MASTER IN BUSINESS ENGINEERING PROFESSIONAL FOCUS IN DATA SCIENCE

Evolution of the Business Intelligence industry between 2006 and 2020
A Study on Gartner reports using Text Mining Analytics

BLANCHY, Thomas

Award date:
2022

Awarding institution:
University of Namur

Evolution of the Business Intelligence industry between 2006 and 2020. A Study on Gartner reports using Text Mining Analytics

Thomas BLANCHY

Directrice: Prof. I. LINDEN

Mémoire présenté
en vue de l'obtention du titre de
Master 120 en ingénieur de gestion, à finalité spécialisée
en data science

ANNEE ACADEMIQUE 2021-2022

Université de Namur, ASBL
Faculté des Sciences économiques, sociales et de gestion – Département des Sciences de gestion

Rempart de la Vierge 8, B-5000 Namur, Belgique, Tel. +32 [0]81 72 49 58/48 41

Evolution of the Business Intelligence industry between 2006 and 2020. A Study on Gartner reports using Text Mining Analytics

Abstract
Nowadays, data and information are becoming more and more important for companies whether it concerns their collection, storage, processing, or reporting. This thesis aims to highlight how technologies have evolved with market needs and how providers have succeeded in adapting or not to these challenges. Text mining was used to extract relevant information on this subject from Gartner reports on Business Intelligence & Analytics Platforms. It turned out to be effective and brought out insights that correspond to other papers related to the subject, which confirms the truthfulness of the topics that the tool brings out. The conclusion is that the market has evolved from segmented offers including traditional OLAP tools towards cloud, augmented capacity and artificial intelligence technologies, provided mainly by large players such as IBM, Microsoft, or Tableau.

Keywords: Business Intelligence ; Data ; Text Mining ; Topic Modeling ; Evolution ; Gartner

---

Evolution de l'industrie de la business intelligence entre 2006 et 2020. Une étude sur les rapports Gartner avec des techniques de Text Mining

Résumé
De nos jours, les données et informations deviennent de plus en plus importantes pour les entreprises, qu'il s'agisse de leur collecte, de leur stockage, de leur traitement ou de leur présentation. Ce mémoire vise à mettre en lumière comment les technologies ont évolué avec les besoins du marché et comment les fournisseurs ont réussi à s'adapter ou non à ces enjeux. Le text mining a été utilisé pour extraire des informations pertinentes des rapports Gartner sur les plateformes de business intelligence et d'analyse. La méthodologie s'est avérée efficace et a fait ressortir des idées qui correspondent à d'autres articles liés au sujet, ce qui confirme la véracité des sujets que l'outil met en évidence. La conclusion est que le marché a évolué d'offres segmentées comprenant des outils OLAP traditionnels vers des technologies de cloud, de capacité augmentée et d'intelligence artificielle, fournies principalement par de grands acteurs tels qu'IBM, Microsoft ou Tableau.

Mots-clés: Business Intelligence ; Données ; Text Mining ; Topic Modeling ; Evolution ; Gartner

Foreword

This work is the result of five years of study at the University of Namur in Business Engineering option Data Science. These years have been very enriching and I would like to thank the people who accompanied me throughout this journey and the realization of this thesis.

First of all, Mrs. Linden for accompanying me during the writing of this thesis as my promoter and for having always been there when I had questions,

My family for the moral support they gave me during these years of study,

Myriam for having been kind enough to proofread my work in depth and to provide me with useful recommendations,

And my girlfriend Marie for taking the time to support me and help me throughout the realization of the work and without whom it would not have been possible.

Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
BI	Business Intelligence
DTM	Document per Term Matrix
ETL	Extract, Transform and Load
GDPR	General Data Protection Regulation
IoT	Internet of Things
KPI	Key Performance Indicator
LDA	Latent Dirichlet Allocation
OLAP	OnLine Analytical Processing

Contents

List of Figures

List of Tables

Introduction

Today's companies value information above anything else to drive their activities towards success. As stated by Romero et al. (2021), information is the most valuable resource of a company. Managing it in a senseful way is thus required to stay competitive in a particular industry. The field of Business Intelligence emerged to solve this problem by providing capabilities to handle the data available to a company. As the field evolved over time, it became more complex and providers in the industry started to look like each other, making it even harder for BI customers to choose between them.

Gartner, a consulting company in the technology industry, publishes yearly reports to assess the extent of the BI field in terms of its providers. The insights that they provide are useful to choose between the competitors but are composed of dozen-page reports, hard to understand by non-initiated readers. In this thesis, we, therefore, use Text Mining to extract the useful information made available by Gartner in their reports. Thanks to this methodology, we are able to assess the evolution of the field of Business Intelligence.

This thesis attempts to identify trends in terms of technologies and solution providers in the business intelligence market using Text Mining. The final result of this project will give us a visual overview of the evolution of the field as well as the confirmation that Text Mining is capable of producing concrete results adapted to a certain context on the basis of a corpus of textual documents.

In Chapter 1, we define the key concepts of our thesis. In Chapter 2, we review the literature on Business Intelligence in terms of providers, process & methodology, and technology & infrastructure. In Chapter 3, we deepen the background on BI by discussing the evolution of raw data, meaningful information, and decision-making. In Chapter 4, we detail the theoretical background of Text Mining. In Chapter 5, we go through the methodology of the analysis. In Chapter 6, we present the Gartner reports and their content. Finally, in Chapter 7, we unveil the results gotten from the Text Mining analysis in terms of evolution in the BI industry. We finish this thesis with a conclusion and some clues about the limitations of the work as well as ideas for future research.

Part I

Background

The background part is structured in four chapters. Chapter 1 is used to define key terms of our problem and then to determine the scope of the work. Chapter 2 highlights the articles corresponding to our scope while Chapter 3 focuses on literature that also deals with the evolution of BI but in a more general way. This gives us insights to interpret the results of our text mining analysis. Finally, Chapter 4 details the methodology used to apply text mining to our data.

Chapter 1

Definitions

The title of the thesis is "Evolution of the Business Intelligence industry between 2006 and 2020. A Study on Gartner reports using Text Mining Analytics". It is therefore important to understand the terms "Evolution" and "Business Intelligence" in general but also specifically. We, therefore, review the scope of these terms to describe precisely on which aspects this thesis focuses. We are only going to select a subset, not a subject that is too broad.

