Carla Bonato Marcolin

Text Analytics in Business Environments: a Managerial and Methodological approach

Brasil

Março, 2018

Carla Bonato Marcolin

Text Analytics in Business Environments: a Managerial and Methodological approach

Tese apresentada ao Programa de Pós-Graduação como requisito parcial para obtenção do título de doutor.

Universidade Federal do Rio Grande do Sul

Escola de Administração

Programa de Pós-Graduação em Administração (PPGA)

Supervisor: João Luiz Becker

Brasil

Março, 2018

CIP - Catalogação na Publicação

Marcolin, Carla Bonato
Text Analytics in Business Environments: a Managerial and Methodological approach / Carla Bonato Marcolin. -- 2018.
127 f.
Orientador: João Luiz Becker.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Escola de Administração, Programa de Pós-Graduação em Administração, Porto Alegre, BR-RS, 2018.

1. Business Analytics. 2. Dados em Texto. 3. LSA. 4. Pesquisa Operacional. I. Becker, João Luiz, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Carla Bonato Marcolin

Text Analytics in Business Environments: a Managerial and Methodological approach

Tese apresentada ao Programa de Pós-Graduação como requisito parcial para obtenção do título de doutor.

João Luiz Becker
Orientador

Profa^a. Dr^a. Viviane Pereira Moreira
(INF/UFRGS)

Prof. Dr. Marcirio Silveira Chaves
(PUCRS)

Prof. Dr. Luciano Ferreira
(EA/UFRGS)

Brasil
Março, 2018

Acknowledgements

As a recently master graduate student, I was encourage and felt motivated to face the challenge to work in Operational Research area. Looking back to the first year, I remember to constantly have a "it will not be possible" sensation. However, several people helped to quickly overcome this feeling, and I know that it would not be possible to finish this work without them. First, my colleagues from 329, that helped not only with academic knowledge, but also with their support and friendship. A special thanks to all professors from OR area, for the knowledge and patience. My research colleagues, Ariel, Fernanda and Giovana, for all advices and for the opportunity to still be on their team. To all professors members of the examination board of this dissertation for all time dedicated to my work and all contributions made. But none of this would have been possible without my advisor, Prof. Becker. Every time he turned my questions into another questions I realized that he was pushing me further, and I certainly owe him too much.

My family was a solid support in every moment. My parents, my brother and sister-in-law and even my little goddaughter were constantly comprehensive about my absences and about my not-so-interesting subjects in every family meeting. My friends, that not only accepted my faults but also listened to all my talk about algorithms, and even so are still my friends. A special thanks to Bibi and to Bentinho (*in memorian*), that were faithful companions. And I dedicate this work to my husband, Felipe, that told me uncountable times that I could do anything I wanted, and constantly sacrificed his leisure moments just to stay with me while I was working.

Additionally, I must thank two important academic figures in my own academic journey. First, Prof. Fridolin Wild, from Oxford Brookes University, and all Performance Augmentation Lab team (Alla and Will) that warmly received me in their Lab, taught me so much and in my last day told me that it was just the beginning, which I truly agree! Second, thanks to Prof. Henrique Freitas, that helped me to understand the academic career, always available to guide me for the best.

Finally, I'd like to thank CAPES, Coordenação de Aperfeiçomento de Pessoal de Nível Superior, for the scholarship granted that made this research possible.

Resumo

O processo de tomada de decisão, em diferentes ambientes gerenciais, enfrenta um momento de mudança no contexto organizacional. Nesse sentido, Business Analytics pode ser visto como uma área que permite alavancar o valor dos dados, contendo ferramentas importantes para o processo de tomada de decisão. No entanto, a presença de dados em diferentes formatos representa um desafio. Nesse contexto de variabilidade, os dados de texto têm atraído a atenção das organizações, já que milhares de pessoas se expressam diariamente neste formato, em muitas aplicações e ferramentas disponíveis. Embora diversas técnicas tenham sido desenvolvidas pela comunidade de ciência da computação, há amplo espaço para melhorar a utilização organizacional de tais dados de texto, especialmente quando se volta para o suporte à tomada de decisões. No entanto, apesar da importância e disponibilidade de dados em formato textual para apoiar decisões, seu uso não é comum devido à dificuldade de análise e interpretação que o volume e o formato de dados em texto apresentam. Assim, o objetivo desta tese é desenvolver e avaliar um framework voltado ao uso de dados de texto em processos decisórios, apoiando-se em diversas técnicas de processamento de linguagem natural (PNL). Os resultados apresentam a validade do framework, usando como instância de demonstração de sua aplicabilidade o setor de turismo através da plataforma TripAdvisor, bem como a validação interna de performance e a aceitação por parte dos gestores da área consultados.

Palavras-chave: Business Analytics. Dados em Texto. LSA. Pesquisa Operacional.

Abstract

The decision-making process, in different management environments, faces a moment of change in the organizational context. In this sense, Business Analytics can be seen as an area that leverages the value of data, containing important tools for the decision-making process. However, the presence of data in different formats poses a challenge. In this context of variability, text data has attracted the attention of organizations, as thousands of people express themselves daily in this format in many applications and tools available. Although several techniques have been developed by the computer science community, there is ample scope to improve the organizational use of such text data, especially when it comes to decision-making support. However, despite the importance and availability of textual data to support decisions, its use is not common because of the analysis and interpretation challenge that the volume and the unstructured format of text data presents. Thus, the aim of this dissertation is to develop and evaluate a framework to contribute with the expansion and development of text analytics in decision-making processes, based on several natural language processing (NLP) techniques. The results presents the validity of the framework, using as a demonstration of its applicability the tourism sector through the TripAdvisor platform, as well as the internal validation of performance and the acceptance by managers.

Keywords: Business Analytics. Text data. LSA. Operational Research.

List of Figures

List of Tables

Contents

Introduction

In a managerial environment, decision-making processes find different challenges, including risk, time and information. Several techniques and models have historically presented themselves as mediators, capable of obtaining better information, in less time and in a more precise way. However, different organizations are facing a change in the informational scenario that can transform the way companies do business that is being compared to the process re-engineering movement in the early 1990s. The trend now has often been called business analytics (MCAFEE et al., 2012; BARTON; COURT, 2012).

In this context, the concept and practice of business analytics had a significant growth in the last decade, attracting the attention of researchers and managers from different areas (MORTENSON; DOHERTY; ROBINSON, 2015). Business analytics can be seen as an emerging phenomenon reflecting the exponential growth of data in terms of volume, variety and velocity. Inside this concept is the extensive use of data driven quantitative methods, specially statistics and mathematics. In a widely discussion, Mortenson, Doherty e Robinson (2015) states that business analytics has seen significant growth, defining it in terms of its related disciplines, namely technologies, decision making and quantitative methods, although claiming a "lack of any clear consensus about analytics' precise definition, and how it differs from related concepts" (MORTENSON; DOHERTY; ROBINSON, 2015, p. 584).

Besides this lack of consensus, there is evidence that data-driven decisions tend to influence organizational performance (MCAFEE et al., 2012), and that developing data skills is an important strategic asset in order to foster or maintain competitive advantage (PROVOST; FAWCETT, 2016). Allowing to leverage value from data, business analytics can be considered an important tool for the decision-making process (ACITO; KHATRI, 2014). In this sense, business analytics have the power to analyze data, from different data sources, making it possible to improve companies performance and identify business opportunities (BAYRAK, 2015).

However, there are plenty of companies unsure about how to effectively use available data into decision-making process (DAVENPORT; DYCHÉ, 2013), for a couple of reasons. The hype from past technology movements that came to drastically change market structures, calling themselves "revolutions" may cause suspicion, specially after some high-technology and expensive platforms previously developed (BARTON; COURT, 2012). Another point is data volume, since traditional methods can no longer meet organization's need to analyze, in a daily basis, new large amounts of data, bringing the need for companies to leverage their data analytics skills in order to embrace opportunities offered by different

data sources (HE et al., 2017), specially those external to organization walls.

Data-driven strategies can already be seen as an increasingly important point of competitive differentiation (BARTON; COURT, 2012), thus, in order to create value from data the focus in analytics skills is an important capability to obtain. In addition, the challenge arises not only from data volume, but also in its variety. In recent years, the presence of data in different formats has posed a new challenge for companies and analysts. In addition to dealing with large volumes of data, they now need to be able to handle new data types, such as voice, text, log files, images and videos (DAVENPORT; DYCHÉ, 2013).

Given this, a singular format comes to attention. Especially after the emergence of social networks and mobile technologies, the volume of text data has been multiplied, being estimated in at least 80% from all organizational data (KOBAYASHI et al., 2017). Thousands of people comment, give their opinions, write and publish all the time on the Internet through various devices such as tablets, smartphones and notebooks, citing just a few. All of these expression formats ends up producing an environment with an information overload: any simple Google search can produce thousands of results in a few seconds (AGGARWAL; ZHAI, 2012). In this sense, text data is expected to become more and more present, specially in management processes, highly related to people, that have in text a free, and therefore more authentic, way of expressing themselves (GORRY; WESTBROOK, 2011).

Text data have not emerged within the analytics era. Disciplines like information science have already addressed issues of text data indexing and organizing for quite long. More recently, computer science tools and advances have supported specific techniques and models in this and others tasks in information retrieval (IR), such as document relevance (MANNING; RHAGAVAN; SCHUTZE, 2009). At the same time, data mining has been increasingly modernized to meet the large volume of existing data, working on how to solve manipulation and analysis issues in order to keep up with the increasing business dynamics (AGGARWAL; ZHAI, 2012).

Therefore, while just a few would argue about text data richness for business, broad approaches for its analysis still remain mostly unsolved (ASHTON; EVANGELOPOULOS; PRYBUTOK, 2014). Reports of applied application and of successful practices to manipulate and use relevant text data like social media are still largely unknown (SCHUCKERT; LIU; LAW, 2015; ASHTON; EVANGELOPOULOS; PRYBUTOK, 2014; LEUNG et al., 2013), making it difficult to find these methods in business processes. Research that helps to uncover how to handle the amount of data generated and how to filter this data in order to obtain a reliable picture of the authors (i. e., customers) perspective is still needed, specially in domain specific business areas (VALDIVIA; LUZÓN; HERRERA, 2017; CANTALLOPS; SALVI, 2014).

In addition, specially for managers, this kind of data impose some difficulties. One among them is how to deal with the unstructured nature of text, that demands new tools that are not widely available inside companies. The direct usage of raw text data to understand, for example, customer opinion, is a challenging task not only due to the lack of methods and procedures to deal with text, but also because of issues like typographical errors and ambiguity, unavoidable in real-life datasets (AGUWA; OLYA; MONPLAISIR, 2017). Additionally, common tools like standard statistical and econometric techniques, consolidated among organizations, do not handle text directly (HAN et al., 2016).

Those could be reasons why still several business leaders learn about their customers from sales, operations and customer service reports than by customers themselves (GORRY; WESTBROOK, 2011). In this sense, there is an opportunity to contribute by providing not only the possibility to increase the volume of data collected with analytics tools, but also to overcome a step in text analysis, that implies a costly and quite subjective task of reading. Thus, it is possible to avoid bias that might bring different conclusions from the same data, once different people with different experiences might see different structures when reading data in text (ASHTON; EVANGELOPOULOS; PRYBUTOK, 2014).

There is also a lack of research considering unstructured data and decision-making process. Although Operational Research (OR) associations like INFORMS and UK OR Society may have included Analytics not only as a conference track, but also as an event subject and even offering an official certification, the amount of research related with analytics is surprisingly low (MORTENSON; DOHERTY; ROBINSON, 2015). In contrast with the quickly adoption from organizations of this term, that is already in departments' name and in many executives agenda (BARTON; COURT, 2012), there is room for research connecting OR and business analytics. Even with the recognition that analytics is a multidisciplinary field, the role of decision-making is evident (MCAFEE et al., 2012).

Mortenson, Doherty e Robinson (2015) brought several opportunities for the OR field in what the authors called the "analytics age" (MORTENSON; DOHERTY; ROBINSON, 2015, p. 592). This dissertation hopes to contribute with the unstructured data issue, derived from the variety aspect from this analytics age, due specially to the exponential growth of user-generated content. Bringing a managerial and methodological approach fully based in text data, the goal is to advance in research directions indicated by the authors, such as pre-processing methods and the use of real-world data for decision-making support.

Having a quantitative approach, it might be argued that text is distant from OR methods, considering the qualitative nature of this data and the dynamic human process of creating a sentence. In order to consider text suitable for methods based on statistics and mathematics, it is important to recall George Zipf studies (ZIPF, 1949; ZIPF, 1935). Zipf was one of the first researchers to demonstrate that language had a structure, that

could be describe as an inverse proportion between frequency and the rank-position given a frequency table. In addition, he presented evidence that this rule was true not only in English, but also in Chinese and in different languages with roots in Latin (ZIPF, 1949).

While many linguistics were concerned about "mind theories" in order to explain language, Zipf believed that the examinations of the facts of the language were more effective in order to better understand its process and, ultimately, spread light into the theories of mind. In fact, statistical approaches for language have been used to understand different cognitive processes (WILD, 2016; LANDAUER, 2007). Although language is a social process, enabling us to free express ourselves, it can also benefit from a numerical structure, allowing to look behind a great amount of words and reveal relations that would hardly be found through reading. Therefore, even that no one selects the words for the "sake of preserving or restoring any imaginable condition of equilibrium in the resultant frequency-distribution of the elements of the speech" (ZIPF, 1935, p. xiii), Zipf constructed strong evidence for what is called the Zipf law, or Zipf curve, focusing in how people do express themselves, and not how they should.

Just as thermometers are sensors that represent real-world data (the temperature) and GPS devices can spot a real-world location (through latitude and longitude), text data can be used as a sensor for measuring perception at individual, a community or even a regional level (ZHAO, 2013). Therefore, text data can be seen as an important source, once it allows to analyze opinions, thoughts and perceptions through raw text, just like it was conceived in the mind of the manager, employee, user or consumer, among other interest groups.

Several literature therefore brought evidence about text data importance and unavoidable growing presence in business environments (KOBAYASHI et al., 2017; PROVOST; FAWCETT, 2016; MORTENSON; DOHERTY; ROBINSON, 2015; RANYARD; FILDES; HU, 2015; ASHTON; EVANGELOPOULOS; PRYBUTOK, 2014; MCAFEE et al., 2012). Despite of that, previous research focused in develop methods to support and analyze text data (XIANG et al., 2015; LEE; HAN; SUH, 2014; CARRASCO et al., 2012; MIRKIN, 2011; MANEVITZ; YOUSEF, 2007), and did not proposed a single integrated set of solutions, developed in this dissertation as a framework.

Given this context, the research question presented is: how to connect text data to decision-making process for managerial analysis by organizations? This dissertation aims, therefore, to contribute with the expansion and development of text analysis techniques and tools applied in decision-making processes through a framework development and evaluation. Supported by design science principles for artifact construction (DRESCH; LACERDA; JR, 2014; AKEN, 2005; HEVNER et al., 2004), the specific objectives are: (1) to develop a tool that automatize data capturing and cleansing; (2) to develop an automated solution to monitor trends in text data; (3) to develop a model to continuously

classify text allowing summarization, pattern identification and market analysis; and (4) to evaluate the developed framework considering pragmatism and usefulness.

In order to accomplish that, this document is organized as follows: after this Introduction, a three-part Theoretical Background presents this dissertation context (Information Retrieval and Text Mining) altogether with the Applied Text Mining Operations and a discussion about the text data source for the framework evaluation, Voice of the Customer, followed by the Methodological Approach. Afterwards, the Results are discussed and finally the Conclusions are presented. In addition, Appendix section (from A to E) presents the functions developed regarding the proposed framework.

Part I

Theoretical Background

1 Information Retrieval and Text Mining

The large volume of text data have leveraged the Information Retrieval (IR) field. Originated from the necessity to retrieve relevant information, especially in academic records and libraries, its purpose is to interpret, in the best way, a user request. After a database search, the objective is to bring those records more closely related with the search purpose, i.e, to retrieve the more relevant ones for the user (MANNING; RHAGAVAN; SCHUTZE, 2009).

There are certainly several challenges in this process. Request interpretation is one of them, since different people express themselves using different terms, although may be referring to the same concept. On the other hand, the same term may be connected with different ideas. In addition, the growth in the amount of data makes the search more complex and difficulties arises from indexing processes in which many IR techniques rely on.

A formal concept of IR can be defined as the action of finding material (usually documents) of an unstructured nature (usually text) that satisfies the need for information from large collections of data (usually stored in computers) (MANNING; RHAGAVAN; SCHUTZE, 2009). The publications in the IR area have keeping pace With the growth in the volume of information available. Indeed, as more data is available, the more sophisticated are the techniques needed to handle it.

Since the collections of documents are stored in computers, one of IR main elements is the matrix of terms and documents, important for the fundamental task of indexing for later retrieval of information. Figure 1 depicts an example of this matrix, with the works of Shakespeare. The rows represent all the terms in all documents forming a collection. The columns, in turn, represent the documents themselves. Each element of the array can be filled with different indexes; in the example, the number 1 represents the presence of the term in the document (independent of frequency) and 0 represents the absence of the term in the document.

Collections of documents, i.e. a set of documents containing a set of related terms, are commonly nominated *corpora*.A traditional approach is to treat a single corpus in the bag-of-words format, that is, a set of terms independent of its ordering in the documents, though there is a loss in working with the words independently of its grammar function in sentences (AGGARWAL; ZHAI, 2012).

Through this simple example it is possible to see the main characteristics of Term-Document Matrices, that are usually sparse and high-dimensional. A simple collection of documents can scale quickly given the large number of terms. In order to contribute

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

Figure 1 – Term-Document Matrix (MANNING; RHAGAVAN; SCHUTZE, 2009, p. 4)

to this issue, there are several pre-processing techniques, fundamental for mining large volumes of text.

A first aspect to be highlighted is the meaning of the information present in each a_{ij} of term-document matrices. The Boolean usage of (0,1), although facilitates interpretation, does not allow much richness in the analysis, omitting the frequency information of the terms in the different documents. Thus, an option is to work with the number of times each term appears in each document, also known as term frequency (tf). However, a a problem with this index is its low discretionary power. For instance, the probability that the word "car" appears in almost all documents in a collection of documents about the automobile industry is surely very high. If we already know that the corpus in question is about car, the term frequency index ends up valuing terms that may not be much helpful for further analysis.

An index that allows to overcomes this issue is $tf - idf$, explained in details in the next section. Briefly, it is a composite index that considers the frequency of the term in documents (tf) combined with the inverse number of documents that have the same term (idf). In this way, it values rare terms, which differentiate documents, and, when present, have a significant frequency with respect to the others. In addition, this index devalues terms that occur in many documents, penalizing them even with the value 0 when they occur in virtually all documents. This is the main weighting scheme adopted in this dissertation.

Another aspect to be highlighted are stopwords. This term refers to words that are very frequent but do not add any meaning in the analytical processes. Lists of stopwords can be obtained externally or generated from the database. In English, stopwords lists include expressions such as "that", "a", "no", "the", "when", "how", among others that do not present any specific meaning, having only the function of linking words in a sentence. This procedure is very important since it allows a significant reduction of the term-document matrix dimensionality.

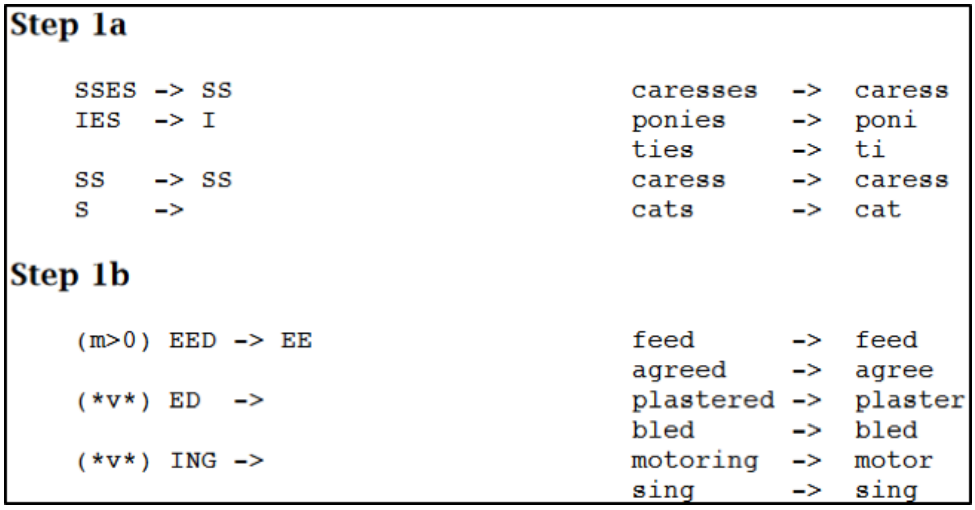

Figure 2 – Porter Stemmer rule sample (adapted from Porter (1980))

Finally, a last aspect to be highlighted is the pre-processing with respect to the suffixes of the words, called stemming. For reasons especially of grammar, the documents use different spellings of words like “manager”, “managing” and “management”, for example, but all with a similar radical, which can denote similar meaning. The process of stemming consists in eliminating the suffix of the words, leaving only the initial characters, so that, in the example above, all could be transformed to "manage". This procedure also acts directly on dimensionality reduction.

The idea of stemming, also known as suffix stripping, goes back to the development of efficient indexing by information science, and Porter (1980) was responsible for unveil and automatize this process through an algorithm. The main impact, and what made this an important publication in the literature, was its simplicity and consequent speed of application. By removing the necessity of a dictionary to remove suffixes, this stemmer demonstrated that the logic of a rule sequence, despite of not having a complete efficiency, achieves good results without needing the support from a list of words, which would require constant update since language is a social, and as such, a very dynamic element.

With a relatively simple function, the Porter stemmer has contributed to highlight the stemming process not only as a step in the IR process, but also as a research field (WILLETT, 2006). A vector composed of vowels and consonants is the main word representation and, with that, a set of rules organized in sequential steps tests whether that vector contains specific combinations of letters, replacing them with another smaller set of letters. Figure 2 shows some of these steps.

Working mostly with collections of documents, the IR field contributes significantly to text mining and thus, with business analytics inside organizations. The origin of structure analysis, however, would not be possible without the contributions of data mining and the related text mining fields, that searches patterns in databases. Still, the IR field is

placed in this dissertation following the understanding that unlike data available within a traditional relational database, text data need to be managed through search engines in a collection, which is the main task of IR.

However, it is not possible to consider text mining as a sub-task of IR. By focusing on locating certain information for a particular user (through specific request, or search query), IR seeks to optimize the search process and retrieve the most relevant information. The area of text mining seeks, through similar techniques, to achieve the same data mining objectives with text: not only to perform the task of relevant information search, but to facilitate the decision-making process by helping to analyze large volumes of text data.

Therefore, having concepts of IR given the unstructured nature of text data, and aiming to extract knowledge just like data mining, text mining can be seen as an area that "go beyond information access to further help users analyze and digest information and facilitate decision making" (AGGARWAL; ZHAI, 2012, p.2). In this sense, text mining aims to contribute with mainly five operations, namely: distance and similarity computing; clustering; dimensionality reduction; classification and topic modeling (KOBAYASHI et al., 2017; AGGARWAL; ZHAI, 2012).

Distance and similarity computing can be a key activity for different tasks such as search and retrieval of relevant documents, given an IR approach, and efficient recommendation, more connected with data mining. The main idea is to compute distance and similarity in vector representations of documents (or sentences, or words) in order to better understand how to explore a corpus in an objectively manner (KOBAYASHI et al., 2017).

For decision-making processes, one important aspect derived from distance and similarity measures is topic coherence. Topic coherence can be seen as a special case of measuring the coherence from a set of statements, applied to models that aim to extract topics from text. This research area emerges from the fact that topic models give no guaranty that its output would be easy to understand (RÖDER; BOTH; HINNEBURG, 2015). The notion of coherence itself have its roots in epistemology and philosophy of science (DOUVEN; MEIJS, 2007), and its structure has inspired developments in many areas, including text analysis, since it can be adapted to better understand the coherence from a set of words (RÖDER; BOTH; HINNEBURG, 2015).

Considering coherence as a notion of "hanging together' (DOUVEN; MEIJS, 2007), the metrics of coherence will try to measure how close and how distant groups of words (i.e., a pair) are positioned considering all the other words. The measures can be divided into three different groups.

First, there are those measures that compare a pair of words with other words in another context. The main metric in this group is PMI (Pointwise Mutual Information),

presented in the study of Newman et al. (2010) as the one more correlated with human judgment, having been used in several studies ever since (STEVENS et al., 2012; RÖDER; BOTH; HINNEBURG, 2015; O'CALLAGHAN et al., 2015). PMI treats an external data source (like Wikipedia) as a single meta-document, and score pairs of top-N words (usually top-10) using term co-ocurrence. PMI of each pair is estimated in that external source, and can be seen as a measure of statistical independence between them. More specifically, given two words, say w_i and w_j,

$$PMI(w_i, w_j) = log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \tag{1.1}$$

where $p(w_i, w_j)$ is the co-occurrence frequency of words w_i and w_j in the external data source and $p(w_i)$ is the frequency of occurrence of word w_i in the external data source, just as $p(w_j)$ is the frequency of occurrence of word w_j in the same database.

Second, there are those measures that compares a pair of words within its own corpus. The main metric in this group is calculated as in Equation 1.2 and was developed by Mimno et al. (2011). Considering other topic modeling methods, it is based on the probability that lower-ranked words in a topic co-occur in documents with higher-ranked words. That can be related to comparing the frequency of low-rank terms in a document with the joint frequency with other high-ranked terms in other documents. Previous studies demonstrates that this measure has higher levels of correlation with human judgment than PMI (MIMNO et al., 2011; O'CALLAGHAN et al., 2015).

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{1.2}$$

In this formulation, $V^{(t)}$ represents a vector $V^{(t)} = (v_1^{(t)}, \dots v_M^{(t)})$ containing a list with the M most likely words given a topic t. In addition, $D(v_l^{(t)})$ represents the number of documents in which the term v_l is present, and $D(v_m^{(t)}, v_l^{(t)})$ is the co-frequency of the terms $v_m^{(t)}$ and $v_l^{(t)}$. Also, the number 1 is placed in order to prevent a logarithm of 0 being evaluated, for computational integrity. It is noteworthy here, according to O'Callaghan et al. (2015), that both measures were developed to work with topics produced by other probabilistic models, having space for similar adaptation and testing with matrix-based models.

At last, a third group encompasses those measures...those measures that compare terms considering their presence on topics, also called topic descriptors, which can be seen as the top-N words from a specific topic, aiming to understand the difference between sets of words. For that, it is possible to work with Jaccard similarity (DEZA; DEZA, 2009), as in Equation 1.3. This measure allows to understand how similar topics are, since it assesses the relationship between the intersection and the union from a set of terms of each topic. In this way, the smaller the distance, the less similar the topics and, consequently, the

more specific in relation to their terms descriptors. The intuition is that pairs of terms that co-occur frequently together or that are close in a vector space are more likely to contribute to the coherence of the topic (O'CALLAGHAN et al., 2015). For that, considering X and Y a set of elements, Jaccard similarity is

$$J = 1 - \frac{|X \cap Y|}{|X \cup Y|} \tag{1.3}$$

O'Callaghan et al. (2015) presents a more formal structure, considering k dimensions and computational implementation, that compares N topic descriptors (TD) with all the other topics, until all k dimensions have been computed two-by-two.

$$MJ = \frac{1}{k} \sum_{j=2}^{k} \sum_{i=1}^{j-1} \frac{|TD_i \cap TD_j|}{|TD_i \cup TD_j|} \tag{1.4}$$

Thus, it is possible to consider TD as the vector containing the terms descriptors of each topic. These are analyzed by pair (i, j) for all k dimensions, resulting in the measure of similarity regarding a set of topics.

