depends on the size of the training corpus, in this sense, the larger the corpus is, the better document representation will be obtained (LAI *et al.*, 2016). For this reason, we employed several word embedding models which were pre-trained in very large corpora.

For Portuguese documents, we used the pre-trained models provided by Hartmann *et al.* (2017) which created a repository called NILC-Embeddings that contains several pre-trained word embedding models for documents in Brazilian Portuguese. In NILC-Embeddings were trained four models: Word2Vec (MIKOLOV *et al.*, 2013b), Wang2Vec (LING *et al.*, 2015), FastText (BOJANOWSKI *et al.*, 2016), and GloVe (PENNINGTON; SOCHER; MANNING, 2014). NILC-Embeddings researchers generated several word embedding representations for different dimensions (50, 100, 300, 600, and 1000). In this Master's work, we evaluated the performance of three word embedding models: Word2Vec, FastText, and GloVe. We also evaluated the different dimensions found in such repository. Table 9 shows the performance of our embedding-based network. In this table, we only show the results of the measurements that achieved acceptable ROUGE-1 scores. Measurements like concentric metrics, clustering coefficient or betweenness were not included because of their poor performance in previous experiments. We also show the vector size and percentage of edge removal which was used in order to generate better summaries. For MDS we show the ARD method which achieved a better performance for this network type.

According to Table 9, for Portuguese language, we observe that Word2Vec embedding achieved the best scores for both SDS and MDS. In the case of SDS, 1000-dimension vectors outperformed the others models with different vector sizes; while networks with the reduction of 40% of least weighted edges yielded best scores. On the other hand, for MDS case, vectors of size 600 and 50% of edge removal are the ideal parameters to be used. Most of the measurements achieved a better performance when the AR1 method for anti-redundancy treatment was considered.

For Portuguese SDS, all measurements had a similar performance with the top baseline, in most cases a little superior to this system. The best measurement was the generalized

Table 9 – Embedding-based network performance for Portuguese SDS-MDS. The parameters which we considered for network creation were the following: vector size (k), percentage of least weighted edge removal (%), and anti-redundancy detection methods (ARD). Results in blue represent the methods which achieved the best ROUGE-1 scores for each network measurement. Additionally results marked in green represent the best score for each category, while results marked in red represent the worst score.

			SDS			MDS			
	Meas.	**Emb.**	**k**	**%**	**RG-1**	**k**	**%**	**ARD**	**RG-1**
1	dg	w2v	1000	50	0.4768	1000	50	AR1	0.5512
		gloVe	100	10	0.4740	300	10	AR1	0.5374
		fastT	600	50	0.4751	600	50	AR1	0.5347
2	stg	w2v	1000	50	0.4717	600	40	AR1	0.5586
		gloVe	1000	40	0.4733	300	30	AR1	0.5394
		fastT	1000	40	0.4728	100	50	AR2	0.5326
3	sp	w2v	1000	50	0.4770	1000	50	AR1	0.5503
		gloVe	100	10	0.4740	600	40	AR2	0.5368
		fastT	1000	10	0.4751	600	50	AR2	0.5331
4	sp_w	w2v	600	30	0.4747	600	50	AR1	0.5510
		gloVe	1000	40	0.4706	300	20	AR1	0.5415
		fastT	1000	40	0.4723	100	50	AR2	0.5293
5	pr	w2v	1000	40	0.4760	1000	50	AR1	0.5511
		gloVe	1000	50	0.4712	1000	50	AR2	0.5363
		fastT	300	40	0.4707	600	50	AR2	0.5434
6	pr_w	w2v	1000	50	0.4732	1000	10	AR2	0.5572
		gloVe	1000	50	0.4739	300	20	AR1	0.5369
		fastT	1000	40	0.4717	600	50	AR2	0.5390
7	access_h2	w2v	600	20	0.4636	100	50	AR2	0.4778
		gloVe	300	40	0.4640	300	40	AR1	0.4982
		fastT	300	10	0.4676	600	50	AR2	0.4950
8	gAccess	w2v	1000	10	0.4778	1000	50	AR1	0.5543
		gloVe	100	10	0.4713	300	10	AR1	0.5366
		fastT	1000	10	0.4743	1000	50	AR2	0.5340
9	sym_h2	w2v	1000	10	0.4766	600	10	AR1	0.5400
		gloVe	100	10	0.4694	300	10	AR1	0.5331
		fastT	1000	10	0.4740	300	40	AR2	0.5242
10	absT	w2v	1000	30	0.4717	1000	10	AR2	0.5661
		gloVe	1000	40	0.4705	300	20	AR1	0.5381
		fastT	1000	40	0.4737	600	50	AR1	0.5303

accessibility, while the accessibility ($h = 2$) achieved the poorest performance, far to the top baseline, but superior to random baseline. For MDS case, the absorption time measurement obtained the best performance, where the majority of network measurements outperformed the top baseline performance, except the accessibility and symmetry ($h = 2$), which achieved the worst performance (inferior to top baseline).

Table 10 – Embedding-based network performance for English SDS-MDS. The parameters which were considered for network creation were the following: percentage of least weighted edge removal (%), and anti-redundancy detection methods (ARD). All results were evaluated over word vectors whose length is 300 features. Results in blue represent the methods which achieved the best ROUGE-1 scores for each network measurement. Additionally results marked in green represent the best score for each category, while results marked in red represent the worst score.

			SDS		MDS		
	Meas.	**Emb.**	**%**	**RG-1**	**%**	**ARD**	**RG-1**
1	dg	w2v	10	0.4598	30	AR1	0.3861
		gloVe	50	0.4586	40	AR1	0.3879
		fastT	10	0.4632	30	AR2	0.3900
2	stg	w2v	20	0.4602	30	AR2	0.3908
		gloVe	50	0.4542	10	AR1	0.3902
		fastT	10	0.4593	10	AR1	0.3910
3	sp	w2v	10	0.4598	30	AR1	0.3854
		gloVe	40	0.4586	20	AR1	0.3839
		fastT	10	0.4632	30	AR2	0.3900
4	sp_w	w2v	10	0.4586	30	AR2	0.3911
		gloVe	50	0.4536	10	AR2	0.3894
		fastT	50	0.4583	30	AR2	0.3905
5	pr	w2v	50	0.4535	30	AR1	0.3849
		gloVe	50	0.4549	40	AR1	0.3854
		fastT	50	0.4544	50	AR1	0.3901
6	pr_w	w2v	10	0.4610	30	AR2	0.3921
		gloVe	50	0.4561	20	AR1	0.3900
		fastT	50	0.4598	50	AR2	0.3918
7	access_h2	w2v	50	0.4208	50	AR1	0.3544
		gloVe	50	0.4203	50	AR1	0.3552
		fastT	50	0.4157	50	AR2	0.3435
8	gAccess	w2v	10	0.4583	30	AR1	0.3855
		gloVe	50	0.4578	30	AR2	0.3860
		fastT	10	0.4598	30	AR2	0.3908
9	sym_h2	w2v	10	0.4582	30	AR1	0.3877
		gloVe	10	0.4561	20	AR2	0.3814
		fastT	10	0.4621	30	AR2	0.3852
10	absT	w2v	10	0.4605	30	AR2	0.3904
		gloVe	50	0.4549	10	AR1	0.3903
		fastT	10	0.4587	10	AR1	0.3910

For the summarization of English documents, we also used pre-trained word embedding corpus for such language. We found pre-trained models for Word2Vec, GloVe, and FastText algorithms. We used the Word2Vec Google News vectors (GOOGLE, 2013), which includes word vectors for a vocabulary of 3 million words. Google researchers trained on 100 billion words from a Google News dataset. They generated vectors of length 300. We also used the models

found in a repository provided by Pennington, Socher and Manning (2014) and Bojanowski *et al.* (2016) for GloVe and FastText embeddings. These repositories contains different word embedding vectors which were generated by training large corpora for several languages. The word embedding vectors we found have 300 dimensions. Table 10 shows the performance we achieved for English documents by using these word embedding representations.

Table 10 shows that Word2Vec model obtained good results for SDS, while FastText method was the best word embedding algorithm for MDS. The best parameters for the elimination of least weighted edges was 50% and 30% for SDS and MDS respectively. Both ARD methods for MDS achieved a very similar performance for English documents.

For the summarization of English documents, the performance of embedding-based methods decreased. For both SDS and MDS all network measurements achieved a performance inferior to the top baseline system. The degree and shortest path metrics obtained the best scores for SDS, while the weighted version of Page Rank algorithm was the best for MDS. For both summarization types, the accessibility measurement ($h = 2$) yielded the worst performance, even inferior to the random baseline.

Despite the fact that word embedding models were trained over large amounts of information, and therefore they could capture more semantic information of sentences, they did not achieve the expected results. We used word embedding models because they had good performances for other NLP tasks such as text classification or machine translation (JIN *et al.*, 2016; ZOU *et al.*, 2013). It seems that word embedding algorithms work better when the generated vectors are used in a more explicit way, for example, for text classification, they are used as features for a supervised classifier (JIN *et al.*, 2016). Instead, our models only considered the cosine similarity of such vector representations for network construction.

5.3 Multilayer Network for MDS

The main idea of using a multilayer network approach for MDS is to make a distinction between edges which connect sentences from the same documents and edges connecting sentences from different documents. We believe that the discrimination between intra- and inter-layer edges will help to improve the characterization of documents because a sentence which is connected to several sentences from other documents could indicate a high relevance of the approached topics.

For its construction, the multilayer network approach (MLN) is based on the Tf-Idf based network, in this sense, edges are calculated according to the cosine similarity between the vector representation of sentences. We made several tests by considering the noun and Tf-Idf approaches, however, in this section, we only show the performance of the Tf-Idf based network approach because this model outperformed the noun-based approach. We did not consider the embedding-based network because its poor performance in previous tests. A relevant parameter

for this multilayer network approach is the establishment of importance weights for edges of inter-layer type. We considered the parameter α, which gives a higher ($\alpha > 1$) or lower ($\alpha < 1$) relevance for inter-document relationships. In this work, we considered different values for α, which were in the range $0.5 < \alpha < 1.9$.