1.1 Evolution

The term “evolution” can be applied to many fields and is hard to define, mainly because the definition of the term itself has evolved over time, especially in biology (Wilkins, 2001) but also because it can be applied to almost any aspect of life.

Based on the Collins dictionary (Collins English Dictionary, 2012), evolution refers to “a process of gradual development in a particular situation or thing over a period of time”. Basically, it is going from one state to another by using new methodologies, new tools or new technologies. Initially, the concept of evolution applies to biology, representing the development of the human being. However, the same process applies to businesses too.

In the case of companies, evolution is a must-have in the sense that their success is based on their ability to evolve and face new trends or adversities to stay in the competition (Capron, 2015). Companies that keep their current activities without trying to evolve are bound to fail and disappear as was the case for companies like Kodak or Blockbuster who terminated their business due to their lack of adaptation to new business trends (Antioco, 2011).

Technologies have also evolved fast over the past decades, not a hundred years ago we did not have what we use in our everyday lives like the Internet, smartphones, the cloud or any new technology. All those inventions that made our lives easier stem from a process of evolution that is still running.

In this thesis, we will thus focus on the evolution of technology. Indeed, it takes into account both the tools (software and software providers) and the purpose of those tools (businesses evolve to stay competitive and technologies evolve with changing demand). We can therefore define the "evolution" in this work as *the development of specialized tools for and by companies in order to remain competitive.*

1.2 Business Intelligence

Business Intelligence has been defined many times as it was shown by Chee et al. (2009). Each definition approaches the subject differently way but overall they all talk about the same topic to reach the same end. Therefore, for the purpose of this work, we base ourselves on the following definition proposed by Vaisman & Zimányi (2014):

BI in companies *"comprises a collection of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information for decision making"*.

The methodologies and tools of BI relate to the different stages in a BI project and the applications or platforms that we use to lead such projects (Petrini & Pozzebon, 2009). A typical BI project includes the collection, storing, processing, and communication of data related to the business problem (Burstein, 2008).

In this work, we will therefore base ourselves on the definition given above. It is important to delineate each concept of this definition to have a complete understanding of business intelligence. Indeed, we are going to do a Topic Modeling analysis on the Gartner reports. We need to interpret the meaning of the terms that will emerge from our analysis by deepening our knowledge of business intelligence. This is why we will individually define Raw data, Meaningful information & Decision making.

1.3 Raw Data, Meaningful Information & Decision Making

Raw data refers to any untransformed data that is provided to the company in order to be stored, analyzed or transformed (Hiter, 2021). It ranges from textual data to images and videos and comprises both structured and unstructured data. Structured data are formatted into tables and follow a particular structure that makes them easier to analyze whereas unstructured data don't follow a particular model and can differ from one another in all aspects. Typically, any type of document that is hard to transform into a table can be considered as unstructured data, such as pdf documents.

However, once the raw data is processed and the results of the analysis delivered to the final user, we have what we call meaningful or useful information, for instance, the evolution of the popularity of a company or their behavior in case of a change in its name. Understanding a business requires a lot of experience or accurate information and their quantity explodes nowadays (Rekatsinas et al., 2015). They are generally transformed into measures reflecting a summary of multiple pieces of information, explaining to the managers the current state of the company. It ranges from the average value of sales to delivery time experienced by the suppliers, basically it covers every aspect of the business with some meaningful indicators, generally referred to as KPI (Pan & Wei, 2012).

Finally, decision-making, the fundamental step for the success of companies, is the final one of the BI process. It is based on the useful information derived from the raw data and allows the business managers to evolve the state of the company. The final goal depends on the objectives, but usually it consists of staying competitive in the market, increasing profit or any other aspect used to keep the company running. Moreover, good decision making establishes trust between the management and employees (Kinsey & Seidel, 2019), further improving the business activities.

In this thesis, we see **business intelligence** as a set of methodologies and technologies that transform raw data into meaningful information that help to make decisions. We can therefore set the scope of our work to the **Evolution of Business Intelligence's methodologies and technologies, focusing on the providers of BI solutions**. The majority of companies use BI tools to analyze their data but some of those instruments have evolved over time or disappeared. The search for the right tool is difficult and can sometimes result in a failure of the BI process implementation. Furthermore, documentation about those tools can be quite extensive and long to read for casual business users, making it even more difficult to make a choice and stay competitive.

Chapter 2

Related Work

Some papers have already dealt with the subject of the evolution of BI from the point of view of technology and methodology. However, for the evolution of solution providers, we have only found managerial reports or blog articles but nothing specific so we refer to the Chapter 3 for more information on the subject. To go over the literature, we start with a brief review of the evolution of providers, then review the specialized papers on methodologies and processes, and technologies and infrastructures.

2.1 Business Intelligence Providers

The beginning of BI solutions was composed of individual software, developed specifically for the purpose of the company ordering it. Each element of the BI solution was independent and the information was spread over the entire company without connection between elements of the system (Raj et al., 2016). Later on, big tech companies started to take interest in such solutions, therefore developed completely integrated applications and provided them in the form of Software as a Service (SaaS). We know them under the name of Microsoft, Oracle, IBM, SAP, etc. These companies have evolved too, but literature on the subject is rare and, based on our research, no analysis of the field of business intelligence providers has been done.

The following two sections talk about methodologies, processes, technologies, infrastructures, and their evolution. We decided to group the first two elements and the two following ones together due to their similarities. The methodology and process part discusses the "how", thus which steps an analysis goes through to perform business intelligence whereas the part on technology and infrastructure discusses the elements supporting those steps (Marjamäki, 2017).

2.2 Process and Methodology

The processes and methodologies of business intelligence were described by Park et al. (2010) and Panian (2012) as a succession of steps. It begins with the collection of data[1] inside the company or in its environment. After the gathering of the needed information, companies have to store that data in infrastructures such as databases and more recently developed data warehouses. Then, this information is processed through different technologies[2]. It consists of transforming that data into a structured format, making

[1]Refer to the section on raw data to see its evolution

[2]Technologies are detailed in the following section

the investigation easier for data analysts. The process called ETL is usually used to perform this task consisting of extracting, transforming, and loading the data (Vaisman & Zimanyi, 2014; Obeidat et al., 2015). The final step of the process consists of communicating the results to business users often untrained in technologies. It can take various formats such as graphs, tables, dashboards, or balanced scorecards (Kaplan & Norton, 1996) and those means of communication have evolved through time. Those same processes and methodologies have been discussed by Panian (2012), Obeidat et al. (2015), and Larson & Chang (2016).

