UNIVERSITY OF THESSALY

DIPLOMA THESIS

Utilizing unstructured data for business intelligence

Author:
Petros LEMONOPOULOS

Supervisor:
Manolis VAVALIS

A thesis submitted in fulfillment of the requirements for the degree of Diploma

in the

Department of Electrical and Computer Engineering

August 10, 2020

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Διπλωματική Εργασία

Αξιοποίηση αδόμητων δεδομένων για επιχειρηματική ευφυΐα

Συγγραφέας:
Πέτρος Λεμονόπουλος

Επιβλέπων:
Μανώλης Βάβαλης

Μια διπλωματική εργασία που υποβλήθηκε για την συμπλήρωση των προϋποθέσεων για την απόκτηση Διπλώματος

στο

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

10 Αυγούστου 2020

Declaration of Authorship

I, Petros LEMONOPOULOS, declare that this thesis titled, "Utilizing unstructured data for business intelligence" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Abstract

Petros LEMONOPOULOS

Utilizing unstructured data for business intelligence

Nowadays, the volume of data is increasing rapidly. In addition, unstructured data comprises the vast majority of data. Accordingly, the large volume of information, the highly hectic business environment and the fact that most enterprises cannot utilize the unstructured data pave the avenue for further research. Specifically, our thesis examines effective ways to search and derive insights from non-structured data in the context of Business Intelligence (BI) for enterprises, especially for Small and Medium Enterprises (SMEs). This includes data acquisition, cleansing, formatting, text extraction, mining, natural language processing, automatic text summarization and sentiment analysis. Moreover, our research focuses on information from various online platforms and social media such as SEC Edgar database, TripAdvisor, YouTube and Hellenic Chamber of Hotels. In particular, our thesis mainly deals with text summarization of financial disclosures and web data analysis of Greek restaurants from the region of Thessaly. Finally, the purpose of this study is to harness valuable unstructured data from various web sources in order to support the modern needs of enterprises.

Περίληψη

Πέτρος Λεμονόπουλος

Αξιοποίηση αδόμητων δεδομένων για επιχειρηματική ευφυΐα

Σήμερα, ο όγκος των δεδομένων αυξάνεται ραγδαία. Επιπλέον, τα μη δομημένα δεδομένα αποτελούν τη συντριπτική πλειονότητα των δεδομένων. Κατά συνέπεια, ο μεγάλος όγκος πληροφοριών, το ιδιαίτερα έντονο επιχειρηματικό περιβάλλον και το γεγονός ότι οι περισσότερες επιχειρήσεις δεν μπορούν να χρησιμοποιήσουν τα μη δομημένα δεδομένα ανοίγουν το δρόμο για περαιτέρω έρευνα. Ειδικότερα, η διατριβή μας εξετάζει αποτελεσματικούς τρόπους αναζήτησης και απόκτησης πληροφοριών από μη δομημένα δεδομένα στο πλαίσιο της Επιχειρηματικής Ευφυΐας για επιχειρήσεις, ειδικά για Μικρές και Μεσαίες Επιχειρήσεις. Αυτό περιλαμβάνει την απόκτηση δεδομένων, τον καθαρισμό, τη μορφοποίηση, την εξαγωγή κειμένου, την εξόρυξη, την επεξεργασία φυσικής γλώσσας, την αυτόματη σύνοψη κειμένου και την ανάλυση συναισθημάτων. Επιπλέον, η έρευνά μας εστιάζει σε πληροφορίες από διάφορες διαδικτυακές πλατφόρμες και μέσα κοινωνικής δικτύωσης, όπως η βάση δεδομένων SEC Edgar, το TripAdvisor, το YouTube και το Ξενοδοχειακό Επιμελητήριο Ελλάδος. Συγκεκριμένα, η διατριβή μας ασχολείται κυρίως με την περίληψη κειμένων των οικονομικών γνωστοποιήσεων και την ανάλυση δεδομένων ιστού ελληνικών εστιατορίων από την περιοχή της Θεσσαλίας. Τέλος, ο σκοπός αυτής της μελέτης είναι να αξιοποιήσει πολύτιμα μη δομημένα δεδομένα από διάφορες πηγές Ιστού για να υποστηρίξει τις σύγχρονες ανάγκες των επιχειρήσεων.

Acknowledgements

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα της διπλωματικής μου εργασίας, τον κύριο Εμμανουήλ Βάβαλη για τις έμπειρες συμβουλές του και την καθοδήγηση του. Επίσης, θα ήθελα να ευχαριστήσω τους αγαπητούς συναδέλφους Χριστόδουλο Παππά και τον Αθανάσιο Ζουμπέκα για τη συνεργασία και την πολύτιμη συμβολή τους στην ολοκλήρωση της εργασίας. Τέλος, ευχαριστώ τους γονείς μου και τα αδέρφια μου για την αγάπη και την στήριξη τους.

Contents

List of Figures

Chapter 1

Introduction

The most frequently used data in Business Intelligence (BI) and machine learning tasks are structured data. Specifically, most tasks deal with data which is in tabular form in either spreadsheets or relational databases [1] [2]. It is fact that structured data offers us some important insights in numerical form. However, nowadays, the volume of data is increasing rapidly. Also, most of this data is unstructured and comes from social media, websites, comments and financial documents. Consequently, as Paul Hoffman, CTO of Space-Time Insight mentioned "If you want to understand people, especially your customers then you have to be able to possess a strong capability to analyze text" [3].

Accordingly, the possibility of utilizing unstructured data is significant and necessary in order to support and enhance enterprises, especially Small and Medium Enterprises (SMEs). Thus, the first step in exploiting data is to find ways to extract, collect, clean and preprocess them so that to extract useful information and conclusions. For this reason, our research presents methods of web crawling and scraping data from different web sources such as SEC (Securities and Exchange Commission) filings from EDGAR database, TripAdvisor, YouTube and Hellenic Chamber of Hotels. Moreover, our thesis presents Natural Language processing (NLP) methods in order to clean and format data and give us the ability to harness them.

Furthermore, automatic text summarization is one of the most challenging fields of NLP. The amount of textual insights is often unmanageable by humans. This problem began to concern researchers as early as 1958, when Hans Peter Luhn published a paper titled "The automatic creation of literature abstracts". Since then until today there has been a lot of research in this area. For instance, the experimental research conducted in 2019 by E. Cardinaels, S. Hollander and B.J. White titled "Automatic summarization of earnings releases: attributes and effects on investors' judgments". In our thesis, we applied various techniques of text summarization on financial disclosures that came from SEC filings in order to extract significant textual information.

In addition, another challenging aspect in business intelligence is the utilization of customer's feedback from social media and websites. In particular, our thesis examines TripAdvisor's insights for Greek restaurants. For this purpose, we extract both detailed information and customer's reviews for restaurants. Thus, analyzing the sentiment of these comments can lead us both to understand the opinion of customers and to deeper insights to improve the weaknesses of the enterprise.

In conclusion, the aim of this thesis is to examine effective ways to search, acquire and derive non-structured data from various web sources in the context of Business Intelligence for enterprises.

The rest of this thesis is organized into two main parts.Specifically, the first part includes Chapter 2 and 3. In Chapter 2 we present the required background knowledge about unstructured data for BI, the definition of data management problem for

enterprises as well as the specification of the basic methods we are going to use in order to handle this problem. In Chapter 3, we provide an overview of the previous studies related to the subject of this research, while we refer to the top performing methods identified up to now in the field. On the other hand, the second part presents the contribution of our research and consists of Chapters 4, 5, 6, 7 and 8. In Chapter 4 we provide web data collection methods for non-structured data from various web data sources. Then, data preprocessing and formatting techniques are presented in Chapter 5. In addition, Chapter 6 contains automatic text summarization models for 10-K financial disclosures. Moreover, in Chapter 7, we present data visualizations as well as opinion mining techniques from customer reviews of Thessaly based restaurants. Finally, in Chapter 8 we make an overall conclusion of the thesis and we suggest future research and development prospects.

Chapter 2

Unstructured Data and Business Intelligence

As mentioned in [4] structured data offer quick and achievable insights to businesses in numerical form. These data are easily manageable and quickly with algorithms, instantaneously turned into stunning data visualizations, and can help guide your business decisions with smart predictive insights that they provide. In addition, structured data is easier to capture than unstructured data. However, solely capturing and analyzing structured data can only provide you with a small glimpse into the world of data. Also, it is true that unstructured data account at least 80% of stored information [5].

Unstructured data (or Non-structured information) is data that is not organized in a predetermined way or does not have a predetermined data model [6]. Particularly, unstructured data is commonly text-heavy, but may contain information such as dates, numbers, and facts as well. For this reason, this leads to irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in structured form in databases or annotated (semantically tagged) in documents [6].

Figure 2.1 illustrates some features for both structured and unstructured data.

	Structured Data	Unstructured Data
Characteristics	• Pre-defined data models • Usually text only • Easy to search	• No pre-defined data model • May be text, images, sound, video or other formats • Difficult to search
Resides in	• Relational databases • Data warehouses	• Applications • NoSQL databases • Data warehouses • Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	• Airline reservation systems • Inventory control • CRM systems • ERP systems	• Word processing • Presentation software • Email clients • Tools for viewing or editing media
Examples	• Dates • Phone numbers • Social security numbers • Credit card numbers • Customer names • Addresses • Product names and numbers • Transaction information	• Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery

FIGURE 2.1: Structured and Unstructured Data [7]

As Merrill Lynch claimed in 1998 "unstructured data comprises the vast majority of data found in an organization, some estimates run as high as 80 %" [5]. This number is generally accepted, but it's unclear what is the source [8]. Additionally, there are other sources which support similar or higher percentages of unstructured data [6] [7] [9] [10].

According to [11] it is predicted that data will grow to 40 zettabytes by 2020, resulting in a 50-fold growth from the beginning of 2010. More recently, it has been claimed that the global volume of data will grow to 163 zettabytes by 2025 and majority of that will be unstructured [12] [11]. In addition, the magazine "Computerworld" argues that unstructured data might account more than 80% of all data in organizations [5] [6].

Figure 2.2 illustrates the company data as an iceberg. Particularly, above the water is visible only the tip of the mass. The mass above the water represents structured data, these data accounts for 20% of all enterprise information. On the other hand, unstructured data accounts for 80% and is underneath the surface and constitutes the bulk of the iceberg. Also unstructured information is the most underutilized resource of a company. Majority of companies focus on mining structured data, because it is readily visible and accessible. On the contrary, unstructured information is a bit trickier on mining [13].

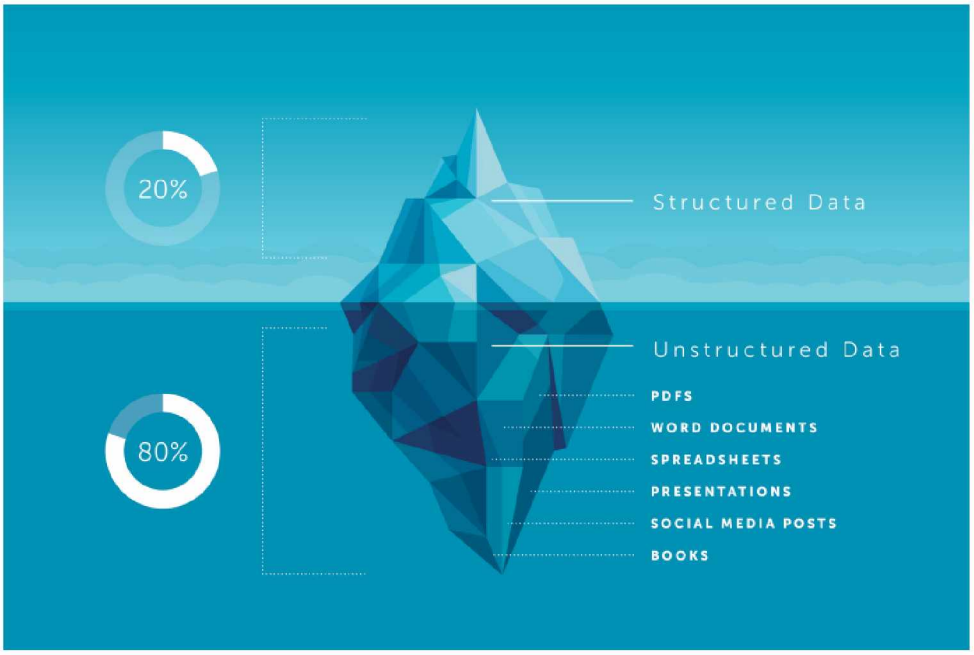

FIGURE 2.2: Distribution of a company's data [13].

Therefore, according to [14] as corporate disclosures over time they lengthen more and more, it is possible investors with limited attention to find it hard to process all of the data [15] [16]. As a result, it is significant to study how summaries help or hinder individual investors in making investment-related judgments, policymakers also claim that this issue deem to be relevant [17] [18]. Moreover, the Securities and Exchange Commission (SEC) does not offer any guidance on summary size and the items that a summary should cover, so it is important to examine how summaries affect investors' investment-related judgments [18].

Furthermore, as seen in [14] given the large volume of information and the fact

that individual investors in most cases are irrational [19] [20], companies often provide summaries of key disclosures and earnings releases. Nevertheless, they do not present a balanced picture of the information disclosed in the underlying document. Also, managers may be more optimistic than they should be, selectively highlight information that is favorable to the firm [21] [22] [23] [24]. In this context, there may be a role for algorithm-based summaries of earnings releases. These algorithms utilize statistical heuristics for sentence extraction to summarize large texts without human intervention. As a result, automatic summaries have the potential to reduce both data overload and bias of managers [14].

Moreover, as mentioned in [4] it is perceived that unstructured information might not be well-organized or easy to access, but they provide their key to getting a comprehensive answer to the most important business questions. Especially, these questions can include who your customers are, what customer preferences are, what goals and successes your competitors are experiencing, attitudes of top industry participants, and so much more. On the other hand, only with structured data, it's difficult to ask the right business questions and difficult to find the right answers. Also, unstructured data have a crucial role in improving customer experience. Particularly, firms with the help of unstructured data can come to understand about the trends on social media, the opinions of consumers. Consequently, the contribution of unstructured data can improve customer satisfaction rates and Return on Investment (ROI). Also, enterprises can innovate with new features according to customers and industry needs, because of information about what customers want and how current products, services or technologies doesn't satisfy them. Finally, unstructured data can give valuable insights to enterprises to fill gaps in industry and provide services that meet new customer needs [4].

In addition, according to [25] the unstructured data might play an important role for Small and Medium Enterprises (SMEs). Specifically, SMEs represent 95% of all enterprises, 66% of total employees are employed in SMEs and 55% of the total production comes from them [26]. Moreover, recent research shows that the current business environment is highly unstable and influenced by modern information, social media, globalization and employee mobility [27] [28]. In fact, the competition of SMEs has increased due to the large number of companies. This results to force SMEs to experience more severe challenges in maintaining their existence and expanding their business. Also, the majority of SMEs don't have the financial capacity of large companies to be supported by financial analysts and managers. Additionally, as seen in [25] researchers have been analyzing growth and success factors of SMEs for decades in order to support SMEs at improving their competitiveness. Thus, because of big data researchers focused on applying data mining techniques to build novel risk and growth prediction models. Nevertheless, existing models don't include a large number of data types, e.g. financial or operational data [29]. Thus, analyzing unstructured data from SMEs can be very helpful for their viability, growth and problem solving.

To sum up, nowadays, as seen in [25] web mining (WM) has emerged as a new approach towards obtaining valuable business data. Also, WM gives the opportunity for an automated and large scale collection and analysis of significant information from the web such as the national commercial register and websites of enterprises. However, WM methods for SMEs support is still scarce. Taking into consideration that as time goes on the amount of data increases, WM provides the ability in revealing valuable data hidden in websites and in social media which is valuable for building a SME business support model. In particular, our thesis utilizes data from

10-K Forms, the TripAdvisor, the Hellenic Chamber of Hotels and YouTube. Specifically, the 10-K Form is an annual report required by the SEC [30]. To conclude, the purpose of this thesis is the utilization of effective ways to search, acquire and derive insights from non-structured data in the context of Business Intelligence mainly for SMEs.

Chapter 3

State of the Art

According to [31] [32] the first research about business intelligence concerned with textual unstructured information, rather than numerical data [31]. Specifically, as early as 1958, H.P. Luhn and other computer science researchers were particularly focused on the extraction and classification of unstructured textual data [31]. Actually, according to [32] the majority of the available textual information in the SEC EDGAR database is weakly structured in technical terms [33] [34] [35] especially until 2002 when the usage of markup languages was rarely [36]. Also, a limited number of elements with label formatting errors and other inconsistencies results in difficulties in accurately identifying and analyzing common text issues in multiple filings [34] [37] [38]. Thus, these problems directly affect the ability to automatically text data extraction from SEC Submissions [37] [39] [40]. Moreover, the business information providers offer expensive commercial products (e.g. Academic EDGAR and Edgar Pro). Today, it is fact that research in the field of text mining and analysis increases [41] [42] [43]. Accordingly, the questions arises as to what specific financial statements and disclosures are available to the public free of charge, how to retrieve these corporate files and how to decode embedded textual data to be incorporated into investment decisions, trading strategies and financial research studies [32] [44].