Even tough the measures described here are non-exhaustive, they can be seen as complementaries, once the first uses an external database for validation, the second is more concerned with the terms co-occurrence within documents from the corpus, and the latter focus in comparing topic descriptors in k dimensions.

An implementation of Jaccard similarity measure considering term-document matrices was developed in the context of this dissertation, and is presented in Appendix D.

Clustering, in turn, involves organizing documents in similar groups among themselves and different from others. From an IR perspective, the main usage is to facilitate search and retrieval; in a data mining perspective, to automatically classify documents given a certain label (KOBAYASHI et al., 2017). Usually, clustering can be performed following two approaches: hierarchical and partitional. The main difference between those groups of techniques lies in the definition of the number of clusters to be considered. In hierarchical approaches, clusters are defined given a certain rule, and usually rely on dendrograms to present different ways that groups can be formed. In partitional approaches, the number of cluster is specified *a priori* by the analyst.

Some well-known techniques, like k-means or general-purpose hierarchical methods, can be extended to any kind of data, including text. However, as text data have unique properties such as the number of word in each document, high-dimensionality, and consequently less number of principal components than features (words), there is a necessity to explore pre-processing and feature selection previously to algorithm application, as well

as to work with specific algorithms that have the capacity to handle with this special structure (AGGARWAL; ZHAI, 2012).

Dimensionality reduction, text classification and topic modeling are the text mining operations discussed in this dissertation. The next chapter discuss topic modeling and one of its models, LSA, focusing on dimensionality reduction effects. Text classification is also presented in details regarding neural networks. This group of operations compose the proposed framework thus achieving our main objective..

2 Applied Text Mining Operations

In this section, three main text mining operations are discussed: topic modeling and the model LSA, dimensionality reduction, discussed in the LSA context; and text classification, regarding neural networks implementation. Together with clustering and distance and similarity computing, they compose the main set of text mining operations in organizational research (KOBAYASHI et al., 2017).

Topic modeling, a growing research area, works with the main idea that there exist a non-observable structure, behind documents and terms. This structure is capable to better represent the main connexions among text data. This semantic structure idea finds its roots in Zipf law (ZIPF, 1949), describing the inverse relationship among frequency and rank-position. Arguing that words frequencies tables are good descriptors for several different languages, (ZIPF, 1935) concludes that there are similar structures for any set of documents. For that, it is important to acknowledge that the same concept, or idea, can be expressed using several different words, and conversely the same term can have other meanings depending on the context (AGGARWAL; ZHAI, 2012). Working with this two issues (namely synonymy and polysemy) is a core task for finding a hidden structure among documents and terms, topic modeling models main objective.

This became an important process given text data growth. Being able to deliver latent relations that allows to better understand, in less time, a group of documents, is crucial to improve text data usage in organizations, specially in decision-making processes. Two main groups of models have emerged to deal with topic modeling: probabilistic and non-probabilistic. Probabilistic models work with distributions of probable topics among documents given their words. The main models in this area Probabilistic Latent Semantic Indexing (PLSI) (HOFMANN, 2001) and Latent Dirichlet Allocation (LDA) (BLEI; NG; JORDAN, 2003). Non-probabilistic methods work with term weights among documents in order to discover topics, and include models such as Latent Semantic Analysis (LSA) (DEERWESTER et al., 1990) and Non-Negative Matrix Factorization (NMF) (LEE; SEUNG, 2001).

In order to uncover this latent structure beyond a simple term count, and to be effective specially in business environments, one of the main challenges given topic modeling is to deal with an increasing volume of text data. For that, a constant analysis between computational efficiency and information extraction has to be made, in order to keep a balance among this two important elements (KULKARNI; APTE; EVANGELOPOULOS, 2014a). For probabilistic models, it is difficult to deal with a large number of topics since that implies simultaneously updating the Term-Document Matrix to meet the probability

distribution assumption for topics, which can be seen as a scalability challenge. On the other hand, since for many non-probabilistic models it is necessary to calculate important matrix information, and sometimes with orthogonality assumption, some high-dimensional problems might be hard to solve with the timing that some business decision demands (WANG et al., 2012; WANG et al., 2013).

Since both approaches have strengths and weaknesses, in this dissertation LSA is the model explored. Given the development of different methods to deal with dimensionality reduction in the literature and the diversity of forms implemented in open-source tools, there are alternatives to deal with high-dimensional databases. In addition, word embedding models have improved significantly when using LSA mathematical structure in order to incorporate more semantic meaning for word vectors representation (PENNINGTON; SOCHER; MANNING, 2014), demonstrating that LSA can not only deliver a reliable document latent structure but also have still room to contribute with different tasks in the deep learning movement (MANNING, 2015).

Giving this context, LSA is one of the techniques developed in response to the different needs of the IR area, and more recently, have been supporting text analytics activities (VISINESCU; EVANGELOPOULOS, 2014). When proposed by Deerwester et al. (1990), its main objective was to face the synonymy and polysemy challenge, related to working with texts in unstructured format. The authors looked for a tool that recovered more relevant documents, focusing on difficulties regarding compatibility among terms.

Its purpose was to address the fact that it is not possible to use only term frequency or raw data for text indexing. Since text data is produced directly from users (thus, it is straightforward), it is important to considerate that there are different ways of communicating a single concept. Therefore, one user might be expressing the same idea of another, although using different terms. On the other hand, two users may choose the same word to express different opinions. This two main problems, namely synonymy and polysemy, were LSA main concern (DEERWESTER et al., 1990).

In order to connect ideas (or topics) besides differences among words, LSA uses singular-value decomposition (SVD). This mathematical structure allows to discover a latent semantic structure hidden between terms presented in a set of documents. SVD is a decomposition solution to deal with non-square matrices, that is, matrices that have a greater number of rows than columns, or vice versa. The Term-Document Matrices are generally non-square, as there will hardly be a same amount of term and documents in a corpus. This decomposition is based on vector space models, an application of linear algebra. These models arose from the acknowledge of the limitation of the boolean model (only with 0 and 1), that tended to simplify too much the information from a set of documents, only considering the presence or absence of a term in a document. From this vector space approach, term weights and the representation of documents as vectors in a

space were possible, allowing the application of concepts such as measures, distances and similarities between documents (BAEZA-YATES; RIBEIRO-NETO et al., 2011).

The LSA model works with a particular application of vector space models to create a semantic space. The input to create this space is the Term-Document Matrix. Thus, a corpus (i.e., a set of documents in a *bag-of-words* representation) containing n documents and m terms can be represented by a matrix X, of order m x n. After X is created, it is possible to represent its terms and documents in a vector space through orthogonal decomposition that will form other three matrices, U, Σ and V. Orthogonal transformations can maintain the properties of the original matrix, like the length and distance of X rows and columns vectors (MARTIN; BERRY, 2011). In order to better understand this structure, it is important to describe the importance of orthogonality. An orthogonal matrix, resulting from a decomposition or transformation, has the fundamental property $Q^tQ = L$, where Q is the orthogonal matrix, Q^T is the transpose of Q, and L is a diagonal matrix

$$L = \begin{bmatrix} q_{11} & 0 & \dots & 0 & 0 \\ 0 & q_{22} & \dots & 0 & 0 \\ \vdots & \vdots & q_{33} & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & 0 & q_{nn} \end{bmatrix}$$

Thus, the n column vectors that form Q, which can be represented by $[q_1, q_2, \dots, q_n]$, are orthogonal, since for any pair (q_i, q_j), we have:

$$\begin{cases} q_i^T q_j = 0, & i \neq j \\ q_i^T q_j \neq 0, & i = j \end{cases}$$

Being orthogonal, these vectors (i.e., the columns of the matrix Q) are positioned in different directions, and form angles of 90° to each other. In this way, the vectors $[q_1, q_2, \dots, q_n]$ form a linearly independent set, and therefore serve as a basis for a vector space, being able to form any other vector, in this same space, from a linear combination of its terms.

One advantage of this structure is to be able to organize and relate documents not by appearance or absence of a group of words, but by connecting them in a topic, represented by left and right eigenvectors (U and V matrices) corresponding to the singular values of X forming the diagonal matrix Σ. This can lead to an important aspect when dealing with high-volume data: dimension reduction, or rank lowering. Dimension reduction main objective is to reduce noise in vector space (in LSA context, latent semantic space), by retaining the main dimensions, which are related with the highest singular values.

This can lead to a richer relationship structure that reveals latent relations presented between documents and terms (BERGAMASCHI; PO, 2014). The U, Σ and V matrices are truncated to k dimensions and the X matrix can be approximated (Equation 2.1).

$$X \approx \hat{X} = U_{mxk}\Sigma_{kxk}V^T_{kxn} \tag{2.1}$$

Through this decomposition process, it is possible to obtain three other matrices, U, Σ and V. Σ represents the singular values of X, being the square root of the eigenvalues of XX^T and X^TX, both square matrices. Also, U and V represents the eigenvectors associated with each of these eigenvalues, also called the right and left eigenvectors of X. In addition, the r columns of U are the r orthonormal eigenvectors associated with the r nonzero eigenvalues of XX^T and in the same way the r columns of V (rows of V^T) are r orthonormal eigenvectors associated with the r nonzero eigenvalues of X^TX. Usually the nonzero singular values of the matrix X are represented by the greek letters $\lambda_1, \lambda_2, \ldots, \lambda_r$. Without loss of generality, it is possible to assume that the singular values are put in descending order, being $\lambda_1 > \ldots > \lambda_r$ (CRAIN et al., 2012; MARTIN; BERRY, 2011). In general, a great portion of the singular values are very small, near zero, and thus one can restrict the decomposition using only the first k singular values, as in Equation 2.1.

Altogether, it is possible to relate the documents in V with the terms in U by the k dimensions retained during the decomposition, which in turn are the k topics from the corpus analyzed. Therefore, the goal is to obtain, from the X Term-Document Matrix, a set of linearly independent vectors, which form the basis of that set. In this way, it is possible to discover the latent semantic structure, hidden between the documents and the terms that compose the corpus analyzed. Figure 3 presents the decomposition structure.

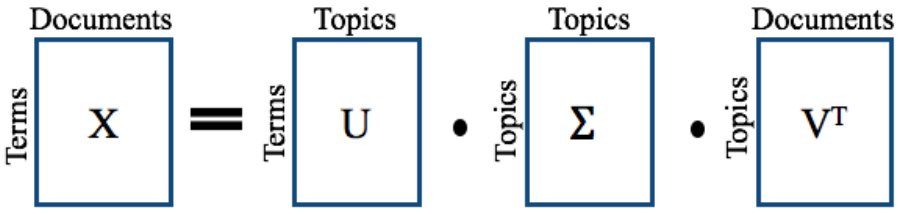

Figure 3 – SVD within LSA context (ASHTON; EVANGELOPOULOS; PRYBUTOK, 2014, p. 2281)

Another key point for LSA structure is the choice of the value of k. Term-Document Matrices are usually sparse, and an optimal k can allow to work with dimension reduction, as in Equation 2.1. The main objective is to reduce noise in latent semantic space, retaining the main dimensions that are related with the highest singular values. This can lead to a richer relationship structure that reveals latent relations presented between documents and terms (BERGAMASCHI; PO, 2014). However, the optimal k is still a challenge.

Different authors have proposed a set of solutions (WILD et al., 2005; KULKARNI; APTE; EVANGELOPOULOS, 2014a; BERGAMASCHI; PO, 2014), most of them based in a percentage or ratio from singular values to be kept during the decomposition process. By reducing dimensionality, one needs to choose to remove from the analysis a series of terms and documents that are more connected to less representative singular values. The expression "truncated SVD" refers to this point. In Figure 4 is possible to see that the original size of the matrices U, Σ and V are all related to the r non-zero singular values, but for later analysis only the first k are retained thus reducing the dimension of U, Σ and V, obtaining an approximation of X, as in Equation 2.1.

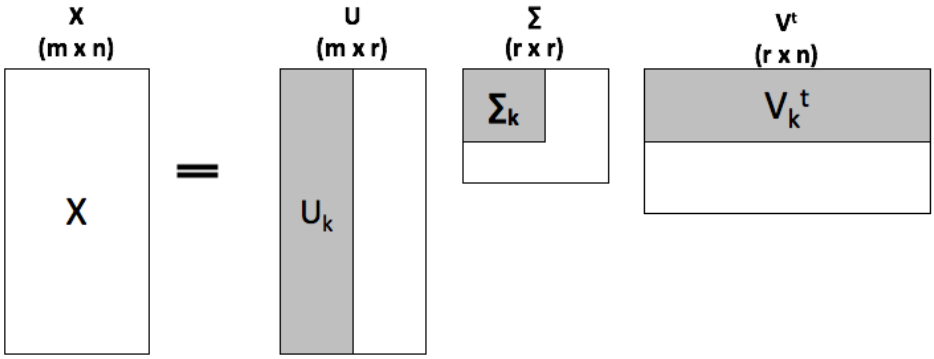

Figure 4 – Truncated SVD (MARTIN; BERRY, 2011, p. 41)

A simple visual demonstration allows to understand the dimension reduction effect. For that, an X Term-Document matrix was built, with boolean operators, from 10 documents (BECKER, 2016). The left of Figure 5 presents a heat map of X, that have the original information distribution, with each a_{ij} highlighted when a certain document contains a specific term. Subsequently, it is presented the recomposition result, after retaining 3 ($k = 3$) and 5 ($k = 5$) highest singular values and multiplying matrices U, Σ and V^T. It is possible to see that when $k = 5$, representing 50% from all singular values, most part of the information is preserved, and the original matrix can be approximated. For high-dimensional data application, this is an important element, bringing confidence that LSA will keep the core latent structure, capturing most part of the information contained in the dataset, even though neglecting some dimensions of the representation matrices.

Wild et al. (2005) developed and tested four different options for dimension choice, discussing pros and cons among them. The first was called share, and represented the percentage of cumulated singular values, calculated using a normalized vector that is used to sum up singular values until a pre-defined threshold. They suggested values between 30% and 50%. Another one was calculated considering that the absolute value of cumulated singular values should be equal to the number of documents. In other words, the sum of the first k singular values have to be equal to n, the number of documents in the original

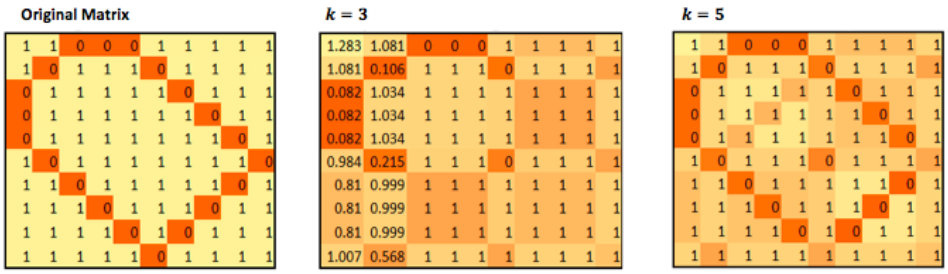

Figure 5 – Different k values effect (BECKER, 2016)

Term-Document matrix. Next, a fraction of all terms was also tested, with values suggested between 1/30 and 1/50. Less sophisticated and last approach is to use just a fixed amount of dimensions. This option can be useful when working with previously known or desired number of topics.

Kulkarni, Apte e Evangelopoulos (2014a) brought a log-likelihood ratio (LLR) test approach, that seeks to quantitatively estimate an elbow point given the eigenvalues plot. This can also be seen as the point where adding another dimension will bring little marginal contribution in the ability to explain the total variance in the X matrix. This can be done with a bootstrap-based simulation from different sets of eigenvalues, although the authors stated that most of the time an elbow point is easily visualized seeking through a plot of singular values. Notwithstanding, many of them agree with Bergamaschi e Po (2014), referring that this point should be defined empirically for each collection. In this dissertation, the approach followed Bergamaschi e Po (2014) and Kulkarni, Apte e Evangelopoulos (2014a), as the k value was defined based in the elbow point through singular values scree plot.

There are also different weighting schemes that can be adopted when building the Term-Document matrix. Some of the most well-known are boolean operators, term-frequency (*tf*) and inverse document frequency times term frequency (*tf-idf*). Boolean operators, as seen before, might not be suitable for several business purposes since it only works with presence or absence of a word in a given document (0 for absence and 1 for presence). Term-frequency, on the other hand, provides more information by assigning for each a_{ij} entry a weight that is equal to the number of occurrences of the term t in a document d (MANNING; RHAGAVAN; SCHUTZE, 2009).

The weighting scheme adopted in this work is the index most commonly used in text analytics area, *tf-idf* (Equation 2.2). The first part is *tf*, that relates the frequency of a term ($\text{freq}_{t,d}$) compared with the high-frequency term in the same document ($\max_{t,d}$). The second part is the inverse of document frequency (*idf*), that is a relation between all N documents and those documents that have a given term (n_t). In this way, $w_{t,d}$ represents the entry a_{ij} in a Term-Document Matrix, valuing rare terms, which have the power

to differentiate documents, and that, when present, have a significant higher frequency compared to the other terms in the same document. In addition, this index devalues terms that occur in many documents, penalizing them even with the value 0 when they are present in virtually all documents (CRAIN et al., 2012; MANNING; RHAGAVAN; SCHUTZE, 2009).

$$w_{t,d} = (tf_{t,d})(idf_t), \tag{2.2}$$

where

$$tf_{t,d} = \frac{\text{freq}_{t,d}}{\text{max}_{t,d}} \tag{2.3}$$

and

$$idf_t = \log_2 \frac{N}{n_t} \tag{2.4}$$

Beyond computer science and statistics research about LSA mechanics and implementation, there are applications in areas like business and education. For example, Thorleuchter e Poel (2012) demonstrated the relation between textual information in e-commerce websites and companies performance, and another research applied LSA to uncover main skills from a set of job offers (O'LEARY et al., 2002). Considering educational studies, some topics that rely on LSA for investigation are essay grading, student development and distance learning suport (OLMOS et al., 2016; WILD et al., 2005; TINKLER; WOODS, 2013; EVANGELOPOULOS, 2011; WIEMER-HASTINGS; WIEMER-HASTINGS; GRAESSER, 1999)

Although the model has been contributing in the literature for more than two decades, the LSA approach still presents barriers to large-scale adoption given its complexity, lacking support from for well-known tools like SPSS and Microsoft Excel (TANEV; LIOTTA; KLEISMANTAS, 2015). There is therefore room for uncovering and improvement of the model's adoption. In addition, its ability to determine the topics of a large set of documents poses LSA as a potential contributor to the decision making process by quantifying information in text objectively and allowing the use of this data, sometimes neglected given the analytical challenge it presents (ASHTON; EVANGELOPOULOS; PRYBUTOK, 2014).

All computations presented in this dissertation were done using R through RStudio. R is an open-source software that allows working with a high diversity of tools, methods and techniques, being already considered one of the main tools related with analytics (ZHAO, 2013). In order to work with LSA, a suite of packages were used, with highlight to lsa (WILD, 2015) and tm (text mining) (FEINERER, 2017) packages.

Mainly, lsa package supports all forms of weighting schemes implemented through tm package, as well as provides different forms to define the number of dimensions to keep.

In addition, it provides a fold in option, dedicated to add documents into an existing latent semantic space without having to re-calculate it. Figure 6 presents the package workflow.

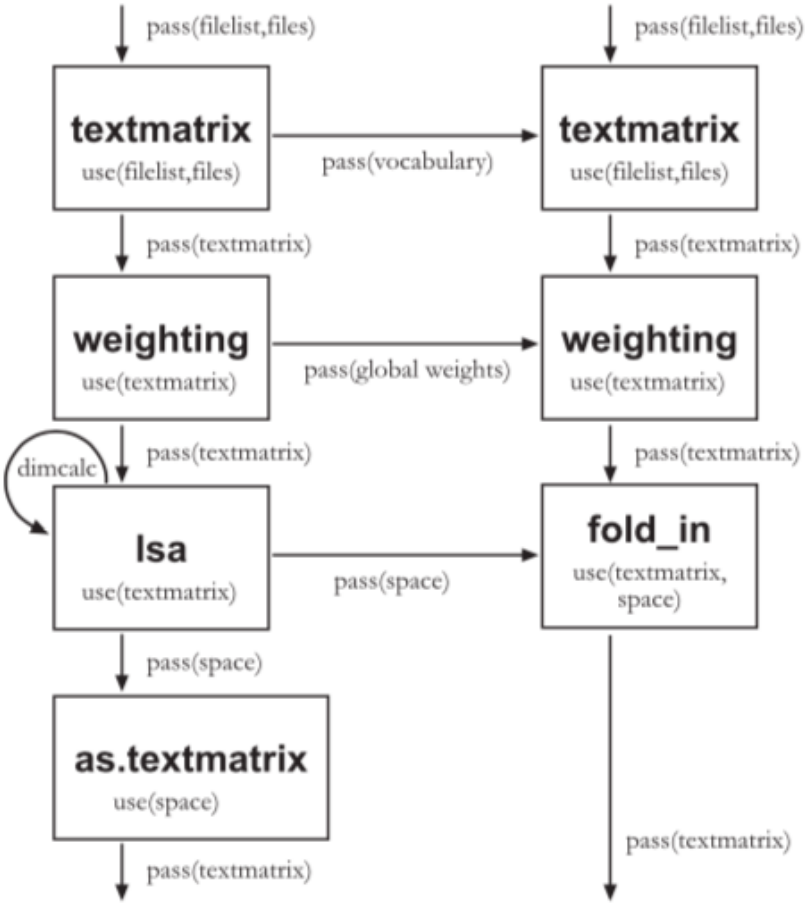

Figure 6 – lsa package workflow (WILD, 2016, p. 79)

The different options given k dimensions settings in R are present in the parameter *dimcalc*, expression from "Dimensionality Calculation", as presented in Figure 7 with package default values. All options were previously validated in the literature (WILD et al., 2005; WILD, 2016). The first option is the share, which works with a fraction of sum the of the singular values to be retained in relation to the sum of all singular values. That is, a numerical proportion that allows to retain the highest singular values in relation to the others. The second option is ndocs. In this, the singular values are sorted in descending order, and those whose sum reaches or exceeds the number of documents in the corpus are returned. Thus, we have the first k singular values with sum greater than or equal to n documents. The third option is kaiser, which calculates the number of singular values from the Kaiser criterion, returning only the singular values greater than 1. Finally, one can not fail to mention the fourth and last option, the fraction, which works with a certain fraction of all the indexed terms, returning their singular values. The option raw, as specified in this package, is present only to complete the options, since it is the one that returns all

singular values.

```
dimcalc_share(share=0.5)
dimcalc_ndocs(ndocs)
dimcalc_kaiser()
dimcalc_raw()
dimcalc_fraction(frac=(1/50))
```

Figure 7 – dimcalc parameter options and default values (WILD, 2015, p. 7)

Additionally to the options provided with the package, and in order to facilitate output visualization in RStudio environment, a parallel function was developed specifically to contribute with the package. Among others available as this dissertation outputs (MARCOLIN, 2017), the function "ordered.lsa" was built focusing in the analyst work while using RStudio, delivering a prompt visualization of the U matrix with all terms from each topic in descending order, indicating the top-terms, with its corresponding weight, for each factor. This function is detailed in Appendix B.

Another important technique that is implemented in the proposed framework is text classification, a growing research area. Specially since the increasing volume of digital content, to classify data in different categories (such as 'complaints', 'compliments', 'staff issues', for example) have the power to transform raw data into input knowledge for decision-making in organizations pursuing improved performance (CHOI; LEE, 2017). Altogether, this area has facing a challenge to change its mainstream, in order to go from 'bag-of-words' to 'bag-of-concepts', that is, to go beyond word frequency and input an approximation of human language interpretation models to the algorithm (CAMBRIA; WHITE, 2014). Contributing to this movement, in this dissertation we applied neural networks as the supervised technique to perform text classification.

Besides the use regarding text classification, neural networks are a mathematical structure that lead the Deep Learning movement, mainly related with finding patters in unstructured data as images and videos (LECUN; BENGIO; HINTON, 2015) and text (MANNING, 2015). In addition, neural networks are also relevant in OR discipline (KRAUSS; DO; HUCK, 2017).

Neural networks can be considered one of the techniques that allows to go beyond syntactic representation, getting closer to human process of understanding a given text data (CAMBRIA; WHITE, 2014). This is possible since neural networks have the capacity to deal with high dimensionality, finding patterns that allows to separate data (documents) according to its multiple features (words, sentences, paragraphs, etc.).

Neural networks have successfully been used to text classification tasks (KIM et al., 2017; JACOBS et al., 2017). The approach chosen was to work with feed-forward neural

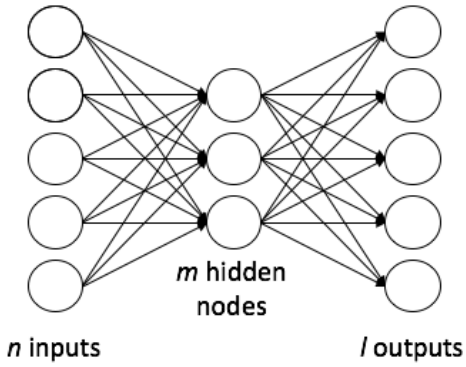

Figure 8 – General Representation from a Feed-Forward Neural Network with single-hidden layer

networks with a hidden layer as a bottleneck (MANEVITZ; YOUSEF, 2007). Specifically, having m hidden layers and n inputs, with $m < n$.

Feed-forward neural networks provide a general framework for representing non-linear functional mappings between a set of input and output variables, as in Figure 8. This flexibility is important since it allows to represent a non-linear function of many variables as a composition of non-linear functions of single variables, a structure that is called the activation function (BISHOP, 1995). In addition, it is also possible, in the same network, to transform the output variables in order to improve interpretation, which is done with a different activation function in the output layer.