In a more recent paper, Andoh-Baidoo (2022) discusses the evolution of the methodologies with the final results obtained after analysis. It started with simple descriptive analysis to evolve towards predictive ones allowing business users to predict outcomes or events. Later on, prescriptive analysis was developed, giving insights about what to do to improve their activities. Finally, we see more and more new types of analysis qualified as augmented ones relying on artificial intelligence and machine learning, two technologies described in the below section.

2.3 Technology and Infrastructure

Technologies and infrastructure in business intelligence have evolved through different stages from the nineties to today. The literature review on the subject is quite extensive and based on the work of Panian (2012), Marjamäki (2017), and Andoh-Baidoo (2022). We can extract four periods in this time slot, beginning in the 90s with data warehousing, OLAP tools and dashboards. Then, in the new millennium emerged new technologies such as NoSQL databases, Hadoop, Map Reduce, and Machine Learning. Ten years later (from 2000) we still have improvements in this area with Mobile analytics and visualizations, Location data, and improvements in Machine Learning. Finally, nowadays we see emerging technologies improving the field of BI with AI and machine learning as a service, deep learning, robotics process automation, chatbots and intelligent assistants, semantic natural language processing and natural language generation.

Overall, we can highlight four main categories of technologies that have evolved with business intelligence: technologies related to data collection, data storage, data processing, and reporting. These categories are detailed in the following sections.

2.3.1 Data Collection

In terms of technology, the Internet has made it possible to track various information that was unattainable in the past (Patel, 2021). In the same way, the extent of data capable of being analyzed has increased with the emergence of analytics techniques for unstructured data (Obeidat et al., 2015). Indeed, the different types of data that companies have encountered have also evolved over the years as will be explained in Chapter 3. These changes in terms of data have obviously influenced the evolution of technologies used to collect information. Examples of such technologies include web mining, text mining, etc. Later on, new technologies that people use on an everyday basis became capable of

generating huge amounts of data. Among these technologies, we can count IoT devices, mobile phones or any connected device generating some information (Ahmed et al., 2017).

2.3.2 Data Storage

In terms of storage, the possibilities of the database management systems had to be extended due to a huge increase in the generated amount of data (Bataweel, 2015). Initial technologies relied on the data warehouse technology (Andoh-Baidoo et al., 2022), merging different sources of data into an integrated way but it had reached its limits. New technologies had to be developed, so NoSQL databases were created in order to handle more data types (Han et al., 2011). However, the problem of the quantity of data was still not answered. Technologies such as Big Data, Hadoop, and Map Reduce thus emerged, providing decentralized capabilities for data storage, which allowed companies to store huge amounts of data on different servers while still keeping everything integrated (Dittrich & Quiané-Ruiz, 2015).

2.3.3 Data Processing

The process of doing business intelligence has also evolved since the nineties. In the past, companies relied mainly on data mining techniques, trying to identify patterns in the data without the help of advanced techniques that we know now (Bataweel, 2015). Then, came OLAP tools allowing to analyze some types of data up to 10 dimensions (Panian, 2012). Instead of analyzing the sales of a company through one lens like the products, companies could now compare those results based on customer types, time, suppliers, locations and so on. It really improved the field of analytics but was just the beginning. Artificial intelligence techniques were developed to automate some processes, and with the help of machine learning, companies were starting to predict the outcomes of typical business problems (Andoh-Baidoo et al., 2022). Now, this field is even more advanced and such services are offered to common business users without deep knowledge of the data industry (Romero et al., 2021).

2.3.4 Data Reporting

The final step of the business intelligence ecosystem is the reporting part, providing KPIs and data visualizations to communicate to the final users the information needed for the related business problem (Vaisman & Zimányi, 2014). Initial reporting techniques only consisted of long and incomprehensible reports that only a people were capable of analyzing (Kateeb et al., 2014). Dashboards and balanced scorecards were then developed to improve the provided insights and allowed more users to interpret the results of the analysis (Kaplan & Norton, 1996). Today, with the improvements in technology and the fact that almost everyone has a mobile phone, reporting techniques follow us everywhere and mobile visualization with real-time information is now the norm (Marjamäki, 2017).

Overall, technologies evolve quickly towards united, immediate, and easy-to-understand information. Business users want information centralized (united) and know in real-time

(immediate) what is happening in their company to be able to respond quickly (easy-to-understand) as corporations are always on the verge of being outdated by competing firms. The race for information is, therefore, the current trend and, as Romero et al. (2021) stated, "Information has been identified as the most valuable asset of a company". It justifies, even more, the use of BI to support the firms' activities.

In conclusion of this part, we collected a lot of information on methodologies and technologies but no paper or article concerning the BI solution providers convinced us. Therefore, the technologies and methodologies will serve as a basis to confront the evolution of BI developed here with the one extracted from our data. We, however, seek more insights to understand why players (providers) enter or leave the market in the next chapter. We detail the evolution of BI concerning raw data, meaningful information, and decision-making. This chapter can help us better interpret the outputs of the textual analysis and have leads concerning the behaviors of the providers.

Chapter 3

Business Intelligence Background

The scientific literature on the specific subject of the BI providers is lacking relevant content and new improvements in the field are not immediately implemented by those vendors as the most promising one are still under development (BCG, 2022). So, in order to be more precise and have more insights to compare our analysis, this part discusses the field of business intelligence regarding the definition proposed by Vaisman & Zimányi (2014), which is, to remember, *"a collection of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information for decision making"*. It is, therefore, divided into three points: Raw data, meaningful information, and decision-making. For each concept we trace its evolution through scientific articles and make connections with their influence on BI solution providers.