Nowadays, as mentioned in [32] only a small amount of specific literature is currently available for extracting textual data from financial statements submitted to the SEC and the EDGAR system with the except of [45] [46] [47] [48] [49] [50] [51] [52] . Notably, EDGAR is the Electronic Data Gathering, Analysis, and Retrieval system [53]. Additionally, EDGAR performs automated collection, validation, indexing, acceptance, and promotion of submissions by enterprises and others who are required by law to submit forms to the U.S. Securities and Exchange Commission(SEC) [53]. Also, the database contains a large amount of data about the Commission and the securities industry which is publicly available via the Internet [53] [54]. This thesis, for downloading company filings from the SEC EDGAR database uses a Python package named "sec-edgar-downloader" and searches can be conducted either by stock ticker or Central Index Key (CIK) [55]. Finally, my research will serve as a guide on how to extract financial statements documents from the SEC and how to decode the embedded text data provided by the EDGAR system for business and research purposes.

According to [14], although the acknowledging that text summary may be useful in the area of corporate disclosures, until a few years ago, there has been no systematic research on how the summary affects individual investor crisis [56]. However, as demonstrated in [14], this research provides data on this significant field in two ways. On the one hand, researchers in [14] test the viability of using automated summary techniques in practice, creating automated summaries of the actual corporate earnings circulations and comparing them to management summaries in various dimensions. On the other hand, they control the impact of summaries on investors'

judgments. My research is based on [14] survey, so we utilize algorithms for text summarization on 10-K Forms and suggest the use of data from social media for more accurate conclusions in the summaries.

Moreover, as mentioned in [25] the bibliography for business development models started in 1967 and has since been evaluated in various streams related to specific industries and business sizes. In particular, Lippitt and Schmidt [57] created a general growth model for all types of enterprises looking at how personality development theories affect the creation, development and maturation of an enterprise. Steinmetz [58] dealt with small businesses by qualitatively analyzing their development, breaking down the growth curve of small businesses into different stages and evaluating the characteristics of each stage. Also, a qualitative research carried out by Scott and Bruce [59] proposed a model for small business development which supports managers to plan for future development. More specific, the suggested model isolates five developmental stages characterized by a unique combination of characteristics. As time goes on, small enterprises go through various stages of development, this leads to changes in characteristics such as management style, organizational structure and the use of technology. While stage models are accepted by most researchers, however they are criticized in some respects [60]. Specifically, empirical surveys conducted only on small sample sizes and clearly defined types of enterprises through questionnaire studies and therefore, stage models are specialized.

Additionally, according to [25] the web is a popular and interactive medium with a large volume of public available data. Particularly, it is a collection of files and multimedia information [61]. The abundance of available data on the Web makes it a key source of data and enables enterprises to export useful information. Consequently, WM research is growing rapidly as WM has proven to be valuable in both the business world and e-commerce [62] [63] [64] [65] [66]. For example, [64] recommended a system that extracts data from a set of web pages in order to extract informative content. Moreover, [65] made an effort to increase sales volume and introduced a system for finding product fame on the Internet to support marketing and customer relationship management. As well, [66], with the use of data from the top 500 worldwide enterprises, they analyzed how text data from e-commerce companies' websites affect commercial success.

On the contrary, as seen in [25] surveys for SME development with WM methods are very limited. Specifically, the research of [67] focused on the relationship between long-term development of SMEs and a web presentation. Additionally, [62] examined micro-level characteristics and the influence of external relations, governmental or academic relations in the development of SMEs. Nevertheless, these studies utilize a limited amount of data dealing only with the information available on the company's websites and consequently this creates significant research gaps and weaknesses. For this reason, further research is needed to analyze and develop methods for SMEs in order to exploit the large volume of information from web and social media.

Chapter 4

Data Collection

This chapter focuses on web data collection. Ordinary types of Web data sources are blogs, business websites, social media, forums, news websites, governmental sites [68]. The present chapter investigates how to collect publicly accessible unstructured data for enterprises. Also, in my thesis I implemented two basic methods for the acquisition of web information. First, web crawling for retrieving publicly available web documents. Second, Application Programming Interface (API) for web data collection. Additionally, it is remarkable that these methods return unstructured data in the form of web documents such as HTML, XML or XBRL files [68]. In conclusion, this chapter presents the process of data collection from SEC filings, TripAdvisor, Hellenic Chamber of Hotels and YouTube.

4.1 SEC Filings - 10-K Forms collection

As mentioned in [69] the SEC filing is a formal financial statement submitted to the U.S. Securities and Exchange Commission (SEC). Further, public companies, certain insiders, and broker-dealers are obliged to make regular SEC filings. These filings are used by investors and financial professionals to extract insights about enterprises they are evaluating for investment purposes. In SEC's EDGAR database the majority of SEC filings are available online.

Additionally, the most frequently SEC forms are the 10-K and the 10-Q. According to [69] these forms are composed of four main sections: The business section, the F-pages, the Risk Factors, and the MD&A.

In particular:

- The business section gives an overview of the enterprise.
- The F-pages provide financial statements evaluated by an independent auditor.
- The Risk Factors provide a list the potential risks that exist for the company.
- The MD&A provides a narrative about the financial outcomes of the enterprise, also this narrative is accompanied by management's expectations for the upcoming year.

The structure of 10-K Forms contains 4 parts and 15 schedules [70]:

- **Part 1**
  Item 1 - "Business"
  Item 1A - "Risk Factors"
  Item 1B - "Unresolved Staff Comments"

Item 2 - "Properties"
Item 3 - "Legal Proceedings"
Item 4 - "Mine Safety Disclosures"

- **Part 2**
  Item 5 - "Market"
  Item 6 - "Consolidated Financial Data"
  Item 7 - "Management's Discussion and Analysis of Financial Condition and Results of Operations"
  Item 7A - "Quantitative and Qualitative Disclosures about Market Risks" Item 8 - "Financial Statements"
  Item 9 - "Changes in and Disagreements with Accountants on Accounting and Financial Disclosure"
  Item 9A- "Controls and Procedures"
  Item 9B- "Other Information"

- **Part 3**
  Item 10 – "Directors, Executive Officers and Corporate Governance"
  Item 11 – "Executive Compensation"
  Item 12 – "Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters"
  Item 13 – "Certain Relationships and Related Transactions, and Director Independence"
  Item 14 – "Principal Accounting Fees and Services"

- **Part 4**
  Item 15 – "Exhibits, Financial Statement Schedules Signatures"

In my thesis, I utilize the "sec-edgar-downloader" python package to collect 10-K Forms. From the downloaded 10-K Forms we extract, preprocess and summarize Item 1A, Item 7 and Item 7A. In addition, after contacting Mr. Eddy Cardinaels, he sent us the data used in the paper "Automatic summarization of earnings releases: attributes and effects on investors' judgments".

More details for items 1A, 7 and 7A:

Item 1A - "Risk Factors" "Includes information about the most significant risks that apply to the company or to its securities. Companies generally list the risk factors in order of their importance. In practice, this section focuses on the risks themselves, not how the company addresses those risks. Some risks may be true for the entire economy, some may apply only to the company's industry sector or geographic region, and some may be unique to the company" [70].

Item 7 - "Management's Discussion and Analysis of Financial Condition and Results of Operations" "Gives the company's perspective on the business results of the past financial year. This section, known as the MD&A for short, allows company management to tell its story in its own words. The MD&A presents: The company's operations and financial results, including information about the company's liquidity and capital resources and any known trends or uncertainties that could materially affect the company's results. This section may also discuss management's views of key business risks and what it is doing to address them" [70].

Item 7A - "Quantitative and Qualitative Disclosures about Market Risk" "Requires information about the company's exposure to market risk, such as interest rate risk, foreign currency exchange risk, commodity price risk or equity price risk. The company may discuss how it manages its market risk exposures" [70].

My system returns 10-K Forms as raw text version, this text is a SEC document as XML Technical Specification. Thus, it is necessary to process these files with text mining techniques in order to extract useful information.

4.2 Web Data Source "TripAdvisor.com"

TripAdvisor, according to [71] is the world's largest travel platform, that serves 463 million travelers each month for their trip. TripAdvisor founded in early 2000, nowadays covers more than 860 million reviews and opinions of 8.7 million accommodations, restaurants, experiences, airlines and cruises and is available in 49 markets and 28 languages. Specifically, users can post reviews and opinions for hotels, restaurants and attractions. In addition, they have the chance to add multimedia elements like photos and videos, travel maps of previous trips and take part in discussion forums. Further, this platform gives the opportunity to tourists to rate restaurants in a 5-star marking system from four separate criteria: food, service, value and atmosphere. Also, these four aspects have a crucial role to consumers' restaurant decision-making [72] [68].

In my thesis I collect information about Greek restaurants in Thessaly. Specifically, to collect information from TripAdvisor I used Python's library BeautifulSoup to extract the web links from restaurants [73] and Python's library urllib2 which is used to access the detailed data from restaurants [74]. My crawler deployed to collect detailed information about every restaurant.

Notably, the collected data is stored in Excel files and includes records of 591 Thessaly's restaurants, which covers the majority of restaurants enterprise from the area of Larissa, Volos, Trikala and Karditsa. In particular, the collected data consist of insights about the restaurant's name, location, street, cuisine type, number of Cuisines, price category, low price(Price-1), high price(Price-2), ranking, number of reviews, total ratings, ratings of the four criteria(food, service, value and atmosphere) and url of restaurant on TripAdvisor. In fact, these insights are extracted using methods from Python's library BeautifulSoup and urllib2.

In conclusion, in addition to the data mentioned above for each restaurant, we also created a system for collecting and storing reviews published by customers of restaurants.

Figure 4.1 presents an example of a restaurant on the publicly accessible and viewable TripAdvisor website [68].

FIGURE 4.1: Restaurant's details on TripAdvisor [68]

4.3 Web Data Source "Hellenic Chamber of Hotels"

According to [75] the "Hellenic Chamber of Hotels" has operated since 1935 as a Legal Entity of Public Law. Additionally, it is the institutional consultant of the Government as far as tourism and hospitality issues are concerned. Moreover, its members are, by law, all the hotels and camping sites of Greece.

In addition, as mentioned in [76] the administration of "Hellenic Chamber of Hotels" belongs to an Administrative Council of elected representatives of hotels and camping sites as well as of representatives of the State. Also, it is a member of the Confederation of National Associations of Hotels, Restaurants and Cafeterias of EU member-states (HOTREC). As well, the Chamber's membership is about 10.000 and classical hotels are the most numerous. It is remarkable that the highest number of hotels-members are in Crete, followed by Macedonia and Central Greece regions.

In my research, I follow a similar approach as described for the TripAdvisor's data collection using BeautifulSoup and urllib2 libraries. However, despite the similarity of data collection methods the website structure of "Hellenic Chamber of Hotels" is significantly different from the one of TripAdvisor. For this reason, I developed a separate system to collect data about all Greek hotels. Specifically, I stored the data of the Hotels in csv files. These files contain information about hotels across the country and are broken down by geographical region. These regions are Attica, Central Greece, Central Macedonia, Crete, Eastern Macedonia and Thrace, Ioanian Islands, Ipeiros, Northen Aegean, Pelloponissos, Southern Aegean, Thessalia, Western Greece, Western Macedonia. Finally, the collected data provides insights about hotel's name, distance from hospital, stars of hotel, website url, e-mail, phone number, phone number 2, alternate phone number, community, city, address and zip code.

Figure 4.2 presents an example of a hotel in "Hellenic Chamber of Hotels" platform.

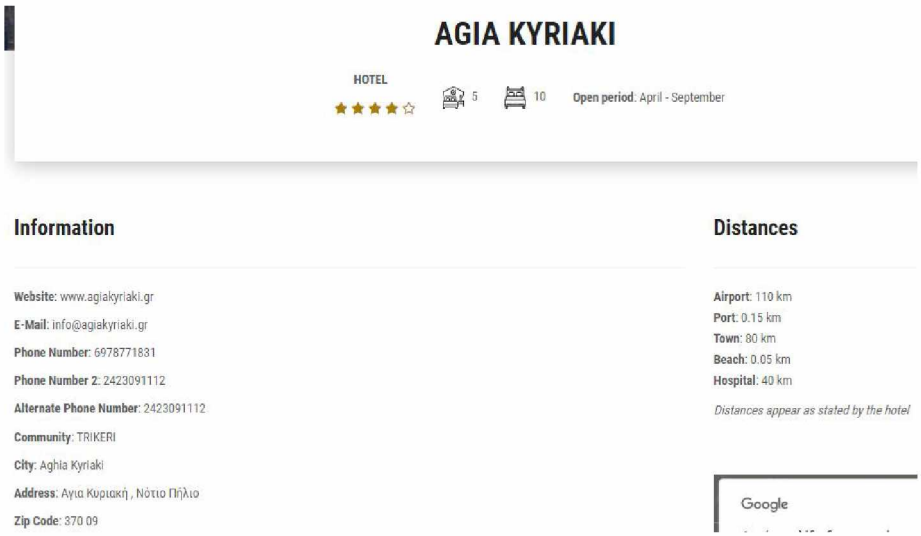

FIGURE 4.2: Hotel's details example [77]

4.4 Web Data Source "YouTube"

A very important source of knowledge and public opinion, wouldn't be something else than Youtube. As mentioned in [78] YouTube is an American online video-sharing platform which has its registered office in San Bruno, California. Particularly, YouTube enables users to upload, view, rate, share, add to playlists, report, comment on videos, and subscribe to other users. Nowdays, Youtube provides a massive volume of different forms of unstructured data. It succesfully combines video, audio but also text data. Huge amount of text data can be found in the comments of each video. These comments often express some kind of feedback or critisism about the video or its content that it attempts to issue. Based on this realization I decided to create a script API based on the official Youtube API. Using this app, after entering the necessary credentials needed, I enter the Youtube video's search identifier (Which is basically the title of video(s) I want to search for) from which I want to extract the comments , and the number of videos, the Youtube search engine extracted, based on their popularity. Finally, I collect the comments and save them into a file. Except from the raw text of each comment I also hold its number of replies, likes and also the title of the video where the comment came from.

Chapter 5

Data Preprocessing

This chapter focuses on the techniques used to extract useful insights from the unstructured SEC documents, especially in 10-K Forms. More specifically, because of SEC document syntax preprocessing techniques need to be applied before conventional Natural Language Preprocessing (NLP) methods can be employed. Therefore, at first I process the characteristics of SEC documents, followed by common methods such as regular expressions(regex), BeautifulSoup and NLP to find useful financial information in 10-K Forms. In particular, I extract text from Items 1A, 7, and 7A of 10-K Forms. Consequently, I describe common text mining techniques which prepare the data for oncoming machine learning tasks and text summarization.

5.1 SEC Document Preprocessing

In general, web documents are very heterogeneous and differ in several ways, for instance the format in which the document is displayed, the type of information it contains, the extent to which it is structured and whether it contains metadata [68]. In addition, according to [79] a web document can be divided into three levels. Specifically, these layers are content, structure and layout. The content related to the actual insights of a document such as the form of textual, numerical or visual data. Additionally, structure refers to the organization of the document, i.e. the links, paragraphs, headings and elements of visual communication such as lists or tables. Finally, layout describes the style and visual representation of the document and includes the size, position, color and font of the structural part. Further, web documents might contain metadata that supplies insights about the document itself [68].

In my thesis I applied regex and BeautifulSoup to find useful financial insights in 10-K Forms. In fact, I extracted text from Items 1A, 7, and 7A of 10-K Form. At first I used regex to find the 10-K Section from the SEC document. Also, I identify patterns of Items 1A, 7, and 7A of a 10-K document.

As shown in Figure 5.1 the items are pretty messy and they contain HTML tags, unicode characters, etc.

```
1 item_1a_raw[0:1000]
```

```
'>Item 1A.</font></div></td><td style="vertical-align:top;"><div style="line-height:120%;text-align:justify;font-siz
e:9pt;"><font style="font-family:Helvetica,sans-serif;font-size:9pt;font-weight:bold;">Risk Factors</font></div></td>
</tr></table><div style="line-height:120%;padding-top:8px;text-align:justify;font-size:9pt;"><font style="font-famil
y:Helvetica,sans-serif;font-size:9pt;">The following discussion of risk factors contains forward-looking statements.
These risk factors may be important to understanding other statements in this Form 10-K. The following information sh
ould be read in conjunction with Part II, Item 7, &#8220;Management&#8217;s Discussion and Analysis of Financial
Condition and Results of Operations&#8221; and the consolidated financial statements and related notes in Part II, It
em 8, &#8220;Financial Statements and Supplementary Data&#8221; of this Form 10-K.</font></div><div style="line-
height:120%;padding-top:16px;text-align:justify;font-size:9pt;"><f'
```

FIGURE 5.1: Raw data of Item 1A

Thus, it is necessary to clean these data before we can do Natural Language Processing. As a result, this means we need to remove all HTML tags and unicode characters. For this reason, I used 'lxml' BeautifulSoup's parser to remove html tags and see the content of items 1A, 7 and 7A. Thus, the combination of regex and BeautifulSoup lead us to extracting and scraping content from 10-K documents of SEC filings.