In this work, a Multilayer Perceptron Network was used, being possible to represent its structure regarding the input and the output layer, connecting them with the single hidden layer, although it its possible to have several hidden layers and any number of hidden neurons in each layer (SAMARASINGHE, 2016). The neural network transforms values associated with input neurons into input values associated with hidden neurons by linear transformations with weights defined for each combination of input and hidden neurons plus a bias weight (feed-forward networks). More specifically, if x_i represents the value associated with the ith input neuron, $i = 1, \ldots, n$, if a_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$, represents the weight associated with the ith input neuron and the jth hidden neuron, and if a_{0j}, $j = 1, \ldots, m$, represents the bias weight associated with the jth hidden neuron, then the input value associated with the jth hidden neuron, u_j, $j = 1, \ldots m$, is given by Equation 2.5.

$$u_j = a_{0j} + \sum_{i=1}^{n} a_{ij} x_i \tag{2.5}$$

The input value u_j, $j = 1, \ldots, m$, is then transformed inside the jth hidden neuron by an activation function (say, f), usually non-linear, into output values y_j, $j = 1, \ldots, m$, of the

jth hidden neuron, as in Equation 2.6

$$y_j = f(u_j) \tag{2.6}$$

The logistic sigmoid function is often used for the hidden output units in multilayer networks. However, as sigmoid functions are limited in the $(0, 1)$ interval, hyperbolic tangent, or 'tanh' activation functions, might present computational advantages, as its range is $(-1, 1)$ (BISHOP, 1995). In this work, 'tanh' functions are used, having the form as in Eq. 2.7.

$$f(u) = tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}} \tag{2.7}$$

After transforming values associated with input neurons into output values associated with hidden neurons, the feed forward neural network goes on and transforms output values associated with hidden neurons into input values associated with output neurons (or the next layer of hidden neurons, if the network has multiple layers) using once more linear transformations with weights defined for each combination of hidden and output neurons plus a bias weight. More specifically, if y_j represents the output value associated with the jth hidden neuron, $j = 1, \ldots, m$, if b_{jk}, $j = 1, \ldots, m$, $k = 1, \ldots, l$, represents the weight associated with the jth hidden neuron and the kth output neuron, and if b_{0k}, $k = 1, \ldots, l$, represents the bias weight associated with the kth output neuron, then the input value associated with the kth output neuron, v_k, $k = 1, \ldots l$, is given by Equation 2.8.

$$v_k = b_{0k} + \sum_{j=1}^{m} b_{jk} y_j \tag{2.8}$$

The input value v_k, $k = 1, \ldots, l$, is then transformed inside the kth output neuron by an activation function (say, g), usually non-linear, into output values z_k, $k = 1, \ldots, l$, of the kth output neuron, as in Equation 2.9.

$$z_k = g(v_k) \tag{2.9}$$

In order to improve interpretation in the output nodes, in this work we use the *softmax* function, as in Eq. 2.10 (BISHOP, 1995).

$$g(v_l) = \frac{exp(v_l)}{\sum_{k=1}^{l} exp(v_l)} \tag{2.10}$$

Finally, towards a more complete picture, sentiment analysis was also implemented. Sentiment analysis, also known as opinion mining, is a growing research area, and has as its main input online reviews from customers (AGUWA; OLYA; MONPLAISIR, 2017).

For that, a lexical approach was chosen due to the popularity and good performance of the models (VALDIVIA; LUZÓN; HERRERA, 2017; DUAN et al., 2016). In addition, lexical approaches are computational efficient and easy to implement, indicated to short text with consistent sentiment words (CHOI; LEE, 2017), as is the case with user reviews. Aiming to improve text use in decision-making process, those are suitable characteristics.

3 Text Data: Voice of the Customer

This section is presented for the sake of completeness, since VOC is here used with the special purpose of validating the proposed framework, which is to be seen as much more general. Toward a framework that allows text data usage for managers decision-making process, a key text data source was choose for validation: customer feedback, also referred as "Voice of the Customer" (VOC) (ASHTON; EVANGELOPOULOS; PRYBUTOK, 2014; SPANGLER; KREULEN, 2007)..

Understanding a customer perspective is vital for business growth. Many organizations, however, still rely on sales pipelines, shopping metrics, or website traffic data for decision-making (GORRY; WESTBROOK, 2011). These kinds of data usually are structured and already fine-tuned for business analytics, as they were originally constructed for tracking customer behavior. Although such data have the potential to reveal insights to executives, truly understanding of how customers feel can only be established with the help of qualitative data, such as through self-provided texts (ZHAO, 2013).

Textual data are available from multiple sources inside and outside the organization. Spangler e Kreulen (2007) identified five different contexts from which text data emerge within the organizational environment, see Figure 9. We can rely on text sources to analyze what is happening inside the organization ("Voice of the Employee", VOE) and between the organizations' front-office and the customer ("Customer Interaction", CI). Outside of the organization, it is also possible to understand what the customers are talking about ("Voice of the Customer", VOC), to keep up with the communications between the organization and its partners ("Business Partners Interaction", BPI), as well as to capture data in order to predict market trends ("Mining to see the Future", MF). All five areas establish a conversational context, in which text data naturally emerges, for example, in the form of email communication or on-line feedback ratings.

Formally, VOC is a set of customers' perspectives about a specific product or service aspect, where each has a priority assignment, indicating the customer position about that aspect (GRIFFIN; HAUSER, 1993). Parasuraman, Zeithaml e Berry (1988) made one of the first efforts to measure quality in services from the perspective of the customer. The authors demonstrated that to listen to customers and understand their perspective is an appropriate approach for assessing quality, given unique characteristics of service – like intangibility and inseparability of production and consumption. To quantify service quality, the authors propose a five-category questionnaire called SERVQUAL. The five categories are: *tangibles*, related to physical facilities like equipment and appearance; *reliability*, which represent the ability to perform the promised service dependably and accurately;

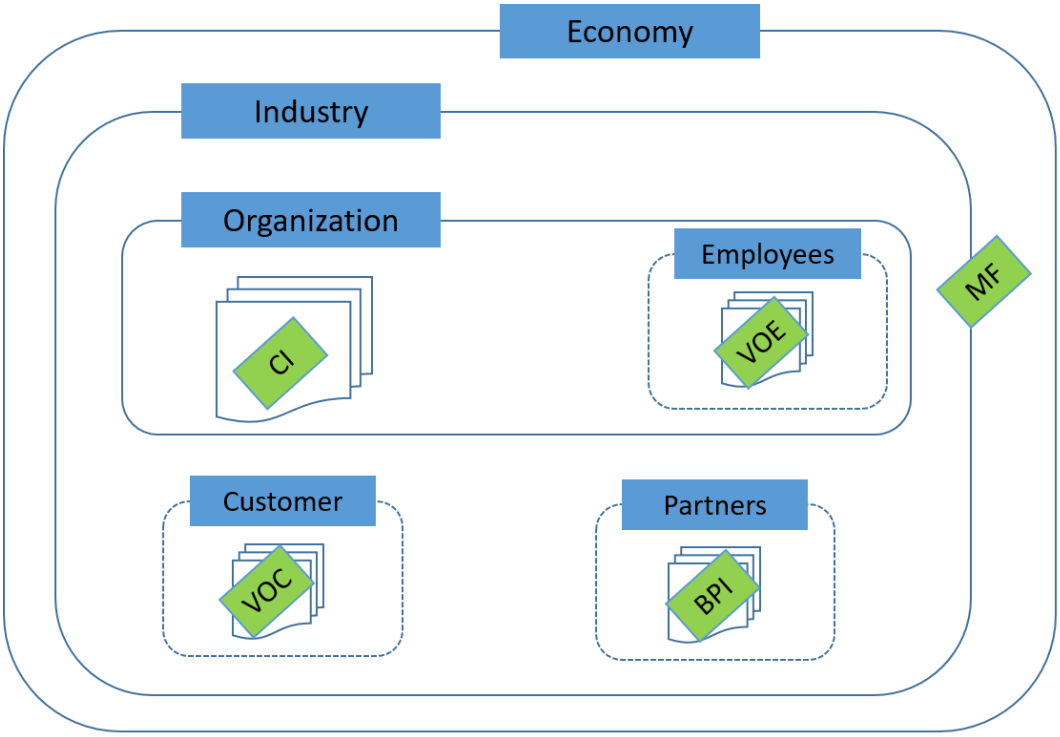

Figure 9 – Text data in organization environment (Adapted from Spangler e Kreulen (2007))

responsiveness, connected with the perception of willingness to help customers and provide prompt service; *assurance*, representing the knowledge and courtesy of employees and their ability to inspire trust and confidence; and *empathy*, which represents the perception of caring and individualized attention for the customer (PARASURAMAN; ZEITHAML; BERRY, 1988).

Similarly, Griffin e Hauser (1993) were one of the first to use the term "Voice of the Customer", also referring to authentic customer text in the context of determining quality. Together with Knutson et al. (1990), who confirmed that the SERVQUAL scale was also valid and reliable to measure quality perception in the lodging industry, those seminal works provide already evidence that utilizing the customer point-of-view in the formulation of strategy has a positive return for the service industries.

Table 1 – Seminal Articles Timeline

Publication	**Title**	**Description**
Parasuraman, Zeithaml and Berry (1988)	*SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality*	Introduces SERVQUAL and validates its categories
Knuston et al. (1990)	*Lodgserv: A Service Quality Index for the Lodging Industry*	Confirms the five generic categories of SERVQUAL in the hotel experience
Griffin and Houser (1993)	*The Voice of the Customer*	Introduces VOC concept

Although questionnaires like SERVQUAL can be very useful to receive and analyze customer feedback, one can argue that analyzing data provided by the customer in unstructured form would be more suitable for hearing the unfiltered VOC (AGUWA; OLYA; MONPLAISIR, 2017). To work with text data imposes some challenges, nonetheless. Though a variety of platforms exist where customers express, in their own words, thoughts, opinions, and their affective attitude, still most business, especially large ones, analyze their customers through sales reports and feedback summaries (GORRY; WESTBROOK, 2011). Reasons for this can be found in the difficulty to combine results from text analysis with findings from numerical data analysis, given the challenge to work with mixed data formats. Another issue may be identified in the level of precision (AGUWA; OLYA; MONPLAISIR, 2017).

Although some differences among this concept can be found in the literature (LEE; HAN; SUH, 2014; MADDULAPALLI; YANG; XU, 2012), we considered "Voice of the Customer" any form of raw text data provided by the customer directly. We choose to work with this kind data due to two main reasons: its exponential growth (VALDIVIA; LUZÓN; HERRERA, 2017; CHEN; ZHANG, 2014; AGGARWAL; ZHAI, 2012) and the potential to become more used in decision-making process (TANEV; LIOTTA; KLEISMANTAS, 2015; ASHTON; EVANGELOPOULOS; PRYBUTOK, 2014; CODY et al., 2002).

In order to listen to VOC, different business have a powerful and open source: the Electronic Word-of-Mouth, or eWOM, also know as consumer e-recommendation. With growing adoption of tools and platforms that allow customers to easily share opinions about previous experiences, eWOM is being increasingly used as a decision tool for service choice (TANG; GUO, 2015; LEUNG et al., 2013). Furthermore, the social media data provided in a diversity of platforms by customers can be seen not only as input for decision about hiring any given service, but also for decision-making processes for organization operations and management (LEUNG et al., 2013).

Consumer perceptions are a major concern in different sectors. Since platforms that facilitate experience sharing have become more and more popular, there are more consumers wiling to rely in eWOM as an important step before a decision (SPARKS; BROWNING, 2011), since there are eWOM platforms to compare service providers from several areas, from food delivery to hotels. As this information is available in text format, effective ways to analyze and transform it into valuable and useful resource is one of the challenges that connects business sectors to text mining issues (TANG; GUO, 2015), since that eWOM provides genuine self-provided consumer information.

The importance of eWOM to management can be related to performance. Social media rating is a significant predictor to explain organizations performance metrics like used capacity and revenue (KIM et al., 2017). Moreover, eWOM is related with impacts in customer retention aspects like loyalty, since it facilitates an online reputation comparison

(CANTALLOPS; SALVI, 2014). In addition, there is evidence of positive correlation between hiring intentions and the valence of the reviews (SCHUCKERT; LIU; LAW, 2015).

Besides, this data is considered trustworthy for customers. eWOM specialized platforms were shown to influence decision towards service contract, being considered more important than editorial and marketing channels (LADHARI; MICHAUD, 2015; DICKINGER, 2011). One main reason is that customers tend to trust more their peers than experts opinion or advertising and, additionally, one main motivation to share opinions is to enable others to make a good decision (LADHARI; MICHAUD, 2015; CANTALLOPS; SALVI, 2014).

Part II

Method

4 Design Science Approach

The framework proposed in this dissertation, designed to allow the use of text data in managerial decision making processes, arises from de perceived lack of approaches that make that connection possible, from data capturing to market analysis. To develop such a structure, several steps had to be followed, such as the identification of a necessity as well as the deployment and test of the framework. For that reasons, the intellectual process of making this dissertation was supported by the design science approach.

Design science concept was first discussed in Simon (1996), that criticized natural science methods so far adopted for applied problem research and solution proposal. In his seminal work, 'The Sciences of the Artificial', the author proposed that in order to develop relevant artifacts it was important to consider different approaches, since that some research focus needs to be a new solution for a specific problem. The main argument is the fact that "natural sciences are concerned with how things are" (SIMON, 1996, p.114), and being so, are not suitable to build and evaluate a to-be-developed solution for any applied problem.

Given this, it could be argued that natural science may not always contribute to diminish the distance between intellectual reasoning and applied solutions, also called in this context as artifacts (SIMON, 1996). The artifact, in this dissertation proposed as the framework, aims to solve an applied question, and for that it can be better built relying in design science structure, more focused on the solution than in the problem study *per se*.

Design can be seen as a concept related with how things should be, given any real-life situation, and alternatives can rarely be searched for; rather, they should be developed considering the problems characteristics and the research objective (DRESCH; LACERDA; JR, 2014), and Simon's initial discussion was important to motivate the design science development considering artifact construction in different research fields. Since than, it has been supporting several application in areas like engineering, health sciences and architecture, having the potential to contribute with business administration given its naturally applied focus in organizations and management problems. In this sense, design science incorporates pragmatic research characteristics such as problem-focus, consequence-oriented and practical question answering (SORDI; AZEVEDO; MEIRELES, 2015). Table 2 presents the literature timeline contribution for this concept.

Table 2 – Design Science timeline (adapted from Dresch, Lacerda e Jr (2014))

Author	**Proposition**
Simon (1996)	Criticizes the exclusive use of the analytical or reductionist method. Argues that the project of knowledge is more important than the object of knowledge. Proposes the use of design sciences.
Takeda et al. (1990)	Make a first attempt to formalize a research method based on the concepts of design.
Nunamaker et al. (1991)	Seek to formalize a method for research based on design science. Exhibit some research products backed by design science. Advocate the use of design science concepts for conducting research.
Walls, Wyidmeyer e Sawy (1992)	Address the concept of prescriptive theories and their importance for the development of practical and effective solutions to existing problems.
Le Moigne (1994)	Discuss on the new sciences, focused on the conception and not only the analysis of the research object.
Van Aken (2004, 2005, 2011)	States that research carried out should be prescriptive, facilitating its use by organizations, and also generalizable - not serve to solve only a problem in a given situation, but to solve a certain class of problems.

Concerned with an artifact development and evaluation, design science is not related with universal or natural laws that are capable to describe the study objects behavior. Instead, the main concern is with the cognitive process whereby the project was conducted (DRESCH; LACERDA; JR, 2014). Above all, design science seeks to develop solutions for existing problems through artifacts.

However, recognizing that organizations might have too specific problems that could difficult any kind of generalization, design science solutions should be built for a certain group of problems. In this sense, considering managerial environments, it is possible to contribute within a certain field, given validation process, but also to infer contributions to other fields with the same or similar artifact.

imply other fields contribution with the same artifact. Thus, it should be highlighted the pragmatic validity of this artifact, that aims to keep usefulness aspect in sight, ensuring the contribution for the problem to be solved (AKEN, 2005) besides other aspects like cost-benefit and managers needs.

In order to conduct the framework development in a design science approach, a design science research need to be conducted, and for that seven main guidelines should be observed (DRESCH; LACERDA; JR, 2014; HEVNER et al., 2004). The actual development of an artifact that can be presented in different formats like methods, models or frameworks and that is technology-based and business-related composes the first and second proposed guidelines. Likewise, the third guideline concerns about an evaluation step that should allow to measure its usefulness, and providing clear and verifiable contributions to the field, as guideline four. Towards validity assurance, the guideline five proposes that research should be conducted with rigor, being supported by problem and desired outputs knowledge, as guideline six states. Finally, guideline seven argues that the artifact should be presented to all stakeholders involved.

The proposed framework presented in this dissertation is developed based in technologies that allows to analyze text data, regarding decision-making inside organizations, in compliance with guidelines one and two. For guidelines three and five, the detailed description and validation processes presented in Results sections aims to clarify the rigor with what the research was conducted, as well as present framework validation given an instance dataset. In order to satisfy guidelines four and six, a section presenting managerial implications described the main knowledge extracted from the collected text data. Finally, for guideline seven the dissertation main findings and contributions are publicized in academia and with managers.

The development of artifacts, according to design science principles, can be seen as one path to respond to the recurring criticisms about the quality of scientific production, sometimes pointed as very fragmented and therefore difficult to apply to concrete problems of society (AKEN, 2005). The design science approach has been intensively used in countries

with innovative capacity, presenting itself as an interesting format for applied research development. The possibility that the adoption of this approach can foster the production of knowledge directed to practical questions, more easily applied to organizational problems (SORDI; AZEVEDO; MEIRELES, 2015) is a main motivation given this dissertation development.

Part III

Results

5 The Framework

This dissertation objective is to develop a framework to connect text data anaysis decision-making processes. The framework is proposed as an artifact, considering design science approach, that allows to capture, treat and analyze text data. In order to accomplish that, a set of phases are presented in sequence, demonstrating the main related aspects. This section closes with the complete framework design.

As an artifact, the pragmatic validity is closely related with usefulness. In this sense, aiming to make the contributions of each phase more clear by itself and, consequently, given the whole framework, an example instance with hotel comments from a eWOM platform (TripAdvisor) will follow the description. The goal is to facilitate the understanding of its usefulness, though the framework supports any other (internal or external) data source.

5.1 Webscrapping

Figure 10 – WebScrapping Process

Reviews from a website can be collected through a WebScrapper, detailed in Appendix A. In the example used to validate the framework, with this tool it was possible to build an automated routine that collected complete comments from an eWOM platform, as in Figure 10. For that, the source webpage XML or HTML code have to be considered, since it is necessary to inform in each part (more specifically, the nodes where) the comments are stored.

Using as an example the TripAdvisor platform, the Webscrapper works receiving a list of input links, that might be from the organization and its main competitors or all other

organization's links available in any given eWOM. With this list, the following process is repeated for every link: the Webscrapper find the nodes containing text information (in this case, comments, stars and date); captures this data; stores in an integrated database; and jump to the next comment page. The broadness of the analysis can be, them, flexible, since the manager can choose how many other organizations, beside its own, will compose the resulting database.

Being open to external data sources might reveal potential to be explored by managers, specially inside those companies that rely solely on internal databases to perform different analytical tasks (BARTON; COURT, 2012), like customer position, explored in the next chapter.

This development allows to overcome some important issues given the data high-volume organizational scenario, specially related with the cost of capturing this data (CHEN; ZHANG, 2014), since its developed in an open-source platform publicly available. This is also convergent with managers view, that expect applications in low-cost technologies that at the same time have the capacity to deal with external data sources (DAVENPORT; DYCHÉ, 2013) helping to move from single to multiple data types without loosing cost-benefit equilibrium. Therefore, for this framework purpose, the Webscrapper was developed in R (WICKHAM, 2015; MARCOLIN, 2017).

In addition, the Webscrapper also contributes by automatizing not only data capture but also the process structure. By having as only input a list of links, it can overcame previous tasks that traditionally would follow any text search like keyword definition and even API (Application Programming Interface) setting (TANEV; LIOTTA; KLEISMANTAS, 2015).

5.2 Pre-Processing

Text pre-processing refers to a data cleaning process, in order to enhance quality, that has the potential to equally enhances analytics, since the main aspect is to retain only the relevant text elements. With that, it is possible to reduce the size of the vocabulary (KOBAYASHI et al., 2017) and consequently the order of the term-document matrix.

The pre-processing implemented in this framework is embedded in text classification functions, available in Appendix E. The main operations, in order, are depicted in Figure 11. First, the language needs to be unified, specially given further steps that will rely on a single idiom to be performed. After, special character removal is done considering previous known groups (like numbers and punctuation) and any special character element (like *, # and %). Finally, given a list of stopwords (terms that have high frequency but little analytical value), those are removed from the vocabulary, for the final term-document matrix construction.

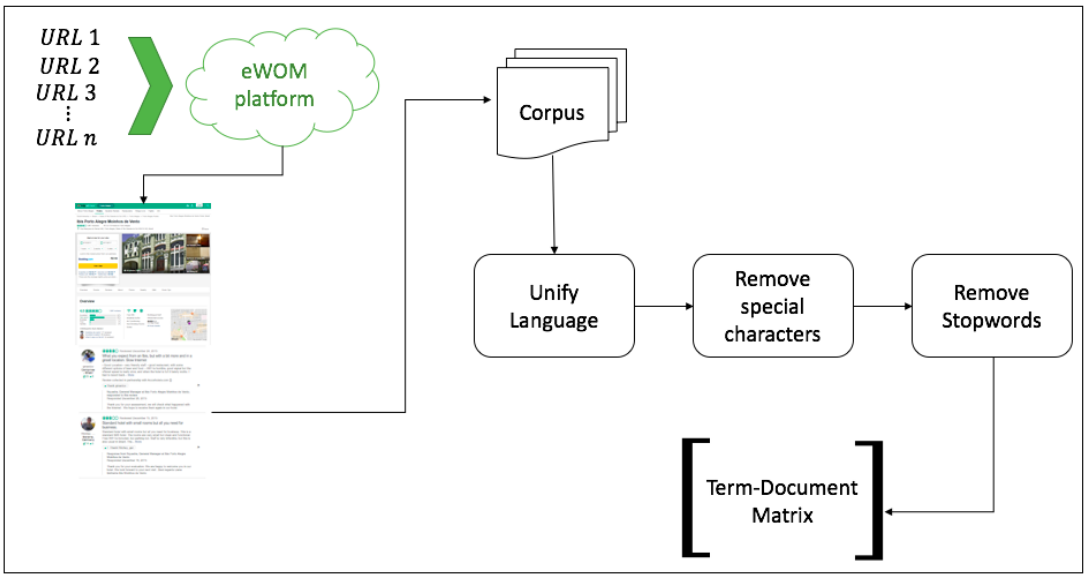

Figure 11 – Pre-processing steps

Some aspects should be highlighted given those choices. Regarding language unification, the main operator that can be chosen is Google Translate API, that demonstrated robustness to deal with over 40,000 registers in exploratory analysis (MARCOLIN; BECKER,). For special characters removal, even though these might represent some information, (like an emoticon representation, i.e "=|") not only it would be punctual but also it would require an extra computational effort for little benefit. In addition, always concerning with manager's analytical flexibility, the stopwords removal might receive new words that, for specific industries, do not aggregate any meaning, although appearing with high-frequency (like the word "car" in a set of cars critics or reviews). Regarding different text data source explorations connected with this dissertation elaboration those three process are suitable for working with raw, real-world text data.

Although previous research presented pre-processing steps (PROVOST; FAWCETT, 2016; MORTENSON; DOHERTY; ROBINSON, 2015; LEE; HAN; SUH, 2014; MCAFEE et al., 2012), none of them actually delivered a ready-to-adopt specification like the one proposed here. With that, it is possible to advance regarding text usage limitation like expertise and resource (KOBAYASHI et al., 2017), not only organizing main processes, but also implemented them in the correct order.

5.3 Topic Modeling

The Term-Document construction allows to have a single and powerful structure to analyze any set of documents. Topic modeling methods can receive this input and deliver the main topics with its related terms. Details are depicted in Figure 12.

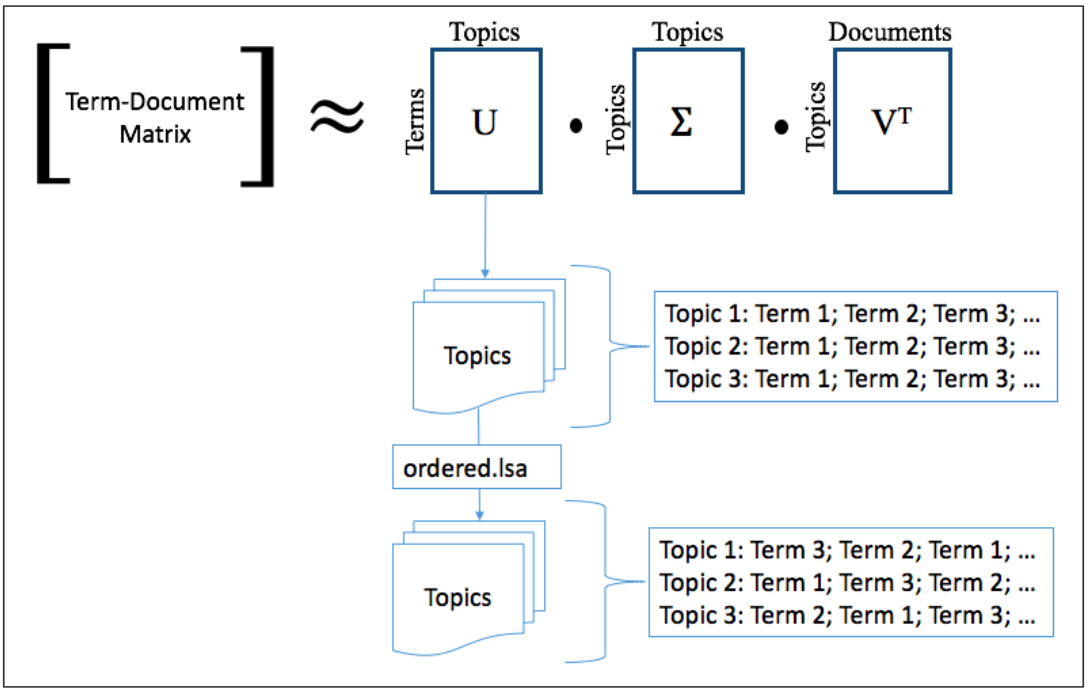

Figure 12 – Topic Modeling process

Topic Modeling concern is essentially allowing the association among documents through a number of latent topics with different weights, that specifies the degree of membership in a given topic, while reducing the dimension space. The desirable result is an understandable representation of topics that can be useful to analyze the themes presented in that set of documents (AGGARWAL; ZHAI, 2012). In this sense, Topic Modeling algorithms have the power to summarize large text dataset by discovering hidden patterns (the topics) found among a set of documents (HAN et al., 2016).

This can be seen as an important aspect to improve text usage in decision-making process. Extracting relevant information from text can be faster and more objectively done through Topic Modeling. Since all models work with dimension reduction, only the core information can be presented, with the advantages of modern visualization tools and techniques. Summarizing a great amount of text is connected with any decision process that can rely on text to improve knowledge about a context. With Topic Modeling, this task can be improved by providing not only the most frequent words, but also the "most-connected" ones, since the output is a set of topics containing strongly related words from that database. Additionally, by using Topic Models this process gains objectivity and consistency, more that it would if it is performed by a human reader (GREENE; CROSS, 2017; KULKARNI; APTE; EVANGELOPOULOS, 2014a; KULKARNI; APTE; EVANGELOPOULOS, 2014b).

LSA, through singular-value decomposition, transform the input matrix into three other matrices. Its purpose is to address the fact that it is not possible to use only term

frequency or raw data for text indexing. Since text data is produced directly from users (thus, it is straightforward), it is important to considerate that there are different ways of communicating a single concept. Therefore, the terms used by users (e.g., consumers) may not correspond to other terms presented in other user text, although it can express the same idea. On the other hand, two users may choose the same word to express different opinions (DEERWESTER et al., 1990). By applying LSA, it is possible to overcome this issue and to actually comprehend the latent structure of meaning in any set of documents.

Here, the matrix U is explored, since it contains, in its rows, the terms, and its columns contains numbers forming a topic representation. These numbers represent different weights in each topic, and simply looking at this output do not allow to understand which are the topic descriptors, essential to compare one with another. For that, a function that orders the terms considering each topic is available in Appendix B. With that, it is possible to deliver a broad view and a first exploratory analysis considering the related corpus.

LSA model is a *bag-of-words* approach. That means that the order of terms and their context in the documents forming the corpus are not taken into account. Despite this simplification, one advantage of this structure is to be able to organize and relate documents not by appearance or absence of a group of words, but by connecting them in a topic, represented by the eigenvectors (U) corresponding to the singular values (Σ).