3.1 Raw Data

3.1.1 Sources of Data

Raw data or primary data is a form of information that hasn't been modified or transformed by the company. It is the data as received, ready to be used for analysis. Inside a company, that kind of data comes from various sources and in multiple formats. Operational and transactional databases are the main sources of data found in a company (Chee et al. 2009). It can take the form of emails, legal contracts, warranties, financial information, and so on (Inmon, 2013) but is not limited to typical documents usually analyzed in such companies. More complicated sources such as charts, images and videos are also important sources of information (Cebotaeran, 2011) and it became increasingly easier to analyze them thanks to the new technologies developed. However, the company is not limited to only the resources available internally. External sources of data are important to take into account too (Chee et al. 2009), companies, therefore, had to merge multiple sources of data to get more information. These sources include market information and evolution, demographics, biographies, economic indicators, or even newspapers (Park et al. 2010). This information is generated through social networks, websites, and libraries among others (Romero et al. 2021), and is coming from open sources or provided by specialized companies in the area. Web Mining and Social Network analysis, among others, thus emerged to fulfill the new needs regarding external sources. We, therefore, noticed that consumer companies and solution provider companies had to adapt to these multitudes of sources. A great amount of data is also generated through new connected IoT devices (Ahmed et al., 2017). Overall, the collection of data is a long process that takes resources and time (Park et al. 2010) due to the heterogeneity of the collected information (Romero et al. 2021). The challenge of the company is thus to balance the

efforts with the outcome in order to be able to leverage everything they collected.

3.1.2 Types of Data

As data can come from various sources, it can also take a lot of different forms. The two main types of data that we can analyze are structured and unstructured ones (Obeidat et al. 2015). In the past, the only type of data that mattered corresponded to the structured one (Inmon, 2013) including simple forms like numbers, statistics, string elements, dates, etc. (Chee et al. 2009) that are easier to analyze thanks to their formatted structure into rows and columns (Larson & Chang, 2016). However, with some evolution and improvements in the data industry, companies began to analyze the full extent of available resources (Inmon, 2013), forcing BI solutions to diversify their capabilities in terms of data analytics. The analysis started to include more complicated forms such as textual information in different formats (Chee et al. 2009; Larson & Chang, 2016), requiring providers to deliver more computational power in their solutions for companies indulging in data activities. These unstructured data correspond to the majority of available information and should be dealt with differently (Cebotarean, 2011) than common structured ones. It is also this type of data that analysts spend most of their time on due to its complexity (Cebotarean, 2011) but get more meaning out of it (Obeidat et al., 2015).

3.1.3 Increased Volume of Data

A parallel trend in the data industry is the increase in volume generated every day by companies (Cebotarean, 2011; Panian, 2012; Obeidat et al. 2015). It is driven by all the new devices generating data such as IoT (Ahmed et al., 2017), sensors, and others as well as the collection of real-time information (Bataweel, 2015) and the increasing capacity for companies to store them. The challenge is now shifting to the selection of meaningful information to analyze (Panian, 2012) because companies cannot blend every source. They have to select the relevant ones for their personal interests and the prosperity of their activities, hence the increase of interest for new processes such as data discovery, which can be considered as the analysis of data in a more visual way to discover hidden patterns. The providers have therefore adapted their solutions to meet these new needs. Previously, users mainly focused on the results of the analysis, considered as meaningful information (see Meaningful Information). Now, they must also pre-select the interesting data to analyze and ultimately obtain useful information.

Nevertheless, despite the development of data collection techniques increasing the quantity of available data, it becomes cheaper and easier (Obeidat et al. 2015) for companies to indulge in data-driven analysis, making it more affordable for smaller firms. This can be explained by several phenomena such as the advancement of techniques and methodologies that can be used, but also the democratization of devices and the free flow of information.

3.1.4 Integration of Data

The majority of businesses did not have the needed resources to analyze all the data they collected (Panian, 2012) except that collecting information from different sources is

important for more complex analysis (Chee et al., 2009). Common data projects were defective, leading to different information from the same request on different systems (Inmon, 2013). This problem resulted in the integration of data across the company. Instead of relying on multiple databases not accessible by everyone in the company, all the information was then centralized in the same place, creating a single version of the truth inside the company. This was possible thanks to emerging architectures like the data warehouse made available by some providers. It gives more meaningful results (Obeidat et al., 2015) and makes it possible to make correct assumptions (Bataweel, 2015).

3.2 Meaningful Information

The information retrieved from data analysis should be clear and concise in order to fulfill the company's goals. It means to be easily accessible and readable (Bataweel, 2015) for the users and provide related context to correctly assess the meaning of the analysis (Park et al., 2010).

The role of meaningful information is to be the sum of multiple other, less sensible data for the user to indulge in the decision-making process. This information can take many forms (Park et al. 2010) such as indicators or graphs and be retrospective or in real-time as modern systems do (Inmon, 2013).

The objectives of this meaningful information have evolved rapidly in companies thanks to technology improvements. Back in time, managers could only reflect on the analysis because the results were immediately out of date due to the delay between the collection of data, the analysis, and the receiving of final reports (Inmon, 2013). Today, employees are expected to make data-driven decisions (Romero et al., 2021) and meaningful information is available everywhere and every time (Inmon, 2013). Overall, the primary goal of the reports remained the discovery of knowledge (Panian, 2012), which is basically the extraction of effective information from the data (Ozgul, 2013).

3.3 Decision Making

Decision-making is the final step of the business intelligence process. In an environment where companies have to stay competitive (Capron, 2015), this step allows them to make informed decisions driving their business activities towards profit or alternative goals such as managing risks (Buchanan & O'Connell, 2006). Moreover, as more and more companies use such technologies, it becomes required for them and their strategic processes (Romero et al. 2021). Decision-making inside companies includes some challenges that business intelligence solves thanks to the benefits provided by the results obtained at the end of the process.

3.3.1 Challenges

As mentioned, the primary goal of companies is either to stay competitive or at least stay relevant to keep on their activities. In order to do that, they have to face multiple

challenges that have evolved with time such as monitoring their operational activities and their environment (Park et al., 2010) to remain aware of the opportunities and threats that can influence them. Now, the world is running faster as illustrated by the VUCA environment principle stating that a globally interconnected world induces more complex and unpredictable situation changes (Khare et al., 2015). It thus requires fast and efficient decision-making (Romero et al., 2021) by analyzing real-time information (Bataweel, 2015) but the problem is that data is initially distributed all over the company and beyond, without context (Vaisman & Zimanyi, 2014). Finally, a lot of valuable information is still set aside due to its complexity to process such as textual data (Inmon, 2013).