Figure 5.2 shows an example of Item 1A content after cleaning.

```
>Item 1A.

Risk Factors

The following discussion of risk factors contains forward-looking statements. These risk factors may be important t
o understanding other statements in this Form 10-K. The following information should be read in conjunction with Pa
rt II, Item 7, "Management's Discussion and Analysis of Financial Condition and Results of Operations" and the cons
olidated financial statements and related notes in Part II, Item 8, "Financial Statements and Supplementary Data" o
f this Form 10-K.

The business, financial condition and operating results of the Company can be affected by a number of factors, whet
her currently known or unknown, including but not limited to those described below, any one or more of which could,
directly or indirectly, cause the Company's actual financial condition and operating results to vary materially fro
m past, or from anticipated future, financial condition and operating results. Any of these factors, in whole or in
part, could materially and adversely affect the Company's business, financial condition, operating results and stoc
k price.
```

FIGURE 5.2: Content of Item 1A

5.2 Natural Language Preprocessing

As mentioned in previous chapters, unstructured data accounts for about 80%, and the majority of this data exists in the textual form which is a highly unstructured format [80]. For this reason, Text Analysis methods are necessary in order to produce meaningful insights from the text. Particularly, the process of deriving meaningful insights from natural language text called Text Mining.

More specifically, we used NLP which deals with human languages and is a part of computer science and artificial intelligence [80]. In fact, NLP is a component of text mining that performs a special kind of linguistic analysis that essentially helps a machine "read" text [80]. Further, NLP uses a different methodology to decipher the ambiguities in human language. These methods are automatic summarization, disambiguation and natural language understanding and recognition.

5.2.1 Main steps in NLP pipeline

The main steps that appear in the NLP pipeline include:

Natural Language Toolkit (NLTK) library

At first it is necessary to install the NLTK library. This library provides an easy to use interface and a toolkit for building Python programs to work with human language information [80] [81].

Tokenization

As mentioned in [80] tokenization is the process of breaking strings into tokens which in turn are small structures or units. Also, tokenization includes three steps: (a) breaking a complex sentence into words, (b) understanding the importance of each word with respect to the sentence and (c) producing structural description on an input sentence.

Stemming

"Stemming usually refers to normalizing words into its base form or root form" [80]. Figure 5.3 gives an example of stemming with words waited, waiting and waits. As seen from these words the root word is "wait". Further, there are two methods in Stemming. First, Porter Stemming which removes common morphological and inflectional endings from words. Second, Lancaster Stemming which is a more aggressive stemming algorithm [80].

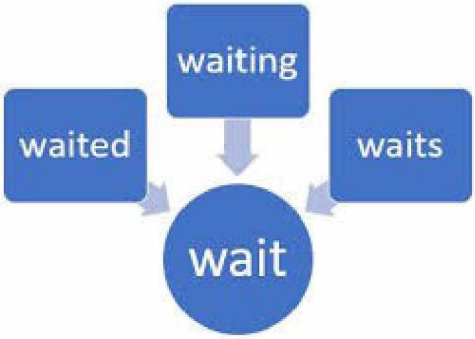

FIGURE 5.3: Stemming example [80]

Lemmatization

Lemmatization is the method of converting a word to its base form. However, the difference between stemming and lemmatization is that lemmatization reflects on the context and converts the word to its meaningful base form. On the contrary, stemming just removes the last few characters, often leading to incorrect meanings and spelling errors. For instance, lemmatization would correctly identify the base form of "caring" to "care" while stemming would remove the 'ing' part and convert it to car. Finally, lemmatization can be used in python by applying Wordnet Lemmatizer, Spacy Lemmatizer, TextBlob, Stanford Core NLP [80].

Stop Words

"Stopwords" are the most usual words in a language (e.g. "the", "a", "at", "for", "above", "on", "is", "all"). In fact, these words do not lead to any meaning and are commonly removed from texts. So, we remove these "stopwords" using the nltk library [80].

5.2.2 Word embeddings for Text

As mentioned earlier, NLP is a process of analysis and interaction with human language, in order to extract useful insights such as sentiment or to summarize the document. In particular, according to [82], word embeddings are a type of word representation that allows words with similar meaning to have a similar representation. More specific, these techniques mapping words and sentences to vectors of real numbers and the basic idea is that the more similar sentences are, the closer the vectors are. In simple words, word embeddings are techniques by which we can understand whether the sentences and words have similar meaning.

Then I will mention three of these methods that I have used in my thesis:

1. **TF-IDF**

   As mentioned in [83] "Tf-idf stands for term frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus". In addition, search engines often utilize this technique as a tool in scoring and ranking a document's relevance given a user query [83]. Also, Tf-idf can be successfully used for stopwords filtering, text summarization and classification [83].

   Term frequency (TF) shows us how frequently a word appears in a document. In fact, it is the number of times the term occurs in a document divided by the total number of words in that document. As a result, TF increases as the number of appearances of that term within the document increases [84].

   $$TF(term) = \frac{\text{Number of times term appears in a document}}{\text{Total number of items in the document}} \tag{5.1}$$

   Inverse Data Frequency (IDF) present the importance of a term in the document. Specifically, the rarer the term, the greater its IDF value. It is true that some stopwords (e.g. "the", "is", "a", "are") are commonly used in documents, so they have high score in TF. Therefore, in order to weigh down the frequent words it is needed to compute IDF [83] [85].

   $$IDF(term) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents with term in it}}\right) \tag{5.2}$$

   TF-IDF is the combination of the above two types, so the score (w) for a word in a document is:

   $$TFIDF(term) = TF(term) * IDF(term) \tag{5.3}$$

2. **Bag-of-Words**

   The bag-of-words model is an algorithm which is used in NLP and information retrieval (IR) [86]. In particular, this technique represents the document as a "bag" of words without caring about grammar or structure of words. In other words, this model is intersted only about whether known words appear in the text. Also, this model is commonly used in classification methods [87].

As mentioned in [88] the bag-of-words model includes two basic steps:

Step 1
A vocabulary of known words: In this step the model collects all the words that occurs in document and creates a vocabulary with unique words of documents.

For instance, if we have the following documents:
D1: He is a lazy boy. She is also lazy.
D2: Smith is a lazy person.

These documents create a unique vocabulary (whithout stopwords)

('He', 'She', 'lazy', 'boy', 'Smith', 'person')

Step 2
Measuring of the presence of known words. In our case the length of words in vocabulary is six and the number of documents is two. Therefore, we will create a matrix of size 2 X 6. Each document is represented by a row and the cells of row give us the amount of the word appearances.

	He	She	lazy	boy	Smith	person
D1	1	1	2	1	0	0
D2	0	0	1	0	1	1

FIGURE 5.4: Bag-of-Words matrix [88]

3. **GloVe**

According to [89] GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Especially, training is carried out on aggregated global word-word co-occurrence statistics from a document, and the resulting representations show interesting linear substructures of the word vector space.

Chapter 6

Text Summarization

As mentioned in [90] summaries are the most efficient way of reducing the length of a document. Particularly, in books, abstracts is a way of representing the condensed form of the document. However, according to the literature summary of a document is a reduced and precise, representation of the text which seeks to convey the exact idea of its content [90]. More specific, automatic summarization is generated by software or an algorithm and is a process of generating a concise and meaningful summary of text from multiple text resources such as books, news articles, blog posts, research papers, emails, and tweets. In fact, Automatic Text Summarization is one of the most significant and interesting problems in the field of NLP [85].

Nowadays, due to the availability of huge amounts of textual data, automatic text summarization systems are necessary [85]. Also, the most of this huge amount of data are documents in non-structured format. Therefore, processing documents is a challenging task [90]. In conclusion, the plethora of data, the lack of manpower and the time that is required to interpret the insights gives to Automatic Text Summarization methods a crucial role in business intelligence.

In general, there are two types of Automatic Text Summarization techniques: (a) abstractive summarization and (b) extractive summarization.

Abstractive summarization: Abstractive methods select words based on semantic understanding. Also, some words may not even appear in the source documents. In addition, it aims at producing important material in a new way. Thus, they interpret and examine the text using advanced natural language techniques in order to generate an entirely new summary that conveys the most significant insights from the original text [91]. However, due to advanced automatic summarization techniques abstraction-based summarization has not been widely developed. In fact, only a few abstract summarization systems have been created and is almost unattainable to develop a genuine automatic text understanding [90] [92].

Extractive summarization: Extractive Summarization: Extractive methods based on selecting several parts of text like phrases and sentences and stack them together to make a summary. Thus, identifying the most important sentences for summarization is crucial in an extractive method [85]. Specifically, the basic idea is to first separate the document into sentences. Then, the algorithm evaluates the sentences using statistical methods and extracts the sentences with the highest scores. So, the algorithm utilizes the most significant points to create the summary [92].

There are several algorithms and techniques that are used to evaluate the sentences and rank them based on significance and similarity among each other. In my thesis, I focused on the extractive summarization technique and I have developed seven algorithms that I used to generate the automatic summaries for the Items 1A, 7 and 7A of 10-K Forms. These algorithms are Luhns Heuristic Method, SumBasic, LexRank, Latent Semantic Analysis, KLSum, Edmundson Heuristic and TextRank. In the following paragraphs I provide an analysis of how each algorithm works.

6.1 Luhns Heuristic Method

Luhn algorithm is one of the first text summarization algorithms was published in 1958 by Hans Peter Luhn, working at IBM research [92]. In particular, this algorithm is based on TF-IDF technique, that generates a summary from given words. According to [93] there are two main phases for its implementation:

1. Defining the words that are most important to the meaning of the text. This is achieved with doing a frequency analysis and finding words which are important, but not stopwords.
2. Finding out the most ordinary words in the text, and then take a subset of those that are not stopwords, but are still significant. This process consists of three steps: (a) Transforming the content of sentences into a mathematical expression or vector and measuring the value of sentence, (b) Evaluating sentences using sentence scoring technique, specifically,

$$Score = \frac{(\text{Number of meaningful words})^2}{(\text{Span of meaningful words})} \tag{6.1}$$

and (c) selection of sentences with the highest overall rankings.

In Figure 6.1 is displayed an example of a sentence with ten words and four of the words are important. As we can observe, the span of meaningful words is six. Therefore, the score of sentence is given by the equation:

$$Score = 4^2/6 = 2.7 \tag{6.2}$$

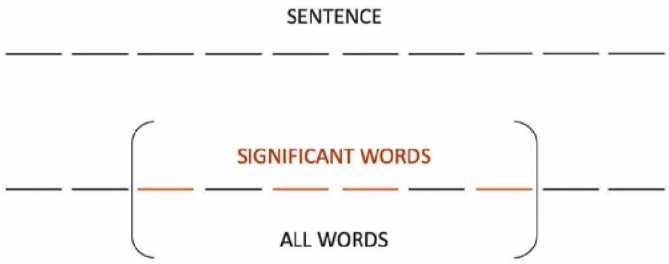

FIGURE 6.1: Sentence example [93]

6.2 SumBasic

The SumBasic algorithm was developed in 2005 and relies only on word probability to calculate significance. More specifically, SumBasic first computes the probability of each content-word $p(w_i)$ by simply counting its frequency in the document set. Then, the average of the probabilities of the words in a sentence computes the weight of a sentence. Subsequently, SumBasic selects the best scoring sentence that contains the word that currently has the highest probability. Accordingly, the selection loop is repeated until the specified summary length is achieved. Furthermore, SumBasic updates probabilities of the words in the selected sentence by squaring them, in order to avoid the likelihood of a word appearance twice in a summary [92] [94] [95].

6.3 Latent Semantic Analysis

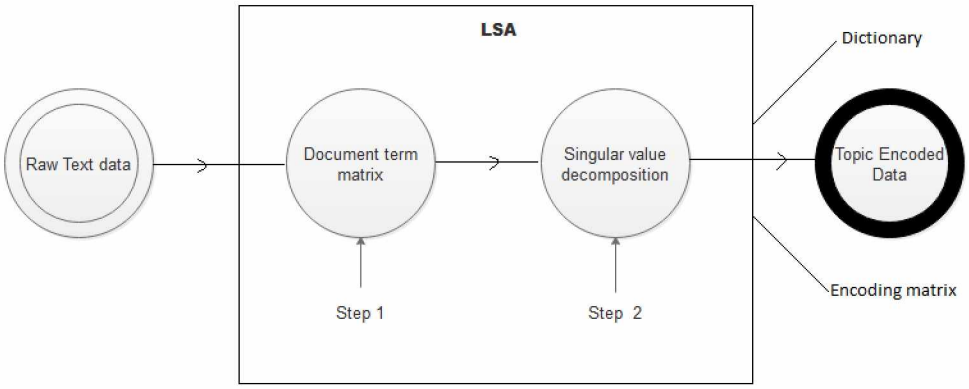

FIGURE 6.2: LSA processing[96]

According to [96] Latent Semantic Analysis (LSA) is an unsupervised method along with the usage of NLP. Notably, is an Algebraic-Statistical method which extracts semantic structures of words and sentences like features that cannot be directly mentioned. Also, these characteristics are essential to data, but are not original characteristics of the dataset.

Particularly, the LSA has three main stages:

1. Creation of input matrix: The input text is represented as a matrix to perform calculations on it. So, the algorithm generates a text term matrix. Additionally, the cells are used to represent the significance of words in sentences. Also, there are some different methods that can be used for filling out the cell values. These methods are: (a) Frequency of word, (b) Binary representation, (c) TF-IDF, (d) Log entropy and (e) Root type [96].

2. Singular Value decomposition (SVD): In this step we perform the decomposition of singular value on the generated document term matrix. Specifically, SVD is an algebraic method that can model relationships among words, phrases and sentences. Also, the main idea of SVD is that the document term matrix can be represented as points in Euclidean space known as vectors. Thus, these vectors are used to show the sentences in this space. However, having the capability of modelling relationships among sentences, SVD has the helps to improve accuracy as well as noise reduction [96].

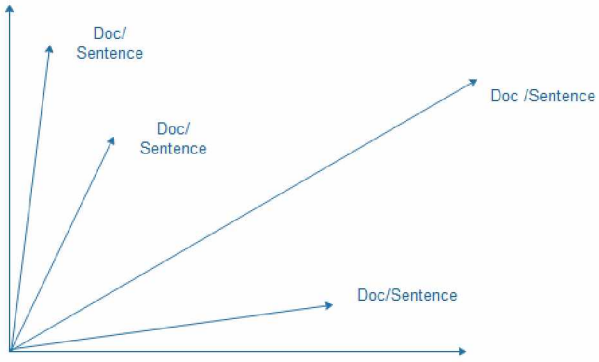

FIGURE 6.3: Euclidean plane[96]

3. Sentence Selection: there are several algorithms to select important sentences. In the above figure we see Topic method to extract concepts from the SVD calculations and are called topics of the input document [96].

6.4 KLSum

The KLSum algorithm (Kullback-Lieber (KL) Sum algorithm) as mentioned in [92] selects a set of sentences from the source text, D, thus the distribution of words in the selected sentences, S, is as close as possible to distribution of words in document D. Also, the algorithm introduces the following criteria for selecting a summary S from given document D:

$$S^* = min(KL(P_D||P_S)) \tag{6.3}$$

$$S : words(S) \leq L \tag{6.4}$$

where $P_S(P_D)$ is the empirical unigram distribution of candidate summary S (document D). To measure similarity across the word distributions, P_S and P_D, the Kullback-Lieber (KL) divergence measure is used. Finally, as mentioned in [97] the definition of KL is:

$$KL(P,Q) = \sum_{i=1}^{n}[p_i * \log \frac{p_i}{q_i}] \tag{6.5}$$

6.5 Edmundson Heuristic

According to [98] Edmundson Heuristic Method for text summarization named after its creator Harold Edmundson in 1969, when he developed his text summarization method. Especially, this algorithm suggests the use of a subjectively weighted combination of features unlike traditionally used feature weights generated using a corpus. As mentioned in this method there are some features which are more important for the summarization. Edmundson considered these features such as Cue Words (C) and Document structure (S) (i.e. headlines, titles, sub-titles etc.) as "bonus words". In total, before 1969 there were two already known features Position (P) and

Word frequency (F) and two new from Edmundson Cue Words (C) Document structure (S).

For scroing, these 4 features we consider:

$$Score = (w_1 * P) + (w_2 * F) + (w_3 * C) + (w_4 * S)$$

Moreover, there are four sub-methods for weighing the words and sentences. These methods are Cue method, Key method, Title method, Location method.