For weighted measurements, the process continues according to the previous explained steps, however, for non-weighted measurements, we required to employ an additional step before the use of such metrics. This step was important because the new establishment of weights for inter-layer edges will no have any effect for such non-weighted metrics. Therefore, we decided to modify the network structure after the inter-layer edge weight establishment was made. This modification included to remove a fixed amount r of the edges with the lowest weights for each value of α ($r = \{0.1, 0.2, 0.3, 0.4, 0.5\}$). Regarding the anti-redundancy detection methods, the AR1 and AR2 techniques displayed similar performances, therefore, we only show the scores with the best results.

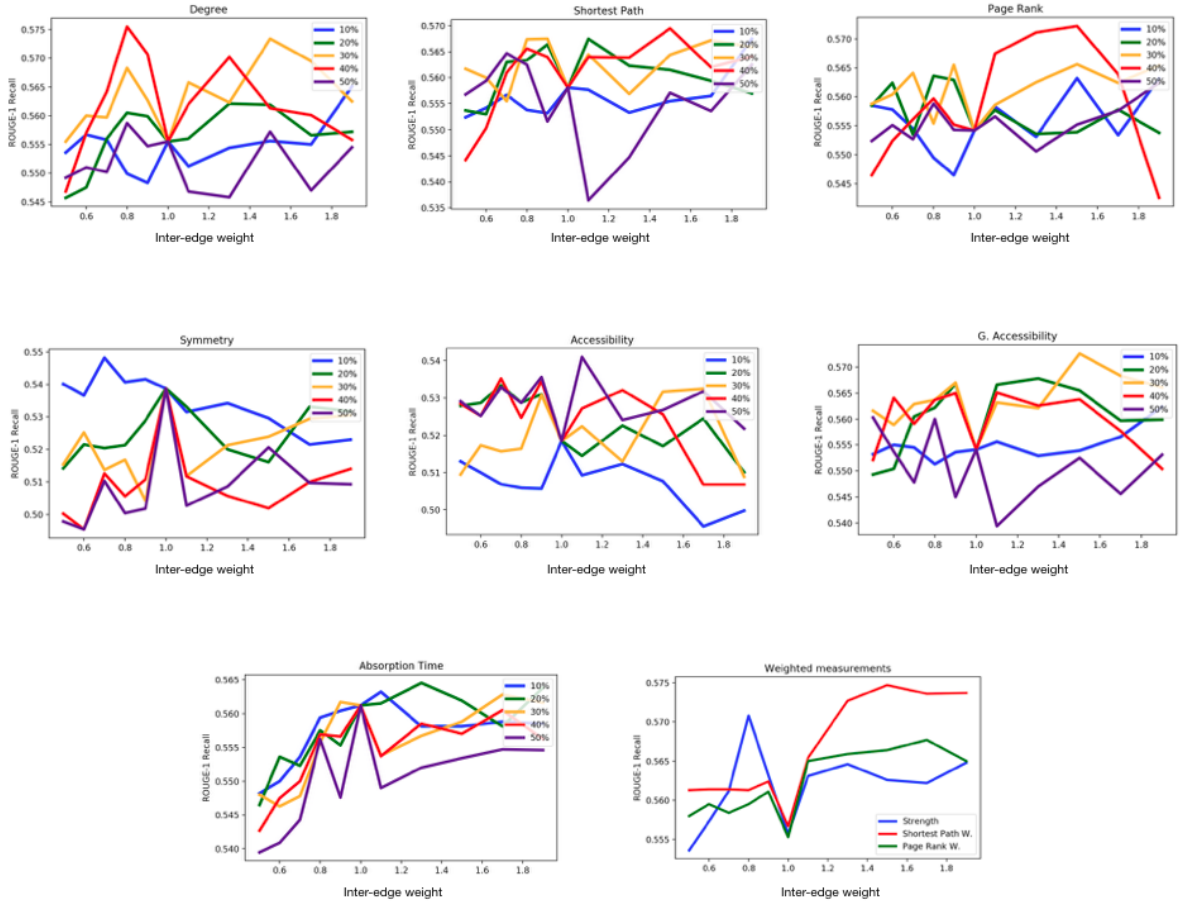


Figure 15 – Performance analysis of the MLN approach for Portuguese MDS. Each subfigure shows the behavior of each network measurement as a function of the inter-layer edge weight parameter (α). x-axis represents the inter-layer edge weight parameter (α) and y-axis is the average ROUGE-1 score (RG-1).

Source: Adapted from Tohalino and Amancio (2018).

In Figure 15 we show the global performance of our multilayer approach (MLN) for Portuguese MDS. For each subfigure, we show the curves by evaluating different values of percentages r of removed edges. We observed, apart from the symmetry metric, an increase in performance when $\alpha \neq 1$. Also, we can observe that the best scores can be achieved in the following two scenarios: (i) $\alpha < 1$, where we assign a higher relevance for intra-document relationships; and (ii) $\alpha > 1$, where we give a higher relevance or importance is assigned for the inter-document relationships. According to the Figure 15, the second scenario could be more important for improving the system performance, since (i) holds just for the degree metric. We can also observe that the best value of r depends on the measurement which is being analyzed.

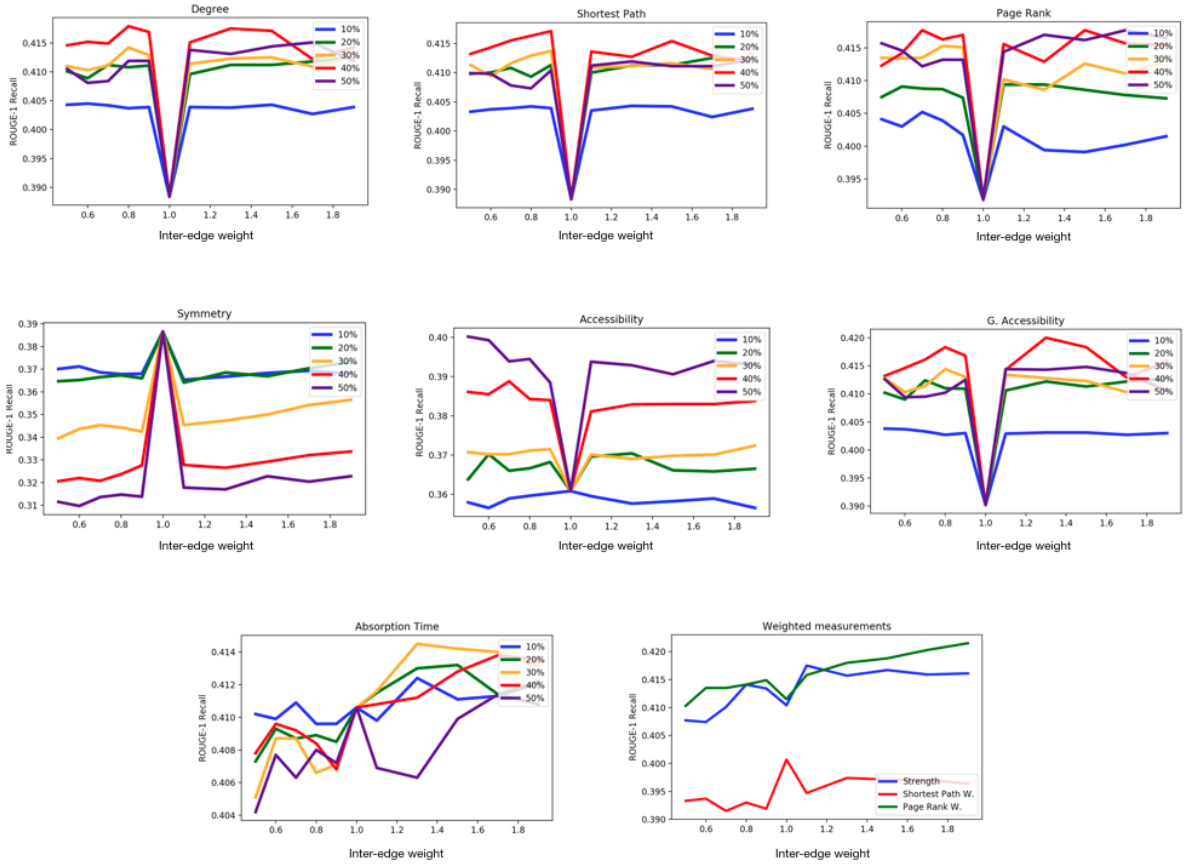


Figure 16 – Performance analysis of the MLN approach for English MDS. Each subfigure shows the behavior of each network measurement as a function of the inter-layer edge weight parameter (α). x-axis represents the inter-layer edge weight parameter (α) and y-axis is the average ROUGE-1 score (RG-1).

Source: Adapted from Tohalino and Amancio (2018).

Figure 16 shows the performance of the multilayer approach for English documents. As was observed in the performance for Portuguese MDS, the value of $\alpha \neq 1$ again increase the system performance, except for the symmetry metric. Differently from earlier analysis, for this case, the intra-document relationships could play an important role for a great number of complex network measurements. We achieved optimized results for $\alpha < 1$ for the degree, shortest path length, Page Rank, accessibility and generalized accessibility measurements. We can also

observe a major improvement in performance for $\alpha > 1$ when the absorption time measurement was considered. Likewise to the behavior observed for Portuguese document summarization, here the best value of r always depends on the adopted network measurement.

Table 11 and Table 12 encapsulate the results we achieved for both Portuguese and English MDS. For Portuguese MDS task, our systems achieved an excellent performance, where the behavior of some network measurements was very superior to the top baseline. Only the symmetry and the accessibility metrics were outperformed by the top baseline system. In the case of English MDS, our MLN system also displayed good results, where almost all network metrics, except symmetry measurement, outperformed the performance of the top baseline method. We also observed that the multilayered approach for MDS surpasses the performance of the previous network models studied in this Master's work.

For the Portuguese case, the best results were achieved with the degree and the weighted version of shortest paths. We see that both measurements considered the AR2 method to address the anti-redundancy problem. We can also observe that the majority of measurements achieved the best performances when the AR2 strategy was considered. For English summarization, the weighted version of Page Rank and the generalized accessibility outperformed the other network measurements. Here, the AR1 strategy for anti-redundancy treatment achieved the best scores.

Table 11 – Best results obtained of MLN approach for Portuguese MDS. We show the performance of each network metric ranked according to its ROUGE-1 (RG-1) score. We show the parameters in which each measurement achieved its best score. The parameters are the following: inter-layer edge weight (α), the threshold for edge removal (r), and the anti-redundancy detection (ARD) method which displayed the best result.