This allows to deal with high-volume data. Applying the dimensionality reduction concept, it is possible to reduce noise in the resulting vector space (in LSA context, latent semantic space) by retaining the main dimensions, which are related with the highest singular values, that can lead to a richer relationship structure that reveals latent relations presented among documents and terms (WILD, 2016; BERGAMASCHI; PO, 2014).

One important aspect of this approach is the fact that it is built to be unsupervised. Different from previous models that worked with raw data, the framework allows to uncover the vocabulary from any industry through the set of documents collected. In this sense, for important processes like text classification and clustering (KOBAYASHI et al., 2017) there is no need to put effort in vocabulary analysis.

This characteristic permits to support language changes over time. As an automated and dynamic process, text analysis should not rely on fixed keywords list, as previously proposed (AGUWA; OLYA; MONPLAISIR, 2017). This obsolete approach makes it necessary to constantly compare new words in order to update any previously built-in list. Text is not structured, and most people do not write considering all grammar rules: misspelling, typographical errors, unknown abbreviations and slangs are just some examples of the challenges to be faced given this kind of data.

In addition, being consistent with design research approach (AKEN, 2005), the

artifact should allow generalization in some degree. A fixed list of keywords makes the approach too specific, while the one proposed here can be applied in any domain, since that the topic descriptors emerge from the set of target documents.

In this direction, it should be stressed the cost-benefit delivered by an automated and unsupervised approach. Although it is possible to rely on specialists judgment to build a domain knowledge or a set of words that composes different topics or categories (CARRASCO et al., 2012), this would imply in a high-cost for organizations, considering human, time and budget resources. In addition, the process could be considered fragile not only due to constraints on the knowledge, but mainly because of unavoidable bias that follows any human-based judgment (GREENE; CROSS, 2017).

5.4 Longitudinal Analysis

Even that text data has become increasingly common, methods for monitoring perceptions from its authors (e.g., customers) are still hard to find, not being available for immediate adoption (ASHTON; EVANGELOPOULOS; PRYBUTOK, 2014). Considering that many business *corpora* are time tagged, a longitudinal analysis could help to monitor trends among subjects presented in documents.

The longitudinal analysis embedded in this framework aims to to provide additional and valuable information in time frames. Given the set of main topic descriptors achieved with LSA, a historical analysis can be delivered by the relative frequency throughout the years, or months, or even weeks. Figure 13 shows this process.

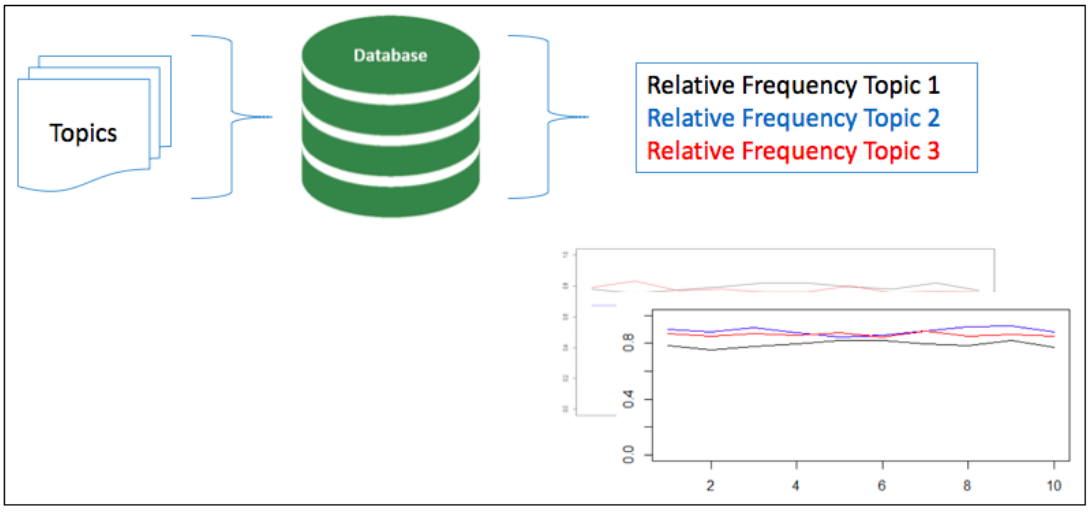

Figure 13 – Longitudinal Analysis process

First, the topics are combined with the database resulting from the Webscrapping and pre-processing phases. Then, the relative frequency of each topic is calculated using

the mean relative frequency of each term. As all terms have weights in all topics, it is suggested that the top-10 words should be considered. The results are a set of graphics that demonstrates each topic fluctuation over the period considered.

This can provide managers with information about the status of each topic, which in turn can provide useful insights considering the organization's market position or point-of-view. This is important to confirm or even contradict the manager's feeling about the document's authors position, going beyond intuition or market rumble by presenting evidence about each trend.

In this sense, not only it is possible to better understand each subject relative power compared to others, but also the topics themselves that have emerged from the dataset, in an unsupervised form. This can lead to a different comprehension about the main trends, making it possible to reveal even the ones that could be out of the organization's radar.

Nevertheless, some set of documents might be more dynamic than others, specially the ones related to opinions (ASHTON; EVANGELOPOULOS; PRYBUTOK, 2014). The framework structure can support new topic discovering based on fold-in options embedded in lsa package in R, as previously detailed in Figure 6. With this option, new documents can be added without the need to recalculate the whole matrix. This can also facilitate new topic discovering, only by adding new relative frequency vectors in the same output.

A monitoring structure should support not only historical analysis, but also combine new data to constantly update trends. In spite of recommendation and importance discussion regarding this process (KOBAYASHI et al., 2017; LEE; HAN; SUH, 2014; MCAFEE et al., 2012), there is a lack of an integrated proposal (ASHTON; EVANGELOPOULOS; PRYBUTOK, 2014), that this framework aims to fill.

5.5 Market and Sentiment Analysis

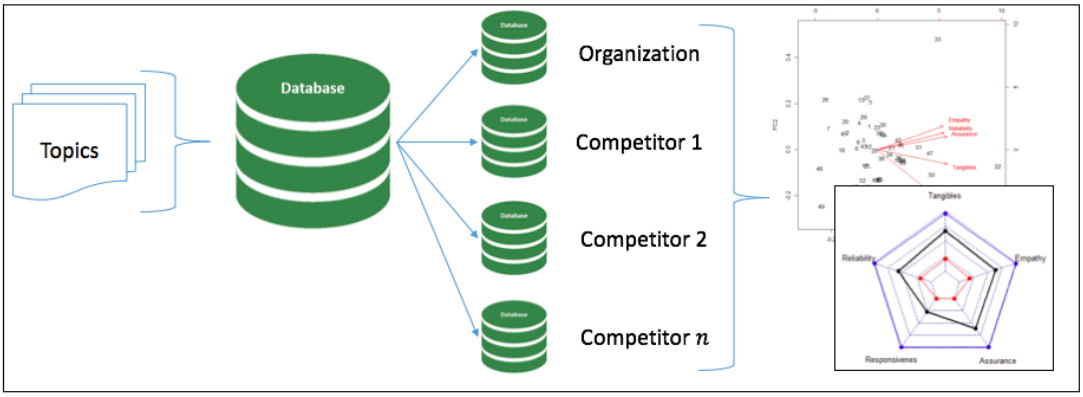

Figure 14 – Market Analysis

Although textual data can be combined with other data sources, text is a valuable source of information on its own. Scope of analysis is typically defined by the target context (customer, employee, etc.) and the actual analytical task. Thereby, three main groups of analytical tasks can be distinguished: text summarization, pattern identification, and market analysis.

Regarding *summarization*, text analytics techniques can help extract relevant information, retaining and presenting only the core arguments expressed in the texts. With different visualization methods, summarization can also save the analyst time in understanding documents. This tasks was accomplished through the webscrapper and the pre-processing sequences, that delivered a single term-document matrix to explore.

Pattern identification allows to find hidden relationships in and between texts. For example, latent semantic relations beyond co-occurrence can be uncovered. Texts often express the author's point-of-view. With pattern identification, not only relations between documents can be made salient, but also the similarity or distance between the authors of documents, e.g. grouped by context like employees, customers, citizens, competitors, among others. This tasks was accomplished given topic modeling sequence.

Finally, considering that web data can also be retrieved or scraped regarding competitors and other market players, its is possible to conduct an inter-organizational *market analysis* (MIRKIN, 2011; BOSE, 2009; MYATT, 2007). This is the final process, depicted in Figure 14. The topics can relate with each organization presented in the database, as well as from selected competitors that the manager would like to compare itself. With this, an integrated market analysis can position, in a vector space, the chosen players by the analyst, regarding the topics that were extracted through an unsupervised format. Beyond that, it is possible to consider the word score, with the use of a dictionary, and to reproduce this analysis regarding not only the main subjects, but also the sentiment involved.

5.6 Framework Final Design

In this dissertation, these three tasks are operationalized combining different techniques, as shown in the next chapter, as an actual example to provide guidance to effective text data use. Beyond business reports and data obtained through questionnaires, it is possible to rely on other external data sources (like consumer-provided text data), aiming to actually "hear" the document's author voices.

The final framework design is depicted in Figure 15. More than being an integrated sum of each part, it allows to automatize a text data analysis, from data collection to analytical overview. Considering that text data is available in several external sources, with a disaggregate structure, the webscrapping and pre-processing phases aim to collect

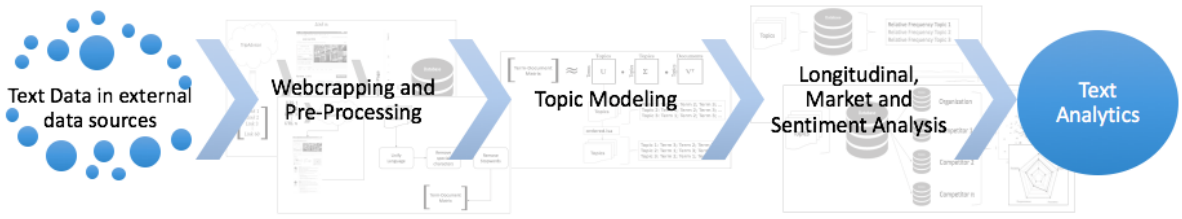

Figure 15 – Complete Framework Design

and treat this data, making it adequately formatted to topic modeling, that will not only uncover latent relations among documents and terms, but also will make the main subjects emerge. With them, a longitudinal analysis have the power to reveal trends in text data. In addition, a market and sentiment analysis will deliver a comparative contrast among the organization and other players. Going beyond text mining, the framework allows to achieve text analytics, that not only can manipulate data in text format, but also "includes identification of core concepts, sentiments and trends in unstructured data, and their use in decision making support" (ASHTON; EVANGELOPOULOS; PRYBUTOK, 2014, p. 2279).

From a flexibility perspective, is possible not only to understand a set of documents, but also to comprehend topic trending in a timeline. In addition, by choosing to analyze only its own business and some selected (or even all) players (like competitors), the framework delivers a comparative market and sentiment analysis. Altogether, the framework achieve this dissertation objective, overcoming text data challenges inside organizations and making the use of this unstructured, but exponentially growing, kind of data suitable for decision-making processes (KOBAYASHI et al., 2017; HAN et al., 2016; MYATT, 2007).

6 Validating the framework: a framework blueprint

6.1 Framework Tuning

Figure 16 summarizes the different NLP (Natural Language Processing) techniques applied to operationalize the developed framework. Supervised machine learning can help classify new data with the help of labeled training data, offering estimates of accuracy as one of the benefits, while costs are determined by the cost of labeling training data. Unsupervised techniques learn structure from the unstructured data provided, allowing for inspection of semantics found. Semantic orientation allows to complement the supervised and unsupervised content analysis with sentiment analysis, often conducted using a lexical approach upon a public dictionary (MANNING, 2015; CAMBRIA; WHITE, 2014).

In order to validate the proposed framework and demonstrate its applicability, an empirical test in the context of "Voice of the Customer" data was conducted. Although the evidence supporting the importance of reviews for other travelers have already been explored, this dissertation approach is different from previous research since it works with raw text data. Rather than experiment or focus group (SPARKS; BROWNING, 2011; HORNER; SWARBROOKE et al., 2016), this study treats and analyze real-world data, extracted directly from websites that provides open-access information, like TripAdvisor (https://www.tripadvisor.com/). This data is originated from texts written directly by consumers, representing the "Voice of the Customer" (VOC) itself, which make it possible to understand what the customers are talking about the organizations (SPANGLER; KREULEN, 2007). Consequently, the opinions expressed by the consumers of tourism and

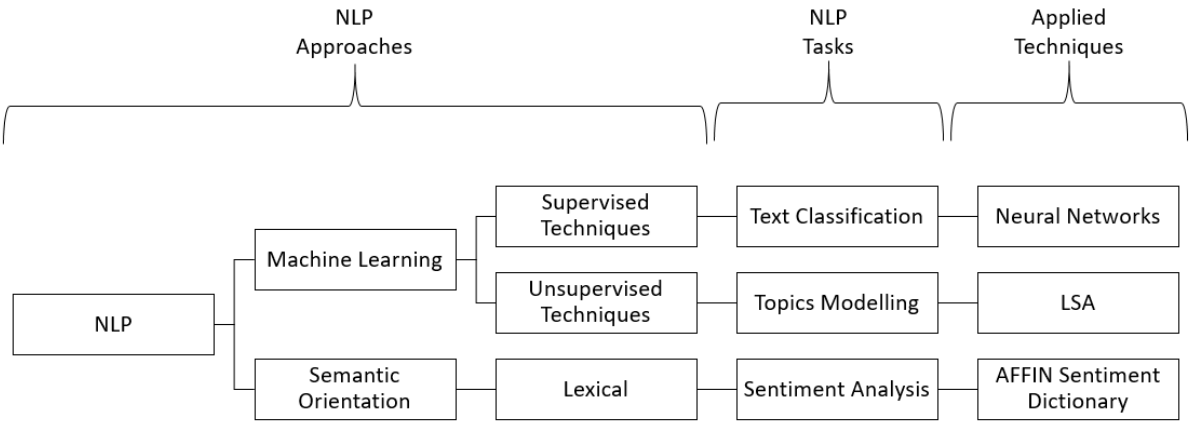

Figure 16 – NLP techniques

hospitality services in natural language are an important source of information for hotel managers (CARRASCO et al., 2012).

The instance chosen to operationalize the framework blueprint is the tourism industry, represented by the accommodation sector, mainly hotels. This choice relies in a couple of evidences regarding this industry social and economical importance, and also the benefit from external data source analysis such as eWOM.

Tourism is a growing service industry, being the most important economic driver in several countries. Given that, understanding more about tourists' perception can help also many kinds of public and private institutions to better focus the travelers wishes. In fact, as a service industry, the quality perception is a crucial point for customer satisfaction, which can have a strong impact in economies that depend on tourism to perform and generate wealth for cities, regions, and even entire countries (VALDIVIA; LUZÓN; HERRERA, 2017; CHAREYRON; DA-RUGNA; RAIMBAULT, 2014).

Previous experience from other customers is a major factor to consider before booking an hotel room on-line. Positive on-line reviews can significantly increase booking rates from hotels as well as negative ones can affect reservations (YE; LAW; GU, 2009). Indeed, tourism and hospitality sector should strongly take into account on-line reviews, especially those posted in public portals apart from the organization's site (YE; LAW; GU, 2009). The review itself also tend to have more importance for consumer perception, being reported to have more effect than ratings alone (SPARKS; BROWNING, 2011).

The importance of previous reviews for decision-making process regarding the hospitality industry has been demonstrated in the literature (SPARKS; BROWNING, 2011; YE; LAW; GU, 2009). Even without knowing the other users behind the screen, one important step in planning a travel, and thus deciding a place to stay, is to access a review from well-know websites and take the information presented in consideration. Therefore, social media and consumer review websites, like TripAdvisor, for example, have changed the tourism and hospitality sectors and the practices of the hotel managers (MOLINILLO et al., 2016).

An important aspect is the strong predictive power of the also called social media review rating and hotel performance metrics. Kim et al. (2017) compared traditional consumer satisfaction made by a hotel with the same data from four different websites. They discovered that not only social media ratings were better predictors from metrics like daily rate average and percentage of occupancy, but also that data from TripAdvisor had closest correlation.

Yen e Tang (2015) analyzed the motivations for posting hotel experiences with the online media chosen and identified the eWOM motivations that are affected by hotel attribute performance. The choice between TripAdvisor and Facebook, for example, is

correlated to different sets of motivation. TripAdvisor is associated with altruism and platform assistance and Facebook is positively associated with extroversion, social benefits, and dissonance reduction. The authors' finding suggests that motivations are not universally equal and the eWOM behaviors may be correlated to different motivations.

In this sense, regarding advances in computer science, specially in NLP (Natural Processing Language), it is possible to work not only with ratings and other numeric metrics, but also with text presented in each review. In fact, the text itself have a strong power regarding consumer decision (LEE et al., 2017), and for that reason should be included in the analysis agenda of hotel managers. Additionally, the research of Perez-Aranda, Anaya-Sanchez e Ruizalba (2017) explores this subject, conducting a study with 301 hotel managers. The main results show that managers are committed to this type of platforms, revealing the importance of analyzing consumer opinions for hotel administration.

Having a strong predictive power and being an important element for consumer decision-making, those evidences reinforces the importance, for hospitality practitioners, of analyzing objectively this kind of data, once it can help to better understand what potential consumers will face while researching about different options. SERVQUAL was shown to be adequate for measuring service quality during a hotel experience (KNUTSON et al., 1990). Since then, in order to measure quality perception, several studies worked with SERVQUAL and other related measures, aiming to provide complete models for hotel managers (ASSAF et al., 2015; DEDEOĞLU; DEMIRER, 2015; CARRASCO et al., 2012).

TripAdvisor is considered an on-line community that is part of the trends that are significantly impacting the tourism industry, being the source to choose hotels from more than one-third of the travelers. In fact, it can be considered the largest travel community, reaching millions of customers worldwide (VALDIVIA; LUZÓN; HERRERA, 2017; LEUNG et al., 2013). There is also literature evidence demonstrating the data reliability, being considered trend shaper regarding consumer behavior in tourism (HORNER; SWARBROOKE et al., 2016).

These comments analysis can bring information for both companies and consumers. Sometimes reading multiple comments may not be possible, and a single bad or good experience expressed in one opinion may not be enough in order to make a decision, which would require reading more comments in an exhausting process. Likewise, any conclusion about a service or a product should not be based on a small number of comments. Thus, for a robust analysis that can support a decision-making process, the same process of reading and analyzing everything that has already been written is required, which would be very costly if it were done through human effort.

6.2 Sample Characteristics

For the framework validation we considered hotels with at least 100 comments registered in TripAdvisor from a state capital in south Brazil, resulting in 60 input links. These were chosen for convenience, with business located in the author home town since that. Even so, the function is based upon TripAdvisor structure, so that it can be used for any other city. All data resulted in 26,141 valid comments (considered as documents). For the purpose of this study, we choose to work with comments from 2011 to 2016, resulting in 23,229 registers, and with 2,542 comments in 2017.

It is interesting to observe that there is a significant adoption of this kind of platform throughout the years. Although the average occupation rate (AOR) (ALEGRE, 2017) had fell slightly, it is possible to observe that the amount of comments grew exponentially, what confirms the high-adoption eWOM applications phenomenon previously evidenced in the literature (YEN; TANG, 2015; KIM et al., 2017) also in this city. Table 3 shows the mean occupation rate by year, which fell from 61.18% in 2011 to 46.01% in 2016. Data was available until April 2017. Figure 17 shows the AOR per month.

Table 3 – Mean Occupation Rate

Year	Mean (%)	Comments
2017	45.20	2,542
2016	46.01	7,530
2015	46.22	5,991
2014	53.67	4,939
2013	53.04	3,630
2012	58.76	743
2011	61.18	396

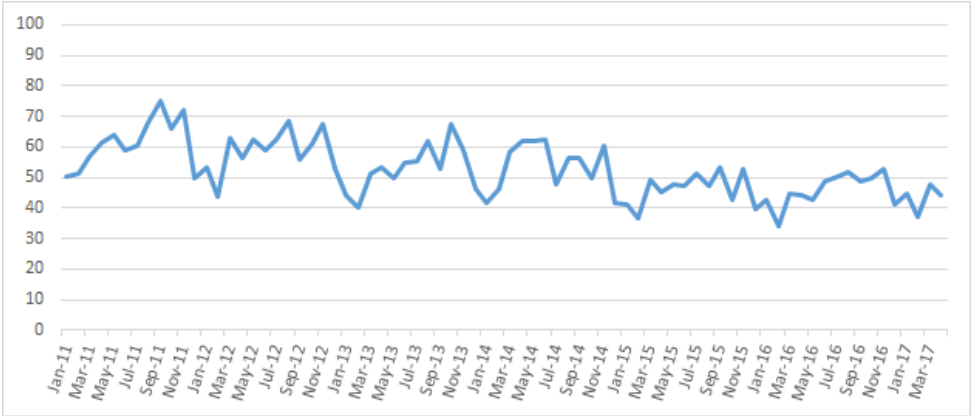

Figure 17 – Average Occupation Rate (ALEGRE, 2017)

Another important aspect is the hotel sample characteristics, depicted in Table 4. From the sample, it is possible to see that in the city there are 7 hotels with two stars,

42 with three stars and 11 with four stars. Most of them are located in downtown area ("Centro"), that concentrate the main touristic attractions and event placement in the city. Hotels name were preserved, but the codes are presented in the market analysis.

Pre-processing steps followed literature recommendation (MANNING; RHAGAVAN; SCHUTZE, 2009), as previously shown in Figure 11. The first procedure was to unify the language of the comments. Records were found in English, Spanish, German and Portuguese. All titles were translated into English, and the translation was done using the "Translate" formula, available in Google Sheets. To confirm the translation process, we checked for accents and special marks, not only to remove them, but also to correct any mistake during translation process. Being an worldwide platform, we found accents like á, é, ü, which were replaced only with the correspondent letter (i.e., a, e, u).

Additionally, we removed special characters like &, *, #, and so on. We considered that these characters may have been resulted from errors in data collection or even some language misuse from the user. Finally, the third procedure within the data pre-processing step was the removal of stopwords, words with high frequency in the database but without significant value.

After these procedures, we began the classification process in order to built an automated classifier that could allow to classify comments in the five different SERVQUAL categories without having to update any keyword list or ask for human judgment from time to time. The classification process aimed then to construct an annotated training database set to not only train the model, but also to validate its capacity and performance.

We perform four classification rounds with a total of seven different human specialists, being two computer scientists, one businessman and four marketing and business graduate students. The first round consisted in free classification, from the first 100 comments, in SERVQUAL categories. For this process, the categories description followed Parasuraman, Zeithaml e Berry (1988), with additional details based on Knutson et al. (1990) and Carrasco et al. (2012). The classification process was non-exclusive, that is, each comment could be classified in more than one category at the same time. This was done because, while analyzing the database it was possible to notice that some users talked about different aspects within the same comment.

After this first round, which was performed by three specialists, two computer scientists and one businessman, the results were discussed in group, in order to align the classification and decide whenever there was divergence. With this, the description about each category was improved, making more clear the difference among them for hotel industry. The second round consisted a new set of the next 125 comments that were classified by the same three specialists. The results were compared and it was achieved a consensus of 89%.

Table 4 – Hotel Sample Characteristics

Code	TripAdvisor Stars	Belongs to Hotel Chain	Neighborhood	Hotel Stars
H1	4.5	No	Moinhos de Vento	3
H2	5	No	Floresta	3
H3	4.5	No	São João	4
H4	4.5	Yes	Bom Fim	3
H5	4	No	Moinhos de Vento	4
H6	4	No	Bela Vista	4
H7	4	No	Centro	4
H8	4	No	São João	3
H9	4	No	São João	3
H10	4	Yes	Praia de Belas	3
H11	4	No	Centro	3
H12	4	No	Centro	3
H13	4	Yes	Moinhos de Vento	4
H14	4	No	Moinhos de Vento	3
H15	4	No	Petrópolis	4
H16	4	Yes	Centro	2
H17	4	No	Centro	2
H18	4	No	Cidade Baixa	2
H19	4	No	Bela Vista	3
H20	3.5	Yes	São João	3
H21	4	Yes	Floresta	3
H22	4,5	No	Centro	2
H23	4	Yes	Navegantes	3
H24	4.5	No	Floresta	3
H25	3.5	No	Floresta	3
H26	4	Yes	Sarandi	3
H27	4	Yes	Independência	3
H28	3.5	No	Centro	3
H29	3.5	No	Centro	2
H30	4	No	Centro	4
H31	4	No	Centro	3
H32	4	No	Floresta	3
H33	4	No	Centro	4
H34	3.5	Yes	São João	3
H35	3.5	No	Centro	3
H36	3.5	No	Floresta	3
H37	3.5	No	Moinhos de Vento	3
H38	3.5	Yes	Centro	3
H39	3.5	Yes	Centro	2
H40	3.5	Yes	Cidade Baixa	3
H41	3.5	No	Centro	3
H42	3	Yes	Navegantes	3
H43	3.5	Yes	São Geraldo	3
H44	3.5	No	Centro	3
H45	3	No	Cidade Baixa	3
H46	3	No	Centro	3
H47	3	No	Centro	3
H48	3	No	Petrópolis	3
H49	3	No	Petrópolis	3
H50	3	No	Centro	3
H51	3	Yes	Cidade Baixa	3
H52	3	No	Centro	3
H53	3	No	Floresta	3
H54	3	Yes	Centro	3
H55	2.5	No	Centro	2
H56	4	No	Centro	4
H57	4	No	Rio Branco	4
H58	4	Yes	Centro	4
H59	4	No	Petrópolis	3
H60	4	Yes	Cidade Baixa	3

Afterwards, a classification protocol was constructed and used to train another four specialists, all graduate students in marketing and business majors, residents in the same city where the hotels are established. They performed the third and fourth classification rounds, that consisted in classify a new set of the next 775 comments. One of this specialists classified all 775 comments, and the same set was also distributed among the other three: one was responsible for 275 and the other two for 250 each. The results were compared and the mean consensus was approximately 82%. For the neural network training and test, we used only the comments where there were complete consensus from at least two different human classifiers. All classified (labeled) data set resulted in 1,000 valid comments (considered as documents) and 1,389 unique terms. The distribution of comments per category and the agreement distribution is on Table 5.

Table 5 – Final Manual Classification

		Labels		
	% Agreement	**0 and 1**	**Only 0**	**Only 1**
Cat 1	80.43%	805	127	678
Cat 2	76.65%	801	606	195
Cat 3	88.62%	894	868	26
Cat 4	75.64%	770	653	117
Cat 5	85.86%	866	837	29
Mean	81.44%	-	-	-

With the labeled data we run the neural network classifier. All processing was made in R. As learning from only positive samples can lead to saturation criteria (MANEVITZ; YOUSEF, 2007), we used 10 nodes in a single-hidden layer as a bottleneck, after testing empirically different values, aiming in the smallest amount due to computation performance. In addition, 10 rounds of random samples were performed for each category, using a ratio of 75/25 for the documents belonging to train and test set.

In order to provide a set of expressions that could better describe each category in the hotel industry context, Topic Modeling was applied to all the comments per category, through LSA method. Figure 18 present the singular value distribution for all database. As it is possible to see, the first dimensions retain great part of the information. Therefore, the top-10 words from the top-10 singular values (or topics) were used to describe each category, using as a measure of representativity the value of each word in the $U_{n\times k}$ matrix weighted by the singular value from that topic.