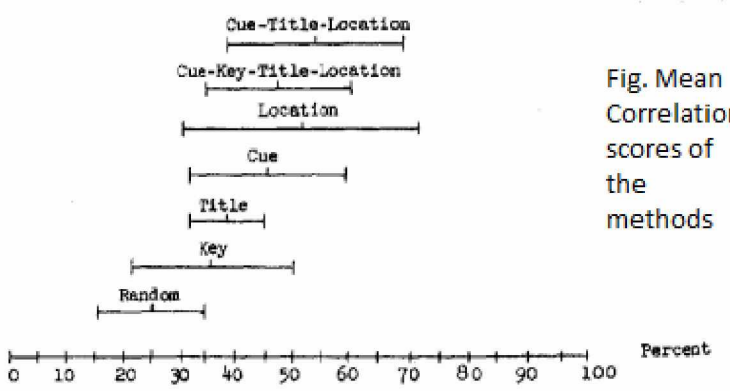

FIGURE 6.4: Mean Correlation scores of the methods[99]

6.6 LexRank

LexRank is an unsupervised graph based method for automatic text summarization. This method is based on the logic of the PageRank algorithm. In fact, PageRank is used to calculate rank of web pages and is used by search engines such as Google [100]. More specific, according to [101] the scoring of sentences is done using the graph method. Also, LexRank for sentence significance computing is based on the concept of eigen vector centrality in a graph representation of sentences. Thereafter, a sentence node is ranked according to its similarity with other nodes. Particularly, S_i is represented as a set of words:

$$S_i = w_1^i, w_2^i, \cdots w_{|S_i|}^i \tag{6.6}$$

Thus, the similarity between two sentences S_i and S_j is defined as:

$$Sim(S_i, S_j) = \frac{|w_k : w_k \in S_i \wedge w_k \in S_j|}{\log(|S_i|) + \log(|S_j|)} \tag{6.7}$$

Moreover, the connectivity matrix is based on intra-sentence cosine similarity which is used as the vicinity matrix of the graph representation of sentences. In other words, the algorithm finds the relative significance of all words in a text and selects the sentences which contain the most of those high-scoring words. Then, the N highest ranked sentences are selected for the summary, where N is the number of sentences we defined for the summary [92].

The following Figure(6.5) presents the graphical approach, which is based on Eigen vector centrality. As we can observe sentences are placed at the vertices of the

graphs and the weight on the Edges are calculated using cosine similarity metric. Also, in the following figure Si are the sentences at the vertices respectively and Wij are weights on the edges [101].

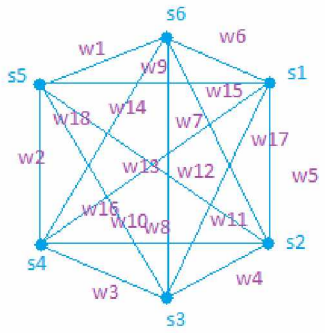

FIGURE 6.5: LexRank graphical approach [101]

6.7 TextRank

In Addition, a similar algorithm with LexRank is the TextRank. In particular,TextRank is an extractive and unsupervised text summarization technique [85].

Figure 6.6 give us the way in which this algorithm works.

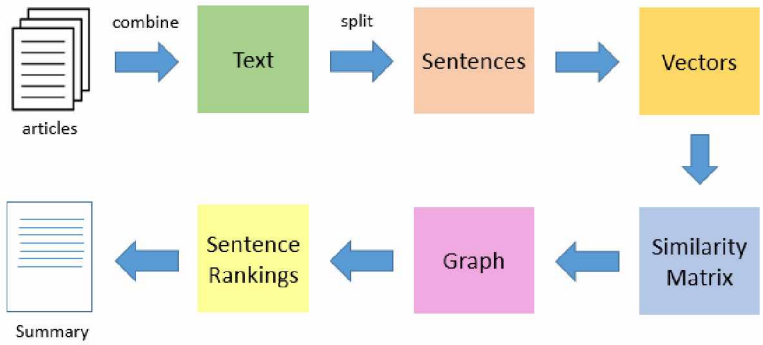

FIGURE 6.6: TextRank processing[85]

As mentrioned in [85], we can observe the main stages of the TextRank algorithm are:

1. Combination of texts that are contained in the documents.
2. Extraction of sentences from text, namely splitting the text into sentences.
3. Vector representation for each sentence.
4. Similarities of vectors are then calculated and create a matrix.
5. A Graph is created from the similarity matrix. Specifically, the sentences represented as nodes and the similarity scores on the edges between two nodes is calculated with a Similarity function such as Cosine Similarity or Jaccard Similarity.

6. At the end, scores of sentences are sorted in a descending order and the top N ranked sentences are selected to create the summary.

6.8 Conclusion

To sum up, according to [92], LexRank technique constitutes the most remarkable algorithm in relation to the other algorithms. In addition, LexRank produces summaries that are consistently rated as equal or superior to management summaries. This result is demonstrated through the experiment they performed to evaluate automated summary algorithms in relation to managers' summaries. It is fact that the management summaries are prone to bias and often have a positive and strategic tone. In contrast, automatic summaries are objective and may lead investors to safer judgments.

Chapter 7

Web Data Analysis

In this chapter I present methods in order to analyze data from TripAdvisor. Specifically, in paragraph 7.1 I provide data visualizations and basic features of my dataset. Then, paragraph 7.2 contains sentiment analysis models for reviews of customers and word frequency analysis.

7.1 Data Visualizations

It is a fact that in order to make my dataset more understandable it is important to create data visualizations. Our TripAdvisor dataset contains insights from 591 Thessaly restaurants region of Greece. Particularly, these data come from Larissa, Volos, Trikala and Karditsa. The pictures below give us answers to questions, such as, how the restaurants are distributed in each area, what cuisines do these restaurants represent, what are the average restaurant ratings, and what is the average cost of restaurants in Thessaly. As a result, these visualizations make the understanding of the dataset clearer and help us to continue in further analysis of the data.

Location of Restaurants

In total I have collected 591 restaurants from Thessaly. Therefore, the restaurants come from four different cities of Thessaly, Larissa, Volos, Trikala and Karditsa as the following figure presents.

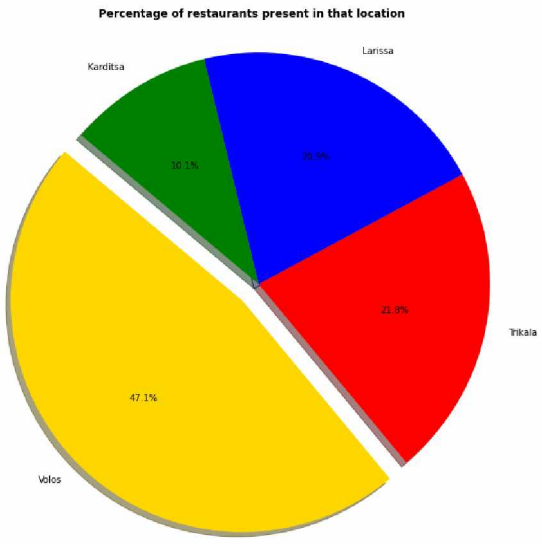

FIGURE 7.1: Percentage of restaurants present in that location

Cuisines

In general, as we can perceive in these restaurants the Greek cuisine excels in comparison to other cuisines.

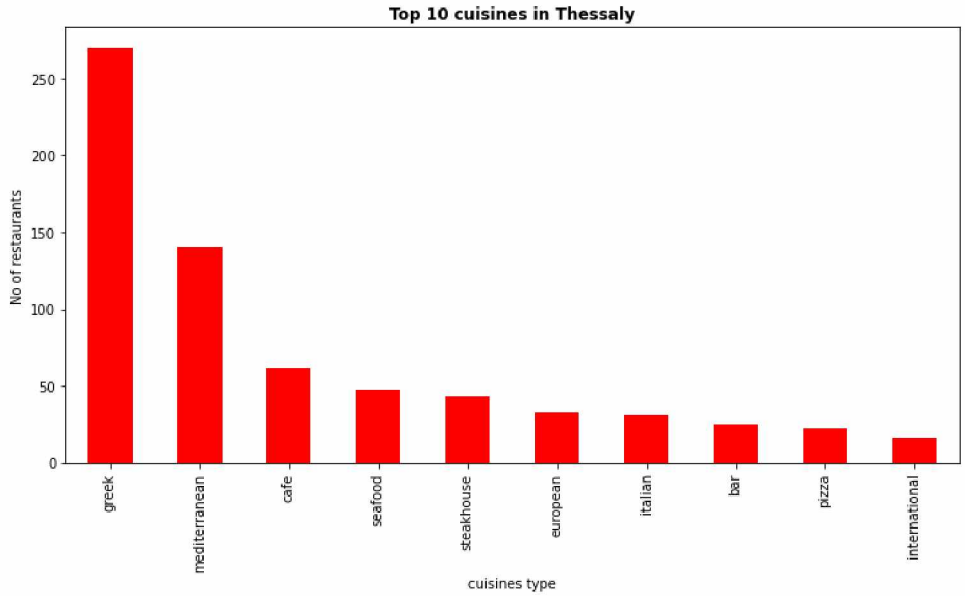

FIGURE 7.2: Top 10 cuisines in Thessaly

Ratings

As we observe most restaurants are rated almost excellent from customers.

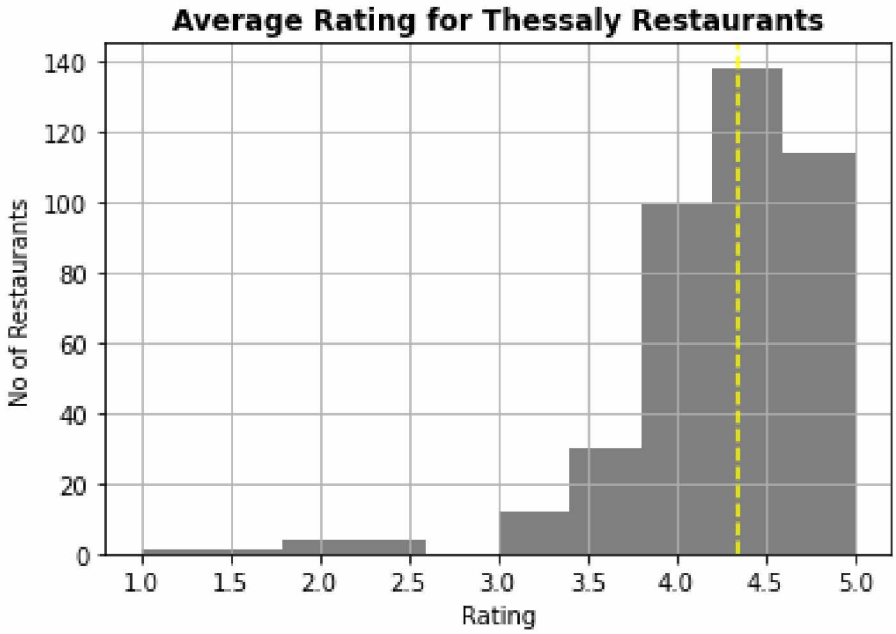

FIGURE 7.3: Average rating for restaurants

Average cost

The following plot presents average costs of restaurants grouped into categories. The cost depends on restaurant type, dishes and cuisine.

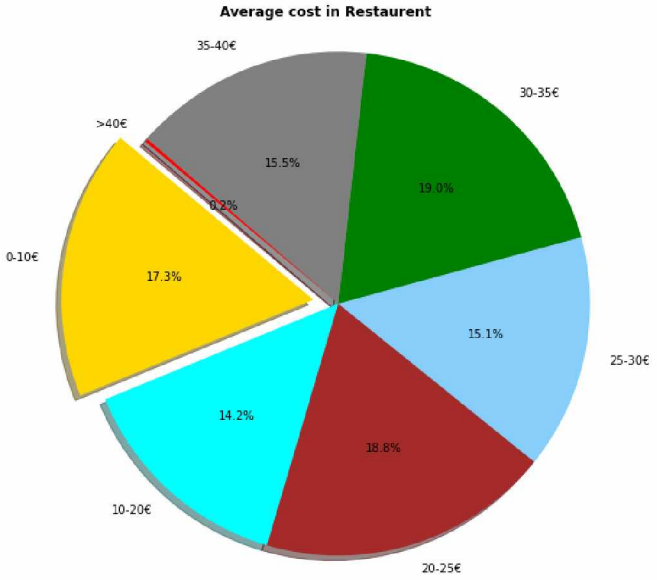

FIGURE 7.4: Average cost

7.2 Sentiment Analysis

This paragraph provides various techniques for sentiment analysis and word frequency analysis in comments of customers. By Applying these methods I predict whether a comment is positive or negative. I also compare the different techniques with each other in order to use the most effective one. Sentiment analysis is a necessary and very useful tool for any enterprise. Especially, restaurant owners can use it to develop their strategies and to understand customers' feelings [102].

7.2.1 Machine Learning models

In my thesis in order to analyze comments from restaurants in TripAdvisor, I applied supervised machine learning algorithms. In particular, supervised learning is the machine learning task of learning a function that maps an input (x) to an output (Y) based on example input-output pairs [103]. The basic concept is to approach the mapping function in order to predict the output variables (Y) well, utilizing input data (x) [88]. Also, these algorithms are learning from the training dataset. Therefore, it is understood that this procedure looks like a teacher who supervises the process. Further, the algorithm is trained by making predictions on data to which we know the answers and thus is corrected by the teacher [88].

Supervised learning methods can be used for both regression and classification problems. More specifically, classification belongs to the problems which the output variable is a category, such as "white" or "black" or in our case "Positive" or "Negative". On the other hand, in regression belongs the problems which the output variable is a real value, such as "dollars" or "weight" [88].

In my case I will examine the classification problems, because I have to classify my results as "Positive" or "Negative". Particularly, my dataset consists of comments posted by people who visited the restaurants. Thus, I split my dataset into training and test-set and I used features from Bag-of-Words and TF-IDF for my machine learning models. Moreover, I applied two different models: (a) Logistic Regression and (b) Decision Trees.

Figure 7.5 give us the basic metrics of confusion matrix i.e. Precision, Recall, Accuracy.

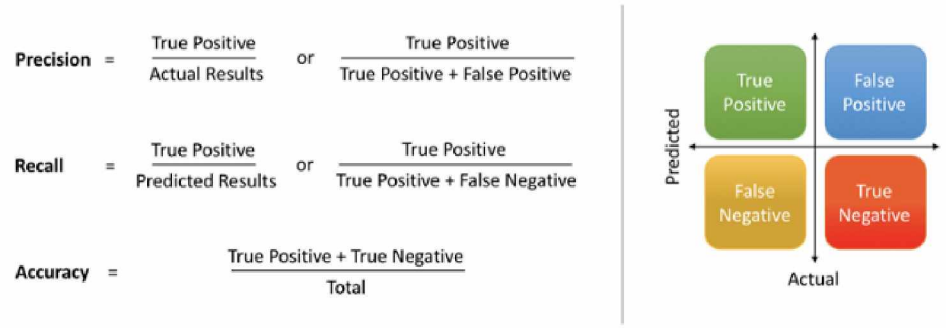

FIGURE 7.5: Precision-Recall-Accuracy metrics[88]

However, in my case I used F1 Score in order to examine the machine learning models. I selected F1 Score because as Figure 7.6 shown, my dataset is extremely imbalanced, as it has a high percentage of positive reviews.

$$\textbf{F1 Score} = 2 * \frac{\text{Percision} * \text{Recall}}{\text{Percision} + \text{Recall}} \tag{7.1}$$

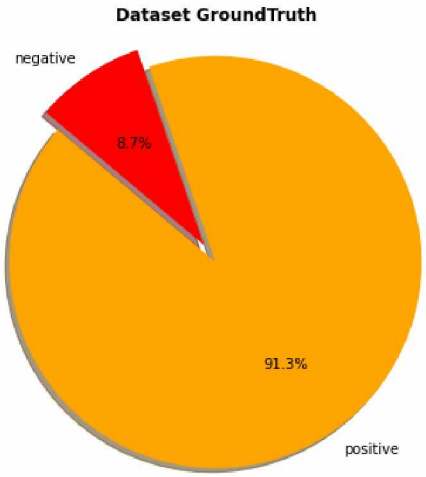

FIGURE 7.6: Percentage of positive and negative comments

As is evident from the above plot, my dataset is quite imbalanced. As we see my dataset has 91.3% of "Positive" comments and 8.7% of "Negative". In particular, the comments which have ratings with a higher score than 3 of 5 bubbles are "Positive" and the others are "Negative". So, because the data is quite imbalanced I will use F1 score, to examine the performance of methods, instead of accuracy. Also, because

the positive comments are much more than the negative ones, it is possible to come across a large number of false positives. Therefore, F1 score will be more reliable in terms of Accuracy metric.