	Meas.	α	r	**ARD**	**RG-1**
1	dg	0.8	50	AR2	0.5755
2	sp_w	1.5	-	AR2	0.5747
3	gAccess	1.5	30	AR1	0.5726
4	pr	1.5	40	AR2	0.5722
5	stg	0.8	-	AR2	0.5705
6	pr_w	1.7	-	AR2	0.5677
7	sp	1.9	10	AR2	0.5675
8	absT	1.3	20	AR2	0.5645
9	sym_h2	0.7	10	AR1	0.5482
10	access_h2	0.9	50	AR1	0.5401

Regarding the measurements based on self-avoiding random walks, the generalized accessibility measurement obtained the best performance. The symmetry measurement obtained low ROUGE-1 scores for both Portuguese and English MDS. It seems to be that such metric mainly quantifies the diversity of links weights and not the prominence of nodes. The displayed results suggest that the most important information to quantify relevance or importance is not captured by such measurement.

The results also suggest that this multilayer approach is relevant to identify the most

Table 12 – Best results obtained of MLN approach for English MDS. We show the performance of each network metric ranked according to its ROUGE-1 (RG-1) score. We show the parameters in which each measurement achieved its best score. The parameters are the following: inter-layer edge weight (α), the threshold for edge removal (r), and the anti-redundancy detection (ARD) method which displayed the best result.

	Meas.	α	r	**ARD**	**RG-1**
1	pr_w	1.9	-	AR1	0.4215
2	gAccess	1.3	40	AR1	0.4200
3	dg	0.8	40	AR1	0.4179
4	pr	1.7	50	AR2	0.4177
5	stg	1.1	-	AR1	0.4175
6	sp	0.9	40	AR1	0.4171
7	absT	1.3	30	AR1	0.4145
8	sp_w	1.0	-	AR2	0.4007
9	access_h2	0.5	50	AR2	0.4002
10	sym_h2	1.0	-	AR1	0.3866

central nodes (sentences) in documents. We can see that the relevance of the intra- and inter-layer relationships seems to be dependent on the dataset, language, and measurement which is being used to rank the sentences (nodes). This premise was clear when we noted that both scenarios ($\alpha > 1$ and $\alpha < 1$) are possible, even when we examined the same language.

Due to the success of this method for document representation, we believe that the multilayer approach could be considered in other related applications where intra- and inter-relationships are important. We could apply these concepts in other NLP tasks such as text mining and the identification of key concepts in scientific areas (SILVA *et al.*, 2016a).

5.4 Machine learning approach for sentence classification

In this approach, we evaluated a machine learning method which used a set of classifiers in order to determine if a sentence belongs to a summary. In this sense, the greater the likelihood of the sentence, the more probability of being included in an extract such sentence (LEITE; RINO, 2006; LEITE; RINO, 2008).

We used the complex network measurements previously studied as features for the training stage. We also employed the most common features used in summarization tasks with machine learning. These traditional summarization features were the following: word frequency, word tf-idf, indicator of main concepts, occurrence of proper names, occurrence of non-essential information, topic modeling, sentence length, sentence to sentence cohesion, sentence to centroid cohesion, and sentence position.

A common issue for machine learning approaches is the feature selection stage (LEITE; RINO, 2008). This stage consists of selecting a set of optimal features which would improve the learning phase. Due to the great number of features, we decided to make several experiments by

using different feature sets. The features proposed in this Master's work were grouped into four sets. Each feature set is described as follows:

- *Group A (10 features):* We only selected the most traditional network measurements and its weighted versions: degree, strength, Page Rank, weighted Page rank, shortest paths, weighted shortest paths, clustering coefficient, weighted clustering coefficient, betweenness, and weighted betweenness.
- *Group B (6 features):* This group only considers the hierarchical and dynamical network measurements: accessibility ($h = 2$, $h = 3$), symmetry ($h = 2$, $h = 3$), generalized accessibility, and absorption time.
- *Group C (16 features):* For this set, we combined the features of Group A and Group B. In other words, we selected all network measurements proposed in this work.
- *Group D:* We selected the feature set which achieved the best performance, and then we combined this set with the traditional summarization features.

Table 13 – Machine learning results for Portuguese SDS-MDS. For each feature set, we show the respective parameters: the network model which is being considered, classifier (Class.), percentage of edge removal (%), and inter-layer edge weight(α). Results in blue represent the best scores obtained for each feature set. Additionally results marked in green represent the best score for each category, while results marked in red represent the worst scores.

		SDS			MDS			
Feature selection	**Network**	**Class.**	**%**	**RG-1**	**Class.**	**α**	**%**	**RG-1**
A	Noun	NB	-	0.4820	NB	-	-	0.5454
	Tf-Idf	NB	-	0.4817	NB	-	-	0.5519
	Emb.	NB	50	0.4817	NB	-	50	0.5512
	MLN	-	-	-	DT	1.9	20	0.5618
B	Noun	DT	-	0.4787	NB	-	-	0.5400
	Tf-Idf	DT	-	0.4801	NB	-	-	0.5454
	Emb.	DT	50	0.4807	NB	-	50	0.5400
	MLN	-	-	-	NB	1.3	30	0.5521
C	Noun	DT	-	0.4831	NB	-	-	0.5400
	Tf-Idf	DT	-	0.4785	NB	-	-	0.5454
	Emb.	DT	40	0.4791	DT	-	30	0.5443
	MLN	-	-	-	DT	1.9	10	0.5660
D	Noun	DT	-	0.4852	NB	-	-	0.5424
	Tf-Idf	DT	-	0.4805	NB	-	-	0.5531
	Emb.	NB	40	0.4806	DT	-	10	0.5515
	MLN	-	-	-	NB	1.7	30	0.5510

Our machine learning approach used the following classifiers: Naive Bayes (NB), Decision Trees (DT), and Support Vector Machines (SVM). We performed a 10-fold cross validation evaluation for training and testing the corpus we used for document summarization. We evaluated all the network models proposed in this work: noun, Tf-Idf, embedding, and MLN based

networks. Table 13 shows the results we achieved for Portuguese documents, while Table 14 presents the obtained scores for English document summarization.

The results in Table 13 suggest that machine learning approaches with complex network concepts have a good performance since the majority of the methods displayed higher results than the top baseline system for Portuguese summarization. For Portuguese SDS case, the combined feature set with traditional summarization features and complex network features (Group D) had the best performance. These network features achieved the best performance when they were generated by evaluating the noun based network. The decision tree classifier outperformed the other classifiers. In the case of Portuguese MDS, the multilayer based network (MLN) outperformed the other network and the best feature set was group C (all network measurements were considered). The naive Bayes classifier in almost all cases generated the best scores for Portuguese MDS.

Table 14 – Machine learning results for English SDS-MDS. For each feature set, we show the respective parameters: the network model which is being considered, classifier (Class.), percentage of edge removal (%), and inter-layer edge weight(α). Results in blue represent the best scores obtained for each feature set. Additionally results marked in green represent the best score for each category, while results marked in red represent the worst score.

		SDS			MDS			
Feature selection	**Network**	**Class.**	**%**	**RG-1**	**Class.**	**α**	**%**	**RG-1**
A	Noun	NB	-	0.4916	DT	-	-	0.3943
	Tf-Idf	DT	-	0.4756	DT	-	-	0.4041
	Emb.	NB	10	0.4858	DT	-	40	0.3826
	MLN	-	-	-	DT	1.7	10	0.4109
B	Noun	NB	-	0.4849	DT	-	-	0.3896
	Tf-Idf	NB	-	0.4842	DT	-	-	0.3994
	Emb.	NB	40	0.4867	DT	-	40	0.3764
	MLN	-	-	-	DT	1.3	50	0.4070
C	Noun	NB	-	0.4863	DT	-	-	0.3937
	Tf-Idf	NB	-	0.4860	DT	-	-	0.4063
	Emb.	NB	50	0.4864	DT	-	20	0.3812
	MLN	-	-	-	DT	1.7	10	0.4112
D	Noun	NB	-	0.4912	DT	-	-	0.3943
	Tf-Idf	NB	-	0.4867	DT	-	-	0.4071
	Emb.	NB	20	0.4863	DT	-	20	0.3780
	MLN	-	-	-	DT	1.7	10	0.4084

Results in Table 14 indicate that our best methods for both English SDS-MDS obtained acceptable results, even, for English SDS, the machine learning approach outperformed the other systems proposed in this work. For SDS case, the noun-based network displayed the best scores by only selecting the traditional network measurements (Group A). The naive Bayes classifier performed better than other classifiers for English SDS. In the case of MDS, the multilayer network approach achieved the best scores, regardless of the feature set that is being used. The MLN approach achieved the best performance when all network measurements were selected as a feature set for learning phase (Group C). For all proposed methods for English MDS, the

decision tree classifier was the best. We can also see that in any case, the SVM classifier achieved the worst results, it was even outperformed in all cases by the other classifiers for both Portuguese and English (SDS and MDS).

We made different experiments by using different feature sets based on network metrics and some traditional summarization features. According to the results we achieved, the feature set which produced the best ROUGE-1 scores was the group C. This group included all the network measurements proposed for this Master's work. We believe that all network measurement had a good performance because they probably described different aspects of the document structure that does not necessarily occurs in the other network metrics. On the other hand, the feature set which only included the dynamical and hierarchical measurements (group B) performed the lowest scores for our machine learning approach. We believed that this poor performance occurred because we took into consideration few features for the learning phase. Regarding the traditional summarization features, we only considered those simple features which did not require complex NLP resources for its computation. We believe if we use more complex features to obtain deeper semantic and linguistic information from sentences, the performance of the proposed classifiers would increase.

For this machine learning approach, the training and testing stages were carried out by using the corpus we used for the validation of our systems. We did not employ or create any additional corpus for the learning phase, for this reason, we used the 10 fold cross-validation method for training and testing our data. In future works, we could construct larger corpus for training stages.

Concerning the feature set creation, we had some issues. Some of the corpus we used do not have reference extracts, in other words, they only included the reference abstracts. In order to construct the feature tuples (feature set, class label) for corpus which does not include reference extracts, we considered sentences to belong to "Present in extract" class those sentences whose lexical similarity is high with any sentence which exists on the reference abstract.