To provide a richer decision tool for managers, a sentiment analysis was also performed. There are different approaches to understand sentiment within human-generated text. One of them is to use a lexical set that contain a score or a label for each word. For that tasks, we used the AFINN dictionary, that in the newest version counts with 2,487 English words and expressions, with a polarity from -5 to 5, implemented in R (SILGE; ROBINSON, 2016; NIELSEN, 2011).

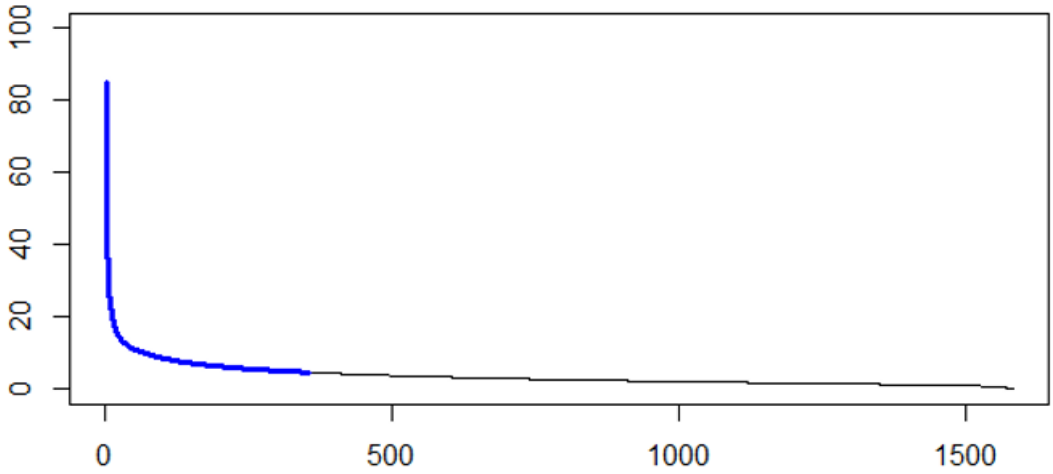

Figure 18 – Singular Values

The next section present the main managerial implications from the applied framework in tourism instance. First, the task to summarize the data is presented, operationalized in this work in two-ways. A topic modeling and a longitudinal analysis are presented, aiming to facilitate trend monitoring using in text data. After, through a supervised-classification method giving SERVQUAL categories, the main descriptors in the specific industry are presented. In the sequence, to find hidden patterns, a topic modeling technique was applied, that uncovered latent relations among expressions in the same category. The advantage of this approach is the possibility to provide a specific description for the analyzed industry (hotels), in contrast with the broad and somehow vague description in SERVQUAL, aiming to achieve actionable value. Finally we performed a market and sentiment analysis, contrasting the competitors and delivering a tool with which performance can be compared against market average, providing insights for a market analysis in a level that is relevant for decision-making.

6.3 Managerial Implications

With more than 20,000 dimensions and a sparsity rate of more than 90%, the full Term-Document Matrix is an example of the high proportions this kind of data can achieve. In order to analyze it, a dimensionality choice regarding the value of k (that is the dimension of the semantic vector space) was needed. An optimal k can allow to work with dimension reduction, which can reduce noise in latent semantic space, retaining the main dimensions that are related with the highest singular values.

This can lead to a richer relationship structure that reveals latent relations presented between documents and terms (BERGAMASCHI; PO, 2014). However, the optimal k is still a challenge. Different authors have proposed a set of solutions (WILD et al., 2005; KULKARNI; APTE; EVANGELOPOULOS, 2014b; BERGAMASCHI; PO, 2014), but

Table 6 – Database

Year	Documents	Unique Terms	Dimensions
2016	7,530	354	140
2015	5,991	365	143
2014	4,939	337	182
2013	3,630	339	181
2012	743	260	119
2011	396	342	122

still many of them refers that this point should be defined empirically for each collection.

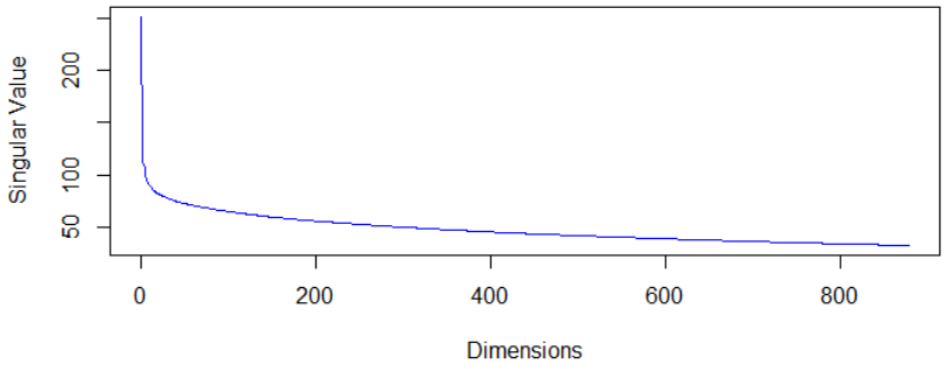

Figure 19 – Singular Values from 2011 subset

The first data exploration was related to the main topics contained inside the database, per year. For each data subset, we choose a k value through a singular-value analysis, as in Figure 19. It is possible to see the decreasing curve of singular-values, indicating that to work with all dimensions would imply more computational cost than information value. For our database, we first removed sparse terms, and after we choose to retain 65% from all singular values, as recomended in previous studies (WILD et al., 2005), ending up with a database structure as in Table 6.

Table 7 presents the five main topics within 2016 comments, that represents the five dimensions with the highest singular values from this subset. We can see that each of them brings a different topic to discussion. T1, that represents Topic One, has united comments that reinforces the proximity from the hotels with the airport, as an important message for other users. T2 brings another discussion, focusing more in attendance, reporting staff strengths like “helpful” and “attentive”. Comments on T3 are more concerned about hotel localization, highlighting the proximity of restaurants or a shopping mall. Although bringing again the word “airport”, it is possible to see that T4 is different from T3, since this word is close together, in LSA space, with “shuttle” and “transfer”, which means that comments on this dimension were more concerned about getting from and to the city airport.

Table 7 – Main Topics - 2016

T1	T2	T3	T4
porto	excellent	located	airport
alegre	staff	center	near
airport	great	old	shuttle
center	helpful	shopping	transfer
near	attentive	restaurant	free

This is just one way to observe this data. After the pre-processing steps, to construct a Term-Document Matrix with 99% of sparsivity (that is, retaining a great amount of data) and an LSA space with 65% of the singular-values it took no more that a few seconds in a notebook with Intel Core i5 2.4Hz processor, 8GB of RAM and with a 128GB SSD storage. After that, the manager can understand, in an objectively way, the main topics that their customers (or cutomers from other hotels) are talking about. We now show other possibilities of analysis with the remaining years.

Another interesting way to look at this data is using modern visual tools, like word clouds. Visualization tools are a challenge within business analytics context, since that we are working not only with large sizes of data, but specially because of the high dimension (CHEN; ZHANG, 2014). To analyze the 2015 data, we choose to construct word clouds from the first four dimensions. Since that each word has a score in each dimension (which is the a_{ij} entry from the U matrix), this can be used as an index to word size into the cloud. In order to be more precise, we calculate a relative index for each word in each dimension, to understand its real contribution for that specific topic. Figure 20 present the word clouds.

The first topic in Figure 20, “room”, represent a set of comments that have mentioned hotel room elements like “bed”, “bathroom” and “shower”, altogether with qualities like “comfortable”, “excellent” and “spacious”, or with complaints like “old” or “bad”. The topic that brings “ibis” word in its center represent a expressive set of comments that mentioned this hotel brand. It is important to note that those are not comments filtered by hotel, but by topic, and the related words - like “standard”, “location” and “neighborhood” - represent the terms that appear in the same dimension. Finally, the topic about “center” (i.e. downtown) connects this part of the city with the main characteristics expressed by the users, like “old”, “simple”, “bad” and “noisy”.

Some interesting conclusions can be draw from this set of word clouds. First, this visualization conveys more information to the manager than a table or a list of words. Also, most topics differ from 2016 topics. In Figure 21 it is possible to see that, besides topic “location”, (which is another representation from T3 in Table 7), the other three topics are distinct. Considering that this are the first four dimensions, i.e, the four eigenvectors

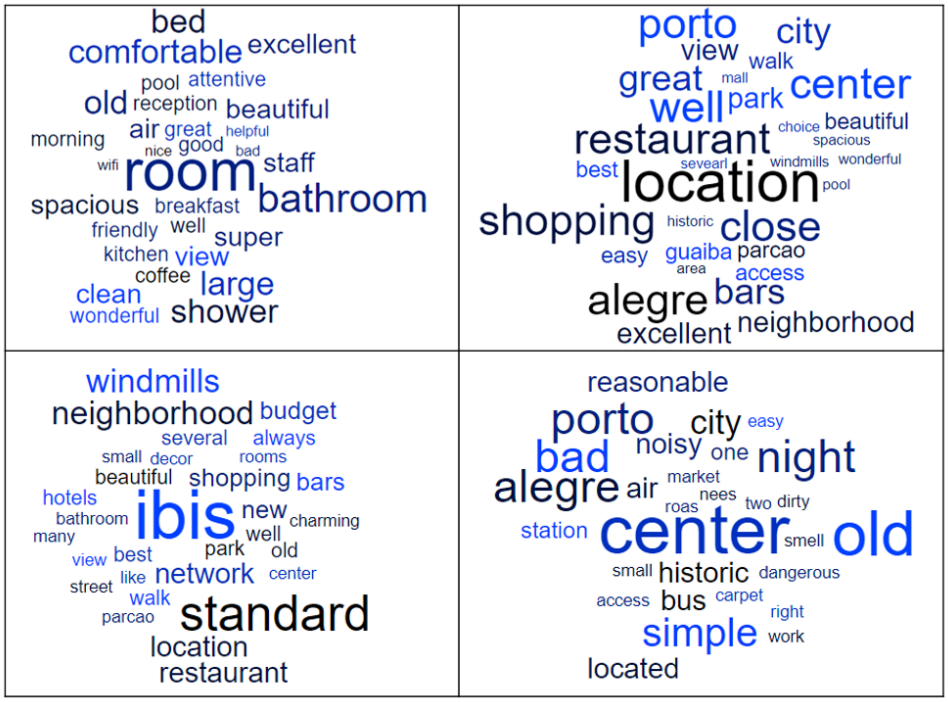

Figure 20 – Main topics from 2015

connected with the highest singular values considering the 2015 subset, we can conclude that "airport" was a concern that have grown up from 2015 to 2016.

Just like "airport" was not among the first dimensions in 2015, "room" and the related terms were not among the first dimensions in 2016. But although there exist some differences between the years, we considered important to compare them with the full-period analysis. In order to do that, we needed to work with the full matrix. That was a challenge that many organizations will have to uncover in order to increase the amount of data for decision-making process: the dimension reduction problem.

This full matrix consisted in 22,062 words distributed in 23,229 documents (i.e., comments). Working with the totality of this data would mean to deal with over 450 millions entries. However, this is not an ideal approach. For example, approximately 60% of the words (13,205 terms) appear only one time across all years and all documents. Figure 22 shows the frequency from all words in the dataset, illustrating this effect. This can illustrate the semantic structure idea, presented by Zipf law (ZIPF, 1949), that states that the inverse relationship among frequency and rank-position given a frequency table for any words is true in different languages, which demonstrate the presence of a similar structure for any set of documents.

We choose to remove these words, in order to reduce our dimension, since this terms would hardly imply some global knowledge about travelers opinion. Another procedure we performed was to recalculate the Term-Document Matrix removing highly sparse terms

Figure 21 – Main topics from 2016

after *tf-idf* computation. With that, we were able to work with a matrix 96% sparse, against the 100% sparsivity that we had before. Finally, when constructing the LSA space, we tested with all dimensions, retaining 50% and 65% of the singular values, choosing to stay with the latter. After that, we ended up with 351 terms distributed in 196 topics.

Since presenting the main topics from all the period with a bar chart can demonstrate another visualization method that can be used by managers, and also can be useful in order to understand the weight from each word considering the whole topic, we choose this visualization method. For that, the top-20 words were used. The main topics are presented in Figure 23.

With this construction, it was possible to obtain the topics and its correspondent terms that describe the main subjects among consumer's comments. In addition, being

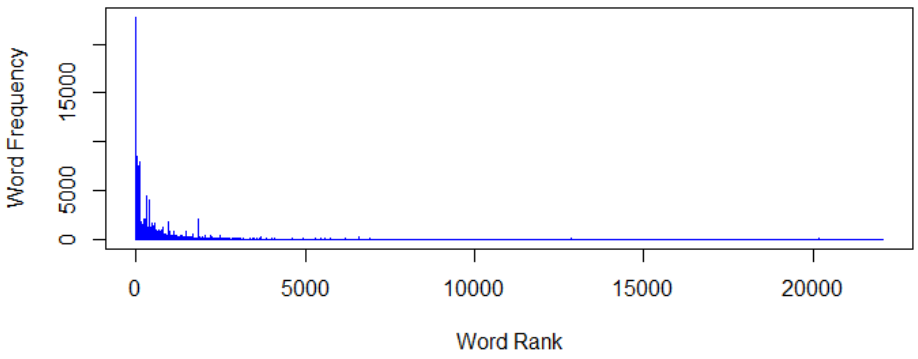

Figure 22 – Word Frequency

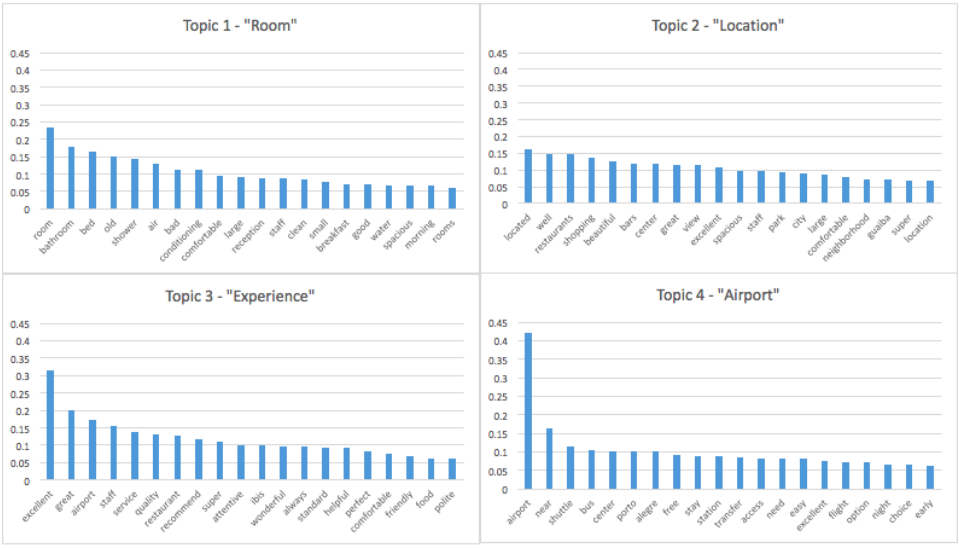

Figure 23 – Main topics from historical data set (2011-2016)

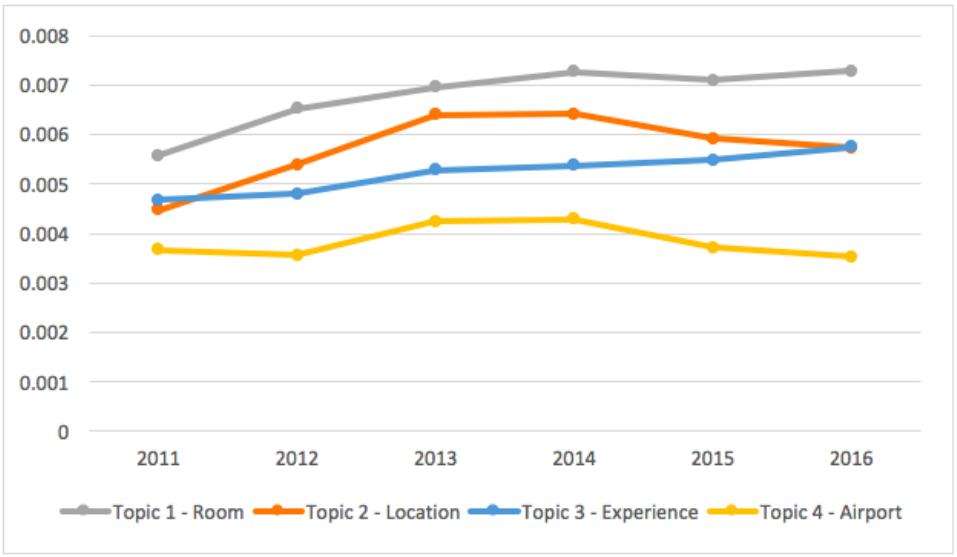

Figure 24 – Longitudinal Analysis

the data base historical, it is possible to observe tendencies regarding these topics, as in Figure 24. For that, we considered the relative frequency of the same top-20 terms in each topic (highlighted in Figure 23) in each year. With that, it is possible to quickly notice topics trends that reflect directly this group of customers concerns.

It is possible to see that the topics have followed different patterns during the years. Although from one year to another it is hard to tell what is trending and what is out of fashion, by analyzing the topics through the years, and specially comparing the trending curves, it is easier to bring new information for decision-making. In addition, managers could consider some topics more important for their market positioning strategy than others, being possible to use the same structure built in this dissertation to monitor the

terms appearance and topic trends in different periods like months or weeks, helping to manage new unstructured information in business environment.

SERVQUAL dimensions descriptions can be hard to adopt for any service industry given its broad description (KNUTSON et al., 1990; CARRASCO et al., 2012). In order to uncover the main expressions from each category in hotel industry, a Topic Modeling analysis was performed through LSA. As the k value is defined to each dataset (BERGAMASCHI; PO, 2014), we perform tests with 100%, 65% and 50% of the singular values. As Figure 18 in page 74 shown, retaining 50% from all the singular values (blue line) allows to keep the majority of the information in the matrix, while being more computer efficient, since the decomposition is simplified.

The main topics uncovered from all data classified are represented with word clouds, aiming to facilitate visualization. The words represent the top-10 words from the top-10 topics from each category, that is, the topics that correspond to the 10 highest singular values for each of the SERVQUAL categories. In addition, the size of the word is proportional with its frequency.

The category 1, "Tangibles", relates with physical facilities. In hotel industry, this is a broad category, having customers talking about interior decoration to room quality, also connected with other elements like restaurants, gym, and pools (CARRASCO et al., 2012; KNUTSON et al., 1990). Figure 25, word cloud 1, shows the main expressions are connected with physical facilities from the hotel (like bed, bathroom, room, shower) and with adjectives to those facilities (like comfortable, large, old, clean). Also, the breakfast was an important aspect, having appeared in four of the first 10 topics.

The category 2, "Reliability", represents the ability to perform the promised service (PARASURAMAN; ZEITHAML; BERRY, 1988). This is a category close-related with expectations, since customers have access not only to pictures from the hotel and the rooms, but also to previous traveler's impressions, what can be related to Carrasco et al. (2012) definition, where the image and the conditions of the hotel receive strong emphasis in this category. Accordingly, the word cloud 2 in Figure 25 shows the importance of the cost-benefit for the customer, connected with the idea of meeting the expectations, since the price itself usually set an expectation of how the promised service will be performed. Another expression that demonstrates this relation is the word "price". Also, there are more adjectives in this category (great, excellent, perfect). This can be seen as the customers expressing their impressions about the service received.

The category 3, "Responsiveness", relates with the willingness to help customers and to provide a prompt service (KNUTSON et al., 1990). When observing the word cloud 3 from Figure 25 it is possible to see that this category relates with problem-solving, where the staff have a central role. This is consistent with previous findings, as the ability to quickly solve any kind of situation, like reservation issues or check-in/check-out matters

Figure 25 – Word Clouds from the five categories considering main topics

(as the word cloud reveals), offers to staff opportunities to positively or negatively impress the customer (CARRASCO et al., 2012).

The two last categories, 4: "Assurance" and 5: "Empathy" are related specifically with the staff. In an industry such as hospitality, to feel welcome is important, and the employees have a considerable responsibility in order to make that happen. The main difference among those categories is that "Assurance" is more related with performing the services, and "Empathy" brings more the perception of caring and special individualized attention (CARRASCO et al., 2012; KNUTSON et al., 1990; PARASURAMAN; ZEITHAML; BERRY, 1988).

In Figure 25 the word cloud 4 demonstrates that most of the topics in category 4 are connected with the words "staff" and "service" together with expressions like great, attentive, need and quality. Most of the comments classified in this category followed the description, that is related with knowledge and courtesy, demonstrating that usually the posture of the staff, even without any specific request, had left some impression in the customer. In contrast, category 5 represented by the word cloud 5 presents more affective words like "beautiful" and "love", together with "staff" and "service" although some dissatisfied comments used mixed words, making this the most vague category, what can also be a result from the broad description that it receives, encompassing from complimentary services to employees sensibility (KNUTSON et al., 1990).

After having developed a latent semantical space related with the five SERVQUAL categories, it is possible to understand the relation from each hotel with this customer-based market perception. That means that each category can be represented as a set of keywords presented, with higher or lower emphasis, in the comments of each hotel. In

order to represent this structure, we formed a matrix containing a belonging-index for each hotel in each category considering the frequency of this set of words (also presented in the word clouds) weighted by the number of comments from each hotel. With that representations, we performed a PCA (Principal Component Analysis) and used the first two dimensions to build a market analysis of this set of hotels. The result is in Figure 26.

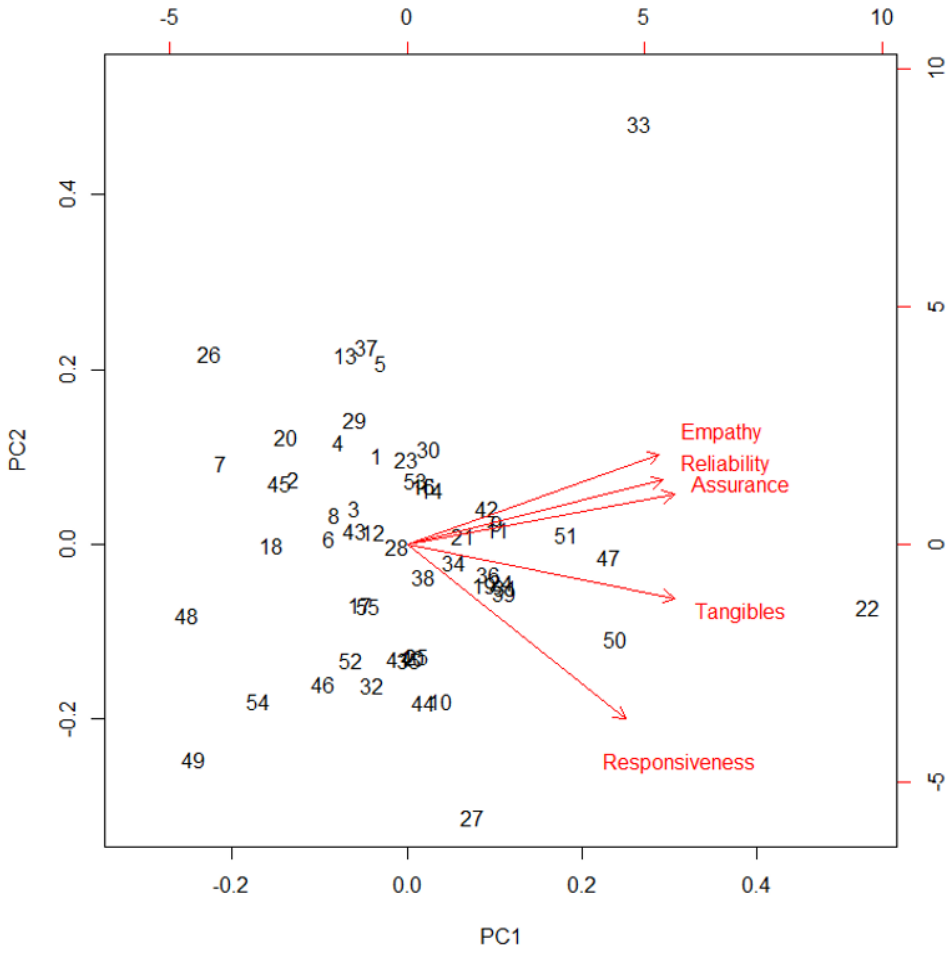

Figure 26 – Hotels (numbers) and the SERVQUAL dimensions (arrows) plotted against their loadings on the first two principal components

The goal with this analysis is to allow the managers to understand their position considering customer perception in different categories. Figure 28 zoom in into categories 2, 4 and 5 (Reliability, Assurance and Empathy) vector space, making more clear the relation of the different competitors in this market. Additionally, with the same data structure, we also can provide a 3D plot using the first three principal components, as in Figure 27.

Some clusters can be spotted from this vector spaces, indicating some similar characteristic that may not be obviously spotted. In Figure 26, hotels 5, 13 and 37 appears very close to each other. Indeed, those three hotels are located in the same neighborhood, and have in their comments a highlight about location. Another example in the same Figure are hotels 32, 46 and 52, that are more corporative style hotels, having in common

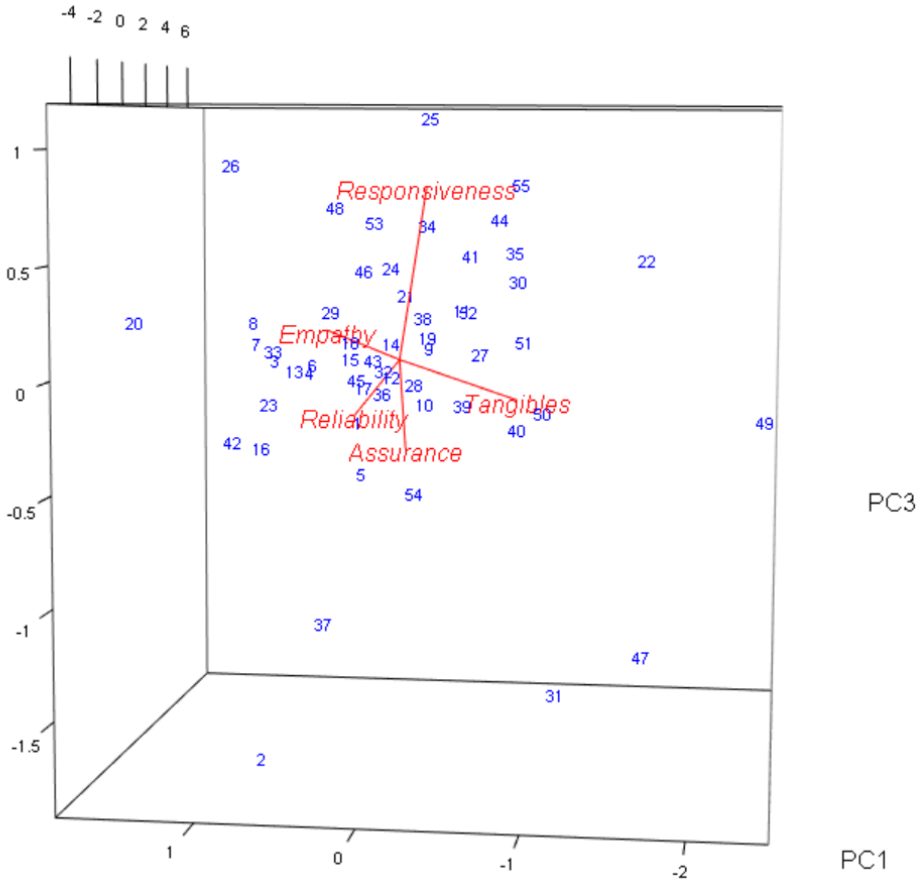

Figure 27 – Hotels (numbers) and the SERVQUAL dimensions (arrows) plotted against their loadings on the first three principal components

comments mentioning about events that have been placed there.