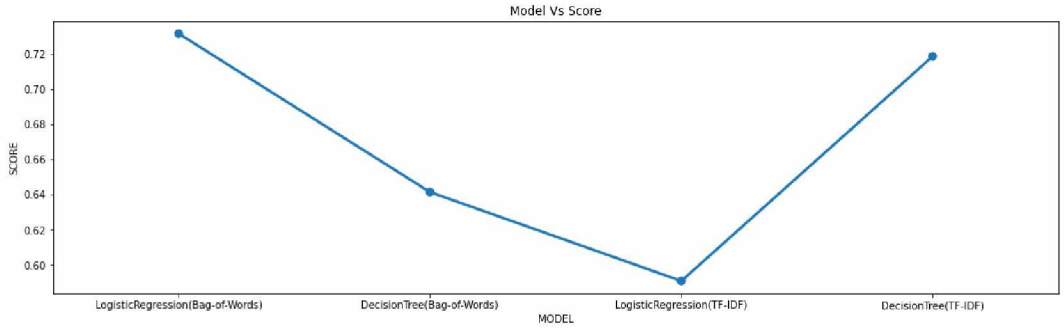

FIGURE 7.7: Supervised models vs Score

	1	2	3	4
Model	LogisticRegression(Bag-of-Words)	DecisionTree(Bag-of-Words)	LogisticRegression(TF-IDF)	DecisionTree(TF-IDF)
F1_Score	0.731707	0.641509	0.590909	0.71875

FIGURE 7.8: Supervised models score

As we observe, I tested the two algorithms for both Bag-of-Words and TF-IDF techniques. As a result, the best possible method from both Logistic Regression and Decision Trees is Logistic Regression using Bag-of-Words features.

7.2.2 Vader sentiment analysis

The way TripAdvisor comments are structured poses serious challenges for successfully analyzing the sentiment of the comments. A fairly successful tool for sentiment analysis in texts coming from social media is VADER (Valence Aware Dictionary and sEntiment Reasoner). More specific, VADER is a lexicon and rule-based sentiment analysis tool. This tool uses a list of lexical features which are classified according to their semantic orientation as either positive or negative [3]. Also, it is significant that this tool not only informs us whether a comment is positive or negative but shows us how positive or negative it is. In addition, other benefits of VADER are its accuracy on social media type text, the non-requirement of training data, the speed and the flexibility [3] [104].

In my research I used the Compound score of VADER. Especially, according to [102] the Compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1 (most extreme positive). To sum up, as it seems in the following figure this method gives us 93% accuracy.

```
In [58]: 1 print(accuracy_score(restaurantDf['GroundTruth'],restaurantDf['comp_score']))
         0.9336538461538462
```

FIGURE 7.9: Vader accuracy

7.2.3 Word frequency Analysis

In this paragraph I calculate word frequency in order to present the most important positive and negative unigrams and bigrams. Hence, I calculate mutual information (MI) and pointwise mutual information (PMI) to better understand the correlations between the words. In this way, enterprises can identify what are the complaints of customers or what restaurant customers like best.

More specifically in this paragraph I will examine 3 methods: (a) word frequency, (b) mutual information (MI) and (c) pointwise mutual information (PMI).

The mutual information (MI) is also named as information gain and is a metric that gives us information about the presence or absence of a word. More specifically, this metric shows us how independent a pair of words is. Also MI is based on PMI, but MI examines the average of all possible events [105] [106].

On the other hand, the Pointwise mutual information (PMI) is a metric that examines single events. More specific, I use PMI in order to find out when a couple of words are independent or are a single expression. For instance, in the expression "social media" both the words can have independent meaning, but, when they are together, they express a different meaning. From the above it is understood that this aspect of NLP is very challenging. Thus, PMI is called upon to face this challenge quantifying the probability of two words coexisting, taking into account the fact that it can be caused by the frequency of individual words [107]. Therefore, the algorithm calculates the (log) probability of coexistence as follows:

$$PMI(a,b) = \log\left(\frac{P(a,b)}{P(a)P(b)}\right) \tag{7.2}$$

Consequently, from the above formula it follows that when 'a' and 'b' are independent the ratio equals 1 and the log equals 0 [107].

The following figures give the results obtained for (a) word frequency, (b) mutual information (MI) and (c) pointwise mutual information (PMI).

The most frequently occured top 10 words in positive reviews

	Word	Count
0	larissa	140
1	amazing	97
2	salad	77
3	just	72
4	recommended	61
5	cooked	61
6	definitely	52
7	visited	51
8	highly	49
9	helpful	48

FIGURE 7.10: Top 10 words in positive reviews

The most frequently occured top 10 words in negative reviews

	Word	Count
0	table	17
1	visited	15
2	informed	14
3	available	14
4	salads	11
5	quite	10
6	went	10
7	don	9
8	places	9
9	low	9

FIGURE 7.11: Top 10 words in negative reviews

The most frequently occured top 10 bigrams in positive reviews

	Word	Count
0	fresh ingredients	33
1	highly recommended	33
2	vegetarian options	15
3	make feel	15
4	days ago	13
5	places larissa	12
6	trip advisor	12
7	amazing totally	11
8	just winebar	11
9	don hesitate	10

FIGURE 7.12: Top 10 bigrams in positive reviews

The most frequently occured top 10 bigrams in negative reviews

	Word	Count
0	table available	14
1	english communication	7
2	speak english	7
3	deserved resort	7
4	immediately informed	7
5	visited friday	7
6	mainly watery	7
7	friday late	7
8	joint students	7
9	satisfaction low	7

FIGURE 7.13: Top 10 bigrams in negative reviews

```
Mutual Information - Unigram
Out[68]:
```

	MI Score
mainly	0.019147
half	0.019147
gives	0.019147
available	0.017446
informed	0.016959
offer	0.016890
table	0.016840
great	0.015808
friendly	0.014277
exhausting	0.013900

FIGURE 7.14: Mutual Information unigram

```
Mutual Information - Bigram
Out[69]:
```

	MI Score
half dishes	0.019147
greeted door	0.016717
deserved restaurant	0.016717
exhausting offer	0.016717
far higher	0.016717
hard exhausting	0.016717
table available	0.016717
soon informed	0.016717
staff speak	0.013900
quality ingredients	0.013900

FIGURE 7.15: Mutual Information bigram

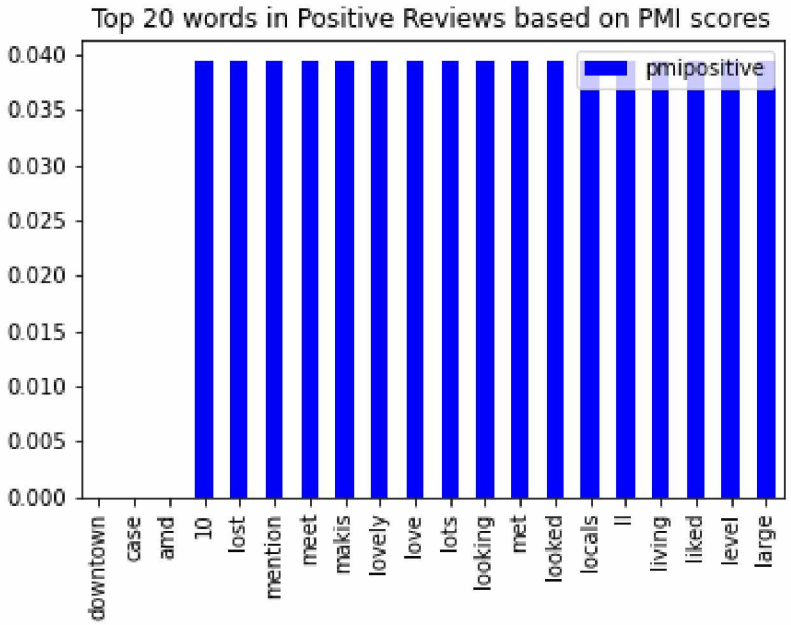

FIGURE 7.16: Top 20 PMI positive words

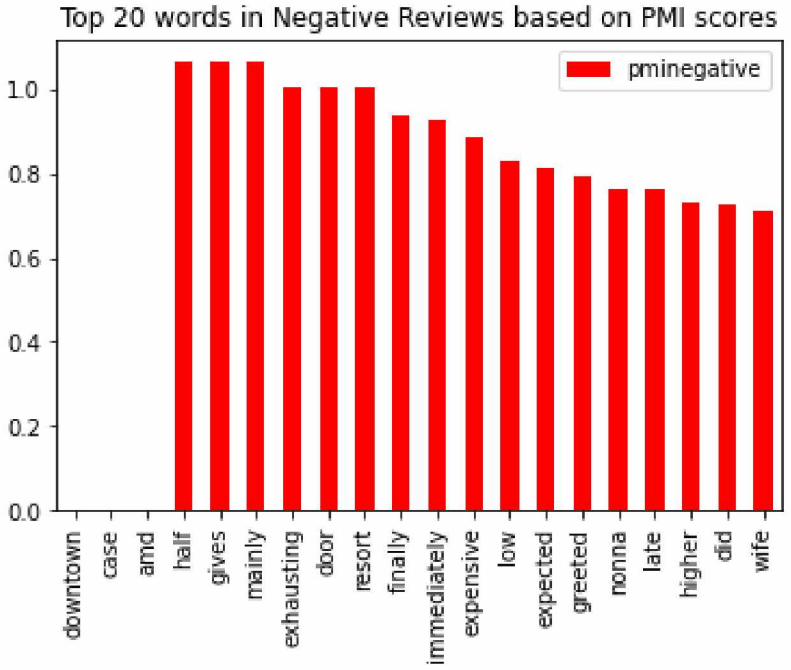

FIGURE 7.17: Top 20 PMI negative words

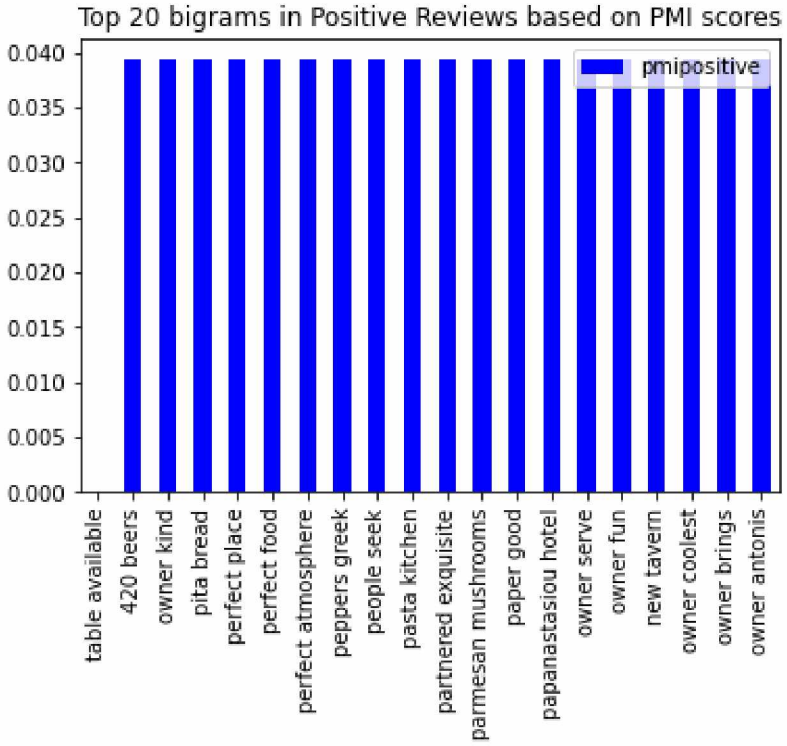

FIGURE 7.18: Top 20 PMI positive bigrams

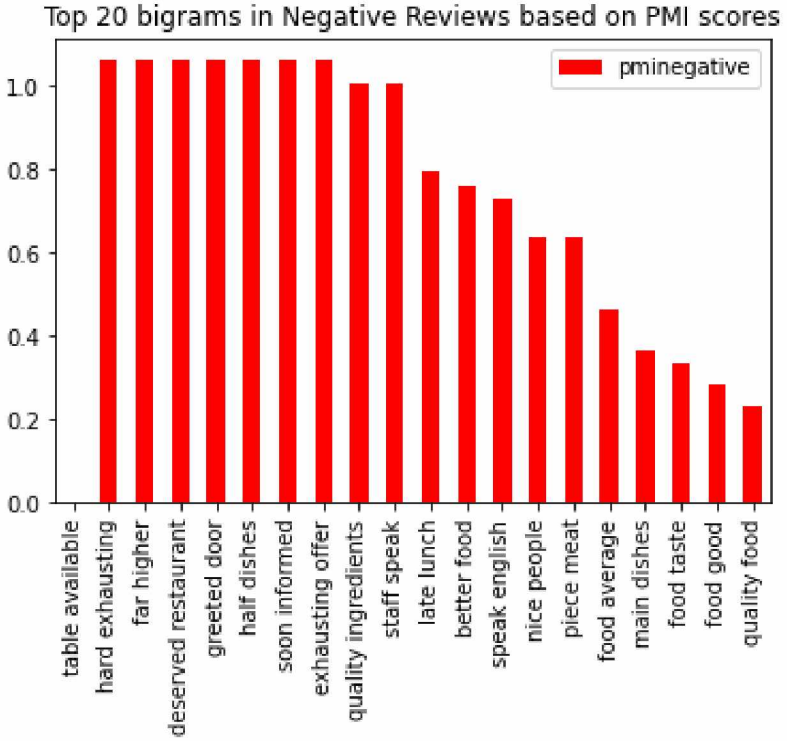

FIGURE 7.19: Top 20 PMI negative bigrams

Chapter 8

Discussion

8.1 Synopsis

Nowadays, firms are called to manage a huge amount of insights every day regarding the sales, customers, products, services that the enterprises deliver. The majority of these insights are unstructured data. This thesis displays how to utilize unstructured data for Business Intelligence. In particular, I describe the methods and tools necessary to collect, preprocess non-structured data from SEC filings, TripAdvisor's restaurants, YouTube and "Hellenic Chamber of Hotels". It is remarkable that the various web data sources give us the advantage to cover a wide range of business factors. Also, this thesis presents text mining components that perform data extraction, tools and processes such as NLP and Word Embeddings. Finally, my research includes methods of automatic text summarization of documents, data visualizations and sentiment analysis of TripAdvisor's comments.

Moreover, automatic summarization of financial documents is a useful tool in various fields. Our research gives seven different algorithms for automatic text summarization. Further, it is fact that management summaries have a bias to be more optimistic for the underlying documents they summarize and often leads investors to false judgments. As mentioned in [14] automatic summaries, especially those which are generated by LexRank can reduce this bias without sacrificing usefulness. Also, experiments show that automatic summaries have less bias and give to investors the confidence for safer decisions [14]. These benefits are due to the fact that automatic summaries are based on sentence extraction from text. Lastly, if we take into account that the volume of business data is constantly increasing to the point that it is often unmanageable by humans, automatic text summarization becomes a realistic solution that gives us significant points of business insights [14].

Furthermore, another important aspect of our research is web data analysis. More specific, I collected useful information from TripAdvisor's restaurants from the region of Thessaly. These insights such as comments, ranking, ratings etc. are related to customer relationships, so they are very important for restaurants. Further, sentiment analysis that applied to comments is a great method for restaurants to have a customer feedback about their satisfaction or the complaints. Accordingly, I applied algorithms to classify the comments into two categories, positive and negative. At first I used Logistic Regression and Decision Trees as traditional machine learning methods and then I used VADER which is a lexicon and rule-based sentiment analysis tool. Eventually, VADER proved to be superior to the other methods, with 93% accuracy.

8.2 Research limitations and development prospects

My thesis is subject to limitations, which pave avenues for future research in unstructured data for BI. First, future research in automatic summarization can examine specialized summaries for other types of documents such emails, reviews, MDA disclosures as well as other files i.e. make summarization of audio or video. Additionally, future study could investigate text summarization using RNN or the utility of abstractive summarization, which requires advanced NLP methods with semantic interpretation. In fact, automatic summaries could give the ability to enterprises to reduce the processing time and search costs in order to identify and manage the most important insights [14].

In addition, further research could investigate other web data sources. Particularly, could examine data from audio, image or video files. Also, future study could explore and preprocess insights from other platforms and social media like Facebook, Twitter, Instagram with the goal to enlarge the input information. In this way as well as by testing new methods of machine learning and by utilizing all the features, further research might achieve more accurate and safe results for firms.

Finally, my thesis is limited to examining non-financial data such as text, comments and reviews from customers. Therefore, further research could harness financial data of enterprises such as growth of revenues, profits and assets etc. in order to combine both non-financial and financial data. As a result, the combination of the above could create a complete BI system that could support enterprises. Lastly, data constantly increases, the research develops and matures, opening new horizons and giving the opportunity for creating useful and intelligent tools.