3.3.2 Benefits

In terms of advantages, decision-making kept providing more and more insights as technologies were improving. It started with benefits for the operational activities, developing into executive summaries designed for strategic decision making at the top level of companies (Buchanan & O'Connell, 2006). Now, business intelligence is so powerful that it can be applied to a lot of different industries (Chee et al., 2009) and companies that successfully implemented those processes have seen their revenue increase substantially (Raj et al., 2016). It helps them reach their goals (Panian, 2012) without too much influence from human judgment (Cebotarean, 2011) by presenting the results in a practical way (Vaisman & Zimanyi, 2014). Overall, it allows companies to become smarter, work smarter and make smarter decisions (Lang & Chang, 2016).

Chapter 4

Methodological Background

This part is used to explain in more depth the Text Mining and the insights that can be drawn from it. The methodology used to code it is detailed in the next chapter.

The greater part of knowledge about business intelligence providers and technologies can be found in the form of unstructured data in managerial documents such as the Gartner reports. The issue is that people interested in such solutions do not always have the time to go through all the reports. One technique that has been developed to tackle such problems is Text Mining.

Text Mining can be defined as a “Process interacting with multiple documents by using analytic tools in order to identify and explore interesting patterns" (Feldman & Sanger, 2007). Basically, it means that, based on statistical methods, the content of a text is analyzed in order to get valuable information without having to read all of it.

Text Mining is still an emerging technology, made useful by the increased quantity of data available everywhere and the need for companies and users to understand large amounts of unstructured information. This technology analyzes a corpus of texts in a limited amount of time, tasks that would have required days or weeks to perform by an analyst without the help of technology.

Moreover, we know that humans are subjective by nature, and by simply reading the reports or documents we might not see all the important topics or their evolution. Text mining, therefore, makes it possible to have objective outputs (e.g. a list of frequent words) from unstructured documents. This will never make it completely objective since the articles are also written and selected by humans, and that biases are always possible to be brought into the analysis if not careful enough, but it gives interesting insights to interpret.

Text Mining can be used for different purposes such as information extraction, categorization, clustering, visualization, or summarization of texts (Gaikwad et al., 2014). The goal that interests us is related to information extraction and more precisely the modeling of topics.

4.1 Topic Modeling

Topic Modeling is used in Text Mining to extract underlying topics available in textual documents by analyzing the co-occurrence of particular words that relate to each other

(Alghamdi & Alfalqi, 2015). The assumption behind Topic Modeling is that documents are composed of topics that are themselves composed of words. So, based on the frequency of words in a particular text, it is possible to detect patterns and thus repetitive topics inside the same document or across multiple ones.

The work of Denter et al. (2019) tells us that Topic Modeling is a well-suited technique to analyze the extent of a technology's applications. Instead of going through the full scientific literature of a specific technology, the approach can be used to summarize the content and give relevant insights to analysts (Alghamdi & Alfalqi, 2015).

The technique uses "Document per Term Matrix", a mathematical technique to transform unstructured data that are all textual information into a dataset, more structured, in the form of a table. The DTM consists of rows (each row corresponds to the report of a year in chronological order in our case) and columns (each column corresponds to one term). An example is shown in Figure 4.1. The values correspond to the number of times a term appears in a specific document. This matrix allows us to apply the Topic Modeling package (referenced as LDA) available in the R language. These different steps are described in the methodology part.

- **Document-1**: Ice creams in summer are awesome
- **Document-2**: I love ice creams in summer
- **Document-3**: Ice creams are awesome all season

	icecream	summer	love	awesome	season
Doc1	1	1	0	1	0
Doc2	1	1	1	0	0
Doc3	1	0	0	1	1

Figure 4.1: Example of a Document per Term Matrix *(Kumar & Paul, 2016)*

DTMs can be optimized by changing the weight of the words in them. We can also make several iterations to find the ideal assemblies for Topic Modeling. We adapted the iterations and the weight of the words in reaction to the outputs given to us by the algorithm.

LDA (Latent Dirichlet allocation) is an algorithm that seeks to find the underlying topics of a document. It is based on the degree of semantic similarity between words to estimate their coherence and the connections between them.

In conclusion, we explained what Text Mining is and how it can be used to perform Topic Modeling. This is our final goal of the code but other manipulations are also performed on the text. However, since the goal of our work is above all to use Text Mining to show the evolution of BI, it is our central focus throughout the analysis and during the writing of the results.

The background part helped us define the scope of our research based on the definitions of evolution and business intelligence. We realized that the literature on providers was underdeveloped but well connected to the literature on the evolution of technologies. We then deepened our research for papers on technologies to understand the current trends that have emerged. After that, we put them in parallel with broader concepts that can help us understand the evolution of needs and challenges influencing the entrance and exit of BI providers. We finished the Background part by detailing the Text Mining and the Topic Modeling process that will be used in the rest of the work. In summary, this part helped us set the boundaries of the research, identify the evolution of technologies in the BI industry, look into certain key points of the discipline that can have an impact on the providers, and finally explain the methodolofy's theory for the analysis.

Part II

Analysis & Results

This part focuses on the presentation of the methodology and the data used for the analysis as well as the results obtained at the end. First, explain the methodology used in Chapter 5, followed by the presentation of the data in Chapter 6. In Chapter 7, we present all our results and interpret them in regard to the literature developed in the previous part.

Chapter 5

Methodology

Our work analyzes multiple documents composed of thousands of words. A precise methodology must then be defined in order to follow a clear plan.

Text mining is a tool that can be applied to each document individually or the entire corpus of documents. We use it to understand the general trends that emerge from the Gartner reports but also to highlight individual results and see their co-implication and evolution over the years. We now describe the methodology used to allow us to produce the results of Chapter 7. Our goal is to practice Topic Modeling but also to visualize frequencies relating to important words in the scope of our research.

To begin the analysis, we first needed to collect the data. We initially gathered the documents related to our analysis, in pdf format. However, going through such documents is harder than simple texts due to the special structure inside each of them. We therefore used a converter in order to transform the pdf documents into text files. The result was still not completely cleaned, thus we continued the process by manually removing some elements of the documents such as unwanted links, headers and footers that do not give any relevant information for the scope of this work.

After that step, we loaded the collection of documents in the software R, useful for analysis. We chose this programming language because it is often used in university work (O'Grady, 2021) and is a very interesting tool for doing statistics and visualization. Moreover, the packages that R offers follow a methodology well suited for Text Mining. The main packages used in the context of this work are described in Annex E.