However, not only the position, but the polarity is important to uncover opportunities for improvement in the customer perspective. Then, aiming to give a broad perspective for a decision-making inside the organization, we performed a sentiment analysis per category. Figure 29 demonstrate the relation, in all comments, of the sentiment from all hotels in each of the categories, that can be seen as the market performance average in the city. The radar chart was built considering the sentiment words in each category, according with the dictionary used (NIELSEN, 2011), weighted by the amount of comments and normalized using the category with the highest score (Empathy).

The chart allows to conclude that Empathy is a strong category in the customers opinion, followed by Tangibles, Assurance and Reliability. Responsiveness, connected with problem-solving, specially in check-in/check-out and reservation issues, was the category that presented the poorest performance. This can be due to customers motivation to express their feeling related with problem-solving: if the problem is solved, it can be seen as the hotel obligation; but if something goes wrong, it can be easily considered a fault.

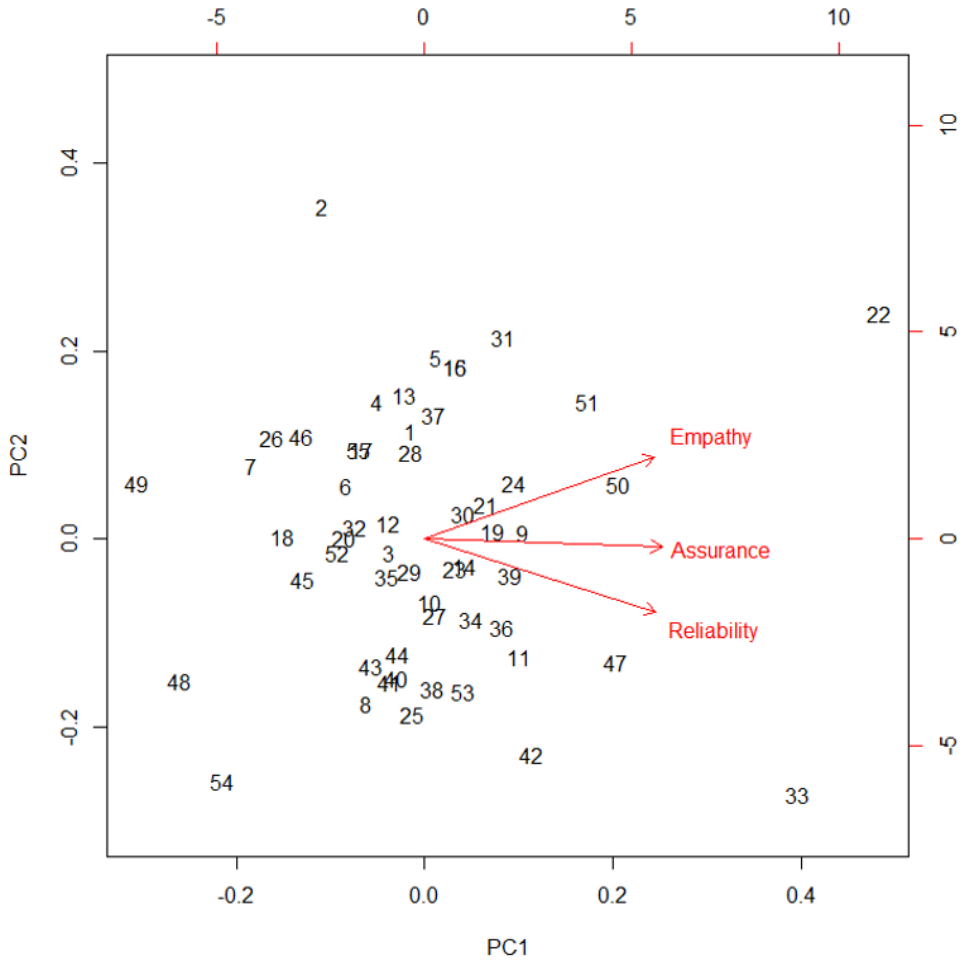

Figure 28 – PCA Analysis - Zoom in

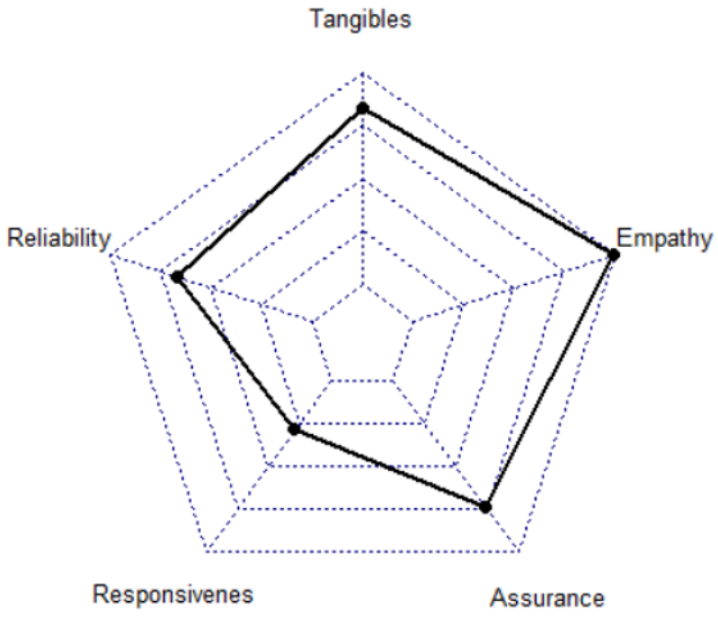

Figure 29 – Sentiment per Category (all units)

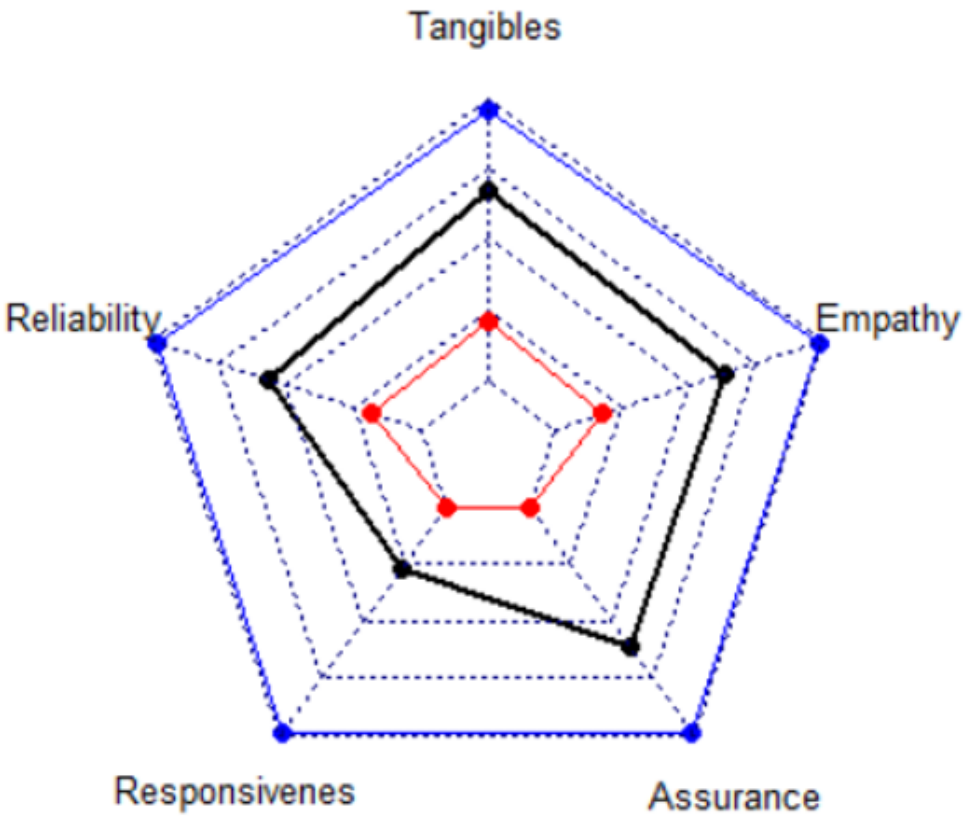

Figure 30 – Compared Sentiment Analysis (unit 2 (blue line), unit 47 (red line) and all units (black line))

While Figure 29 gives a broad view from all the competitors, the same tool allows to compare a group of comments from one single hotel regarding the overall market performance, improving the PCA image with polarity for a single organization. This Figure 29 presents the average of all comments sentiment normalized by overall Empathy, in order to uncover a single structure to contrast each organization given a general customer opinion.

This single structure can be used to analyze customers opinion from one hotel against the market average, or even between two or more units. To demonstrate this feature, we choose two hotels that are more distant from the others, in at least one category, according to Figure 28: Hotel 47 and Hotel 2. Figure 30 compares Hotel 47, represented by the red line, Hotel 2, represented by the blue line, and the market average, represented by the black line, in the same sentiment analysis. The scales are normalized in the Empathy category for Hotel 2.

From Figure 30 it is possible to acknowledge that Hotel 2 have a high satisfaction level in all categories, what makes this almost an outlier, reinforcing the position found in Figure 28. Analyzing the hotel characteristics, it can be seen that this is a small business, with fewer rooms than the average, and not belonging to any multinational hotel chain.

Those uncertainty aspects might easily arise positive surprising feelings, along with the good service. Some comments in Table 8 illustrate this point.

On the other hand, Hotel 47 has a lot to improve, since it has scored below the average. As it was expected regarding overall market performance, Responsiveness was the lowest-rated category, altogether with Assurance. Being both categories related with the professionalism level perceived by the customer, we could point that some inappropriate staff behavior along with unsolved, or not properly solved problems, might be this hotel main issues. Indeed, looking at some of the comments from this hotel, it is possible to deduce the unpleasant surprise, as in Table 8.

Table 8 – Sample Comments

Hotel 2
Beautiful hotel! Beautiful beautiful hotel and surprised positively. Room clean and nicely decorated. Friendly and welcoming staff. Good breakfast. Great location with shopping restaurants all close. Nothing to complain.
Impeccable awesome place clean and organized that captivates the details and the exquisite taste. Very willing staff. Hotel unlike any other city that has really attentive and caring owners. Great location. Worth the stay!
Hotel 47
Stayed with my family in Double Room, air conditioning insufficient and small bathroom. Shower over the bath terrible can shower very bad. Breakfast poor service no courtesy.
HORRIBLE Hotel POOR! Does not have a good room service with no ventilation (air conditioning does not work) beds are sunk! Snacks offered are very bad too! Not to mention the breakfast very badly served.

Although different and sophisticate techniques are being developed due to the combination of strong computer capacity and scientific progress, the actual use of analytics inside organizations can be seen as still modest. One reason for that can be the fact that most of new unstructured data is outside the organizations, since traditional data warehouses still use ETL (Extract, Transform and Load) process that are not up-to-date with unstructured data (CHEN; ZHANG, 2014). Another important aspect is that most of the times real-world data is rarely presented in research, which can also increase the gap between techniques and technologies developed and business applications (HAN et al., 2016; MANNING, 2015).

6.4 Model Performance

Following design science guidelines, to validate the artifact is important to make the contribution clear (HEVNER et al., 2004). In order to understand the data distribution among the categories to operationalize the proposed framweork, a neural network classifier was developed. Figure 31 presents the accuracy results in the training phase regarding the portion of test data. The model training considered, for each round (x-axis in Figure 31) a random sample of 75% of the data for training, and 25% for testing, what caused a

fluctuation in the accuracy measure. In addition, this sample contained only the comments with agreement among at least two human classifiers.

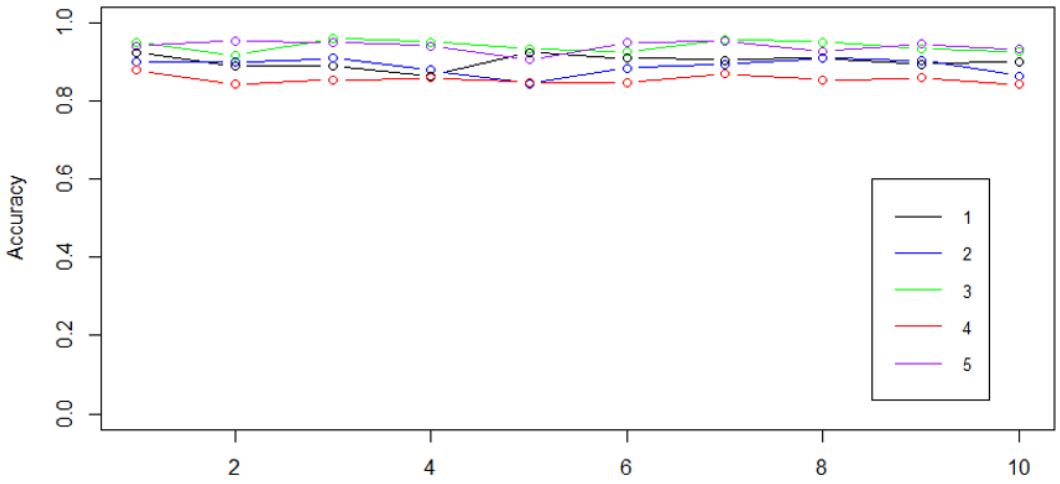

Figure 31 – Text Classification Accuracy

The results showed that most of the time the classifier achieved more than 80% of correct classification. Regarding decision-making inside organizations, being able to deliver the category that each comment belongs with less than 20% error still have room for improvement, but at the same time the automated analysis is quick and much less expensive than human resource. Also, machine learning models can improve with more data, available in a daily basis, and are free from subjectivity, giving a more consistent output (ASHTON; EVANGELOPOULOS; PRYBUTOK, 2014).

Table 9 – Accuracy Results

Round	**Cat 1**	**Cat 2**	**Cat 3**	**Cat 4**	**Cat 5**
1	0.925	0.880	0.952	0.880	0.946
2	0.891	0.875	0.917	0.844	0.955
3	0.891	0.900	0.961	0.854	0.946
4	0.866	0.870	0.952	0.859	0.937
5	0.925	0.850	0.935	0.849	0.928
6	0.910	0.880	0.926	0.849	0.946
7	0.905	0.890	0.957	0.870	0.955
8	0.910	0.905	0.952	0.854	0.946
9	0.896	0.895	0.935	0.859	0.937
10	0.900	0.875	0.926	0.844	0.942

To expand the model performance analysis, another two measures were calculated: Cohen's Kappa (κ) and Gwet's Agreement Coefficient (AC1). The objective was to better understand the model's performance given the possibility that the agreement between a human classifier and the machine can be due to chance, for Cohen's Kappa, and to analyze

the reliability level of agreement, for Gwet's AC1 (see Ullmann (2015) for details). The indicated values in Table 10 shows ten rounds average of those measures in each category. Values above 0.7 for AC1, and 0.4 for κ, are considered good performance (ULLMANN, 2015). All categories in all measures have achieved these thresholds, with exception of Categories 3 and 5 in κ values. This can be due to the fact that those categories had the least amount of comments, even that the values are not so distant from acceptable, considering that all evidence comes from real-world data.

Table 10 – Performance Measures

Category	**Accuracy**	**AC1**	κ
Cat 1	0.902	0.872	0.575
Cat 2	0.882	0.822	0.645
Cat 3	0.941	0.937	0.339
Cat 4	0.856	0.815	0.455
Cat 5	0.943	0.939	0.337

After the classifier was trained, we classified all collected comments from 2017, resulting in 2,582 comments. With those, we applied Topic Modeling, an unsupervised approach, in order to uncover the hidden patterns among comments classified in the same category. Being SERVQUAL a broad model for service industry, the objective was to understand the main words and expressions that belongs to each SERVQUAL category and represented that for the specific industry, in this case, hotels.

Although not all the words from the unlabeled comments were in the classification process, since language follow Zipf's law, that is, few words are used to express most of the opinions (ZIPF, 1949), this did not affect the final results. Even so, we compared the difference between the words in the unlabeled comments and the words in the labeled comments, used to train the classifier. As the results presented in Table 11 shows the new words represents around 5% of the total count in the vocabulary. Therefore 95% of the vocabulary from the unlabeled comments was already present in the labeled comments.

Table 11 – Vocabulary Covered

Category	**Terms Covered**	**Terms Lost**	**Total Frequency Terms Lost**	**% in all *Corpora***
1	3174	1533	1970	5.26%
2	3117	1640	2069	5.52%
3	3237	1520	1845	4.92%
4	3136	1621	2061	5.51%
5	3241	1516	1859	4.96%

From an utilitarian perspective (AKEN, 2005; HEVNER et al., 2004), this is an important aspect. Despite the smaller number of comments used to train the model against

the ones classified by the machine, it is possible to see that over 95% of total frequency have appeared in the training dataset. This implies that the model developed here allows to continuously classify new sets of comments. In a managerial and dynamic environment, this is a key aspect, allowing the framework to work in an automated and human-independent format, saving time for actual analysis.

6.5 External Validation

Considering external validity, three interviews were conducted with managers of one hotel of each category presented in the city, i.e., two, three and four stars, namely hotels H13, H17 and H43 (see Table 72 on page 72). The interview process consisted in presenting the graphics resulting from topic modeling, classification, market and sentiment analysis. All three of them stated that, although eWOM platforms like TripAdvisor delivered some information considering only the comments from each hotel, the model contributed by demonstrating the market perspective, allowing to compare themselves with their main known (and for some, unknown) competitors.

The executive in charge of social media from hotel H17 stated that most Brazilian softwares do not deliver a broad text understanding, and suggested that the analysis could be made in different languages. In addition, this manager commented about the model usefulness for an internal evaluation by the manager considering weaknesses and strengths from his or her hotel, specially given the comparative market analysis. The manager from H43 highlighted the specific points that the model used to compare one hotel with the others. Another aspect mentioned was its summarization capacity, that allowed to understand a great amount of information with an easy visualization (word clouds). Finally, H13 supervisor asked for the hotel position in the graphic, confirming that the staff empathy (see Figure 27 on page 83) was among one of the most commented in his perception. Also, the novelty format from the data visualization, considering previous tools available, and specially the sentiment analysis, were highlighted as positive aspects of the model.

One final aspect about the framework external validity is the demonstrated interest by an investor to develop a commercial product based on the framework capacity to contribute with analytics and the interviewed managers support. In the time of this dissertation deposit, there are parallel meetings aiming to build a startup firm that can give to the framework an user interface. Other e WOM plattforms are being considered to be integrated in the commercial product, as well.

7 Conclusions

The objective of this dissertation was to develop a framework to contribute with the effective use of text data in decision-making process. Understanding this as a multi-disciplinary field, the set of works presented aimed to serve as blocks to build a bridge between several developed techniques and managers' needs given the analytics movement. An important aspect of this movement is the increasing volume of text data, and the urge to use such a rich data to support decisions related with customers, employees, partners, among others stakeholders.

Instead of playing the 'Kaggle game' (MANNING, 2015, p 702), it is important that more research have a problem-based, with real-world data and real business problems, a role that OR area have the possibility to play with success. For that reason, organizing and systematizing a process that allows to uncover value from unstructured real-world data has always been in this dissertation radar. Focusing in using text, a framework was developed and tested in order to foster the usage of this kind of data inside organizations. In this sense, the objective was also to answer Mortenson, Doherty e Robinson (2015) call, addressing some of the points highlighted as a research agenda.

Text data is presented in many business interfaces, from customer relations to human resource management. We choose to work with consumer data mainly for two reasons. The first one is the fast growth that different word-of-mouth platforms are witnessing. Reputation is a fundamental currency whether in on-line or off-line business, which might be attached to the popularity of these platforms (VALDIVIA; LUZÓN; HERRERA, 2017). The second one is the power that this data can deliver to organizations. Being open and public, not only more information about a singular business, but also a broad view about the other players and a comparison is possible, always taking the customer point-of-view, in a semantically and meaningfully format (TANG; GUO, 2015).

Listening the voice of the customer can be leveraged through the framework developed and implemented in this dissertation. Even for small organizations, it is possible to follow the steps uncovering a semantic structure able to enrich the knowledge about a specific market (XIANG et al., 2015). This focus can help organizations to adopt text analysis techniques, since the documents analyzed are highly correlated with their daily-based activities, and that the independence from a human reader can potentially avoid bias and improve knowledge dissipation throughout the entire organization.

Another objectivity that is possible to achieve is from specific situations and the relation with a broad picture. Most of the decision-making process involve executives that are not in the front-office, while three from the five categories from SERVQUAL were

strongly related with the staff, as the analysis of the latent topics revealed. Analyzing the business performance through the voice of the customer can allow managers to better understand whether an aspect in one comment was an exception, is something frequent or is an urgent situation. Having historical data (available in the same data source) it is possible to improve the analysis, mapping the expressions that have grown-up or disappeared being important tools to hotel operators seeking to improve, cost-effectively, their operations (HAN et al., 2016).

Different from previous approaches, we did not seek to build a domain fixed vocabulary (AGUWA; OLYA; MONPLAISIR, 2017; CARRASCO et al., 2012). Understanding that the reviews contain contextualized perceptions (XIANG et al., 2015), a single fixed vocabulary would not help to develop a framework that can be adopted in different contexts. Another issue is that vocabulary changes over the time, as well as regular expressions and the market itself. An example is wi-fi offer, that have changed over the years (from non-existent to a paid option and now an essential service). With recent development in unsupervised methods for NLP (MANNING, 2015), models that work with the main expressions that emerge from the data are more flexible and allow to use the same methods to uncover different latent semantic relations.

Within the information-technology area, it is possible to realize that information has a growing value for any business-related context. Tourism and hospitality sector have always needed customer's feedback in order to improve their services. As an intangible asset, getting to know opinions, positions, what could have been better, among others, is important not only for customer retention, but also because of word-of-mouth effect. With social media explosion, this effect grew exponentially, since that millions of other customers have also gained access to that feedback, that were transformed in public feedback through platforms like the one used to evaluate the framework in this work, TripAdvisor.

Additionally, this public opinion can be interpreted as eWOM, meaning that it will be stored in an online platform, allowing indexing and availability in a much larger scale (CHAVES, 2003). However, despite the relevance of consumer data to support the decision-making process, the analysis of such data is not always performed because of the difficulty of dealing with a large amount of consumer information (HE et al., 2017).

In order to validate the proposed framework, we apply it to analyze an expressive amount of data in the context of the hospitality industry, having collected and treated over 20,000 comments from 2011 to 2017. LSA was used in order to extract the main themes, or topics, from this set of comments. Besides treating and analyzing data, another contribution was the development of an automated script in R that, with just the hotel's URL, can scrap all comments already posted. This code, as well as the data used here, are available also in an online repository (MARCOLIN, 2017). In addition, the presentation of these findings aims to help managers understand the importance of analyzing a large

amount of data to support their decisions.

For this reason, this dissertation contributes with that task in business analytics scenario, demonstrating different forms to extract knowledge from this public available data considering a managerial and a methodological approach (HAN et al., 2016). For that, the three specif objectives were achieved: (1) an automated Webscrapper was developed in R. With just the organization's URL, this tool can scrap all comments already posted in a eWOM platform, with some adjustments. Also, it is already fine-tuned for TripAdvisor. Developed in an open-source code, it is available even for smaller organizations that might not have enough resources to acquire proprietary solutions; (2) a solution based on main terms connected with highest singular-values from LSA was developed, allowing to monitor trends in text data; (3) a model that classifies comments was implemented and tested with TripAdvisor data, allowing to further uses them for deliver a market and sentiment analysis; and (4), the evaluation process shows that over 80% of accuracy was achieved, and that even with a single set of comments belonging to train, the classifiers was able to handle about 95% of term-frequency given new unlabeled data. Besides, the framework was considered useful in many aspects from hotel manager's perspective, and received an investor green light to further development as a startup project.

Having an objective to contribute using real-world data poses not only an opportunity but also brings limitations for a few reasons, and for that, result should be interpreted with care. First, there is no comparison-base given the performance reported, since the data is original and unique, although this is also an opportunity to contribute with practitioners. Second, results are limited to the data collection context, all business analyzed were located in the same hometown from the principal investigator, that have knowledge from local culture and reference points. In addition, the data labeling process was done by a multidisciplinary and international team, also focusing in reducing the bias that might be generated from a single-location analysis. Third, the data source was limited to consumer reviews from hotels, what might have narrowed the results only for this industry. Nevertheless, the SERVQUAL scale can be applied to different service industries that also have consumer reviews in the same platform (like restaurants and touristic points), allowing to expand the methodology to another private and public organizations.

As new techniques and new ways to uncover value from text data arises constantly, the business analytics research field has to be interdisciplinary, as should be all business analytics teams inside organizations. A business background is important to understand what each data source can deliver, but a quantitative background, focused in optimization methods, is also fundamental, since the majority of techniques demands not only to solve, but to structure a problem in cost-effective and informative way (RANYARD; FILDES; HU, 2015). In this sense, OR researchers and practitioners should focus the attention to this area. Given text data, there are a diversity of open-problems, specially when related

with real-world data applications, that should count with OR community participation for further solutions development:

- Even with improved models and new developments, the output from LSA and others Topic Modeling models were found hard to interpret (VISINESCU; EVANGELOPOULOS, 2014; LANDAUER, 2007). This may end up, in a business position, providing less value within text data. From that perspective, to pursue for improved topic coherence is one direction that can still be investigated, in order to approximate the topics from human judgment of interpretability (O'CALLAGHAN et al., 2015; RÖDER; BOTH; HINNEBURG, 2015; MIMNO et al., 2011).

- Machines can only rely on what they see, which is strongly related with word frequencies and co-occurrences. On the other hand, humans can activate several related concepts, memories and sensory experiences. One step to improve this understanding is to approximate algorithms to the human-mind structure. The neural network applied in this work is part of this movement, since it aims to better understand patterns, beyond only word distribution. In this sense, in a context where user-generated content is drawing in itself, OR can help to get to the next NLP curve by improving non-linear models that aim to see beyond than a bag-of-words (CAMBRIA; WHITE, 2014).

- A growing research area that can improve with the participation of the OR community is Deep Learning. Deep Learning can be defined as a group of methods that use multiple levels of representation, and by composing simple (but non-linear) functions transform different layers of inputs into more abstract, high-level representations. With enough layers, in neural networks very complex functions can be learned, allowing to recognize complex patterns such as language (KRAUSS; DO; HUCK, 2017; LECUN; BENGIO; HINTON, 2015). Although state-of-the-art performance and decreasingly low-rates are being achieved by Deep Learning methods in different areas, higher-level NLP problems have not improved in the same rate, what poses language understanding as the next Deep Learning challenge (MANNING, 2015).

- Since text data can be seen as sensor for measuring perception (ZHAO, 2013), not only a high volume of text can be analyzed, uncovering a group perception, but also the other way around can be performed, discovering new groups based on similar perceptions. With this approach, documents can be used as a semantic representation of their authors, and modeling profiling can count not only with socio-demographic characteristics like gender and age, but also with a more rich information about the customer, self-provided. This is convergent with the personalization trend, that can improve service experience to a new level (TANG; GUO, 2015).