Appendix A

Automatic summaries - implementations and results

This Appendix presents automatic text summary techniques. In particular, every section initially includes code implementation and then text summary result. In my case, the original text that I use for summary is an Apple's 10-K Form with accession number 0000320193-18-000145. So, summaries that are contained in subsections A.1.2, A.2.2, A.3.2, A.4.2, A.5.2, A.6.2, A.7.2 and A.8.2 are generated from 10-K Form that you can find online on: https://www.sec.gov/Archives/edgar/data/320193/000032019318000145/a10-k20189292018.htm. More specifically, I have cleaned and extracted this SEC document and I have selected Item 1A for summarization. Section A.1 contains implementation of TextRank using Glove word embedding and the corresponding summary result. Lastly, the rest sections A.1 until A.8 includes quick implementations of summarization methods Luhn, SumBasic, LexRank, LSA, KL-Sum, Edmundson and TextRank, utilizing "sumy" python library, as well as the corresponding summaries resulting from these methods.

In conclusion, according to [92] LexRank technique constitutes the most remarkable algorithm in relation to the other algorithms. However, my personal assessment is that the most efficient algorithms are TextRank with sumy and LexRank with sumy.

A.1 TextRank implementation with Glove

A.1.1 Code

```
item_1a = item_1a_content.get_text()
sentences = []
sentences = sent_tokenize(item_1a)
sentences

# remove punctuations, numbers and special characters
clean_sentences = pd.Series(sentences).str.replace("[^a-zA-Z]", " ")
print(clean_sentences)
# make alphabets lowercase
clean_sentences = [s.lower() for s in clean_sentences]

# Extract word vectors
word_embeddings = {}
f = open('C:\\Users\\glove.6B\\glove.6B.100d.txt', encoding='utf-8')
for line in f:
    values = line.split()
```

```
    word = values[0]
    coefs = np.asarray(values[1:], dtype='float32')
    word_embeddings[word] = coefs
f.close()

sentence_vectors = []
for i in clean_sentences:
    if len(i) != 0:
        v = sum([word_embeddings.get(w, np.zeros((100,)))
        for w in i.split()])/(len(i.split())+0.001)
    else:
        v = np.zeros((100,))
    sentence_vectors.append(v)

# similarity matrix
sim_mat = np.zeros([len(sentences), len(sentences)])

for i in range(len(sentences)):
    for j in range(len(sentences)):
        if i != j:
            sim_mat[i][j] =
            cosine_similarity(sentence_vectors[i].reshape(1,10
            0), sentence_vectors[j].reshape(1,100))[0,0]

import networkx as nx

nx_graph = nx.from_numpy_array(sim_mat)
scores = nx.pagerank(nx_graph)

ranked_sentences = sorted(((scores[i],s) for i,s in
enumerate(sentences)), reverse=True)

# Specify number of sentences to form the summary
sn = 12

# Generate summary
for i in range(sn):
    print(ranked_sentences[i][1])
```

A.1.2 Summary

Any such costs, which may rise in the future as a result of changes in these laws and regulations or in their interpretation, could individually or in the aggregate make the Company's products and services less attractive to the Company's customers, delay the introduction of new products in one or more regions, or cause the Company to change or limit its business practices.

Quality problems could also adversely affect the experience for users of the Company's products and services, and result in harm to the Company's reputation, loss of competitive advantage, poor market acceptance, reduced demand for products and services, delay in new product and services introductions and lost revenue.The Company relies on access to third-party digital content, which may not be available

to the Company on commercially reasonable terms or at all.The Company contracts with numerous third parties to offer their digital content to customers.

Such an incident could, among other things, impair the Company's ability to attract and retain customers for its products and services, impact the Company's stock price, materially damage supplier relationships, and expose the Company to litigation or government investigations, which could result in penalties, fines or judgments against the Company.Although malicious attacks perpetrated to gain access to confidential information, including PII, affect many companies across various industries, the Company is at a relatively greater risk of being targeted because of its high profile and the value of the confidential information it creates, owns, manages, stores and processes.The Company has implemented systems and processes intended to secure its information technology systems and prevent unauthorized access to or loss of sensitive data, including through the use of encryption and authentication technologies.

While the Company has procedures to monitor and limit exposure to credit risk on its trade and vendor non-trade receivables, as well as long-term prepayments, there can be no assurance such procedures will effectively limit its credit risk and avoid losses.The Company could be subject to changes in its tax rates, the adoption of new U.S. or international tax legislation or exposure to additional tax liabilities.The Company is subject to taxes in the U.S. and numerous foreign jurisdictions, including Ireland, where a number of the Company's subsidiaries are organized.

Because the Company relies on single or limited sources for the supply and manufacture of many critical components, a business interruption affecting such sources would exacerbate any negative consequences to the Company.Apple Inc. | 2018 Form 10-K | 15The Company's operations are also subject to the risks of industrial accidents at its suppliers and contract manufacturers.

The success of new product and service introductions depends on a number of factors including, but not limited to, timely and successful development, market acceptance, the Company's ability to manage the risks associated with new product production ramp-up issues, the availability of application software for new products, the effective management of purchase commitments and inventory levels in line with anticipated product demand, the availability of products in appropriate quantities and at expected costs to meet anticipated demand and the risk that new products and services may have quality or other defects or deficiencies.

In addition, certain countries have passed or may propose and adopt legislation that would force the Company to license its digital rights management, which could lessen the protection of content and subject it to piracy and also could negatively affect arrangements with the Company's content providers.The Company's future performance depends in part on support from third-party software developers.The Company believes decisions by customers to purchase its hardware products depend in part on the availability of third-party software applications and services.

To help protect customers and the Company, the Company monitors its services and systems for unusual activity and may freeze accounts under suspicious circumstances, which, among other things, may result in the delay or loss of customer orders or impede customer access to the Company's products and services.In addition to the risks relating to general confidential information described above, the Company may also be subject to specific obligations relating to health data and payment card data.

If developers reduce their use of these platforms to distribute their applications and offer in-app purchases to customers, then the volume of sales, and the commission that the Company earns on those sales, would decrease.The Company relies on

access to third-party intellectual property, which may not be available to the Company on commercially reasonable terms or at all.Many of the Company's products include third-party intellectual property, which requires licenses from those third parties.

In addition, manufacturing or logistics in these locations or transit to final destinations may be disrupted for a variety of reasons including, but not limited to, natural and man-made disasters, information technology system failures, commercial disputes, military actions, economic, business, labor, environmental, public health or political issues, or international trade disputes.Apple Inc. | 2018 Form 10-K | 10The Company has invested in manufacturing process equipment, much of which is held at certain of its outsourcing partners, and has made prepayments to certain of its suppliers associated with long-term supply agreements.

The Company could also experience a significant increase in payment card transaction costs or lose the ability to process payment cards if it fails to follow payment card industry data security standards, which would materially adversely affect the Company's reputation, financial condition and operating results.While the Company maintains insurance coverage that is intended to address certain aspects of data security risks, such insurance coverage may be insufficient to cover all losses or all types of claims that may arise.The Company's business is subject to a variety of U.S. and international laws, rules, policies and other obligations regarding data protection.The Company is subject to federal, state and international laws relating to the collection, use, retention, security and transfer of PII.

Further, such an outcome could result in significant compensatory, punitive or trebled monetary damages, disgorgement of revenue or profits, remedial corporate measures or injunctive relief against the Company that could materially adversely affect its financial condition and operating results.While the Company maintains insurance coverage for certain types of claims, such insurance coverage may be insufficient to cover all losses or all types of claims that may arise.Apple Inc. | 2018 Form 10-K | 12The Company is subject to laws and regulations worldwide, changes to which could increase the Company's costs and individually or in the aggregate adversely affect the Company's business.The Company is subject to laws and regulations affecting its domestic and international operations in a number of areas.

A.2 Luhn's summary

A.2.1 Code

```
#Luhn
from sumy.summarizers.luhn import LuhnSummarizer
# For Strings
parser=PlaintextParser.from_string(filtered_item_1a,Tokenizer("english"))
# Using Luhn
summarizer = LuhnSummarizer()
#Summarize the document with 12 sentences
summary = summarizer(parser.document,12)

for sentence in summary:
    print(sentence)
```

A.2.2 Summary

The following information read conjunction Part II, Item 7, "Management's Discussion Analysis Financial Condition Results Operations" consolidated financial statements related notes Part II, Item 8, "Financial Statements Supplementary Data" Form 10-K.The business, financial condition operating results Company affected number factors, currently known unknown, including limited described below, could, directly indirectly, cause Company's actual financial condition operating results vary materially past, anticipated future, financial condition operating results.

Any factors, part, materially adversely affect Company's business, financial condition, operating results stock price.Because following factors, factors affecting Company's financial condition operating results, past financial performance considered reliable indicator future performance, investors use historical trends anticipate results trends future periods.Global regional economic conditions materially adversely affect Company's business, results operations, financial condition growth.The Company international operations sales outside U.S. representing majority Company's total net sales.

These economic factors materially adversely affect Company's business, results operations, financial condition growth.Global markets Company's products services highly competitive subject rapid technological change, Company unable compete effectively markets.The Company's products services offered highly competitive global markets characterized aggressive price competition resulting downward pressure gross margins, frequent introduction new products services, short product life cycles, evolving industry standards, continual improvement product price/performance characteristics, rapid adoption technological advancements competitors price sensitivity consumers businesses.The Company's ability compete successfully depends heavily ability ensure continuing timely introduction innovative new products, services technologies marketplace.

There assurance Company able continue provide products services compete effectively.To remain competitive stimulate customer demand, Company successfully manage frequent introductions transitions products services.Due highly volatile competitive nature industries Company competes, Company continually introduce new products, services technologies, enhance existing products services, effectively stimulate customer demand new upgraded products services successfully manage transition new upgraded products services.

The success new product service introductions depends number factors including, limited to, timely successful development, market acceptance, Company's ability manage risks associated new product production ramp-up issues, availability application software new products, effective management purchase commitments inventory levels line anticipated product demand, availability products appropriate quantities expected costs meet anticipated demand risk new products services quality defects deficiencies.

While arrangements help ensure supply components finished goods, outsourcing partners suppliers experience severe financial problems disruptions business, continued supply reduced terminated recoverability manufacturing process equipment prepayments negatively impacted.The Company's products services affected time time design manufacturing defects materially adversely affect Company's business result harm Company's reputation.The Company offers complex hardware software products services affected design manufacturing defects.

These U.S. foreign laws regulations affect Company's activities areas including, limited to, labor, advertising, digital content, consumer protection, real estate,

billing, e-commerce, promotions, quality services, telecommunications, mobile communications media, television, intellectual property ownership infringement, tax, import export requirements, anti-corruption, foreign exchange controls cash repatriation restrictions, data privacy data localization requirements, anti-competition, environmental, health safety.By way example, laws regulations related mobile communications media devices jurisdictions Company operates extensive subject change.

Other factors include, limited to, Company's ability manage costs associated retail store construction operation; manage relationships existing retail partners; manage costs associated fluctuations value retail inventory; obtain renew leases quality retail locations reasonable cost.Apple Inc. | 2018 Form 10-K | 13Investment new business strategies acquisitions disrupt Company's ongoing business present risks originally contemplated.The Company invested, future invest, new business strategies acquisitions.

Such incident could, things, impair Company's ability attract retain customers products services, impact Company's stock price, materially damage supplier relationships, expose Company litigation government investigations, result penalties, fines judgments Company.Although malicious attacks perpetrated gain access confidential information, including PII, affect companies industries, Company relatively greater risk targeted high profile value confidential information creates, owns, manages, stores processes.The Company implemented systems processes intended secure information technology systems prevent unauthorized access loss sensitive data, including use encryption authentication technologies.

The Company experience significant increase payment card transaction costs lose ability process payment cards fails follow payment card industry data security standards, materially adversely affect Company's reputation, financial condition operating results.While Company maintains insurance coverage intended address certain aspects data security risks, insurance coverage insufficient cover losses types claims arise.The Company's business subject variety U.S. international laws, rules, policies obligations data protection.The Company subject federal, state international laws relating collection, use, retention, security transfer PII.

Experienced personnel technology industry high demand competition talents intense, especially Silicon Valley, Company's key personnel located.The Company's business impacted political events, international trade disputes, war, terrorism, natural disasters, public health issues, industrial accidents business interruptions.Political events, international trade disputes, war, terrorism, natural disasters, public health issues, industrial accidents business interruptions harm disrupt international commerce global economy, material adverse effect Company customers, suppliers, contract manufacturers, logistics providers, distributors, cellular network carriers channel partners.International trade disputes result tariffs protectionist measures adversely affect Company's business.

Therefore, Company realized significant losses cash, cash equivalents marketable securities, future fluctuations value result significant realized losses material adverse impact Company's financial condition operating results.The Company exposed credit risk trade accounts receivable, vendor non-trade receivables prepayments related long-term supply agreements, risk heightened periods economic conditions worsen.The Company distributes products third-party cellular network carriers, wholesalers, retailers resellers.

A.3 SumBasic summary

A.3.1 Code

```
from sumy.summarizers.sum_basic import SumBasicSummarizer

parser=PlaintextParser.from_string(filtered_item_1a,Tokenizer("english"))
summarizer_sum_basic = SumBasicSummarizer()
summary_sum_basic =summarizer_sum_basic(parser.document,12)
for sentence in summary_sum_basic:
    print(sentence)
```

A.3.2 Summary

These risk factors important understanding statements Form 10-K.

Many components, including available multiple sources, times subject industry-wide shortages significant commodity pricing fluctuations materially adversely affect Company's financial condition operating results.

A significant concentration manufacturing currently performed small number outsourcing partners, single locations.

Other content owners, providers distributors seek limit Company's access to, increase cost of, content.

There assurance third-party developers continue develop maintain software applications services Company's products.

Such changes include, others, restrictions production, manufacture, distribution use devices, locking devices carrier's network, mandating use devices carrier's network.

Compliance applicable U.S. foreign laws regulations, import export requirements, anti-corruption laws, tax laws, foreign exchange controls cash repatriation restrictions, data privacy data localization requirements, environmental laws, labor laws anti-competition regulations, increases costs business foreign jurisdictions.

The Company exposed credit collectibility risk trade receivables customers certain international markets.

Tariffs Company's products expensive customers, Company's products competitive reduce consumer demand.

Additionally, new product introductions significantly impact net sales, product costs operating expenses.

The Company believes stock price reflect expectations future growth profitability.

In addition, Company prepayments associated long-term supply agreements secure supply inventory components.

A.4 LexRank summary

A.4.1 Code

```
#LexRank

from sumy.summarizers.lex_rank import LexRankSummarizer

parser=PlaintextParser.from_string(filtered_item_1a,Tokenizer("english"))
```

```
summarizer = LexRankSummarizer()
#Summarize the document with 12 sentences
summary = summarizer(parser.document,12)
for sentence in summary:
    print(sentence)
```

A.4.2 Summary

Any factors, part, materially adversely affect Company's business, financial condition, operating results stock price.Because following factors, factors affecting Company's financial condition operating results, past financial performance considered reliable indicator future performance, investors use historical trends anticipate results trends future periods.Global regional economic conditions materially adversely affect Company's business, results operations, financial condition growth.The Company international operations sales outside U.S. representing majority Company's total net sales.

In addition adverse impact demand Company's products, uncertainty about, decline in, global regional economic conditions significant impact Company's suppliers, contract manufacturers, logistics providers, distributors, cellular network carriers channel partners.

The success new product service introductions depends number factors including, limited to, timely successful development, market acceptance, Company's ability manage risks associated new product production ramp-up issues, availability application software new products, effective management purchase commitments inventory levels line anticipated product demand, availability products appropriate quantities expected costs meet anticipated demand risk new products services quality defects deficiencies.

In addition, manufacturing logistics locations transit final destinations disrupted variety reasons including, limited to, natural man-made disasters, information technology failures, commercial disputes, military actions, economic, business, labor, environmental, public health political issues, international trade disputes.Apple Inc. | 2018 Form 10-K | 10The Company invested manufacturing process equipment, held certain outsourcing partners, prepayments certain suppliers associated long-term supply agreements.

While arrangements help ensure supply components finished goods, outsourcing partners suppliers experience severe financial problems disruptions business, continued supply reduced terminated recoverability manufacturing process equipment prepayments negatively impacted.The Company's products services affected time time design manufacturing defects materially adversely affect Company's business result harm Company's reputation.The Company offers complex hardware software products services affected design manufacturing defects.

Quality problems adversely affect experience users Company's products services, result harm Company's reputation, loss competitive advantage, poor market acceptance, reduced demand products services, delay new product services introductions lost revenue.The Company relies access third-party digital content, available Company commercially reasonable terms all.The Company contracts numerous parties offer digital content customers.

Failure obtain right third-party digital content available, content available commercially reasonable terms, material adverse impact Company's financial condition

operating results.Some third-party digital content providers require Company provide digital rights management security solutions.

There assurance third-party developers continue develop maintain software applications services Company's products.