First, we removed all extra uninteresting elements such as punctuation, whitespace, symbols, numbers, and common English words (stop words). Furthermore, we made sure not to have any repetitions of words that would create different instances and then distort the results such as plurals or proper names. To do this, we transformed all uppercase letters into lowercase (*to_lower*) and plurals into the singular thanks to lemmatization.

Lemmatization uses an English dictionary to bring words back to their roots (e.g. "data mining" becomes "data mine"). This offers a lot of advantages but has created some problems, for example, we had to change "datum, 3 and 2" back to "data, third and second". Those were spotted during an initial check of the data and had to be corrected by hand to have clean outputs. Another crucial step is to check the spelling of words. We use in this work the method of "Rasmus Bååth's" (Bååth, 2014) to correct the words. The Gartner reports, unsurprisingly, did not contain any misspellings, but the lemmatization step could have created some. A check was important to make sure it did not.

The methodology used for the descriptive analysis of the data is also the part where we thoughtfully transformed unstructured data into structured ones to facilitate the application of topic modeling packages as much as possible. The goal of this part is to realize the extent of the data and to find out how to exploit them to have interesting insights for the general Gartner Report Analysis as such. A sample of the data in a structured form can be seen in Annex C.

At first, we created histograms displaying the content of the different documents. This allowed us to understand that it was important to normalize the word count per document in order to obtain frequencies for the analysis of the reports and make scaled evolution plots.

Then, we tried to find the words common to the corpus which could have an important weight in the analysis whereas they bring nothing to the comprehension of the topics since they are present everywhere such as “data” or “bi”. During this step, we also pulled out the top 10 of the most present words in each document and therefore each year, but this did not give us conclusive results since they were almost similar everywhere, only their order changed. To confirm our intuition that many words were common and repeated frequently in each document we made a visual output of the "Zipf's law" that you can see in the descriptive part of the results which proves that the data tend toward the same patterns of common words.

The next step was therefor to deal with these common words and their weight, which is why we applied the TF-IDF method, giving more weight to frequent words in a document but absent from others and reduces the weight of common words in the entire corpus.

It was then time to create a "Document per Term Matrix" which facilitates the computation time and is necessary for the application of the Topic Modeling packages. The DTM intended for Topic Modeling had been relieved of words common to the corpus that would have been too prominent in the outputs if they had been left. This way, the topics of all the documents were not centered around these concepts. We, therefore, speak of "Weighted DTM" for the unigram[1].

However, for the analysis of the evolution of the frequency of words about technologies and providers in the reports, we used unmodified DTM since we needed words like "data" which are very common but especially very present in the names of technologies. We created these "non-weighted" DTMs centered on bigrams[2] and unigrams in the descriptive part of the code.

As for the analysis part as such, we started by creating individual graphs representing the evolution of the frequency of the same word (about technology and providers) from one year to another in the reports. This allowed us to check the calculations and the understanding of the code. Since we had applied the lemmatization it was important to

[1]Unigram is when we analyze word-by-word
[2]Bigrams are about two words by two words

verify that each individual output was indeed based on the root of the word of interest (e.g. "machine learning" was now written "machine learn"). After this exploration, we decided to produce graphs with several evolutions of technology words at once in order to compare them on the same scale. These graphs are interpreted in Chapter 7 and allow us to verify that the trends in the literature are also present in the Gartner reports and that the Text Mining tool is fully capable of detecting them.

When we started Topic Modeling, we created 3 to 10 topics (using unigrams, bigrams, and trigrams) per document. However, the outputs were not interpretable because the topics all looked the same and were repeated without any central tendencies emerging. We then decided to work with an approach that would highlight the main topics of the corpus and give them a weight per document according to their presence. To find the optimal number of topics, we followed the method of Murzintcev (2020) which is based on the metric of Griffiths & Steyvers (2004). The resulting graph showed us that 11 topics were enough (neither too little nor too much) to capture the essence of the corpus.

After 1000 iterations based on Gibbs sampling which optimizes the output (and therefore the links between the words composing a topic) we found the 11 main topics of all the Gartner reports. This results have been exported and are visible in Table 7.1.

We then created a heat map that included 11 topics for the whole corpus but with different weights according to their frequency in each of the reports. This way of displaying is inspired from the work of Schweinberger (2022) and has multiple advantages:

- A global view is offered
- The trends according to the years are clearly emerging
- The evolution 11 topics related to BI is clearly highlighted

In conclusion, we started the analysis with the collection of raw data in the form of pdf files that we cleaned to have a clear overview of the information we had. Then, we analyzed the evolution of key words related to our subject as well as the names of the BI providers, outlining potential trends in the industry. Finally, as our final result, we used Topic Modeling to create a heat map representing the evolution of the discussed topics in the documents year by year.

Chapter 6

Data Presentation

Analyzing Business Intelligence Platforms can be done in various ways but, in the scope of this work, we decided to focus on Gartner and their reports about Business Intelligence & Analytics. First, the company Gartner is presented with its mission and what it delivers. Following, as Gartner publishes multiple reports on a lot of topics each year, they are broadly presented, focusing on their Magic Quadrants. Finally, the reports that specifically interest us, Business Intelligence & Analytics are detailed to have an overview of the structure and its content.

6.1 Gartner Company

Gartner is a consulting firm with more than 40 years of experience in the area of technology. They are implanted in more than 100 countries providing counseling and research results to companies, helping them make faster and smarter decisions (Gartner, 2022).

Their strength is to provide immediately actionable insights delivered by former experts in the field. Moreover, they provide companies' managers with tools to tackle the fast-changing environment of the IT-industry. More than 15,000 companies are customers of Gartner, including 73% of the Fortune Global 500 index (senhasegura, 2021).

They have two main products for their clients, the Hype Cycle and the Magic Quadrants. The former helps its customers evaluate the maturity of a technology and its potential future evolution (Gartner, 2022). The latter, positions the players of a market based on specific criteria developed in the following part.

6.2 Magic Quadrant Reports

The Magic Quadrants of Gartner are designed to evaluate the position of a company in a particular industry. The two criteria used to classify those firms are their ability to execute and the completeness of their vision, detailed in Table 6.1. Based on those principles, four categories of companies emerge: the leaders, the challengers, the visionaries, and the niche players (Carter, 2017). Each provider's category is tracked in Annex B and the description of each quadrant as stated by Gartner can be seen in Figure 6.1.