The machine learning approach performed well for the SDS methods. However, although the results were acceptable, its performance decreased for MDS. The process for MDS was the following: first, we selected all sentences from the entire document set, and then, they were trained by a supervised classifier. After training stage, all sentences were ranked according to their likelihood to be included in a final summary. Finally, the best ranked sentences were selected to belong to the multi-document summary. We believe the decrease of performance of the MDS system arises because the texts which belong to a document cluster could include many sentences which are similar or even they are the same. In this sense, many of these similar sentences could be classified as "Present in summary" class, however, when they are included in the final summary, most of them are removed because of the redundancy treatment. In this way, our algorithm could select sentences that are less likely to be included in the final summary. In future works, we intend to construct a smarter algorithm for the treatment of such sentences with

a machine learning approach for MDS.

5.5 Final comparisons

Here we make a comparative analysis of all the network models and measurements used in this Master's work. In Table 15 and Table 16 we show a summary of the four network models and the machine learning approach studied in this work for Portuguese and English document summarization respectively. We show the measurements which achieved the best ROUGE-1 scores for each model and summarization method. In Table 17 we show the results achieved by other works and the performance of top and random baseline (similar to Table 6). In this table, we add the results of our systems that achieved the best performances for this Master's research.

Table 15 – Best results for Portuguese SDS-MDS. For each network model, we show the top five measurements with the best performance. We also show the scores for each network model which displayed the best results for the machine learning approach. Results in blue represent the best systems for each summarization method. Additionally results marked in green represent the methods which outperformed all the proposed methods for this Master's research.

	Noun		Tf-Idf		Embedding		MLN		CN + ML	
SDS	**Meas.**	**RG-1**	**Meas.**	**RG-1**	**Meas.**	**RG-1**	**Meas.**	**RG-1**	**Meas.**	**RG-1**
	pr_w	0.4842	pr_w	0.4807	gAccess	0.4778	-	-	Noun	0.4852
	dg	0.4826	dg	0.4796	sp	0.4770	-	-	Tf-Idf	0.4817
	stg	0.4822	stg	0.4796	dg	0.4768	-	-	Emb.	0.4817
	gAccess	0.4821	gAccess	0.4792	sym_h2	0.4766	-	-	-	-
	sp_w	0.4820	sym_h2	0.4779	pr	0.4760	-	-	-	-
MDS	**Meas.**	**RG-1**	**Meas.**	**RG-1**	**Meas.**	**RG-1**	**Meas.**	**RG-1**	**Meas.**	**RG-1**
	gAccess	0.5523	absT	0.5612	absT	0.5661	dg	0.5755	MLN	0.5660
	sp_w	0.5518	sp	0.5581	stg	0.5586	sp_w	0.5747	Tf-Idf	0.5531
	sp	0.5502	sp_w	0.5567	pr_w	0.5572	gAccess	0.5726	Emb.	0.5515
	dg	0.5499	dg	0.5555	gAccess	0.5543	pr	0.5722	Noun	0.5454
	pr_w	0.5474	stg	0.5553	dg	0.5512	stg	0.5708	-	-

We compared our methods with the top-ranked works that addressed the extractive document summarization task. The results shown in previous tables suggest that the network-based approaches in most cases displayed good performances, because their ROUGE-1 scores were close to the results of other works, even in some cases, our systems outperformed the scores of such works.

For both languages, we generally achieved prominent results. For Portuguese SDS, our systems yielded good scores, but these scores did not outperform the results of other works. However, we achieved excellent results for Portuguese MDS case, where the networked systems obtained better results in comparison to the scores of other works. For English document summarization, our systems achieved good results for English SDS, outperforming some other works, but the performance decreased for the MDS case, where the networked systems yielded acceptable results, superior to baseline system, but very distant from the results of other works.

Table 16 – Best results for English SDS-MDS. For each network model, we show the top five measurements with the best performance. We also show the scores for each network model which displayed the best results for the machine learning approach. Results in blue represent the best systems for each summarization method. Additionally results marked in green represent the methods which outperformed all the proposed methods for this Master's research.

	Noun		Tf-Idf		Embedding		MLN		CN + ML	
SDS	**Meas.**	**RG-1**	**Meas.**	**RG-1**	**Meas.**	**RG-1**	**Meas.**	**RG-1**	**Meas.**	**RG-1**
	dg	0.4712	absT	0.4734	sp	0.4632	-	-	Noun	0.4916
	pr	0.4692	pr_w	0.4723	dg	0.4632	-	-	Tf-Idf	0.4867
	gAccess	0.4689	stg	0.4699	sym_h2	0.4621	-	-	Emb.	0.4867
	sp	0.4677	sym_h2	0.4690	pr_w	0.4610	-	-	-	-
	pr_w	0.4671	dg	0.4688	absT	0.4605	-	-	-	-
MDS	**Meas.**	**RG-1**	**Meas.**	**RG-1**	**Meas.**	**RG-1**	**Meas.**	**RG-1**	**Meas.**	**RG-1**
	sp	0.4033	pr_w	0.4115	pr_w	0.3921	pr_w	0.4215	MLN	0.4112
	pr	0.4024	absT	0.4106	sp_w	0.3911	gAccess	0.4200	Tf-Idf	0.4071
	dg	0.4023	stg	0.4104	absT	0.3910	dg	0.4179	Noun	0.3943
	pr_w	0.4021	sp_w	0.4007	stg	0.3910	pr	0.4177	Emb.	0.3826
	stg	0.4015	pr	0.3918	gAccess	0.3908	stg	0.4175	-	-

Table 17 – List of works for Portuguese and English Document Summarization with the respective ROUGE-1 Recall (RG-1) sores. The best results of the approaches evaluated in this Master's research are highlighted.

Portuguese Summarization				**English Summarization**			
SDS		**MDS**		**SDS**		**MDS**	
System	**RG-1**	**System**	**RG-1**	**System**	**RG-1**	**System**	**RG-1**
Supor2-LogistRegr	0.5316	GistSumm	0.6643	*HITS$_A$B.*	0.5023	DUC-best	0.4986
Supor-2	0.5227	**Our system**	**0.5755**	ntt.duc02	0.5013	BSTM	0.4881
PageRank B.	0.5121	*Top B.*	*0.5497*	PageRank B.	0.5008	FGB	0.4850
HITS$_A$B.	0.5002	Bushy Path	0.5397	**Our system**	**0.4916**	LexPR	0.4796
Our system	**0.4852**	Depth-first Path	0.5340	ULeth131m	0.4911	**Our system**	**0.4215**
Top B.	*0.4757*	CSTSumm	0.5065	*Top B.*	*0.4860*	*Top B.*	*0.3928*
Random B.	*0.4565*	*Random B.*	*0.4622*	*Random B.*	*0.4258*	*Random B.*	*0.3623*

We believe that such decrease of performance for English MDS is due to the fact that the analysis at the structural level of English documents is probably not enough, and it requires to be combined with more sophisticated NLP methods. This fact can be confirmed because our methods are compared with other works which employed advanced NLP tools for English documents.

Regarding to the performance of the network models proposed in this work, for Portuguese SDS task, the machine learning approaches and the noun-based networks were the best systems. In the case of English SDS, the machine learning approach with the Tf-Idf based network yielded the best performance. We can see that machine learning approaches achieved an excellent performance for SDS.

For the MDS task, for both Portuguese and English, the multilayer network approach (MLN) always obtained the best performances, even they far exceeded the results of most other

works for Portuguese case. We can see that the different weighting schemes for inter-layer edges can make a better representation of the document set for multi-document summarization. On the other hand, the machine learning approaches did not perform well for MDS, because these methods were outperformed by the other systems.

Both noun and Tf-Idf based networks had a similar performance, achieving a slightly better performance the Tf-Idf based network. They generally got acceptable results for the summarization task with complex networks. However, the embedding-based networks, in most cases, obtained low scores compared to the other network models, especially when they were evaluated for the summarization of English documents. Despite the fact that embedding methods had a low performance, their ROUGE-1 scores were not so distant for the noun and Tf-Idf models. We observed that the performance of the sentence embedding based networks had a performance slightly lower than the other models. At the beginning, we believed that sophisticated methods for document representation like word embeddings would improve the system performance. Surprisingly, methods based on the number of common words between sentences performed better than the embedding-based methods.

In this work, we used a myriad of network measurement for node weighting and sentence selection for extractive summarization. Those traditional metrics like degree, Page Rank, shortest paths and their weighted versions achieved the best scores regardless of the language and summarization methods which are being used. Also, we used novel dynamical network measurements, such as absorption time and generalized accessibility, which is most cases outperformed several network metrics proposed in this work.

Finally, we report the network models and measurements with the best performances for each case. For Portuguese SDS, the machine learning approach for noun-based network had the best score, while for MDS case, the best result was obtained by using the multilayer approach with the node degree. For English documents, the machine learning approach for noun-based network again outperformed the other methods. The MLN approach with the weighted version of Page Rank was the best for English MDS.

We finally remark that our measurements could be used to improve the characterization of networks for the extractive summarization task. The success of these network measurement also depended on the network model in which they were generated. We remark that the Tf-Idf and the multilayer-based networks achieved the best performances, in this sense, they represented in a more convenient way the source documents. For the MLN approach, we observe that it always outperformed the other systems. In future works, it would be interesting to develop and employ novel network measurements whose focus is the characterization of multilayer networks. We also see that our machine learning approach achieved an excellent performance for SDS.

The main advantage of the proposed methods is that they do not depend on any source of external information. In other words, they do not use lexical datasets to obtain linguistic information or complex NLP resources for the treatment and processing of documents. We found

that documents could be represented as networks, and we concluded that some complex network measurements are adequate tools for the extraction of most relevant sentences for an extractive summary. Results showed that such models help to increase the informativeness of the generated summaries.

CHAPTER

6

CONCLUSIONS AND FUTURE WORKS

In recent years, several works employed graph-based methods and network theory in order to generate extractive summaries for the document summarization task. These methods represented the documents as networks, where nodes are words, sentences or paragraphs; and then they applied a set of network measurements with the aim of providing a relevance weight to each node. This relevance weight allows ranking sentences according to their importance. Despite the relative success of these methods for document summarization, several recent network tools and concepts have not yet been widely employed for the summarization problem, including the use of dynamical measurements and the multilayer representation.