Bibliography

ACITO, F.; KHATRI, V. *Business analytics: Why now and what next?* [S.l.]: Elsevier, 2014. 565-570 p. Citado na página 17.

AGGARWAL, C.; ZHAI, C. *Mining Text Data*. [S.l.]: Springer Science & Business Media, 2012. Citado 7 vezes nas páginas 18, 25, 28, 31, 33, 47, and 60.

AGUWA, C.; OLYA, M. H.; MONPLAISIR, L. Modeling of fuzzy-based voice of customer for business decision analytics. *Knowledge-Based Systems*, Elsevier, v. 125, p. 136–145, 2017. Citado 5 vezes nas páginas 19, 43, 47, 61, and 92.

AKEN, J. E. V. Management research as a design science: Articulating the research products of mode 2 knowledge production in management. *British journal of management*, Wiley Online Library, v. 16, n. 1, p. 19–36, 2005. Citado 4 vezes nas páginas 20, 53, 61, and 88.

ALEGRE, P. M. de P. *BEMTUR - Boletim Estatístico Municipal do Turismo em Porto Alegre*. 2017. "<http://www2.portoalegre.rs.gov.br/turismo/default.php?p_secao=336>". Accessed: 2017-07-01. Citado 2 vezes nas páginas 11 and 70.

ASHTON, T.; EVANGELOPOULOS, N.; PRYBUTOK, V. Extending monitoring methods to textual data: a research agenda. *Quality & Quantity*, Springer, v. 48, n. 4, p. 2277–2294, 2014. Citado 12 vezes nas páginas 11, 18, 19, 20, 36, 39, 45, 47, 62, 63, 65, and 87.

ASSAF, A. G. et al. The effects of customer voice on hotel performance. *International Journal of Hospitality Management*, Elsevier, v. 44, p. 77–83, 2015. Citado na página 69.

BAEZA-YATES, R.; RIBEIRO-NETO, B. et al. *Modern information retrieval*. [S.l.]: ACM press New York, 2011. v. 463. Citado na página 35.

BARTON, D.; COURT, D. Making advanced analytics work for you. *Harvard business review*, v. 90, n. 10, p. 78–83, 2012. Citado 4 vezes nas páginas 17, 18, 19, and 58.

BAYRAK, T. A review of business analytics: a business enabler or another passing fad. *Procedia-Social and Behavioral Sciences*, Elsevier, v. 195, p. 230–239, 2015. Citado na página 17.

BECKER, J. L. Estatística aplicada a administração volume 2. *Notas de Aula*, 2016. Citado 3 vezes nas páginas 11, 37, and 38.

BERGAMASCHI, S.; PO, L. Comparing lda and lsa topic models for content-based movie recommendation systems. In: SPRINGER. *International Conference on Web Information Systems and Technologies*. [S.l.], 2014. p. 247–263. Citado 6 vezes nas páginas 36, 37, 38, 61, 74, and 80.

BISHOP, C. M. *Neural networks for pattern recognition*. [S.l.]: Oxford university press, 1995. Citado 2 vezes nas páginas 42 and 43.

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. Citado na página 33.

BOSE, R. Advanced analytics: opportunities and challenges. *Industrial Management & Data Systems*, Emerald Group Publishing Limited, v. 109, n. 2, p. 155–172, 2009. Citado na página 64.

CAMBRIA, E.; WHITE, B. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, IEEE, v. 9, n. 2, p. 48–57, 2014. Citado 3 vezes nas páginas 41, 67, and 94.

CANTALLOPS, A. S.; SALVI, F. New consumer behavior: A review of research on ewom and hotels. *International Journal of Hospitality Management*, Elsevier, v. 36, p. 41–51, 2014. Citado 2 vezes nas páginas 18 and 48.

CARRASCO, R. A. et al. A linguistic multicriteria decision-making model applied to hotel service quality evaluation from web data sources. *International Journal of Intelligent Systems*, Wiley Online Library, v. 27, n. 7, p. 704–731, 2012. Citado 8 vezes nas páginas 20, 62, 68, 69, 71, 80, 81, and 92.

CHAREYRON, G.; DA-RUGNA, J.; RAIMBAULT, T. Big data: A new challenge for tourism. In: IEEE. *Big data (Big data), 2014 IEEE international conference on*. [S.l.], 2014. p. 5–7. Citado na página 68.

CHAVES, M. S. Um estudo e apreciação sobre algoritmos de stemming para a língua portuguesa. *IX Jornadas Iberoamericanas de Informática*, 2003. Citado na página 92.

CHEN, C. P.; ZHANG, C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, Elsevier, v. 275, p. 314–347, 2014. Citado 4 vezes nas páginas 47, 58, 76, and 86.

CHOI, Y.; LEE, H. Data properties and the performance of sentiment classification for electronic commerce applications. *Information Systems Frontiers*, Springer, p. 1–20, 2017. Citado 2 vezes nas páginas 41 and 44.

CODY, W. F. et al. The integration of business intelligence and knowledge management. *IBM systems journal*, IBM, v. 41, n. 4, p. 697–713, 2002. Citado na página 47.

CRAIN, S. P. et al. Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond. In: *Mining text data*. [S.l.]: Springer, 2012. p. 129–161. Citado 2 vezes nas páginas 36 and 39.

DAVENPORT, T. H.; DYCHÉ, J. Big data in big companies. *International Institute for Analytics*, 2013. Citado 3 vezes nas páginas 17, 18, and 58.

DEDEOĞLU, B. B.; DEMIRER, H. Differences in service quality perceptions of stakeholders in the hotel industry. *International Journal of Contemporary Hospitality Management*, Emerald Group Publishing Limited, v. 27, n. 1, p. 130–146, 2015. Citado na página 69.

DEERWESTER, S. et al. Indexing by latent semantic analysis. *Journal of the American society for information science*, American Documentation Institute, v. 41, n. 6, p. 391, 1990. Citado 3 vezes nas páginas 33, 34, and 61.

DEZA, M. M.; DEZA, E. Encyclopedia of distances. In: *Encyclopedia of Distances*. [S.l.]: Springer, 2009. p. 1–583. Citado na página 29.

DICKINGER, A. The trustworthiness of online channels for experience-and goal-directed search tasks. *Journal of Travel Research*, SAGE Publications Sage CA: Los Angeles, CA, v. 50, n. 4, p. 378–391, 2011. Citado na página 48.

DOUVEN, I.; MEIJS, W. Measuring coherence. *Synthese*, Springer, v. 156, n. 3, p. 405–425, 2007. Citado na página 28.

DRESCH, A.; LACERDA, D. P.; JR, J. A. V. A. *Design science research: a method for science and technology advancement*. [S.l.]: Springer, 2014. Citado 5 vezes nas páginas 13, 20, 51, 52, and 53.

DUAN, W. et al. Exploring the impact of social media on hotel service performance: A sentimental analysis approach. *Cornell Hospitality Quarterly*, Sage Publications Sage CA: Los Angeles, CA, v. 57, n. 3, p. 282–296, 2016. Citado na página 44.

EVANGELOPOULOS, N. Citing taylor: Tracing taylorism's technical and sociotechnical duality through latent semantic analysis. *Journal of Business and Management*, Journal of Business and Management, v. 17, n. 1, p. 57, 2011. Citado na página 39.

FEINERER, I. Text mining package. In: *CRAN R Project*. [S.l.: s.n.], 2017. p. 62. Citado na página 39.

GORRY, G. A.; WESTBROOK, R. A. Can you hear me now? learning from customer stories. *Business horizons*, Elsevier, v. 54, n. 6, p. 575–584, 2011. Citado 4 vezes nas páginas 18, 19, 45, and 47.

GREENE, D.; CROSS, J. P. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, Cambridge University Press, v. 25, n. 1, p. 77–94, 2017. Citado 2 vezes nas páginas 60 and 62.

GRIFFIN, A.; HAUSER, J. R. The voice of the customer. *Marketing science*, INFORMS, v. 12, n. 1, p. 1–27, 1993. Citado 2 vezes nas páginas 45 and 46.

HAN, H. J. et al. What guests really think of your hotel: Text analytics of online customer reviews. 2016. Citado 6 vezes nas páginas 19, 60, 65, 86, 92, and 93.

HE, W. et al. Application of social media analytics: a case of analyzing online hotel reviews. *Online Information Review*, Emerald Publishing Limited, v. 41, n. 7, p. 921–935, 2017. Citado 2 vezes nas páginas 18 and 92.

HEVNER, R. et al. Design science in information systems research. *MIS quarterly*, Springer, v. 28, n. 1, p. 75–105, 2004. Citado 4 vezes nas páginas 20, 53, 86, and 88.

HOFMANN, T. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, Springer, v. 42, n. 1-2, p. 177–196, 2001. Citado na página 33.

HORNER, S.; SWARBROOKE, J. et al. *Consumer behaviour in tourism*. [S.l.]: Routledge, 2016. Citado 2 vezes nas páginas 67 and 69.

JACOBS, A. D. et al. Word2vec inversion and traditional text classifiers for phenotyping lupus. *BMC medical informatics and decision making*, BioMed Central, v. 17, n. 1, p. 126, 2017. Citado na página 41.

KIM, W. G. et al. Social media review rating versus traditional customer satisfaction: Which one has more incremental predictive power in explaining hotel performance? *International Journal of Contemporary Hospitality Management*, Emerald Publishing Limited, v. 29, n. 2, p. 784–802, 2017. Citado 4 vezes nas páginas 41, 47, 68, and 70.

KNUTSON, B. et al. Lodgserv: A service quality index for the lodging industry. *Hospitality Research Journal*, ICHRIE Sage CA: Los Angeles, CA, v. 14, n. 2, p. 277–284, 1990. Citado 5 vezes nas páginas 46, 69, 71, 80, and 81.

KOBAYASHI, V. B. et al. Text mining in organizational research. *Organizational Research Methods*, p. 1–33, 2017. Citado 10 vezes nas páginas 18, 20, 28, 30, 33, 58, 59, 61, 63, and 65.

KRAUSS, C.; DO, X. A.; HUCK, N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. *European Journal of Operational Research*, Elsevier, v. 259, n. 2, p. 689–702, 2017. Citado 2 vezes nas páginas 41 and 94.

KULKARNI, S. S.; APTE, U. M.; EVANGELOPOULOS, N. E. The use of latent semantic analysis in operations management research. *Decision Sciences*, Wiley Online Library, v. 45, n. 5, p. 971–994, 2014. Citado 4 vezes nas páginas 33, 37, 38, and 60.

KULKARNI, S. S.; APTE, U. M.; EVANGELOPOULOS, N. E. The use of latent semantic analysis in operations management research. *Decision Sciences*, Wiley Online Library, v. 45, n. 5, p. 971–994, 2014. Citado 2 vezes nas páginas 60 and 74.

LADHARI, R.; MICHAUD, M. ewom effects on hotel booking intentions, attitudes, trust, and website perceptions. *International Journal of Hospitality Management*, Elsevier, v. 46, p. 36–45, 2015. Citado na página 48.

LANDAUER, T. K. Lsa as a theory of meaning. *Handbook of latent semantic analysis*, Mahwah, NJ: Lawrence Erlbaum Associates, p. 3–34, 2007. Citado 2 vezes nas páginas 20 and 94.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, Nature Research, v. 521, n. 7553, p. 436–444, 2015. Citado 2 vezes nas páginas 41 and 94.

LEE, D. D.; SEUNG, H. S. Algorithms for non-negative matrix factorization. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2001. p. 556–562. Citado na página 33.

LEE, H.; HAN, J.; SUH, Y. Gift or threat? an examination of voice of the customer: The case of mystarbucksidea. com. *Electronic Commerce Research and Applications*, Elsevier, v. 13, n. 3, p. 205–219, 2014. Citado 4 vezes nas páginas 20, 47, 59, and 63.

LEE, M. et al. Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review website: A text mining approach. *International Journal of Contemporary Hospitality Management*, Emerald Publishing Limited, v. 29, n. 2, p. 762–783, 2017. Citado na página 69.

LEUNG, D. et al. Social media in tourism and hospitality: A literature review. *Journal of Travel & Tourism Marketing*, Taylor & Francis, v. 30, n. 1-2, p. 3–22, 2013. Citado 3 vezes nas páginas 18, 47, and 69.

MADDULAPALLI, A. K.; YANG, J.-B.; XU, D.-L. Estimation, modeling, and aggregation of missing survey data for prioritizing customer voices. *European Journal of Operational Research*, Elsevier, v. 220, n. 3, p. 762–776, 2012. Citado na página 47.

MANEVITZ, L.; YOUSEF, M. One-class document classification via neural networks. *Neurocomputing*, Elsevier, v. 70, n. 7, p. 1466–1481, 2007. Citado 3 vezes nas páginas 20, 42, and 73.

MANNING; RHAGAVAN; SCHUTZE. *An Introduction to Information Retrieval*. [S.l.]: Cambridge University Press, 2009. Citado 7 vezes nas páginas 11, 18, 25, 26, 38, 39, and 71.

MANNING, C. D. Last words: Computational linguistics and deep learning. *Computational Linguistics*, MIT Press, v. 41, n. 4, p. 701–707, 2015. Citado 7 vezes nas páginas 34, 41, 67, 86, 91, 92, and 94.

MARCOLIN, C. *GitHub Project*. 2017. "<https://github.com/carlamarcolin/topicmodelling>". Created: 2017-05. Citado 3 vezes nas páginas 41, 58, and 92.

MARCOLIN, C.; BECKER, J. Exploring latent semantic analysis in a big data (base). In: . [S.l.: s.n.]. Citado na página 59.

MARTIN, D. I.; BERRY, M. Mathematical foundations behind latent semantic analysis. In: LANDAUER, T. K. et al. (Ed.). *Handbook or Latent Semantic Analysis*. Oxfordshire: Routledge, 2011. cap. 2, p. 35–56. Citado 4 vezes nas páginas 11, 35, 36, and 37.

MCAFEE, A. et al. Big data: The management revolution. *Harvard Bus Rev*, v. 90, n. 10, p. 61–67, 2012. Citado 5 vezes nas páginas 17, 19, 20, 59, and 63.

MIMNO, D. et al. Optimizing semantic coherence in topic models. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. [S.l.], 2011. p. 262–272. Citado 2 vezes nas páginas 29 and 94.

MIRKIN, B. *Core concepts in data analysis: summarization, correlation and visualization*. [S.l.]: Springer Science & Business Media, 2011. Citado 2 vezes nas páginas 20 and 64.

MOLINILLO, S. et al. Hotel assessment through social media: The case of trip advisor. *Tourism & Management Studies*, Universidade do Algarve, v. 12, n. 1, 2016. Citado na página 68.

MORTENSON, M. J.; DOHERTY, N. F.; ROBINSON, S. Operational research from taylorism to terabytes: A research agenda for the analytics age. *European Journal of Operational Research*, Elsevier, v. 241, n. 3, p. 583–595, 2015. Citado 5 vezes nas páginas 17, 19, 20, 59, and 91.

MYATT, G. J. *Making sense of data: a practical guide to exploratory data analysis and data mining*. [S.l.]: John Wiley & Sons, 2007. Citado 2 vezes nas páginas 64 and 65.

NEWMAN, D. et al. Automatic evaluation of topic coherence. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. [S.l.], 2010. p. 100–108. Citado na página 29.

NIELSEN, F. Å. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011. Citado 2 vezes nas páginas 73 and 83.

O'CALLAGHAN, D. et al. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, Elsevier, v. 42, n. 13, p. 5645–5657, 2015. Citado 3 vezes nas páginas 29, 30, and 94.

O'LEARY, B. S. et al. Selecting the best and brightest: Leveraging human capital. *Human resource management*, Wiley Online Library, v. 41, n. 3, p. 325–340, 2002. Citado na página 39.

OLMOS, R. et al. Transforming lsa space dimensions into a rubric for an automatic assessment and feedback system. *Information Processing & Management*, Elsevier, v. 52, n. 3, p. 359–373, 2016. Citado na página 39.

PARASURAMAN, A.; ZEITHAML, V. A.; BERRY, L. L. Servqual: A multiple-item scale for measuring consumer perc. *Journal of retailing*, New York University, v. 64, n. 1, p. 12, 1988. Citado 5 vezes nas páginas 45, 46, 71, 80, and 81.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Empirical Methods in Natural Language Processing (EMNLP)*. [s.n.], 2014. p. 1532–1543. Disponível em: <http://www.aclweb.org/anthology/D14-1162>. Citado na página 34.

PEREZ-ARANDA, J.; ANAYA-SANCHEZ, R.; RUIZALBA, J. Predictors of review sites usage in hotels. *Tourism & Management Studies*, Universidade do Algarve, v. 13, n. 2, 2017. Citado na página 69.

PORTER, M. F. An algorithm for suffix stripping. *Program*, MCB UP Ltd, v. 14, n. 3, p. 130–137, 1980. Citado 2 vezes nas páginas 11 and 27.

PROVOST, F.; FAWCETT, T. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. Alta Books, 2016. ISBN 9781449374280. Disponível em: <https://books.google.com.br/books?id=4ZctAAAAQBAJ>. Citado 3 vezes nas páginas 17, 20, and 59.

RANYARD, J. C.; FILDES, R.; HU, T.-I. Reassessing the scope of or practice: The influences of problem structuring methods and the analytics movement. *European Journal of Operational Research*, Elsevier, v. 245, n. 1, p. 1–13, 2015. Citado 2 vezes nas páginas 20 and 93.

RÖDER, M.; BOTH, A.; HINNEBURG, A. Exploring the space of topic coherence measures. In: ACM. *Proceedings of the eighth ACM international conference on Web search and data mining*. [S.l.], 2015. p. 399–408. Citado 3 vezes nas páginas 28, 29, and 94.

SAMARASINGHE, S. *Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition*. [S.l.]: CRC Press, 2016. Citado na página 42.

SCHUCKERT, M.; LIU, X.; LAW, R. Hospitality and tourism online reviews: Recent trends and future directions. *Journal of Travel & Tourism Marketing*, Taylor & Francis, v. 32, n. 5, p. 608–621, 2015. Citado 2 vezes nas páginas 18 and 48.

SILGE, J.; ROBINSON, D. tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, The Open Journal, v. 1, n. 3, 2016. Disponível em: <http://dx.doi.org/10.21105/joss.00037>. Citado na página 73.

SIMON, H. A. *The sciences of the artificial*. [S.l.]: MIT press, 1996. Citado na página 51.

SORDI, J. O. D.; AZEVEDO, M. Carvalho de; MEIRELES, M. A pesquisa design science no brasil segundo as publicações em administração da informação. *JISTEM: Journal of Information Systems and Technology Management*, Universidade de São Paulo, v. 12, n. 1, 2015. Citado 2 vezes nas páginas 51 and 54.

SPANGLER, S.; KREULEN, J. *Mining the Talk: Unlocking the Business Value in Unstructured Information (Adobe Reader)*. [S.l.]: Pearson Education, 2007. Citado 4 vezes nas páginas 11, 45, 46, and 67.

SPARKS, B. A.; BROWNING, V. The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management*, Elsevier, v. 32, n. 6, p. 1310–1323, 2011. Citado 3 vezes nas páginas 47, 67, and 68.

STEVENS, K. et al. Exploring topic coherence over many models and many topics. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. [S.l.], 2012. p. 952–961. Citado na página 29.

TANEV, S.; LIOTTA, G.; KLEISMANTAS, A. A business intelligence approach using web search tools and online data reduction techniques to examine the value of product-enabled services. *Expert Systems with Applications*, Elsevier, v. 42, n. 21, p. 7582–7600, 2015. Citado 3 vezes nas páginas 39, 47, and 58.

TANG, C.; GUO, L. Digging for gold with a simple tool: Validating text mining in studying electronic word-of-mouth (ewom) communication. *Marketing Letters*, Springer, v. 26, n. 1, p. 67–80, 2015. Citado 3 vezes nas páginas 47, 91, and 94.

THORLEUCHTER, D.; POEL, D. V. D. Predicting e-commerce company success by mining the text of its publicly-accessible website. *Expert Systems with Applications*, Elsevier, v. 39, n. 17, p. 13026–13034, 2012. Citado na página 39.

TINKLER, S.; WOODS, J. The readability of principles of macroeconomics textbooks. *The Journal of Economic Education*, Taylor & Francis, v. 44, n. 2, p. 178–191, 2013. Citado na página 39.

ULLMANN, T. D. *Automated detection of reflection in texts. A machine learning based approach*. Tese (Doutorado) — The Open University, 2015. Citado na página 88.

VALDIVIA, A.; LUZÓN, M. V.; HERRERA, F. Sentiment analysis in tripadvisor. *IEEE Intelligent Systems*, IEEE, v. 32, n. 4, p. 72–77, 2017. Citado 6 vezes nas páginas 18, 44, 47, 68, 69, and 91.

VISINESCU, L. L.; EVANGELOPOULOS, N. Orthogonal rotations in latent semantic analysis: An empirical study. *Decision Support Systems*, Elsevier, v. 62, p. 131–143, 2014. Citado 2 vezes nas páginas 34 and 94.

WANG, Q. et al. Group matrix factorization for scalable topic modeling. In: ACM. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. [S.l.], 2012. p. 375–384. Citado na página 34.

WANG, Q. et al. Regularized latent semantic indexing: A new approach to large-scale topic modeling. *ACM Transactions on Information Systems (TOIS)*, ACM, v. 31, n. 1, p. 5, 2013. Citado na página 34.

WICKHAM, H. *RVest Package Demonstration*. 2015. "<https://github.com/hadley/rvest/blob/master/demo/tripadvisor.R>". Accessed: 2016-10-30. Citado na página 58.

WIEMER-HASTINGS, P.; WIEMER-HASTINGS, K.; GRAESSER, A. Improving an intelligent tutor's comprehension of students with latent semantic analysis. In: AMSTERDAM: IOS PRESS. *Artificial intelligence in education*. [S.l.], 1999. v. 99. Citado na página 39.

WILD, F. An lsa package for r. In: *CRAN R Project*. [S.l.: s.n.], 2015. p. 12. Citado 3 vezes nas páginas 11, 39, and 41.

WILD, F. *Learning analytics in R with SNA, LSA, and MPIA*. [S.l.]: Springer, 2016. Citado 4 vezes nas páginas 11, 20, 40, and 61.

WILD, F. et al. Parameters driving effectiveness of automated essay scoring with lsa. *Proceedings of the 9th CAA*, © Loughborough University, 2005. Citado 5 vezes nas páginas 37, 39, 40, 74, and 75.

WILLETT, P. The porter stemming algorithm: then and now. *Program*, Emerald Group Publishing Limited, v. 40, n. 3, p. 219–223, 2006. Citado na página 27.

XIANG, Z. et al. What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management*, Elsevier, v. 44, p. 120–130, 2015. Citado 3 vezes nas páginas 20, 91, and 92.

YE, Q.; LAW, R.; GU, B. The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, Elsevier, v. 28, n. 1, p. 180–182, 2009. Citado na página 68.

YEN, C.-L. A.; TANG, C.-H. H. Hotel attribute performance, ewom motivations, and media choice. *International Journal of Hospitality Management*, Elsevier, v. 46, p. 79–88, 2015. Citado 2 vezes nas páginas 68 and 70.

ZHAO, Y. *R and data mining: Examples and case studies*. [S.l.]: Academic Press, 2013. Citado 4 vezes nas páginas 20, 39, 45, and 94.

ZIPF, G. K. *The psycho-biology of language*. [S.l.]: Cambridge: MIT Press, 1935. Citado 3 vezes nas páginas 19, 20, and 33.

ZIPF, G. K. *Human behavior and the principle of least effort*. [S.l.]: Addison-Wesley Press, 1949. Citado 5 vezes nas páginas 19, 20, 33, 77, and 88.