If developers reduce use platforms distribute applications offer in-app purchases customers, volume sales, commission Company earns sales, decrease.The Company relies access third-party intellectual property, available Company commercially reasonable terms all.Many Company's products include third-party intellectual property, requires licenses parties.

Compliance applicable U.S. foreign laws regulations, import export requirements, anti-corruption laws, tax laws, foreign exchange controls cash repatriation restrictions, data privacy data localization requirements, environmental laws, labor laws anti-competition regulations, increases costs business foreign jurisdictions.

Violations laws regulations materially adversely affect Company's brand, international growth efforts business.The Company significantly affected risks associated international activities including, limited to, economic labor conditions, increased duties, taxes costs, political instability international trade disputes.

Gross margins Company's products foreign countries, products include components obtained foreign suppliers, materially adversely affected international trade regulations, including duties, tariffs antidumping penalties.

A.5 LSA summary

A.5.1 Code

```
## LSA
from sumy.summarizers.lsa import LsaSummarizer

parser=PlaintextParser.from_string(filtered_item_1a,Tokenizer("english"))
summarizer_lsa = LsaSummarizer()
summary_lsa =summarizer_lsa(parser.document,12)
for sentence in summary_lsa:
    print(sentence)
```

A.5.2 Summary

The Company believes unique designs develops nearly entire solution products, including hardware, operating system, numerous software applications related services.

The Company faces substantial competition markets companies significant technical, marketing, distribution resources, established hardware, software digital content supplier relationships.

Many components, including available multiple sources, times subject industry-wide shortages significant commodity pricing fluctuations materially adversely affect Company's financial condition operating results.

Component suppliers suffer poor financial conditions, lead business failure supplier consolidation particular industry, limiting Company's ability obtain sufficient quantities components commercially reasonable terms.

In addition, manufacturing logistics locations transit final destinations disrupted variety reasons including, limited to, natural man-made disasters, information technology failures, commercial disputes, military actions, economic, business, labor,

environmental, public health political issues, international trade disputes.Apple Inc. | 2018 Form 10-K | 10The Company invested manufacturing process equipment, held certain outsourcing partners, prepayments certain suppliers associated long-term supply agreements.

For example, technology patent-holding companies frequently assert patents seek royalties enter litigation based allegations patent infringement violations intellectual property rights.

The plaintiffs actions frequently seek injunctions substantial damages.Regardless merit particular claims, litigation expensive, time consuming, disruptive Company's operations distracting management.

These U.S. foreign laws regulations affect Company's activities areas including, limited to, labor, advertising, digital content, consumer protection, real estate, billing, e-commerce, promotions, quality services, telecommunications, mobile communications media, television, intellectual property ownership infringement, tax, import export requirements, anti-corruption, foreign exchange controls cash repatriation restrictions, data privacy data localization requirements, anti-competition, environmental, health safety.By way example, laws regulations related mobile communications media devices jurisdictions Company operates extensive subject change.

Such endeavors involve significant risks uncertainties, including distraction management current operations, greater expected liabilities expenses, inadequate return capital unidentified issues discovered Company's diligence.

In addition reputational impacts, penalties include ongoing audit requirements significant legal liability.The Company's success depends largely continued service availability key personnel.Much Company's future success depends continued availability service key personnel, including Chief Executive Officer, executive team highly skilled employees.

A substantial majority Company's outstanding trade receivables covered collateral, third-party bank support financing arrangements, credit insurance.

The Company's exposure credit collectibility risk trade receivables higher certain international markets ability mitigate risks limited.

A.6 KLSum summary

A.6.1 Code

```
#KL algorithm
from sumy.summarizers.kl import KLSummarizer

# For Strings
parser=PlaintextParser.from_string(filtered_item_1a,Tokenizer("english"))
# Using KL
summarizer_kl = KLSummarizer()
#Summarize the document with 12 sentences
summary_kl = summarizer_kl(parser.document,12)

#printing the summarized sentences
for sentence in summary_kl:
    print(sentence)
```

A.6.2 Summary

Item 1A.Risk FactorsThe following discussion risk factors contains forward-looking statements.

These risk factors important understanding statements Form 10-K. A significant concentration manufacturing currently performed small number outsourcing partners, single locations.

Quality problems adversely affect experience users Company's products services, result harm Company's reputation, loss competitive advantage, poor market acceptance, reduced demand products services, delay new product services introductions lost revenue.The Company relies access third-party digital content, available Company commercially reasonable terms all.The Company contracts numerous parties offer digital content customers. No assurance given agreements obtained acceptable terms litigation occur.

Such changes include, others, restrictions production, manufacture, distribution use devices, locking devices carrier's network, mandating use devices carrier's network.

Global climate change result certain types natural disasters occurring frequently intense effects.

Additionally, new product introductions significantly impact net sales, product costs operating expenses.

The Company subject unexpected developments, lower-than-anticipated demand Company's products, issues new product introductions, information technology failures network disruptions, failure Company's logistics, components supply, manufacturing partners.The Company's stock price subject volatility.The Company's stock price experienced substantial price volatility past continue future.

dollar–denominated sales operating expenses worldwide.

Due economic political conditions, tax rates jurisdictions subject significant change.

There assurance outcome examinations.

A.7 Edmundson's summary

A.7.1 Code

```
##Edmundson Heuristic Method for text summarization
from sumy.summarizers.edmundson import EdmundsonSummarizer

# For String type documents
parser=PlaintextParser.from_string(filtered_item_1a,Tokenizer("english"))

summarizer1 = EdmundsonSummarizer(cue_weight=1, key_weight=1, title_weight=0, location_weig
summarizer1.bonus_words = ( "Discourse", "GitHub","internship")
summarizer1.stigma_words = ("positivity", "innovation","organisation","generation")

summary = summarizer1(parser.document,12)
for sentence in summary:
    print(sentence)
```

A.7.2 Summary

Item 1A.Risk FactorsThe following discussion risk factors contains forward-looking statements.

These risk factors important understanding statements Form 10-K.

The following information read conjunction Part II, Item 7, "Management's Discussion Analysis Financial Condition Results Operations" consolidated financial statements related notes Part II, Item 8, "Financial Statements Supplementary Data" Form 10-K.The business, financial condition operating results Company affected number factors, currently known unknown, including limited described below, could, directly indirectly, cause Company's actual financial condition operating results vary materially past, anticipated future, financial condition operating results.

Any factors, part, materially adversely affect Company's business, financial condition, operating results stock price.Because following factors, factors affecting Company's financial condition operating results, past financial performance considered reliable indicator future performance, investors use historical trends anticipate results trends future periods.Global regional economic conditions materially adversely affect Company's business, results operations, financial condition growth.The Company international operations sales outside U.S. representing majority Company's total net sales.

In addition, majority Company's supply chain, manufacturing assembly activities, located outside U.S. As result, Company's operations performance depend significantly global regional economic conditions.Adverse macroeconomic conditions, including inflation, slower growth recession, new increased tariffs, changes fiscal monetary policy, tighter credit, higher rates, high unemployment currency fluctuations materially adversely affect demand Company's products services.

In addition, consumer confidence spending adversely affected response financial market volatility, negative financial news, conditions real estate mortgage markets, declines income asset values, changes fuel energy costs, labor healthcare costs economic factors.

In addition adverse impact demand Company's products, uncertainty about, decline in, global regional economic conditions significant impact Company's suppliers, contract manufacturers, logistics providers, distributors, cellular network carriers channel partners.

Potential effects include financial instability; inability obtain credit finance operations purchases Company's products; insolvency.A downturn economic environment lead increased credit collectibility risk Company's trade receivables; failure derivative counterparties financial institutions; limitations Company's ability issue new debt; reduced liquidity; declines fair value Company's financial instruments.

These economic factors materially adversely affect Company's business, results operations, financial condition growth.Global markets Company's products services highly competitive subject rapid technological change, Company unable compete effectively markets.The Company's products services offered highly competitive global markets characterized aggressive price competition resulting downward pressure gross margins, frequent introduction new products services, short product life cycles, evolving industry standards, continual improvement product price/performance characteristics, rapid adoption technological advancements competitors price sensitivity consumers businesses.The Company's ability compete successfully depends heavily ability ensure continuing timely introduction innovative new products, services technologies marketplace.

The Company believes unique designs develops nearly entire solution products, including hardware, operating system, numerous software applications related services.

As result, Company significant investments RD.

There assurance investments achieve expected returns, Company able develop market new products services successfully.The Company currently holds significant number patents copyrights registered, applied register, numerous patents, trademarks service marks.

A.8 TextRank summary

A.8.1 Code

```
#TextRank
from sumy.summarizers.text_rank import TextRankSummarizer

# For Strings
parser=PlaintextParser.from_string(filtered_item_1a,Tokenizer("english"))
# Using TextRank
summarizer_textRank = TextRankSummarizer()

#Summarize the document with 12 sentences
summary_textRank = summarizer_textRank(parser.document,12)
for sentence in summary_textRank:
    print(sentence)
```

A.8.2 Summary

Any factors, part, materially adversely affect Company's business, financial condition, operating results stock price.Because following factors, factors affecting Company's financial condition operating results, past financial performance considered reliable indicator future performance, investors use historical trends anticipate results trends future periods.Global regional economic conditions materially adversely affect Company's business, results operations, financial condition growth.The Company international operations sales outside U.S. representing majority Company's total net sales.

These economic factors materially adversely affect Company's business, results operations, financial condition growth.Global markets Company's products services highly competitive subject rapid technological change, Company unable compete effectively markets.The Company's products services offered highly competitive global markets characterized aggressive price competition resulting downward pressure gross margins, frequent introduction new products services, short product life cycles, evolving industry standards, continual improvement product price/performance characteristics, rapid adoption technological advancements competitors price sensitivity consumers businesses.The Company's ability compete successfully depends heavily ability ensure continuing timely introduction innovative new products, services technologies marketplace.

The financial condition resellers weaken, resellers stop distributing Company's products, uncertainty demand Company's products cause resellers reduce ordering marketing Company's products.Apple Inc. | 2018 Form 10-K | 9The Company faces substantial inventory asset risk addition purchase commitment cancellation risk.The

Company records write-down product component inventories obsolete exceed anticipated demand, cost exceeds net realizable value.

Because Company's markets volatile, competitive subject rapid technology price changes, risk Company forecast incorrectly order produce excess insufficient amounts components products, fully utilize firm purchase commitments.Future operating results depend Company's ability obtain components sufficient quantities commercially reasonable terms.Because Company currently obtains certain components single limited sources, Company subject significant supply pricing risks.

While arrangements help ensure supply components finished goods, outsourcing partners suppliers experience severe financial problems disruptions business, continued supply reduced terminated recoverability manufacturing process equipment prepayments negatively impacted.The Company's products services affected time time design manufacturing defects materially adversely affect Company's business result harm Company's reputation.The Company offers complex hardware software products services affected design manufacturing defects.

Quality problems adversely affect experience users Company's products services, result harm Company's reputation, loss competitive advantage, poor market acceptance, reduced demand products services, delay new product services introductions lost revenue.The Company relies access third-party digital content, available Company commercially reasonable terms all.The Company contracts numerous parties offer digital content customers.

Further, outcome result significant compensatory, punitive trebled monetary damages, disgorgement revenue profits, remedial corporate measures injunctive relief Company materially adversely affect financial condition operating results.While Company maintains insurance coverage certain types claims, insurance coverage insufficient cover losses types claims arise.Apple Inc. | 2018 Form 10-K | 12The Company subject laws regulations worldwide, changes increase Company's costs individually aggregate adversely affect Company's business.The Company subject laws regulations affecting domestic international operations number areas.

The Company implemented policies procedures designed ensure compliance applicable laws regulations, assurance Company's employees, contractors, agents violate laws regulations Company's policies procedures.The Company's business subject risks international operations.The Company derives majority revenue earnings international operations.

Such incident could, things, impair Company's ability attract retain customers products services, impact Company's stock price, materially damage supplier relationships, expose Company litigation government investigations, result penalties, fines judgments Company.Although malicious attacks perpetrated gain access confidential information, including PII, affect companies industries, Company relatively greater risk targeted high profile value confidential information creates, owns, manages, stores processes.The Company implemented systems processes intended secure information technology systems prevent unauthorized access loss sensitive data, including use encryption authentication technologies.

The Company experience significant increase payment card transaction costs lose ability process payment cards fails follow payment card industry data security standards, materially adversely affect Company's reputation, financial condition operating results.While Company maintains insurance coverage intended address certain aspects data security risks, insurance coverage insufficient cover losses types claims arise.The Company's business subject variety U.S. international laws, rules, policies obligations data protection.The Company subject federal, state international laws relating collection, use, retention, security transfer PII.

The Company subject unexpected developments, lower-than-anticipated demand Company's products, issues new product introductions, information technology failures network disruptions, failure Company's logistics, components supply, manufacturing partners.The Company's stock price subject volatility.The Company's stock price experienced substantial price volatility past continue future.

Gross margins Company's products foreign countries products include components obtained foreign suppliers materially adversely affected foreign currency exchange rate fluctuations.Weakening foreign currencies relative U.S. dollar adversely affects U.S. dollar value Company's foreign currency–denominated sales earnings, generally leads Company raise international pricing, potentially reducing demand Company's products.

Bibliography

[1] P. NATARAJA and B. RAMESH, "MACHINE LEARNING ALGORITHMS FOR HETEROGENEOUS DATA: A COMPARATIVE STUDY", *INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING AND TECHNOLOGY*, vol. 10, no. 3, May 2019, ISSN: 0976-6367. DOI: 10.34218/ijcet.10.3.2019.002.

[2] P. Logdanidis, "Heterogeneous data for machine learning: The case of load forecasting", Tech. Rep., 2019.

[3] *Simplifying Sentiment Analysis using VADER in Python (on Social Media Text) | by Parul Pandey | Analytics Vidhya | Medium*. [Online]. Available: https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f.

[4] *Why Unstructured Data Collection is So Essential to Businesses*. [Online]. Available: https://fuelcycle.com/blog/why-unstructured-data-collection-is-so-essential-to-businesses/.

[5] M. Lynch, "In-depth Report United States Server and Enterprise Software Enterprise Information Portals Move Over Yahoo!; the Enterprise Information Portal Is on Its Way Reason for Report: Industry Overview", Tech. Rep., 1998.

[6] *Unstructured data - Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Unstructured_data.

[7] *Structured vs. Unstructured Data*. [Online]. Available: https://www.datamation.com/big-data/structured-vs-unstructured-data.html.

[8] *Unstructured Data and the 80 Percent Rule – Breakthrough Analysis*. [Online]. Available: http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/.

[9] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics", *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, Apr. 2015, ISSN: 02684012. DOI: 10.1016/j.ijinfomgt.2014.10.007.

[10] *The biggest data challenges that you might not even know you have - Watson Blog*. [Online]. Available: https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/.

[11] *New Digital Universe Study Reveals Big Data Gap Less Than 1 of World s Data is Analyzed Less Than 20 is Protected | Dell Technologies US*. [Online]. Available: https://corporate.delltechnologies.com/en-us/newsroom/announcements/2012/12/20121211-01.htm.

[12] *DataAge 2025 - The Digitization of the World | Seagate UK*. [Online]. Available: https://www.seagate.com/gb/en/our-story/data-age-2025/.

[13] *Unstructured Data: Elevating Your Business With One of Its Most Valuable Hidden Resources*. [Online]. Available: https://www.epam.com/insights/blogs/elevating-your-business-with-unstructured-data.

[14] E. Cardinaels, S. Hollander, and B. J. White, "Automatic summarization of earnings releases: attributes and effects on investors' judgments", *Review of Accounting Studies*, 2019, ISSN: 15737136. DOI: 10.1007/s11142-019-9488-0.

[15] T. Dyer, M. Lang, and L. Stice-Lawrence, "The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation", *Journal of Accounting and Economics*, 2017, ISSN: 01654101. DOI: 10.1016/j.jacceco.2017.07.002.

[16] J. Francis, K. Schipper, and L. Vincent, *Expanded disclosures and the increased usefulness of earnings announcements*, 2002. DOI: 10.2308/accr.2002.77.3.515.

[17] U. Securities and E. Commission, "Report on Review of Disclosure Requirements in Regulation S-K", Tech. Rep.

[18] ——, "SECURITIES AND EXCHANGE COMMISSION", Tech. Rep. [Online]. Available: http://www.sec.gov/rules/interim-final-.

[19] D. Hirshleifer and S. H. Teoh, "Limited attention, information disclosure, and financial reporting", *Journal of Accounting and Economics*, 2003, ISSN: 01654101. DOI: 10.1016/j.jacceco.2003.10.002.

[20] W. B. Elliott, J. L. Hobson, and B. J. White, "Earnings Metrics, Information Processing, and Price Efficiency in Laboratory Markets", *Journal of Accounting Research*, 2015, ISSN: 1475679X. DOI: 10.1111/1475-679X.12080.