Businesses use those analyses to understand how the market is shaped in order to assess some possibilities for investment. Each provider of the analysis fills a specific need and the final choice of Gartner's customers will depend solely on their goals, thus choosing

Figure 6.1: Magic Quadrant's Categories Description

Source: https://www.gartner.com/en/research/methodologies/magic-quadrants-research

the leader is not necessarily the best option (Gartner, 2022).

However, even if Gartner is a renowned company, some argue that these reports have their limits. Indeed, in the B2B market trust is an important element to take into account and those reports, while providing social proof, are not a solid base to establish trust (Sullivan, 2020). Furthermore, the selection criteria to be part of the Magic Quadrants have been modified upwards, thus reducing the chance of small players entering the reports (Carter, 2017).

Ability to Execute	How well a company is doing with their activities
Completeness of Vision	The ability for the company to understand where the market is going. It is dependent of the last three years of activity, the innovations, and their success

Table 6.1: Gartner Ranking Criteria, *(Gartner, 2022)*

6.3 Magic Quadrant for Business Intelligence & Analytics Platforms

Our work focuses on the report "Magic Quadrants for Business Intelligence & Analytics Platforms" as it gives a meaningful overview of the business intelligence industry. We first talk about the process to collect those reports, then discuss their objectives, and finally present their structure.

6.3.1 Collect of Data

Gartner reports on business intelligence and analytics are posted annually against a fee to the firm. However, the current year's report is not free. The former ones are however available on the internet and that is also why our scope focuses on the time period from 2006 to 2020. This gives us 15 reports to analyze.

6.3.2 Goal of the Reports

The goal of these reports is to give a quick overview of the different vendors in the competitive market of BI solutions as well as their capacity to meet users' needs. We also get to have a look at how the vendors are positioned compared to the competition and the strategies or technologies they are using to differentiate themselves from each other. Those reports are useful both for companies trying to find a BI tool to implement in their activities and the companies providing those tools in order for them to compare their place, forces and challenges on the market with those of others.

6.3.3 Structure of the Reports

A typical Garner report on BI Platforms starts with a definition or description of the Market in general. Then, we can have a look at the Magic Quadrant with the different vendors followed by the strengths and cautions of each of them more specifically. After that, we have the different inclusion and exclusion criteria explaining how each vendor is selected and how each of them is placed on the quadrant, the graphical summary.

Chapter 7

Results

This part focuses on the communication of the results of the analysis conducted with the text mining process. We start with a descriptive analysis going through the reports in general to have an overview of the data, followed by the results we got in terms of evolution for the tools and technologies of business intelligence and then the related vendors of BI solutions. Finally, we go through the results of the topic modeling technique, discussing the extracted topics and how they evolved using an original visualization. Thanks to that, we can see the evolution of the industry across the years, which corresponds with the literature in terms of technologies and gives us interesting insights to grasp the evolution of the providers.

7.1 Descriptive Analysis

The analysis begins with an overview of data, showing us the content we have to perform the Text Mining process on. We start with an overview of the number of words in each document shown in Figure 7.1, followed by the analysis of the top terms in the corpus.

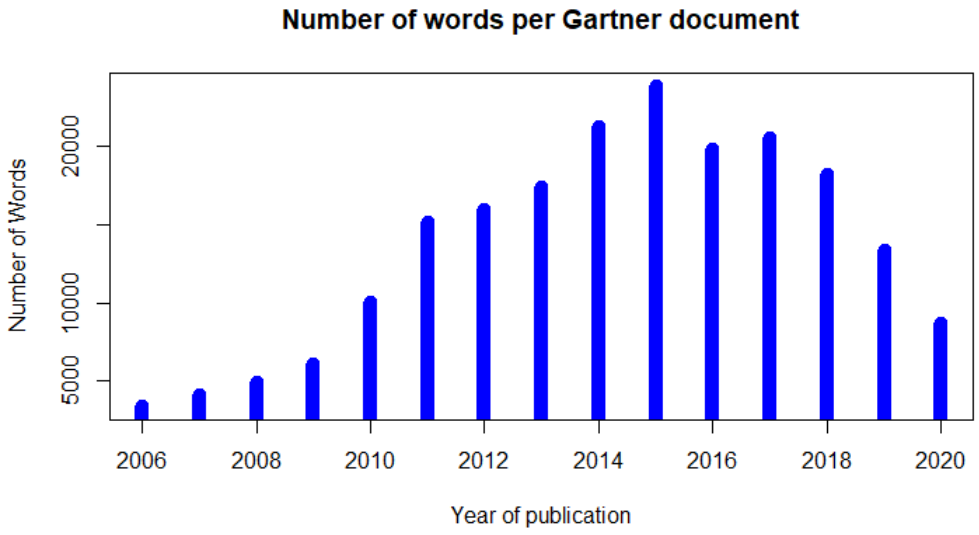


Figure 7.1: Number of words per Gartner report

The first results show us that the reports kept increasing the size of their content from 2006 and reached the maximum in 2015 to end up experiencing a decrease after the peak.

It could be explained by the integration of Gartner's new guidelines for the selection of providers in 2017 (Carter, 2017). Moreover, people using those reports use more and more alternative sources of information (company website, online reviews of products, etc.) to make their choice of business intelligence solution provider (Sullivan, 2020). Gartner could have then focused on the essential information and thus reduced the length of their reports.

The graph on the distribution of the number of words in Figure 7.2 shows us that the majority of reports oscillates from 15,000 to 25,000 words, with some composed of 0 to 10,000. The majority of smaller ones correspond to the reports in the first years. As it was the beginning of such analysis, the content was less furnished. Then, as the field became more interesting, the length increased as well. This gives us the information that for our analysis, we have to take those differences into account in order to avoid giving too much weight to one report over the others.