Appendix

APPENDIX A – Webscrapper

The Appendix full section presents all functions developed in order to capture, treat and analyze text data given the TripAdvisor platform. The first one presented is WebScrapper: WebScrapper to collect comments from TripAdvisor. Scrap comments considering the following parameters:

- List of main URL from each hotel.
- List of sequence of comments (pages available x comments per page).

```
#Webscrapping TripAdvisor

# load libraries
library(RCurl)
library(XML)
library(lubridate)

options(stringsAsFactors=FALSE)
getOnePage=function(urllink){

# get html page content
doc=htmlTreeParse(urllink,useInternalNodes=TRUE)

## get node sets
# review id
ns_id=getNodeSet(doc,"//div[@class='quote isNew' or @class='quote ' or
    @class='quote']/a[@href]")
# top quote for a review
ns_topquote=getNodeSet(doc,"//div[@class='quote isNew' or @class='quote '
     or @class='quote']/a[@href]/span")
# get partial entry for review that shows in the page
ns_partialentry=getNodeSet(doc,"//div[@class='col2of2']//p[@class='
    partial_entry'][1]")
# date of rating
ns_ratingdt=getNodeSet(doc,"//div[@class='col2of2']//span[@class='
    ratingDate relativeDate' or @class='ratingDate']")
# rating (number of stars)
ns_rating=getNodeSet(doc,"//div[@class='col2of2']//span[@class='rate
    sprite-rating_s rating_s']/img[@alt]")

# get actual values extracted from node sets
# review id
id=sapply(ns_id,function(x) xmlAttrs(x)["id"])
```

```
29 # top quote for the review
30 topquote=sapply(ns_topquote,function(x) xmlValue(x))
31 # rating date (couple of formats seem to be used and hence a and b below)
32 ratingdta=sapply(ns_ratingdt,function(x) xmlAttrs(x)["title"])
33 ratingdtb=sapply(ns_ratingdt,function(x) xmlValue(x))
34 # rating (number of stars)
35 rating=sapply(ns_rating,function(x) xmlAttrs(x)["alt"])
36 # partial entry for review
37 partialentry=sapply(ns_partialentry,function(x) xmlValue(x))
38
39 # get rating date in date format
40 ratingdt.pick=ratingdta
41 ratingdt.pick[is.na(ratingdta)]=ratingdtb[is.na(ratingdta)]
42 ratingdt=mdy(gsub("Reviewed ","",ratingdt.pick))
43
44 # put all the fields in a dataframe
45 dfrating=data.frame(id=id,topquote=topquote,ratingdt=ratingdt,rating=
      rating,partialentry=partialentry)
46 dfrating$ratingnum=as.numeric(substr(dfrating$rating,1,1),1,1)
47 return(dfrating)
48 }
49
50 #Sample URL list from TripAdvisor
51 urlmainlist=c(
52 lagvivmon = "https://www.tripadvisor.com.br/Hotel_Review-g303546-d5006148
      -Reviews-Hotel_Laghetto_Viverone_Moinhos-Porto_Alegre_State_of_Rio_
      Grande_do_Sul.html",
53 porretbou = "https://www.tripadvisor.com.br/Hotel_Review-g303546-d7153353
      -Reviews-Porto_Retro_Flat_Boutique-Porto_Alegre_State_of_Rio_Grande_
      do_Sul.html",
54 devpripoa = "https://www.tripadvisor.com.br/Hotel_Review-g303546-d306290-
      Reviews-Hotel_Deville_Prime_Porto_Alegre-Porto_Alegre_State_of_Rio_
      Grande_do_Sul.html",
55 ibiscentro = "https://www.tripadvisor.com.br/Hotel_Review-g303546-
      d7392609-Reviews-Ibis_Styles_Porto_Alegre_Centro-Porto_Alegre_State_
      of_Rio_Grande_do_Sul.html",
56 sheraton = "https://www.tripadvisor.com.br/Hotel_Review-g303546-d304505-
      Reviews-Sheraton_Porto_Alegre_Hotel-Porto_Alegre_State_of_Rio_Grande_
      do_Sul.html",
57 radisson = "https://www.tripadvisor.com.br/Hotel_Review-g303546-d300916-
      Reviews-Radisson_Porto_Alegre-Porto_Alegre_State_of_Rio_Grande_do_Sul
      .html",
58 ritter = "https://www.tripadvisor.com.br/Hotel_Review-g303546-d7179012-
      Reviews-Ritter_Hotel-Porto_Alegre_State_of_Rio_Grande_do_Sul.html"
59 )
60
61 #Sequence to navigate through each comment page (X comments per page, X
```

```
      total comments, X to update the link)
morepglist=list(
lagvivmon = seq(5,1190,5),
porretbou = seq(5,95,5),
devpripoa = seq(5,2105,5),
ibiscentro = seq(5,595,5),
sheraton = seq(5,595,5),
radisson = seq(5,660,5),
ritter = seq(5,730,5)
)

#Use these lists to scrap the comments
#url link for first search page and sequence counter
urllinkmain=urlmainlist[pickhotel]
morepg=as.numeric(morepglist[[pickhotel]])

urllinkpre=paste(strsplit(urllinkmain,"Reviews-")[[1]][1],"Reviews",sep="
    ")
urllinkpost=strsplit(urllinkmain,"Reviews-")[[1]][2]
urllink=rep(NA,length(morepg)+1)

#Test if the URL is retrieving the data
urllink[1]=urllinkmain
for(i in 1:length(morepg)){
urllink[i+1]=paste(urllinkpre,"-or",morepg[i],"-",urllinkpost,sep="")
}
head(urllink)

#Create a Data Frame to store the data
DF <- as.data.frame(matrix(ncol = 4,nrow=1))
colnames(DF) <- c("id","quote","review","date")

for(j in 1:length(morepg)){
urllink[j+1]=paste(urllinkpre,"-or",morepg[j],"-",urllinkpost,sep="")
url <- urllink[j]

reviews <- url %>%
read_html() %>%
html_nodes("#REVIEWS .innerBubble")

id <- reviews %>%
html_node(".quote a") %>%
html_attr("id")

quote <- reviews %>%
html_node(".quote span") %>%
html_text()
```

```
107
108 date <- reviews %>%
109 html_node(".relativeDate") %>%
110 html_attr("title") #%>%
111
112 review <- reviews %>%
113 html_node(".entry .partial_entry") %>%
114 html_text()
115
116 newRow <- data.frame(id,quote,review,date,stringsAsFactors = FALSE)
117 DF <- rbind(DF,newRow)
118 }
119
120 DF_full <- rbind(DF, DF_full)
```

APPENDIX B – LSA Function

LSA function to be distributed within LSA package in R, that ordered Term-Topic matrix (contribuitor: Rodrigo Heldt). Parameters:

- decomp.matrix = the Term-Topic matrix (suggested: tk from lsa package).

```
#Function to be distributed within LSA Package
ordered.lsa <-
function(decomp.matrix){

lsa.ordered <- as.data.frame(matrix(0, nrow = nrow(decomp.matrix), ncol =
    2*ncol(decomp.matrix)))

j <- 0
for(i in seq(2, (2*ncol(decomp.matrix)), 2))
{
j <- j+1
lsa.ordered[ ,i-1] <- names(decomp.matrix[order(decomp.matrix[ ,i-j],
    decreasing = T),i-j])
lsa.ordered[ ,i] <- decomp.matrix[order(decomp.matrix[ ,i-j], decreasing
    = T),i-j]
}

colnames(lsa.ordered) <- rep(1:ncol(decomp.matrix), each = 2)
return(lsa.ordered)
}
```

APPENDIX C – SVM and Naive Bayes trials

Trials with SVM and NaiveBayes classifiers in R, to be sure about best performance metrics from Neural Networks.

```
#SVM and NaiveBayes Classification

#######SVM
library(RTextTools)

#Take off zeros
#SentTest[is.na(SentTest)] <- 0
#SentTest$`1` <- as.factor(SentTest$`1`)

#Set dimension in analysis
original <- Cat1Class
original <- AllClass

##Create Document Term Matrix
DTM <- create_matrix(original$Comm, language="english", removeNumbers=
    TRUE,
removePunctuation = TRUE, removeStopwords = TRUE, toLower = TRUE,
stemWords=TRUE, removeSparseTerms=.998)

freqs <- rowSums(as.matrix(DTM))
lower <- which(freqs>1)
Alower <- DTM[-lower,]
DTM <- DTM[lower,]
freqs <- rowSums(as.matrix(DTM))
upper <- which(freqs<ncol(DTM)/4)
DTM <- DTM[,upper]
empty <- as.integer(which(colSums(as.matrix(DTM))<14))
DTM <- DTM[,-empty]
DTM <- as.matrix(DTM)

#Parameter tunning:
tune_out <- tune.svm(x = container@training_matrix,
y = container@training_codes,
gamma = 10^(-10:10),
cost = 10^(-1:1)
)
```

```
36
37 ##Create container object that prepares to train data in different
       algorithms
38 #Total Size = rows from DTM, i.e., the amount of documents
39 #virgin = false, we dont have virgin docs yet
40
41 #Object to store analytics values
42 analytics <- array(1:10,dim=c(10,6))
43 ensemble <- array(1:10,dim=c(10,4))
44 alg <- c('SVM','TREE')
45
46 for (i in 1:10){
47 #Set seed to reproduce same example
48 set.seed(100*i)
49 #Cross-fold validation
50 mod <- original[sample(nrow(original)),]
51 #mod <- mod[lower,]
52 #Train size
53 limTrain <- round(nrow(mod)*0.75)
54 #Create object to train
55 container <- create_container(DTM, mod$Cat, trainSize=1:limTrain,
56 testSize = (limTrain+1):nrow(mod), virgin=FALSE)
57
58 #Train SVM
59 model <- train_models(container, alg, kernel="radial", cost=0.1, gamma =
       1e-06)
60 #Test
61 multi_classify <- classify_models(container, model)
62 #Get analytics and store in analytics array
63 multi_analytics <- create_analytics(container,multi_classify)
64 #[,1] = precision; [,2] = recall; [,3] = fscore
65 analytics[i,] <- summary(multi_analytics)
66 ensemble[i,]<- multi_analytics@ensemble_summary
67
68 }
69
70 ###Test with new data
71 #New Unlabeled Data and Cleaning
72 virgin <- read_csv("your_path")
73
74 DTM_new <- create_matrix(virgin$Comm, language="english", removeNumbers=
       TRUE,
75 removePunctuation = TRUE, removeStopwords = TRUE, toLower = TRUE,
76 stemWords=TRUE, removeSparseTerms=.998)
77
78 #trace("create_matrix",edit=T) Change line 42 to "acronym" instead of "
       Acronym" -- original package error fix
```

```
79 #Continer object with the unlabeled and labeled data
80 container_New1 <- create_container(DTM_new, Cat1Class$Cat, trainSize=NULL
       , testSize = 1:1581,
81 virgin=TRUE)
82
83 #Choose algorithms to compare
84 alg <- c('SVM','MAXENT')
85 #Create model object with trainning data
86 model_multi <- train_models(container, alg)
87 #Perform classification
88 Multi_classify <- classify_models(container,model_multi)
89 #Analytis
90 Multi_Analytics <- create_analytics(container,Multi_classify)
91 summary(Multi_Analytics)
92 #Here we can compare both classifiers testd
93 Multi_Analytics_New@document_summary
94
95 #####Other SVM option from e1071 that allows to test with new data##
96 ##1. Data
97 #Set DataSet we are working
98 categorie <- yourData
99
100 #Randomize
101 set.seed(155)
102 categorieRnd = categorie[sample(nrow(categorie)),]
103
104 #Treat the corpus
105 x <- stopwords("en")
106 x <- c(x, "more")
107 corpus <- Corpus(VectorSource(categorie$Comm))
108 corpus <- tm_map(corpus, removePunctuation)
109 corpus <- tm_map(corpus, removeNumbers)
110 corpus <- tm_map(corpus, tolower)
111 corpus <- tm_map(corpus, removeWords, x)
112
113 #DTM Categorie 1
114 DTMC1 <- DocumentTermMatrix(corpus)
115 DTMC1 <- removeSparseTerms(DTMC1,0.998)
116 DTMC1Train <- DTMC1[1:780,]
117 DTMC1Train <- as.matrix(DTMC1Train)
118 DTMC1Test <- DTMC1[781:1045,]
119 DTMC1Test <- as.matrix(DTMC1Test)
120
121 #Use LSA as feature selection to improve SVM results.
122 LSAC1 <- lsa(DTMC1, dimcalc_share())
123 #Multiply back to use as the new DTM
124 DTMLSA <- as.textmatrix(LSAC1)
```

```
125 DTMLSATrain <- DTMLSA[1:780,]
126 DTMLSATrain <- as.matrix(DTMLSATrain)
127 DTMLSATest <- DTMLSA[781:1045,]
128 DTMLSATest <- as.matrix(DTMLSATest)
129
130 categorieRnd$Cat <- as.factor(categorieRnd$Cat)
131 CatTrain <- categorieRnd[1:780,]
132 CatTest <- categorieRnd[781:1045,]
133
134 #For unbalanced categories, set weights
135 weights <- c("0"=0.7,"1"=0.7)
136 #Train Model
137 model <- svm(DTMC1Train, CatTrain$Cat, kernel="radial",cost = 0.1,type = "
      C-classification", gamma = 1e-06,
138 scale = FALSE, probability = TRUE)
139 #class.weights = weights,
140 #Train with LSA
141 model <- svm(DTMLSATrain, CatTrain$Cat, kernel="radial",cost = 0.001,type
      = "C-classification", gamma = 1e-06,
142 probability = TRUE)
143 #class.weights = weights)
144 #Test Model
145 predTest <- predict(model,DTMC1Test)
146 #Test Model with LSA
147 predTest <- predict(model,DTMLSATest)
148 #Model Performance
149 acc <- table(predTest, CatTest$Cat)
150 #Combine results
151 Xx <- cbind(as.character(predTest),as.character(CatTest$Cat),as.character
      (CatTest$Comm))
152
153 #######Naive Bayes Classifier
154 library(e1071)
155 #create 75:25 partitions of the dataframe and document term matrix
156 ##The label collum has to be factor otherwise naiveBayes won't get it!
157 #SentTestT <- yourData
158 #SentTestT$`1` <- as.factor(SentTestT$`1`)
159 #SentTest_train <- SentTestT[1:90,]
160 #SentTest_test <- SentTestT[91:98,]
161
162 DTMSent <- create_matrix(Cat1ClassMod$Comm, language="english",
      removeNumbers=TRUE,
163 removePunctuation = TRUE, removeStopwords = TRUE, toLower = TRUE,
164 stemWords=TRUE, removeSparseTerms=.998, weighting = weightTfIdf)
165
166 #trainSize=1:753, testSize = 754:805
167 Cat1ClassMod$Cat <- as.factor(Cat1ClassMod$Cat)
```

```
168 Cat1ClassMod_train <- Cat1ClassMod[1:753,]
169 Cat1ClassMod_test <- Cat1ClassMod[754:805,]
170 mod <- Cat3Class
171 mod$Class <- as.factor(mod$Class)
172 modTrain <- mod[1:670,]
173 modTest <- mod[671:894,]
174
175 DTMSent_train <- DTMSent[1:753,]
176 DTMSent_train <- as.matrix(DTMSent_train)
177 DTMSent_test <- DTMSent[754:805,]
178 DTMSent_test <- as.matrix(DTMSent_test)
179
180 #trainSize=1:670, testSize = 671:894
181 DTMTrain <- as.matrix(DTM[1:670,])
182 DTMTest <- as.matrix(DTM[671:894,])
183 dim(DTMTrain)
184 dim(DTMSent_train)
185 #Second number = Amount of features (combination of possibilities)
186 #You have to define for each of the features how would be a condition
      True [DTP Training]
187 #For improve, can try to decrease the number of features (vocabulary)
188
189 ##First parameter has to be a Matrix
190 system.time(classifier <- naiveBayes(DTMSent_train, Cat1ClassMod_train$
      Cat, laplace = 1))
191 system.time(pred <- predict(classifier, newdata=DTMSent_test) )
192
193 classifier <- naiveBayes(DTMTrain,modTrain$Class,laplace=1)
194 pred <- predict(classifier,newdata = DTMTest)
195
196 table("Predictions"= pred,  "Actual" = Cat1ClassMod_test$Cat)
197 library(caret)
198 #Confusion Matrix
199 conf_mat <- confusionMatrix(pred, modTest$Class)
200 conf_mat$byClass
201 conf_mat$overall
202 conf_mat$overall['Accuracy']
203
204 ##Since the predictor variables here are all continuous, the Naive Bayes
      classifier
205 #generates three Gaussian (Normal) distributions for each predictor
      variable (each word)
206 classifier$tables$bad
207 #mean (first column) and standard deviation (second column) for each
      class
208 #also, how to access these four values
209 classifier$tables$good[1:4]
```

```
210
211 #Plotting word curves #shower, breakfast and bed (mean and SD pick up
        from previous line with the desired word
212 plot(function(x) dnorm(x, 0.17948718, 0.4514185), 0, 3, col="red",
213 main="Words distribution among 'Tangibles' Dimension", ylab="", xlab="")
214 curve(dnorm(x, 0.35897436, 0.7775528), add=TRUE, col="blue")
215 curve(dnorm(x, 0.4615385, 0.6002698), add=TRUE, col="green")
```

APPENDIX D – Jaccard Similarity

JacSim: Jaccard similarity for text. Analyzes topic descriptors two by two. Parameters:

- nWords = Numeric. Amount of words to be compared (suggested: 10).
- nDim = Numeric. Amount of dimensions (vectors) to be considered.
- matrix = Matrix. The Term-Topic matrix (suggested: tk from lsa package).
- rotate = Logical. Indicate to apply matrix rotation, implemented with promax. For varimax, just perform a simple code editing.

```
###Jaccard Similarity
jac_similarity <- function(nWords, nDim, matrix, rotate=FALSE){

if (missing(rotate)) {
original <- matrix$tk[,1:(nDim)]
orderVec <- ordered.lsa(original)
}
else {
original <- promax(scale(matrix$tk[,1:(nDim)]),m = 2)
orderVec <- ordered.lsa(original$loadings)
}

vector <- 1:length(orderVec)
n <- length(vector)
odd <- vector[seq(n)%%2==1]
Jac <- as.vector(1:(length(odd)))
for (j in 1:length(odd)){
v1 <- orderVec[1:nWords,odd[j]]
for (i in 1:length(odd)){
v2 <- orderVec[1:nWords,odd[i+1]]
I <- length(intersect(v1,v2))
S <- I/(length(v1)+length(v2)-I)
Jac[i] <- S
}
}
MN <- mean(Jac)
return(MN)
}
```

APPENDIX E – Neural Networks and Topic Modeling

Neural Networks and Topic Modeling (NNandTM): Combine Neural Network Classifier with LSA for classify and construct main topics regarding TripAdvisor data.

```
####Neural Network Classifier#######
##Classification taks
#Adapted from: http://tjo-en.hatenablog.com/entry/2016/03/30/233848
#Installation:
cran <- getOption("repos")
cran["dmlc"] <- "https://s3-us-west-2.amazonaws.com/apache-mxnet/R/CRAN/"
options(repos = cran)
install.packages("mxnet",dependencies = T)
require(mlbench)
require(mxnet)
require(tm)
require(lsa)
require(dplyr)

#Words removed from all corpus:
x <- stopwords("en")
x <- c(x, "more")

##For categories with few comments, mix with some machine annotated
    comments
#First, select the main words (with Topic Modeling) from each of the
    categories
#For that, corpus with only the comments that belongs to the category
categorie <- yourData
corpus <- Corpus(VectorSource(categorie$Comm))
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, tolower)
corpus <- tm_map(corpus, removeWords, x)
TDMNew <- TermDocumentMatrix(corpus)
LSANew <- lsa(TDMNew, dimcalc_share())
#Order them to select top-n (function ordered.LSA available in root)
ordLSA <- ordered.lsa(LSANew$tk)
#(opcional) Visualization of the process
#()View(ordLSA[1:20,1:20])
#()findFreqTerms(TDMNew, lowfreq =20)
```

```
35 #DataFrame with top-n words from top-(n*2) topics
36 ft1 <- ordLSA[1:10,1:20]
37 #Just collumns with words (odd columns)
38 ft1<- ft1[,c(1,3,5,7,9,11,13,15,17,19)]
39 #Create a chr vector
40 ft1 <- c(ft1$`2`,ft1$`3`,ft1$`4`,ft1$`5`,ft1$`5`,ft1$`6`,ft1$`8`,ft1$`9`,
       ft1$`10`)
41 #Remove repeated words
42 ft1 <- unique(ft1)
43
44 #Now grab all unlabeled documents and run again to make a DTM with them
45 #HERE RUN AGAIN line 26 to 32 with categorie <- yourDataUnlabeled
46 TDMNewM <- as.matrix(TDMNew)
47 #Only comments that contains at least a % of this ft1 words
48 ##Otherwise it would bring so many confusing things!
49 #Take those words because they don't appear in the new database
50 remove <- c("christmas","")
51 ft1 <- ft1 [! ft1 %in% remove]
52 Docs <- TDMNewM[ft1,]
53 #Make it a logical matrix so that the amount of each word won't affect
54 Docs <- sapply(as.data.frame(Docs), as.logical)
55 #Multiply by one so that you have 0's and 1's
56 Docs <- Docs * 1
57 #Choose only those columns where sum > n (at least n from the m words in
       the same document)
58 Docs <- which(colSums(Docs)>=6)
59 #Use this index to choose only the correspondent documents with those
       words
60 Docs <- NotClass[Docs,]
61
62 #Insert category columm (all 1) information
63 Docs$Cat <- as.numeric(1)
64 #Reorder to rbind
65 Docs <- Docs[c(2,1)]
66 #Put together with the existent CatnClass (0's and 1's comment)
67 #Select the data to work
68 categorie <- yourData
69 categorie <- rbind(CatnClass,Docs)
70 #mx.mlp requires the following parameters:
71 #Training data and label
72 #Number of hidden nodes in each hidden layer
73 #Number of nodes in the output layer
74 #Type of the activation
75 #Type of the output loss
76 #The device to train (GPU or CPU)
77 #Other parameters for mx.model.FeedForward.create
78 mx.set.seed(0)
```

```
#Create a vector to keep track of differences in accuracy during the
    training process
accuracy <- as.vector(1:10)
##Neural Network Feed-Foward Classifier!
for (i in 1:10){
#Randomize the data every time
set.seed(12*i)
categorieRnd = categorie[sample(nrow(categorie)),]
#Build a Corpus
corpus <- Corpus(VectorSource(categorieRnd$Comm))
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, tolower)
corpus <- tm_map(corpus, removeWords, x)
#Build Document-Term Matrix Train and Test
DTM <- DocumentTermMatrix(corpus)
#Optional: Remove Sparse Terms
#DTM <- removeSparseTerms(DTMC1,0.998)
limTrain <- round(nrow(categorieRnd)*0.75)
DTMTrain <- DTM[1:limTrain,]
DTMTrain <- as.matrix(DTMTrain)
DTMTest <- DTM[(limTrain+1):nrow(categorieRnd),]
DTMTest <- as.matrix(DTMTest)
#Build labels Train and Test
CatTrain <- categorieRnd$Cat[1:limTrain]
CatTest <- categorieRnd$Cat[(limTrain+1):nrow(categorieRnd)]
CatTrain <- as.numeric(CatTrain)
CatTest <- as.numeric(CatTest)

#Multi-layer Perceptron, train with Train data:
model <- mx.mlp(DTMTrain, CatTrain, hidden_node=15, out_node=4, out_
    activation="softmax",
num.round=20, array.batch.size=15, learning.rate=0.07, momentum=0.9,
eval.metric=mx.metric.accuracy)

#Predict Test data
preds = predict(model, DTMTest)
pred.label = max.col(t(preds))-1
#Table with rights and weongs in Test Data
table <- table(pred.label, CatTest)
#Keep track of accuracy
accuracy[i] <- (table[1,1]+table[2,2])/(sum(table))
}

plot(accuracy,type="b", ylim=c(0,1),xlab="",ylab="Accuracy")

#########Can you predict new data?
```

```
124 categorie <- yourUnlabeledData
125 #Since that in Docs we used some of this data, we have to remove now
126 #Otherwise it would be duplicated
127 #anti_join() return all rows from x where there are not matching values
       in y, keeping just x columns
128 test <- anti_join(yourUnlabeledData,categorie)
129 categorie <- test
130
131 corpus <- Corpus(VectorSource(categorie$Comm))
132 corpus <- tm_map(corpus, removePunctuation)
133 corpus <- tm_map(corpus, removeNumbers)
134 corpus <- tm_map(corpus, tolower)
135 corpus <- tm_map(corpus, removeWords, x)
136 #Build DTM just to compare the vocabulary
137 NDTM <- DocumentTermMatrix(corpus)
138 #Demonstrate that the vocabulary is similar, so we can classify based on
       this classifier:
139 ft2 <- findFreqTerms(DTM)
140 ft1 <- findFreqTerms(NDTM)
141 #Common terms
142 common.c1c2 <- data.frame(term = character(0), freq = integer(0))
143 for(t in ft1){
144 find <- agrep(t, ft2)
145 if(length(find) != 0){
146 common.c1c2 <- rbind(common.c1c2, data.frame(term = t, freq = length(find
       )))
147 }
148 }
149 #Difference among them
150 same <- common.c1c2[,1]
151 allNDTM <- NDTM$dimnames$Terms
152 outsiders <- setdiff(allNDTM,same)
153 #Frequency of this outside terms
154 outsidersFreq <- NDTM[,intersect(colnames(NDTM), outsiders)]
155 #Percentage:
156 sum(outsidersFreq)
157 outsidersPerc <- (sum(outsidersFreq$v)/sum(NDTM$v))*100
158
159 #Now build DTM with the same vocabulary as the previous data
160 NDTM <- DocumentTermMatrix(corpus,
161 control=list(dictionary=DTM$dimnames$Terms))
162 NDTM <- as.matrix(NDTM)
163 #Run the model in the unlabeled data
164 predsNew <- predict(model,NDTM)
165 predsNew.label <- max.col(t(predsNew))-1
166 Xx <- cbind(as.character(predsNew.label),as.character(categorie$Comm))
167 #Build the dataframe
```

```
168 Xx <- as.data.frame(Xx)
169 #Stay only with the "1"
170 CatnNew <- Xx[which(Xx$V1=="1"),]
171 #Coerce names otherwise rbind won't work
172 names(CatnNew) <- names(CatnClass)
173 #Join with CatNClass (That was labeled by humans)
174 CatnNew <- rbind(CatnNew, CatnClass[which(CatnClass$Cat==1),])
175
176 ##LSA with new data to uncover the topics
177 categorie <- yourData
178 corpus <- Corpus(VectorSource(categorie$Comm))
179 corpus <- tm_map(corpus, removePunctuation)
180 corpus <- tm_map(corpus, removeNumbers)
181 corpus <- tm_map(corpus, tolower)
182 corpus <- tm_map(corpus, removeWords, x)
183 TDMNew <- TermDocumentMatrix(corpus)
184 LSANew <- lsa(TDMNew, dimcalc_share())
185 ordLSA <- ordered.lsa(LSANew$tk)
186 View(ordLSA)
187
188 #Now create some wordclouds based on main topics!
189 library(wordcloud)
190 #Frequency of the words
191 freq <- rowSums(as.matrix(TDMNew))
192 #Name in rows
193 freq <- as.data.frame(freq,rownames(as.matrix(TDMNew)))
194 #Select main words from topics
195 ftw <- ordLSA[1:10,1:22]
196 #Just columns with words (odd columns)
197 ftw<- ftw[,c(1,3,5,7,9,11,13,15,17,19,21)]
198 #Create a chr vector
199 ftw <- c(ftw$`2`,ftw$`3`,ftw$`4`,ftw$`5`,ftw$`5`,ftw$`6`,ftw$`8`,ftw$`9`,
      ftw$`10`,ftw$`11`)
200 #Remove repeated words and put in exclusive variables (per category)
201 ftwn <- unique(ftw)
202 freqn <- freq
203 #Plot Word cloud, min n words (here 15)
204 WCCatn <- wordcloud(ftwn,freqn[ftwn,],min.freq = 15, random.color = TRUE)
205
206 #LSAAll to visualize singular values
207 corpus <- Corpus(VectorSource(yourUnlabeledData$Comm))
208 corpus <- tm_map(corpus, removePunctuation)
209 corpus <- tm_map(corpus, removeNumbers)
210 corpus <- tm_map(corpus, tolower)
211 corpus <- tm_map(corpus, removeWords, x)
212 TDM <- TermDocumentMatrix(corpus)
213 #Generates raw = all singular values; and share = 50%, and plot both to
```

```
    compare
LSAAll <- lsa(TDM, dimcalc_raw())
LSA50 <- lsa(TDM, dimcalc_share())
plot(LSAAll$sk, type="l",xlab="",ylab="",main="Singular Values",ylim=c
    (0,100))
par(new=TRUE)
plot(LSA50$sk, type="l",xlab="",ylab="",xaxt="n",yaxt="n",col="blue",
xlim=c(0,1582),ylim=c(0,100),lwd=3)

######PRCOMP Visualization######
#My Data: Hotels in line and category in collumns
##First: weight each hotel by each category by main topics word frequency
#Initialize a data.frame to keep the values:
##strongAsFactors = FALSE, otherwise problems to inser Hotel Name
CatHotel <- data.frame(Hotel=as.character(),Cat1=as.numeric(), Cat2=as.
    numeric(),Cat3=as.numeric(),
Cat4=as.numeric(),Cat5=as.numeric(),stringsAsFactors = FALSE)
#First, only counting the words and normalizing by the total comments
    from that hotel
hotel <- "H1"
#Here import data frame with all hotels, nicknames to preserve hotel real
     name, and comments
hotelComm <- All2017[which(All2017$Nickname==hotel),]
#Corpus with that hotel words
categorie <- hotelComm
corpus <- Corpus(VectorSource(categorie$Comm))
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, tolower)
corpus <- tm_map(corpus, removeWords, x)
TDMHotel <- TermDocumentMatrix(corpus)
#See frequency of each word for this hotel
freqHotel <- rowSums(as.matrix(TDMHotel))
#Name in rows
freqHotel <- as.data.frame(freqHotel,rownames(as.matrix(TDMHotel)))
#Sum  words for each category, dividing by the number of comments
##Do it for all categories
CatnHotel <- sum(freqHotel[ftwn,],na.rm=TRUE)/nrow(hotelComm)
#Make it a row in a data.frame
CatRow <- c(hotel,CatnHotel,CatmHotel,CatxHotel) #and how many more you
    have
CatHotel[nrow(CatHotel)+1,] <- CatRow
#Hotels in rownames
library(tidyverse)
CatHotel <- CatHotel %>% remove_rownames %>% column_to_rownames(var="
    Hotel")
CatHotel <- as.data.frame(lapply(CatHotel, as.numeric))
```

```
254 ##PCA plot to see where are the hotels with the categories
255 biplot(prcomp(CatHotel, scale = TRUE))
```