[21] E. Henry, "Are investors influenced by how earnings press releases are written?", *Journal of Business Communication*, 2008, ISSN: 00219436. DOI: 10.1177/0021943608319388.

[22] E. Guillamon-Saorin, B. G. Osma, and M. J. Jones, "Opportunistic disclosure in press release headlines", *Accounting and Business Research*, 2012, ISSN: 00014788. DOI: 10.1080/00014788.2012.632575.

[23] X Huang, A Nekrasov, S. T. A. a. SSRN, and u. 2013, "Headline salience and over-and underreactions to earnings", *cba.uh.edu*, [Online]. Available: https://www.cba.uh.edu/departments/accy/research/documents/Teoh-paper.pdf.

[24] X. Huang, S. H. Teoh, and Y. Zhang, "Tone management", *Accounting Review*, 2014, ISSN: 00014826. DOI: 10.2308/accr-50684.

[25] Y. F. Te and I. P. Cvijikj, "Design of a small and medium enterprise growth prediction model based on web mining", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10360 LNCS, Springer Verlag, 2017, pp. 600–607, ISBN: 9783319601304. DOI: 10.1007/978-3-319-60131-1{_}48.

[26] OECD - Organisation for economic Co-Operation and Development, "Small and Medium-Sized Enterprises in Turkey: Issues and Policies", *Small*, 2004.

[27] K. Antlová, "Motivation and barriers of ict adoption in small and medium-sized enterprises", *E a M: Ekonomie a Management*, 2009, ISSN: 12123609.

[28] H. A. Post, "Building a strategy on competences", *Long Range Planning*, 1997, ISSN: 00246301. DOI: 10.1016/s0024-6301(97)00049-6.

[29] A. S. Koyuncugil and N. Ozgulbas, "Financial early warning system model and data mining application for risk detection", *Expert Systems with Applications*, 2012, ISSN: 09574174. DOI: 10.1016/j.eswa.2011.12.021.

[30] *Form 10-K - Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Form_10-K.

[31] *What is text mining (text analytics)? - Definition from WhatIs.com*. [Online]. Available: https://searchbusinessanalytics.techtarget.com/definition/text-mining.

[32] J. Hering, "The Annual Report Algorithm : Retrieval of Financial Statements and Extraction of Textual Information", 2017. DOI: 10.5121/csit.2017.70415.

[33] T. Stümpert, "Extracting Financial Data from SEC Filings for US GAAP Accountants", in *Handbook on Information Technology in Finance*, 2008. DOI: 10.1007/978-3-540-49487-4{_}16.

[34] M. Bovee, A. Kogan, K. Nelson, R. P. Srivastava, and M. A. Vasarhelyi, "Financial Reporting and Auditing Agent with Net Knowledge (FRAANK) and eXtensible Business Reporting Language (XBRL)", *Journal of Information Systems*, 2005, ISSN: 0888-7985. DOI: 10.2308/jis.2005.19.1.19.

[35] S O'Riain, "Semantic Paths in Business Filings Analysis", 2012. [Online]. Available: https://books.google.com/books?hl=el&lr=&id=BiMMfRXKl28C&oi=fnd&pg=PA1&dq=Semantic++Paths++in++Business++Filings++Analysis&ots=DJRCAFx-Bj&sig=pv3thmJ8KQc3PGfheE47VMPOJVU.

[36] T. Loughran and B. Mcdonald, "Measuring readability in financial disclosures", *Journal of Finance*, 2014, ISSN: 15406261. DOI: 10.1111/jofi.12162.

[37] J. Gerdes, "EDGAR-Analyzer: Automating the analysis of corporate data contained in the SEC's EDGAR database", *Decision Support Systems*, 2003, ISSN: 01679236. DOI: 10.1016/S0167-9236(02)00096-9.

[38] A. Kambil and M. Ginsberg, "Public Access Web Information Systems: Lessons from the Internet EDGAR Project", *Communications of the ACM*, 1998, ISSN: 15577317. DOI: 10.1145/278476.278493.

[39] A. K. Davis and I. Tama-Sweet, "Managers' Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A", *Contemporary Accounting Research*, 2012, ISSN: 08239150. DOI: 10.1111/j.1911-3846.2011.01125.x.

[40] T. Loughran and B. Mcdonald, "Textual Analysis in Accounting and Finance: A Survey", *Journal of Accounting Research*, 2016, ISSN: 1475679X. DOI: 10.1111/1475-679X.12123.

[41] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market", *Journal of Finance*, 2007, ISSN: 00221082. DOI: 10.1111/j.1540-6261.2007.01232.x.

[42] T. Loughran and B. Mcdonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks", *Journal of Finance*, 2011, ISSN: 00221082. DOI: 10.1111/j.1540-6261.2010.01625.x.

[43] N. Jegadeesh and D. Wu, "Word power: A new approach for content analysis", *Journal of Financial Economics*, 2013, ISSN: 0304405X. DOI: 10.1016/j.jfineco.2013.08.018.

[44] D. García and Norli, "Crawling EDGAR", *Spanish Review of Financial Economics*, 2012, ISSN: 21731268. DOI: 10.1016/j.srfe.2012.04.001.

[45] Cetinkaya, D. Seese, R. Spöth, and T. Stümpert, "EASE–A SOFTWARE AGENT THAT EXTRACTS FINANCIAL DATA FROM THE SEC'S EDGAR DATABASE",

[46] S. Conlon, "EDGAR Extraction System: An Automated Approach to Analyze Employee Stock Option Disclosures", *Article in Journal of Information Systems*, 2006. DOI: 10.2308/jis.2006.20.2.119. [Online]. Available: https://www.researchgate.net/publication/251005781.

[47] J. Engelberg and S. Sankaraguruswamy, "How to Gather Data Using a Web Crawler: An Application Using SAS to Search Edgar", *SSRN Electronic Journal*, 2011. DOI: 10.2139/ssrn.1015021.

[48] Y. Cong, A. Kogan, and M. A. Vasarhelyi, "Extraction of Structure and Content from the Edgar Database: A Template-Based Approach", *Journal of Emerging Technologies in Accounting*, 2007, ISSN: 1554-1908. DOI: 10.2308/jeta.2007.4.1.69.

[49] V. T. Thai, B. Davis, S. O'Riain, D. O'Sullivan, and S. Handschuh, "Semantically enhanced passage retrieval for business analysis activity", in *16th European Conference on Information Systems, ECIS 2008*, 2008.

[50] V. Chakraborty and M. Vasarhelyi, "Collected Papers of the Nineteenth Annual Strategic and Emerging Technologies Research Workshop Automating the process of taxonomy creation and comparison of taxonomy structures", Tech. Rep., 2010. [Online]. Available: http://www.sec.gov/idea/searchidea/webusers.htm.

[51] M. A. Hernándezhernhernández, H. Ho, G. Koutrika, R. Krishnamurthy, L. Popa, I. R. Stanoi, S. Vaithyanathan, and S. Das, "Unleashing the Power of Public Data for Financial Risk Measurement, Regulation, and Governance", Tech. Rep. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1814232.

[52] V. Desai, J. W. Kim, R. P. Srivastava, and R. V. Desai, "Textual Analysis and Business Intelligence in Big Data Environment : Search Engine versus XBRL", *Journal of Emerging Technologies in Accounting*, 2016, ISSN: 1554-1908. DOI: 10.2308/jeta-51933.

[53] *EDGAR - Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/EDGAR.

[54] *SEC.gov | What We Do*. [Online]. Available: https://www.sec.gov/Article/whatwedo.html.

[55] *sec-edgar-downloader — sec-edgar-downloader 3.0.5 documentation*. [Online]. Available: https://sec-edgar-downloader.readthedocs.io/en/latest/.

[56] M. E. Barth, "Financial Accounting Research, Practice, and Financial Accountability", *Abacus*, 2015, ISSN: 14676281. DOI: 10.1111/abac.12057.

[57] G. Lippitt and W. Schmidt, *Crises in a Developing Organization*, 1967.

[58] L. L. Steinmetz, "Critical stages of small business growth. When they occur and how to survive them", *Business Horizons*, 1969, ISSN: 00076813. DOI: 10.1016/0007-6813(69)90107-4.

[59] M. Scott and R. Bruce, "Five stages of growth in small business", *Long Range Planning*, 1987, ISSN: 00246301. DOI: 10.1016/0024-6301(87)90071-9.

[60] P. N. O'Farrell and D. M. Hitchens, "Alternative theories of small-firm growth: a critical review", *Environment & Planning A*, 1988. DOI: 10.1068/a201365.

[61] R Malarvizhi and K Saraswathi, "Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study", *Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study*, 2013, ISSN: 2231-2803.

[62] Y. Li, S. Arora, J. Youtie, and P. Shapira, "Using web mining to explore Triple Helix influences on growth in small and mid-size firms", *Technovation*, 2018, ISSN: 01664972. DOI: 10.1016/j.technovation.2016.01.002.

[63] S. H. Lin and J. M. Ho, "Discovering informative content blocks from Web documents", in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002. DOI: 10.1145/775107.775134.

[64] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, "Mining product reputations on the web", in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002. DOI: 10.1145/775047.775098.

[65] S. Saini and H. Mohan Pandey, "Review on Web Content Mining Techniques", *International Journal of Computer Applications*, 2015. DOI: 10.5120/20848-3536.

[66] D. Thorleuchter and D. Van Den Poel, "Predicting e-commerce company success by mining the text of its publicly-accessible website", *Expert Systems with Applications*, 2012, ISSN: 09574174. DOI: 10.1016/j.eswa.2012.05.096.

[67] K. Antlová, L. Popelínský, and J. Tandler, "Long term growth of SME from the view of ICT competencies and web presentations", *E a M: Ekonomie a Management*, 2011, ISSN: 1212-3609.

[68] Y. Te and Funk, "Predicting the Financial Growth of Small and Medium-Sized Enterprises using Web Mining", Dec. 2018. DOI: 10.3929/ethz-b-000309271. [Online]. Available: https://doi.org/10.3929/ethz-b-000309271.

[69] *SEC filing - Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/SEC_filing.

[70] *SEC.gov | How to Read a 10-K*. [Online]. Available: https://www.sec.gov/fast-answers/answersreada10khtm.html.

[71] *Tripadvisor: Διαβάστε κριτικέ, συγκρίνετε τιμέ και κάντε κράτηση*. [Online]. Available: https://www.tripadvisor.com.gr/.

[72] V. C. Heung, "American theme restaurants: A study of consumer's perceptions of the important attributes in restaurant selection", *Asia Pacific Journal of Tourism Research*, vol. 7, no. 1, pp. 19–28, Jan. 2002, ISSN: 1094-1665. DOI: 10.1080/10941660208722106.

[73] L. Richardson, "Beautiful Soup Documentation", *media.readthedocs.org*, 2016.

[74] *Web Scraping with Python - Richard Lawson - Βιβλία Google*. [Online]. Available: https://books.google.gr/books?hl=el&lr=&id=V_l_CwAAQBAJ&oi=fnd&pg=PP1&dq=Lawson+Web+scraping+with+Python&ots=G-Ep2wLs_n&sig=PcSDBbuXGgfdihuOphT9dlo2HqM&redir_esc=y#v=onepage&q=Lawson%20Web%20scraping%20with%20Python&f=false.

[75] *Greek Hotels Chamber Sets Out Goals for 2018 | GTP Headlines*. [Online]. Available: https://news.gtp.gr/2018/02/09/greek-hotels-chamber-sets-goals-2018/.

[76] *Identity & Objectives – Hellenic Chamber of Hotels*. [Online]. Available: https://www.grhotels.gr/en/about-us/skopos-antikeimeno/.

[77] *AGIA KYRIAKI – Hellenic Chamber of Hotels*. [Online]. Available: https://www.grhotels.gr/en/listing/agia-kyriaki/.

[78] *YouTube - Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/YouTube.

[79] *Basiswissen Web-Programmierung: XHTML, CSS, JavaScript, XML, PHP, JSP, ASP ... - Heide Balzert - Βιβλία Google*. [Online]. Available: https://books.google.gr/books?hl=el&lr=&id=K3i6l5KjMCEC&oi=fnd&pg=PA2&dq=Basiswissen+Web-Programmierung:+XHTML,+CSS,+JavaScript,+XML,+PHP,+JSP,+ASP.+NET,+Ajax&ots=un03xPwT9s&sig=off5sy4UH8qA52WX0SIQ3dgeY80&redir_esc=y#v=onepage&q=Basiswissen%20Web-Programmierung%3A%20XHTML%2C%20CSS%2C%20JavaScript%2C%20XML%2C%20PHP%2C%20JSP%2C%20ASP.%20NET%2C%20Ajax&f=false.

[80] *Text Mining in Python: Steps and Examples | by Dhilip Subramanian | Towards AI—Multidisciplinary Science Journal | Medium*. [Online]. Available: https://medium.com/towards-artificial-intelligence/text-mining-in-python-steps-and-examples-78b3f8fd913b.

[81] *Natural Language Toolkit — NLTK 3.5 documentation*. [Online]. Available: https://www.nltk.org/.

[82] *What Are Word Embeddings for Text?* [Online]. Available: https://machinelearningmastery.com/what-are-word-embeddings/.

[83] *Tf-idf :: A Single-Page Tutorial - Information Retrieval and Text Mining*. [Online]. Available: http://www.tfidf.com/.

[84] *An Introduction to TF-IDF. What is TF-IDF? | by Roshan Kumar Gupta | Analytics Vidhya | Medium*. [Online]. Available: https://medium.com/analytics-vidhya/an-introduction-to-tf-idf-using-python-5f9d1a343f77.

[85] *An Introduction to Text Summarization using the TextRank Algorithm (with Python implementation) | by Prateek Joshi | Analytics Vidhya | Medium*. [Online]. Available: https://medium.com/analytics-vidhya/an-introduction-to-text-summarization-using-the-textrank-algorithm-with-python-implementation-2370c39d0c60.

[86] *Fundamentals of Bag Of Words and TF-IDF | by Prasoon Singh | Analytics Vidhya | Medium*. [Online]. Available: https://medium.com/analytics-vidhya/fundamentals-of-bag-of-words-and-tf-idf-9846d301ff22.

[87] *A Gentle Introduction to the Bag-of-Words Model*. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-bag-words-model/.

[88] *Social Media Sentiment Analysis using Machine Learning : Part — II | by Deepak Das | Towards Data Science*. [Online]. Available: https://towardsdatascience.com/social-media-sentiment-analysis-part-ii-bcacca5aaa39.

[89] *GloVe: Global Vectors for Word Representation*. [Online]. Available: https://nlp.stanford.edu/projects/glove/.

[90] J. M. Torres Moreno, *Automatic Text Summarization*. 2014, vol. 9781848216686. DOI: 10.1002/9781119004752.

[91] *Understand Text Summarization and create your own summarizer in python | by Praveen Dubey | Towards Data Science*. [Online]. Available: https://towardsdatascience.com/understand-text-summarization-and-create-your-own-summarizer-in-python-b26a9f09fc70.

[92] E. Cardinaels, S. Hollander, and B. J. White, "Automatic Summarization of Corporate Disclosures", Tech. Rep., 2017.

[93] *Luhn's Heuristic Method for text summarization*. [Online]. Available: https://iq.opengenus.org/luhns-heuristic-method-for-text-summarization/.

[94] *SumBasic algorithm for text summarization*. [Online]. Available: https://iq.opengenus.org/sumbasic-algorithm-for-text-summarization/.

[95] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond SumBasic: Task-Focused Summarization with Sentence Simplification and Lexical Expansion", Tech. Rep. [Online]. Available: http://duc.nist.gov.

[96] *Latent Semantic Analysis for text summarization*. [Online]. Available: https://iq.opengenus.org/latent-semantic-analysis-for-text-summarization/.

[97] *KL Sum algorithm for text summarization*. [Online]. Available: https://iq.opengenus.org/k-l-sum-algorithm-for-text-summarization/.

[98] *Edmundson Heuristic Method for text summarization*. [Online]. Available: https://iq.opengenus.org/edmundson-heuristic-method-for-text-summarization/.

[99] H. P. Edmundson, "New Methods in Automatic Extracting", Tech. Rep.

[100] *TextRank for Text Summarization*. [Online]. Available: https://iq.opengenus.org/textrank-for-text-summarization/.

[101] *LexRank method for Text Summarization*. [Online]. Available: https://iq.opengenus.org/lexrank-text-summarization/.

[102] *Python | Sentiment Analysis using VADER - GeeksforGeeks*. [Online]. Available: https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/.

[103] *Supervised learning - Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Supervised_learning.

[104] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text", in *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 2014.

[105] *Mutual information - Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Mutual_information.

[106] *Pointwise mutual information - Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Pointwise_mutual_information.

[107] *Understanding Pointwise Mutual Information in NLP | by Valentina Alto | DataSeries | Medium*. [Online]. Available: https://medium.com/dataseries/understanding-pointwise-mutual-information-in-nlp-e4ef75ecb57a.