In addition to the already known models based on the number of common words between two sentences (Noun and Tf-Idf based networks), this Master's research proposed the creation of two additional network representations based on word embeddings and multilayer networks for multi-document summarization. Regarding the network metrics, in addition to the traditional complex network measurements which have been used with great success in previous works, we employed novel dynamical network measurements for the summarization task. We found that such dynamical measurements were good indicators of sentence importance for both summarization methods (SDS and MDS) and the languages applied in this Master's research. We also developed a machine learning technique which combined all the proposed network measurements with features commonly used for the summarization task. We employed this method with the objective of making a smarter selection of sentences which takes into account all the network measurements addressed in this research. We also used this approach in order to combine the complex networks concepts with basic NLP methods.

This chapter is divided into 3 sections: In Section 6.1 we show the main contributions resulting from this Master's research. Section 6.2 presents the limitations and remarks for future works originated from the analysis of these Master's work. Finally, Section 6.3 lists the research publications derived from this work.

6.1 Contributions and limitations

One of the main contributions of this research is that it does not employ sophisticated NLP resources or tools for text treatment. In this sense, this work is easily adaptable for different languages, as we found its good performance for both Portuguese and English languages. We could have used a deeper tool, but the focus of this research was to evaluate the performance of complex networks and measurements for the extractive summarization task. The proposed summarization method based on complex networks concepts was divided into four stages. Next, we detail the main contributions and limitations for each stage.

- **Stage 1: Text conversion**
  - *Contributions:*
    - We only used simple NLP resources at the lexical or syntactical level such as stopwords lists for the elimination of unnecessary words, text lemmatization, and POS tagging. We did not employ more advanced resources at the semantical level. Although we use simple tools, we obtained excellent results for such pre-processing methods.
    - For document vector representation, we tested several word embedding algorithms which have not yet been used for document summarization.
  - *Limitations:*
    - For text pre-processing phase, co-reference problems could arise because we did not address deeper approaches for document treatment. Co-reference is the relationship between two or more terms referring to the same entity (CRYSTAL, 2011). For example, in the sentence *"Although everyone saw the President, nobody recognized him"*, the pronoun *him* is related with the noun *president*, and both are co-referential, since they point to the same entity of the real world. In this sense, if a noun in a sentence is referenced by other sentences, we will lose these relationships for the network creation stage because those relationships will not be identified.
- **Stage 2: Network creation**
  - *Contributions:*
    - We explored two novel methods for document representation as networks: word embedding-based network and multilayer network (MLN). Although the word embedding network achieved an acceptable performance, it did not overcome simpler network models such as Noun or Tf-Idf based networks. On the other hand, the MLN approach displayed excellent performances, even outperforming all proposed methods in this research. We found that different weighting schemes for inter-layer edges are relevant for the representation of a set of documents.

- *Limitations:*
  - As explained about the limitations of the text pre-processing stage, important relations between sentences could be lost because of the co-reference problem. This issue will decrease the performance of the noun-based network. Also, for such network, we will lose relevant connections about the synonyms of the nouns that are being considered for the edge creation between the sentences.

- **Stage 3: Sentence ranking**
  - *Contributions:*
    - We employed new metrics for document summarization task with complex networks. These metrics, such as generalized accessibility or absorption time generally achieved great performances. We believe that these measurements could be used for the characterization of networks for different NLP tasks.
    - We developed a machine learning approach which considered all network measurements proposed in this research. Such machine learning method displayed very good performances when the most traditional and dynamical measurements of complex network were taken into account. In this sense, all network measurements that we used in this research could be used together with the aim of determining the relevance or importance of a sentence.
  - *Limitations:*
    - In the literature was difficult to find specialized corpus to be used for sentence classification with machine learning methods. Generally, the authors of such researches used the same summarization corpus for the classification task by using 10 fold cross-validation.

- **Stage 4: Summarization**
  - *Contributions:*
    - We employed a very simple sentence selection method achieving good results. However, this approach has many limitations.
    - We used two anti-redundancy detection methods for MDS task. We found that anti-redundancy treatment did not have a big impact on the corpus we used in this research. However, we need to evaluate other corpus and other anti-redundancy detection methods.
  - *Limitations:*
    - For MDS task, co-reference problems could arise when are selected sentences from different documents, because summaries probably have references to other sentences which have not been included in the final extract. This problem also

occurs for SDS task, but it is more frequent for MDS. Therefore, this issue will generate inconsistent extracts. Summaries generated should be readable and relevant to the user.

* Summaries must respect a chronological order. However, for our approach, it is difficult to address this problem. Again, we will probably have coherence problems with the generated summaries.

6.2 Future works

Some future works derived from this Master's work are described below:

- In this research, we employed news corpus for the validation of our methods. We pretend to use other corpus with different domains in order to validate the efficiency of our network-based summarizers. In the same way, we propose to test our methods for other languages such as Spanish. In this sense, we intend to compare the performance between Spanish and Portuguese-based summarizers. We believe that both languages would achieve similar performances because they have great similarities being closely related to each other.

- Regarding the network creation methods, we pretend to use a thesaurus with the aim of identifying more strong relations between sentences. By using thesaurus we could find all words which have the same or a similar meaning with respect to a word. More specifically, for the noun-based network, we could not only consider the common nouns between sentences but also we could consider the synonyms of such nouns for the edge creation stage between two network nodes. We believe that it would increase the relation between two sentences.

- For the multilayer network approach, we applied network measurements which were implemented for single layer networks. Although these metrics displayed good results for this approach, we believe that novel network measurement whose focus is the characterization of multilayer networks would improve the system performance for the multi-document summarization approach.

- We propose to construct a corpus for the training stage for the machine learning approach. In this work, we only used our corpus for training and testing stage by using 10 fold cross-validation. We believe that the accuracy of the classifier would increase if we employ a more specialized corpus for sentence classification for document summarization. Likewise, we pretend to use more sophisticated features which employ more complex NLP resources for its computation. Such features include semantic relations between words, cross-document relations, CST relations, among others.

- Regarding the sentence selection methods, the simplest approach is to use each network measurement and then it is generated a different summary for each metric. In this way, we have to evaluate independently each network metric. In order to overcome this issue, in this work, we used a machine learning method which generated a single summary by using all network metrics. We propose to use another approach which employs all or some metrics together. We pretend to use voting systems, which select the sentences which were ranked by different network measurements to be included in a final summary. It was verified that these voting systems achieved excellent results for the summarization of documents in the Portuguese language.

6.3 Publications

The main contributions of this Master's work are reported in the following research papers:

- Jorge Valverde Tohalino and Diego Raphael Amancio. **Extractive multi-document summarization using dynamical measurements of complex networks.** *In 2017 Brazilian Conference on Intelligent Systems, BRACIS 2017, Uberlândia, Brazil, October 2-5, 2017, pages 366–371, 2017.*

- Jorge V. Tohalino and Diego R. Amancio. **Extractive multi-document summarization using multilayer networks.** *Physica A: Statistical Mechanics and its Applications, pages –, 2018.*

BIBLIOGRAPHY

AGGARWAL, G.; SUMBALY, R.; SINHA, S. Update summarization. **Communications of the ACM**, v. 48, n. 10, 2009. Citation on page 30.

AGOSTINI, V.; PARDO, T. Multi-document summarization using complex and rich features. **Anais do VIII Encontro Nacional de Inteligencia Artificial**, 2011. Citation on page 50.

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. **Reviews of modern physics**, APS, v. 74, n. 1, p. 47, 2002. Citation on page 34.

AMANCIO, D.; JR., O. O.; COSTA, L. da F. On the concepts of complex networks to quantify the difficulty in finding the way out of labyrinths. **Physica A: Statistical Mechanics and its Applications**, v. 390, n. 23–24, p. 4673 – 4683, 2011. ISSN 0378-4371. Available: <http://www.sciencedirect.com/science/article/pii/S0378437111005267>. Citations on pages 39 and 70.

AMANCIO, D. R.; ALTMANN, E. G.; JR, O. N. O.; COSTA, L. da F. Comparing intermittency and network measurements of words and their dependence on authorship. **New Journal of Physics**, IOP Publishing, v. 13, n. 12, p. 123024, 2011. Citation on page 70.

AMANCIO, D. R.; ANTIQUEIRA, L.; PARDO, T. A.; COSTA, L. da F.; JR, O. N. O.; NUNES, M. G. Complex networks analysis of manual and machine translations. **International Journal of Modern Physics C**, World Scientific, v. 19, n. 04, p. 583–598, 2008. Citation on page 27.

AMANCIO, D. R.; NUNES, M. G.; OLIVEIRA, O. N.; COSTA, L. d. F. Extractive summarization using complex networks and syntactic dependency. **Physica A: Statistical Mechanics and its Applications**, v. 391, n. 4, p. 1855–1864, 2012. Available: <http://EconPapers.repec.org/RePEc:eee:phsmap:v:391:y:2012:i:4:p:1855-1864>. Citations on pages 27, 50, and 57.

AMANCIO, D. R.; SILVA, F. N.; COSTA, L. d. F. Concentric network symmetry grasps authors' styles in word adjacency networks. **EPL (Europhysics Letters)**, IOP Publishing, v. 110, n. 6, p. 68001, 2015. Citations on pages 39 and 70.

ANTIQUEIRA, L. **Desenvolvimento de técnicas baseadas em redes complexas para sumarização extrativa de textos**. Phd Thesis (PhD Thesis) — Universidade de São Paulo, 2007. Citations on pages 32, 33, 60, and 78.

ANTIQUEIRA, L.; NUNES, M. d. G. V.; JR, O. O.; COSTA, L. d. F. Strong correlations between text quality and complex networks features. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 373, p. 811–820, 2007. Citation on page 27.

ANTIQUEIRA, L.; OLIVEIRA, O. N.; COSTA, L. d. F.; NUNES, M. d. G. V. A complex network approach to text summarization. **Inf. Sci.**, Elsevier Science Inc., New York, NY, USA, v. 179, n. 5, p. 584–599, 2009. ISSN 0020-0255. Available: <http://dx.doi.org/10.1016/j.ins.2008.10.032>. Citations on pages 25, 26, 27, 32, 48, 49, 50, 51, 53, 56, 57, and 65.

ANTIQUEIRA, L.; PARDO, T. A. S.; NUNES, M. d. G. V.; JR, O. N. O.; COSTA, L. d. F. Some issues on complex networks for author characterization. **Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial**, v. 11, n. 36, p. 51–58, 2007. Citation on page 27.

ARRUDA, G. Ferraz de; BARBIERI, A. L.; RODRIGUEZ, P. M.; MORENO, Y.; COSTA, L. da F.; RODRIGUES, F. A. The role of centrality for the identification of influential spreaders in complex networks. **ArXiv e-prints**, 2014. Citations on pages 38 and 70.

BALINSKY, H.; BALINSKY, A.; SIMSKE, S. J. Automatic text summarization and small-world networks. In: **Proceedings of the 11th ACM Symposium on Document Engineering**. New York, NY, USA: ACM, 2011. (DocEng '11), p. 175–184. ISBN 978-1-4503-0863-2. Available: <http://doi.acm.org/10.1145/2034691.2034731>. Citations on pages 54 and 57.

BARZILAY, R.; ELHADAD, M. Using lexical chains for text summarization. In: **Proceedings of the ACL/EACL 1997 Workshop on Intelligent Scalable Text Summarization**. [s.n.], 1997. p. 10–17. Available: <http://research.microsoft.com/en-us/um/people/cyl/download/papers/lexical-chains.pdf>. Citation on page 51.

BIRD, S. Nltk: the natural language toolkit. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the COLING/ACL on Interactive presentation sessions**. [S.l.], 2006. p. 69–72. Citation on page 62.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003. Citation on page 41.

BOCCALETTI, S.; BIANCONI, G.; CRIADO, R.; GENIO, C. I. D.; GÓMEZ-GARDENES, J.; ROMANCE, M.; SENDINA-NADAL, I.; WANG, Z.; ZANIN, M. The structure and dynamics of multilayer networks. **Physics Reports**, Elsevier, v. 544, n. 1, p. 1–122, 2014. Citation on page 40.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. **arXiv preprint arXiv:1607.04606**, 2016. Citations on pages 45, 84, and 87.

BORGATTI, S. P. Centrality and network flow. **Social Networks**, v. 27, n. 1, p. 55–71, 2005. Citation on page 35.

BRUNN, M.; CHALI, Y.; DUFOUR, B. The university of lethbridge text summarizer at duc 2002. In: CITESEER. **Document Understanding Conferences. Retrieved May**. [S.l.], 2002. v. 19, p. 2003. Citation on page 77.

CARDILLO, A.; GÓMEZ-GARDENES, J.; ZANIN, M.; ROMANCE, M.; PAPO, D.; POZO, F. D.; BOCCALETTI, S. Emergence of network features from multiplexity. **Scientific reports**, Nature Publishing Group, v. 3, p. 1344, 2013. Citation on page 40.

CARDOSO, P. C.; MAZIERO, E. G.; JORGE, M. L. C.; SENO, E. R.; FELIPPO, A. D.; RINO, L. H. M.; NUNES, M. d. G. V.; PARDO, T. A. S. Cstnews - a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In: **Proceedings of the 3rd RST Brazilian Meeting**. Cuiabá, Brazil: [s.n.], 2011. p. 88–105. Citations on pages 49 and 60.

CHOUDHURY, M.; THOMAS, M.; MUKHERJEE, A.; BASU, A.; GANGULY, N. How difficult is it to develop a perfect spell-checker? a cross-linguistic analysis through complex network approach. **arXiv preprint physics/0703198**, 2007. Citation on page 27.

COSTA, L. D. F.; OLIVEIRA JR., O.; TRAVIESO, G.; RODRIGUES, F. A.; BOAS, P. V.; ANTIQUEIRA, L.; VIANA, M. P.; ROCHA, L. C. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. **Advances in Physics**, v. 60, p. 329–412, 2011. Citations on pages 26 and 27.

COSTA, L. D. F.; RODRIGUES, F. A.; TRAVIESO, G.; BOAS, P. R. V. Characterization of complex networks: A survey of measurements. **Advances in Physics**, v. 56, p. 167–242, 2007. Citations on pages 33, 34, and 35.

COSTA, L. da F.; SILVA, F. N. Hierarchical Characterization of Complex Networks. **Journal of Statistical Physics**, v. 125, p. 841–872, 2006. Citations on pages 36, 37, and 70.

CRYSTAL, D. **A dictionary of linguistics and phonetics**. [S.l.]: John Wiley & Sons, 2011. Citation on page 100.

CSARDI, G.; NEPUSZ, T. The igraph software package for complex network research. **InterJournal, Complex Systems**, v. 1695, n. 5, p. 1–9, 2006. Citation on page 70.

DADACHEV, B.; BALINSKY, A.; BALINSKY, H.; SIMSKE, S. J. On the helmholtz principle for data mining. In: **2012 Third International Conference on Emerging Security Technologies, Lisbon, Portugal, September 5-7, 2012**. [s.n.], 2012. p. 99–102. Available: <http://dx.doi.org/10.1109/EST.2012.11>. Citation on page 54.

DAS, D.; MARTINS, A. F. A survey on automatic text summarization. **Literature Survey for the Language and Statistics II course at CMU**, v. 4, p. 192–195, 2007. Citation on page 30.

DEERWESTER, S.; DUMAIS, S. T.; FURNAS, G. W.; LANDAUER, T. K.; HARSHMAN, R. Indexing by latent semantic analysis. **Journal of the American society for information science**, American Documentation Institute, v. 41, n. 6, p. 391, 1990. Citation on page 44.

DUMAIS, S. T. Latent semantic analysis. **Annual review of information science and technology**, Wiley Online Library, v. 38, n. 1, p. 188–230, 2004. Citation on page 41.

EDMUNDSON, H. P. New methods in automatic extracting. **J. ACM**, ACM, New York, NY, USA, v. 16, n. 2, p. 264–285, Apr. 1969. ISSN 0004-5411. Available: <http://doi.acm.org/10.1145/321510.321519>. Citation on page 51.

ERDÖS, P.; RÉNYI, A. On random graphs, i. **Publicationes Mathematicae (Debrecen)**, v. 6, p. 290–297, 1959. Citations on pages 33 and 34.

ERKAN, G.; RADEV, D. R. Lexpagerank: Prestige in multi-document text summarization. In: LIN, D.; WU, D. (Ed.). **Proceedings of EMNLP 2004**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 365–371. Available: <http://www.aclweb.org/anthology/W04-3247>. Citation on page 78.

____. Lexrank: Graph-based lexical centrality as salience in text summarization. **J. Artif. Int. Res.**, AI Access Foundation, USA, v. 22, n. 1, p. 457–479, Dec. 2004. ISSN 1076-9757. Available: <http://dl.acm.org/citation.cfm?id=1622487.1622501>. Citations on pages 53 and 57.

FERREIRA, R.; CABRAL, L. de S.; LINS, R. D.; SILVA, G. P. e; FREITAS, F.; CAVALCANTI, G. D. C.; LIMA, R.; SIMSKE, S. J.; FAVARO, L. Assessing sentence scoring techniques for extractive text summarization. **Expert Syst. Appl.**, v. 40, n. 14, p. 5755–5764, 2013. Available: <http://dx.doi.org/10.1016/j.eswa.2013.04.023>. Citations on pages 25, 29, and 30.

FERREIRA, R.; FREITAS, F.; CABRAL, L. d. S.; LINS, R. D.; LIMA, R.; FRANÇA, G.; SIMSKEZ, S. J.; FAVARO, L. A four dimension graph model for automatic text summarization. In: IEEE COMPUTER SOCIETY. **Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01**. [S.l.], 2013. p. 389–396. Citations on pages 54, 55, and 57.

GAIZAUSKAS, R.; SAGGION, H. Multi-document summarization by cluster/profile relevance and redundancy removal. In: **Proceedings of the HLT/NAACL Document Understanding Workshop**. [S.l.: s.n.], 2004. p. 1–8. Citations on pages 74 and 75.

GOOGLE. **Word2Vec - Google Code Archive**. 2013. Available: <https://code.google.com/archive/p/word2vec/>. Citation on page 86.

GUPTA, V.; LEHAL, G. S. A survey of text summarization extractive techniques. **Journal of emerging technologies in web intelligence**, Citeseer, v. 2, n. 3, p. 258–268, 2010. Citation on page 30.

HALL, M. A. Correlation-based feature selection for discrete and numeric class machine learning. In: **Proceedings of the Seventeenth International Conference on Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000. (ICML '00), p. 359–366. ISBN 1-55860-707-2. Available: <http://dl.acm.org/citation.cfm?id=645529.657793>. Citation on page 51.

HARTMANN, N.; FONSECA, E.; SHULBY, C.; TREVISO, M.; SILVA, J.; ALUÍSIO, S. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: **Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology**. Sociedade Brasileira de Computação, 2017. p. 122–131. Available: <http://aclweb.org/anthology/W17-6615>. Citation on page 84.

HAYKIN, S. S.; HAYKIN, S. S.; HAYKIN, S. S.; HAYKIN, S. S. **Neural networks and learning machines**. [S.l.]: Pearson Upper Saddle River, NJ, USA:, 2009. Citation on page 72.

HIRAO, T.; SASAKI, Y.; ISOZAKI, H.; MAEDA, E. Ntt's text summarization system for duc-2002. In: **Proceedings of the Document Understanding Conference 2002**. [S.l.: s.n.], 2002. p. 104–107. Citation on page 77.

HIRSCHBERG, J. B.; MCKEOWN, K.; PASSONNEAU, R.; ELSON, D. K.; NENKOVA, A. Do summaries help? a task-based evaluation of multi-document summarization. Proceedings of ICASSP, Special Session on Human Language Technology Applications and Challenges for Speech Processing, 2005, 2005. Citation on page 26.

HUANG, E. H.; SOCHER, R.; MANNING, C. D.; NG, A. Y. Improving word representations via global context and multiple word prototypes. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1**. [S.l.], 2012. p. 873–882. Citation on page 42.

JIN, P.; ZHANG, Y.; CHEN, X.; XIA, Y. Bag-of-embeddings for text classification. In: **IJCAI**. [S.l.: s.n.], 2016. v. 16, p. 2824–2830. Citation on page 87.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: **Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (UAI'95), p. 338–345. ISBN 1-55860-385-9. Available: <http://dl.acm.org/citation.cfm?id=2074158.2074196>. Citation on page 51.

JORGE, M. L. d. R. C.; PARDO, T. A. S. Experiments with cst-based multidocument summarization. In: **Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (TextGraphs-5), p. 74–82. ISBN 978-1-932432-77-0. Available: <http://dl.acm.org/citation.cfm?id=1870490.1870502>. Citations on pages 49 and 77.

KHUSHBOO, S.; DHARASKAR, R.; CHANDAK, M. Graph-based algorithms for text summarization. In: **Third International Conference on Emerging Trends in Engineering and Technology**. [S.l.: s.n.], 2010. p. 49–53. Citations on pages 54 and 57.

KIVELÄ, M.; ARENAS, A.; BARTHELEMY, M.; GLEESON, J. P.; MORENO, Y.; PORTER, M. A. Multilayer networks. **Journal of complex networks**, Oxford University Press, v. 2, n. 3, p. 203–271, 2014. Citation on page 40.

KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. **J. ACM**, ACM, New York, NY, USA, v. 46, n. 5, p. 604–632, Sep. 1999. ISSN 0004-5411. Available: <http://doi.acm.org/10.1145/324133.324140>. Citation on page 53.

KUPIEC, J.; PEDERSEN, J.; CHEN, F. A trainable document summarizer. In: **Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: ACM, 1995. (SIGIR '95), p. 68–73. ISBN 0-89791-714-6. Available: <http://doi.acm.org/10.1145/215206.215333>. Citation on page 51.

____. A trainable document summarizer. In: ACM. **Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval**. [S.l.], 1995. p. 68–73. Citation on page 72.

LAI, S.; LIU, K.; HE, S.; ZHAO, J. How to generate a good word embedding. **IEEE Intelligent Systems**, IEEE, v. 31, n. 6, p. 5–14, 2016. Citation on page 84.

LANCASTER, F. W.; GALLUP, E. **Information retrieval on-line**. [S.l.], 1973. Citation on page 41.

LATTANZI, S.; SIVAKUMAR, D. Affiliation networks. In: **Proceedings of the 41st Annual ACM Symposium on Theory of Computing**. [s.n.], 2009. p. 427–434. Available: <http://portal.acm.org/citation.cfm?id=1536414.1536474>. Citation on page 54.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: **International Conference on Machine Learning**. [S.l.: s.n.], 2014. p. 1188–1196. Citation on page 84.

LEE, D. L.; CHUANG, H.; SEAMONS, K. Document ranking and the vector-space model. **IEEE Software**, v. 14, n. 2, p. 67–75, Mar 1997. ISSN 0740-7459. Citation on page 41.

LEE, S.; BELKASIM, S.; ZHANG, Y. Multi-document text summarization using topic model and fuzzy logic. In: **Proceedings of the 9th International Conference on Machine Learning and Data Mining in Pattern Recognition**. Berlin, Heidelberg: Springer-Verlag, 2013. (MLDM'13), p. 159–168. ISBN 978-3-642-39711-0. Available: <http://dx.doi.org/10.1007/978-3-642-39712-7_12>. Citation on page 25.

____. Multi-document text summarization using topic model and fuzzy logic. In: SPRINGER. **International Workshop on Machine Learning and Data Mining in Pattern Recognition**. [S.l.], 2013. p. 159–168. Citation on page 73.

LEITE, D. S.; RINO, L. H. M. Selecting a feature set to summarize texts in brazilian portuguese. In: **Advances in Artificial Intelligence - IBERAMIA-SBIA 2006, 2nd International Joint Conference, 10th Ibero-American Conference on AI, 18th Brazilian AI Symposium, Ribeirão Preto, Brazil, October 23-27, 2006, Proceedings**. [s.n.], 2006. p. 462–471. Available: <http://dx.doi.org/10.1007/11874850_50>. Citations on pages 51, 77, and 91.

____. Combining multiple features for automatic text summarization through machine learning. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2008. p. 122–132. Citations on pages 50, 51, 56, 57, 71, 72, 77, and 91.

LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: **Proc. ACL workshop on Text Summarization Branches Out**. [s.n.], 2004. p. 10. Available: <http://research.microsoft.com/~cyl/download/papers/WAS2004.pdf>. Citations on pages 32 and 47.

LING, W.; DYER, C.; BLACK, A. W.; TRANCOSO, I. Two/too simple adaptations of word2vec for syntax problems. In: **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. [S.l.: s.n.], 2015. p. 1299–1304. Citation on page 84.

LOUIS, A.; JOSHI, A.; NENKOVA, A. Discourse indicators for content selection in summarization. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue**. [S.l.], 2010. p. 147–156. Citation on page 55.

LUHN, H. P. The automatic creation of literature abstracts. **IBM J. Res. Dev.**, IBM Corp., Riverton, NJ, USA, v. 2, n. 2, p. 159–165, Apr. 1958. ISSN 0018-8646. Available: <http://dx.doi.org/10.1147/rd.22.0159>. Citation on page 51.

MAÑA-LÓPEZ, M. J.; BUENAGA, M. D.; GÓMEZ-HIDALGO, J. M. Multidocument summarization: An added value to clustering in interactive retrieval. **ACM Transactions on Information Systems (TOIS)**, ACM, v. 22, n. 2, p. 215–241, 2004. Citation on page 26.

MANI, I. **Automatic summarization**. [S.l.]: John Benjamins Publishing Company, 2001. Citations on pages 25 and 26.

MANI, I.; BLOEDORN, E. Summarizing similarities and differences among related documents. **Information Retrieval**, v. 1, n. 1, p. 35–67, 1999. ISSN 1573-7659. Available: <http://dx.doi.org/10.1023/A:1009930203452>. Citations on pages 52 and 57.

MASUDA, N.; PORTER, M. A.; LAMBIOTTE, R. Random walks and diffusion on networks. **Physics Reports**, Elsevier, 2017. Citation on page 69.

MIHALCEA, R. Language independent extractive summarization. In: **Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (ACLdemo '05), p. 49–52. Available: <http://dx.doi.org/10.3115/1225753.1225766>. Citations on pages 53, 57, and 77.

MIHALCEA, R.; TARAU, P. Textrank: Bringing order into text. In: **Proceedings of the 2004 conference on empirical methods in natural language processing**. [S.l.: s.n.], 2004. Citation on page 55.

MIKOLOV, T.; CHEN, K.; CORRADO, G. S.; DEAN, J. Efficient estimation of word representations in vector space. **CoRR**, abs/1301.3781, 2013. Citation on page 42.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2013. p. 3111–3119. Citations on pages 41, 42, 43, 44, and 84.

MILLER, G. A. Wordnet: a lexical database for english. **Communications of the ACM**, ACM, v. 38, n. 11, p. 39–41, 1995. Citation on page 27.

NENKOVA, A.; MASKEY, S.; LIU, Y. Automatic summarization. In: **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011**. [S.l.]: Association for Computational Linguistics, 2011. (HLT '11), p. 3:1–3:86. Citations on pages 26 and 29.

NENKOVA, A.; MCKEOWN, K. A survey of text summarization techniques. In: **Mining text data**. [S.l.]: Springer, 2012. p. 43–76. Citations on pages 25, 26, 29, 30, and 47.

NETO, J. L.; FREITAS, A. A.; KAESTNER, C. A. Automatic text summarization using a machine learning approach. In: SPRINGER. **Brazilian Symposium on Artificial Intelligence**. [S.l.], 2002. p. 205–215. Citations on pages 72 and 73.

NETO, J. L.; SANTOS, A. D.; KAESTNER, C. A.; FREITAS, A. A. Generating text summaries through the relative importance of topics. In: ____. **Advances in Artificial Intelligence: International Joint Conference 7th Ibero-American Conference on AI 15th Brazilian Symposium on AI IBERAMIA-SBIA 2000 Atibaia, SP, Brazil, November 19–22, 2000 Proceedings**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000. p. 300–309. ISBN 978-3-540-44399-5. Available: <http://dx.doi.org/10.1007/3-540-44399-1_31>. Citation on page 51.

NEWMAN, M. **Networks: an introduction**. [S.l.]: Oxford university press, 2010. Citation on page 34.

NEWMAN, M. E. J. A measure of betweenness centrality based on random walks. **eprint arXiv:cond-mat/0309045**, 2003. Citation on page 36.

OVER, P.; LIGGETT, W. **Introduction to DUC: An Intrinsic Evaluation of Generic News Text Summarization Systems**. 2002. <http://www-nlpir.nist.gov/projects/duc/pubs/2002slides/overview.02.pdf>. Citations on pages 47, 60, and 77.

PADMANABHAN, D.; DESIKAN, P.; SRIVASTAVA, J.; RIAZ, K. Wicer: A weighted inter-cluster edge ranking for clustered graphs. In: **Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence**. Washington, DC, USA: IEEE Computer Society, 2005. (WI '05), p. 522–528. ISBN 0-7695-2415-X. Citation on page 67.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. The pagerank citation ranking: Bringing order to the web. In: **Proceedings of the 7th International World Wide Web Conference**. Brisbane, Australia: [s.n.], 1998. p. 161–172. Available: <citeseer.nj.nec.com/page98pagerank.html>. Citation on page 53.

PAGLIARDINI, M.; GUPTA, P.; JAGGI, M. Unsupervised learning of sentence embeddings using compositional n-gram features. **CoRR**, abs/1703.02507, 2017. Available: <http://arxiv.org/abs/1703.02507>. Citation on page 84.

PARDO, T. A. S.; ANTIQUEIRA, L.; NUNES, M. d. G. V.; OLIVEIRA, O. N.; COSTA, L. da F. Modeling and evaluating summaries using complex networks. In: ____. **Computational**

Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 13-17, 2006. Proceedings. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 1–10. ISBN 978-3-540-34046-1. Available: <http://dx.doi.org/10.1007/11751984_1>. Citation on page 26.

PARDO, T. A. S.; RINO, L. H. M. **TeMário: Um Corpus para Sumarização Automática de Textos**. São Carlos-SP, 2003. 11 p. Citations on pages 47, 49, and 60.

____. Descrição do gei-gerador de extratos ideais para o português do brasil. **Série de Relatórios do NILC NILC-TR-04-07, Núcleo Interinstitucional de Lingüistica Computacional (NILC), Sao Carlos-SP**, v. 8, 2004. Citation on page 49.

PARDO, T. A. S.; RINO, L. H. M.; NUNES, M. d. G. V. Gistsumm: A summarization tool based on a new extractive method. In: ____. **Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003 Faro, Portugal, June 26–27, 2003 Proceedings**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. p. 210–218. ISBN 978-3-540-45011-5. Available: <http://dx.doi.org/10.1007/3-540-45011-4_34>. Citations on pages 49 and 77.

PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: Global vectors for word representation. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543. Citations on pages 44, 84, and 87.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0. Citation on page 51.

____. **C4. 5: programs for machine learning**. [S.l.]: Elsevier, 2014. Citation on page 72.

RADEV, D. R. Experiments in single and multidocument summarization using mead. In: **In First Document Understanding Conference**. [S.l.: s.n.], 2001. Citation on page 49.

RATNAPARKHI, A. *et al.* A maximum entropy model for part-of-speech tagging. In: **Proceedings of the conference on empirical methods in natural language processing**. [S.l.: s.n.], 1996. v. 1, p. 133–142. Citation on page 62.

RESNIK, P. Using information content to evaluate semantic similarity in a taxonomy. **arXiv preprint cmp-lg/9511007**, 1995. Citation on page 55.

RIBALDO, R.; AKABANE, A. T.; RINO, L. H. M.; PARDO, T. A. S. Graph-based methods for multi-document summarization: Exploring relationship maps, complex networks and discourse information. In: **Computational Processing of the Portuguese Language**. [S.l.]: Springer Berlin Heidelberg, 2012. v. 7243, p. 260–271. Citations on pages 25, 32, 33, 49, 56, 57, 65, 73, 74, and 77.

ROBERTSON, S. Understanding inverse document frequency: on theoretical arguments for idf. **Journal of documentation**, Emerald Group Publishing Limited, v. 60, n. 5, p. 503–520, 2004. Citation on page 63.

ROETHLISBERGER, F. J.; DICKSON, W. J. **Management and the Worker**. [S.l.]: Psychology Press, 2003. Citation on page 40.

SALTON, G.; MCGILL, M. J. **Introduction to Modern Information Retrieval**. New York, NY, USA: McGraw-Hill, Inc., 1986. ISBN 0070544840. Citation on page 53.

SALTON, G.; SINGHAL, A.; MITRA, M.; BUCKLEY, C. Automatic text structuring and summarization. **Inf. Process. Manage.**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 33, n. 2, p. 193–207, Mar. 1997. ISSN 0306-4573. Available: <http://dx.doi.org/10.1016/S0306-4573(96)00062-3>. Citations on pages 51, 52, and 57.

SALTON, G.; WONG, A.; YANG, C.-S. A vector space model for automatic indexing. **Communications of the ACM**, ACM, v. 18, n. 11, p. 613–620, 1975. Citation on page 41.

SAMEI, B.; ESTIAGH, M.; KESHTKAR, F.; HASHEMI, S. Multi-document summarization using graph-based iterative ranking algorithms and information theoretical distortion measures. In: **FLAIRS Conference**. [S.l.: s.n.], 2014. Citations on pages 55, 56, and 57.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM computing surveys (CSUR)**, ACM, v. 34, n. 1, p. 1–47, 2002. Citation on page 42.

SILVA, F. N.; AMANCIO, D. R.; BARDOSOVA, M.; COSTA, L. d. F.; JR, O. N. O. Using network science and text analytics to produce surveys in a scientific topic. **Journal of Informetrics**, Elsevier, v. 10, n. 2, p. 487–502, 2016. Citation on page 91.

SILVA, F. N.; COMIN, C. H.; PERON, T. K. D.; RODRIGUES, F. A.; YE, C.; WILSON, R. C.; HANCOCK, E. R.; COSTA, L. d. F. Concentric network symmetry. **Information Sciences**, Elsevier, v. 333, p. 61–80, 2016. Citation on page 70.

SILVA, F. N.; COSTA, L. da F. 2013. Available: <http://cyvision.ifsc.usp.br/concentric/software>. Accessed: 01/04/2016. Citation on page 70.

SPENCER, N. 2016. Available: <https://www.gwava.com/blog/internet-data-created-daily>. Accessed: 08/09/2016. Citation on page 25.

STROGATZ, S. H. Exploring complex networks. **Nature**, Nature Publishing Group, v. 410, n. 6825, p. 268–276, 2001. Citation on page 33.

SYSTEMS, C.; LAB, N. **Network Theory: Multiplex Networks**. 2018. Available: <http://cosnet.bifi.es/network-theory/multiplex-networks/>. Citation on page 40.

TELLEX, S.; KATZ, B.; LIN, J.; FERNANDES, A.; MARTON, G. Quantitative evaluation of passage retrieval algorithms for question answering. In: ACM. **Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval**. [S.l.], 2003. p. 41–47. Citation on page 42.

TOHALINO, J. V.; AMANCIO, D. R. Extractive multi-document summarization using dynamical measurements of complex networks. In: **2017 Brazilian Conference on Intelligent Systems, BRACIS 2017, Uberlândia, Brazil, October 2-5, 2017**. [s.n.], 2017. p. 366–371. Available: <https://doi.org/10.1109/BRACIS.2017.41>. Citations on pages 17, 63, 65, and 66.

____. Extractive multi-document summarization using multilayer networks. **Physica A: Statistical Mechanics and its Applications**, v. 503, p. 526 – 539, 2018. ISSN 0378-4371. Available: <http://www.sciencedirect.com/science/article/pii/S0378437118303212>. Citations on pages 69, 88, and 89.

TOMBROS, A.; SANDERSON, M. Advantages of query biased summaries in information retrieval. In: ACM. **Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval**. [S.l.], 1998. p. 2–10. Citation on page 26.

TRAVENçOLO, B.; COSTA, L. da F. Accessibility in complex networks. **Physics Letters A**, v. 373, n. 1, p. 89 – 95, 2008. ISSN 0375-9601. Available: <http://www.sciencedirect.com/science/article/pii/S0375960108015867>. Citation on page 37.

TURIAN, J.; RATINOV, L.; BENGIO, Y. Word representations: a simple and general method for semi-supervised learning. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 48th annual meeting of the association for computational linguistics**. [S.l.], 2010. p. 384–394. Citation on page 42.

VAPNIK, V. N. **The Nature of Statistical Learning Theory**. New York, NY, USA: Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8. Citation on page 51.

VIANA, M. P.; BATISTA, J. a. L. B.; COSTA, L. d. F. Effective number of accessed nodes in complex networks. **Phys. Rev. E**, American Physical Society, v. 85, p. 036105, Mar 2012. Available: <http://link.aps.org/doi/10.1103/PhysRevE.85.036105>. Citation on page 70.

WANG, D.; ZHU, S.; LI, T.; CHI, Y.; GONG, Y. Integrating clustering and multi-document summarization to improve document understanding. In: ACM. **Proceedings of the 17th ACM conference on Information and knowledge management**. [S.l.], 2008. p. 1435–1436. Citation on page 78.

WANG, D.; ZHU, S.; LI, T.; GONG, Y. Multi-document summarization using sentence-based topic models. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the ACL-IJCNLP 2009 Conference Short Papers**. [S.l.], 2009. p. 297–300. Citation on page 78.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. **nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. Citations on pages 33 and 34.

WEI, F.; LI, W.; LU, Q.; HE, Y. A document-sensitive graph model for multi-document summarization. **Knowledge and information systems**, Springer, v. 22, n. 2, p. 245–259, 2010. Citation on page 67.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123748569, 9780123748560. Citation on page 51.

WU, Z.; PALMER, M. Verbs semantics and lexical selection. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 32nd annual meeting on Association for Computational Linguistics**. [S.l.], 1994. p. 133–138. Citation on page 55.

WUBBEN, S.; BOSCH, A. van den. A semantic relatedness metric based on free link structure. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the Eighth International Conference on Computational Semantics**. [S.l.], 2009. p. 355–358. Citation on page 55.

ZHANG, H. The optimality of naive bayes. **AA**, v. 1, n. 2, p. 3, 2004. Citation on page 72.

ZOU, W. Y.; SOCHER, R.; CER, D.; MANNING, C. D. Bilingual word embeddings for phrase-based machine translation. In: **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2013. p. 1393–1398. Citation on page 87.

APPENDIX

A

STOPWORD LIST FOR ENGLISH AND PORTUGUESE

A.1 English stopword list

i, me, my, myself, we, our, ours, ourselves, you, your, yours, yourself, yourselves, he, him, his, himself, she, her, hers, herself, it, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very, s, t, can, will, just, don, should, now, d, ll, m, o, re, ve, y, ain, aren, couldn, didn, doesn, hadn, hasn, haven, isn, ma, mightn, mustn, needn, shan, shouldn, wasn, weren, won, wouldn.

A.2 Portuguese stopword list

de, a, o, que, e, do, da, em, um, para, com, não, uma, os, no, se, na, por, mais, as, dos, como, mas, ao, ele, das, à, seu, sua, ou, quando, muito, nos, já, eu, também, só, pelo, pela, até, isso, ela, entre, depois, sem, mesmo, aos, seus, quem, nas, me, esse, eles, você, essa, num, nem, suas, meu, às, minha, numa, pelos, elas, qual, nós, lhe, deles, essas, esses, pelas, este, dele, tu, te, vocês, vos, lhes, meus, minhas, teu, tua, teus, tuas, nosso, nossa, nossos, nossas, dela, delas, esta, estes, estas, aquele, aquela, aqueles, aquelas, isto,

aquilo, estou, está, estamos, estão, estive, esteve, estivemos, estiveram, estava, estávamos, estavam, estivera, estivéramos, esteja, estejamos, estejam, estivesse, estivéssemos, estivessem, estiver, estivermos, estiverem, hei, há, havemos, hão, houve, houvemos, houveram, houvera, houvéramos, haja, hajamos, hajam, houvesse, houvéssemos, houvessem, houver, houvermos, houverem, houverei, houverá, houveremos, houverão, houveria, houveríamos, houveriam, sou, somos, são, era, éramos, eram, fui, foi, fomos, foram, fora, fôramos, seja, sejamos, sejam, fosse, fôssemos, fossem, for, formos, forem, serei, será, seremos, serão, seria, seríamos, seriam, tenho, tem, temos, tém, tinha, tínhamos, tinham, tive, teve, tivemos, tiveram, tivera, tivéramos, tenha, tenhamos, tenham, tivesse, tivéssemos, tivessem, tiver, tivermos, tiverem, terei, terá, teremos, terão, teria, teríamos, teriam.


ICMC
USP
SÃO CARLOS