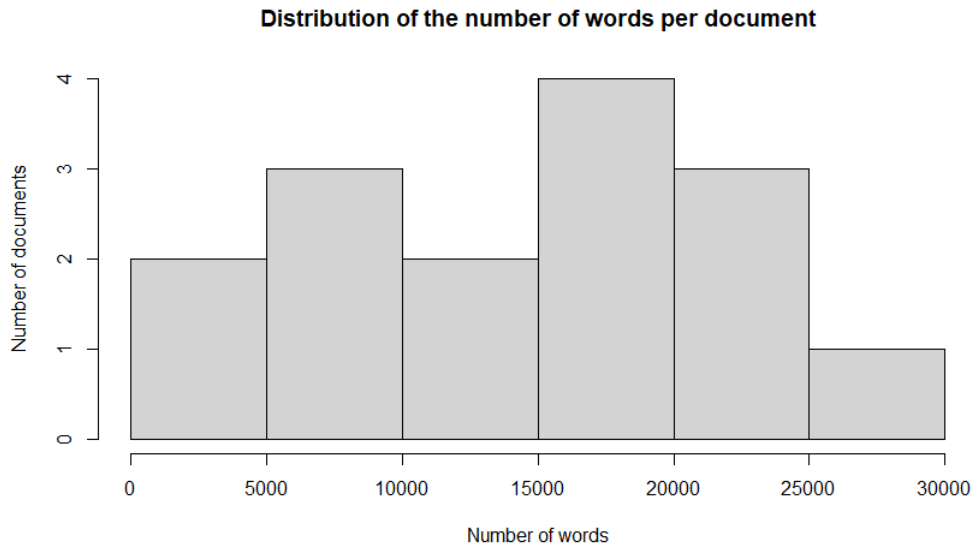


Figure 7.2: Distribution of the number of words per Gartner document

The length of the texts differs from one another but Figure 7.3, a representation of Zipf's Law, shows us the similarity between the reports in terms of word frequency. We see the slope deviating in the left area of the graph, it means that words with a low rank (present in few reports) have a lower frequency of occurrence, which is common. However, the deviation around rank 100 - 1000 is uncommon and represents a higher frequency of common words, especially those related to our field of study and that can be found in the most used terms shown below. This is a problem for the analysis as most common words will take too much weight compared to the others, putting aside some insights relevant for our analysis.

Taking into account these frequencies, we use the measures TF-IDF to be able to remove the influence of common words related to our topic. Our analysis could avoid taking those occurrences into account and get more meaningful results. The given outcome is shown

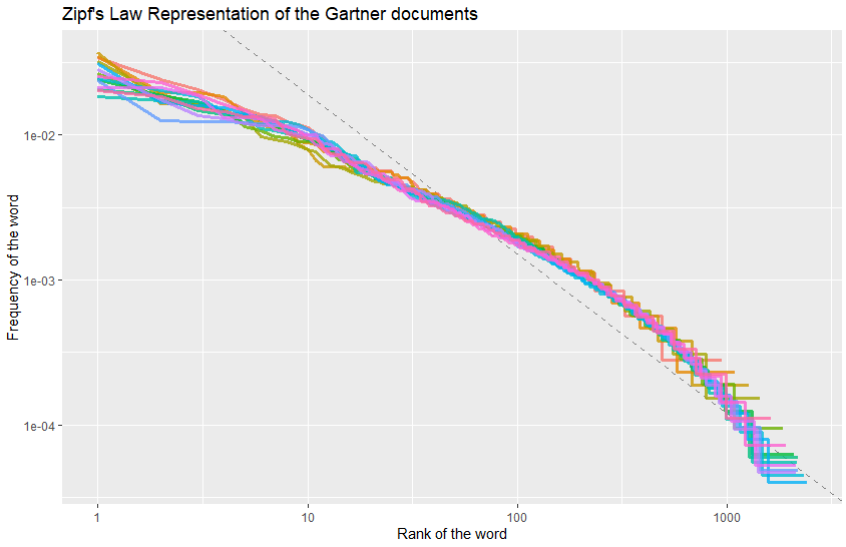


Figure 7.3: Zipf's Law Representation of the Gartner reports

in Annex D.

We see that common words like data and bi, vastly present in our reports, have a weight of 0 for the TF-IDF value. It gives more importance to the words present mainly in one report and decreases the weight of those that can be found in the entire corpus. This allows us to identify underlying topics, which we might not have seen without removing the influence of common words related to our topic. The most discussed words in the corpus can be seen in Figure 7.4.

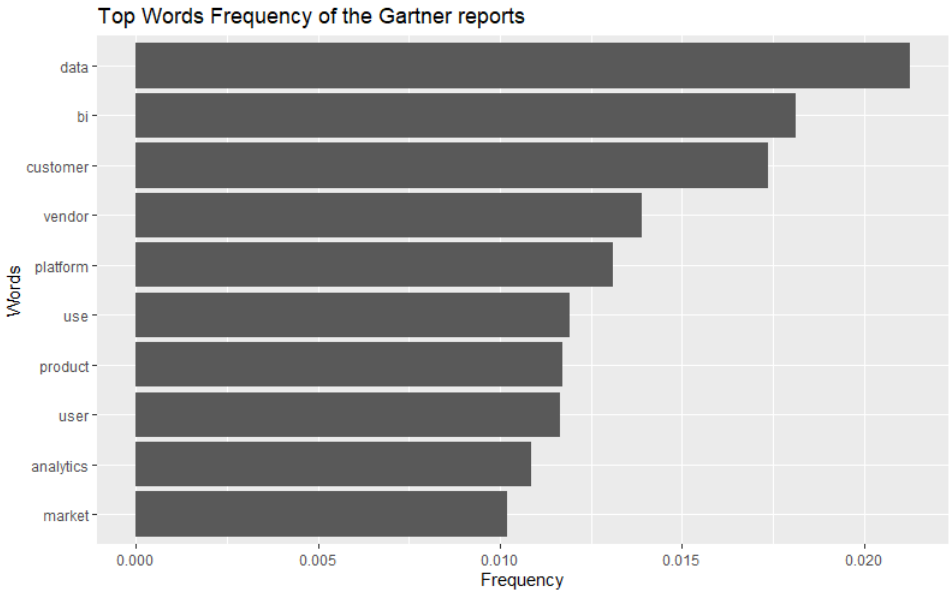


Figure 7.4: Top Words Frequency of the Gartner reports

"Data" and "BI" account for the most part of the words used across the corpus. It relates well to the topic of our analysis. The next most important words discussed are *customer, vendor, platform, use, product, user, analytics, and market.* As the reports talk about analytics platforms in the BI market provided by specific vendors for business customers, this output correctly describes our data.

However, those words are common to the entire corpus. The results of the Topic Modeling could lack meaning and give us only a combination of those words as the final topics. Other measures such as TF-IDF and IDF take into account that problem, allowing us to fix it in the code. In order to still see the evolution of those common words, they are only removed from the Topic Modeling analysis. The results with the ranking based on the TF-IDF value for 2008 are shown in Figure 7.5.

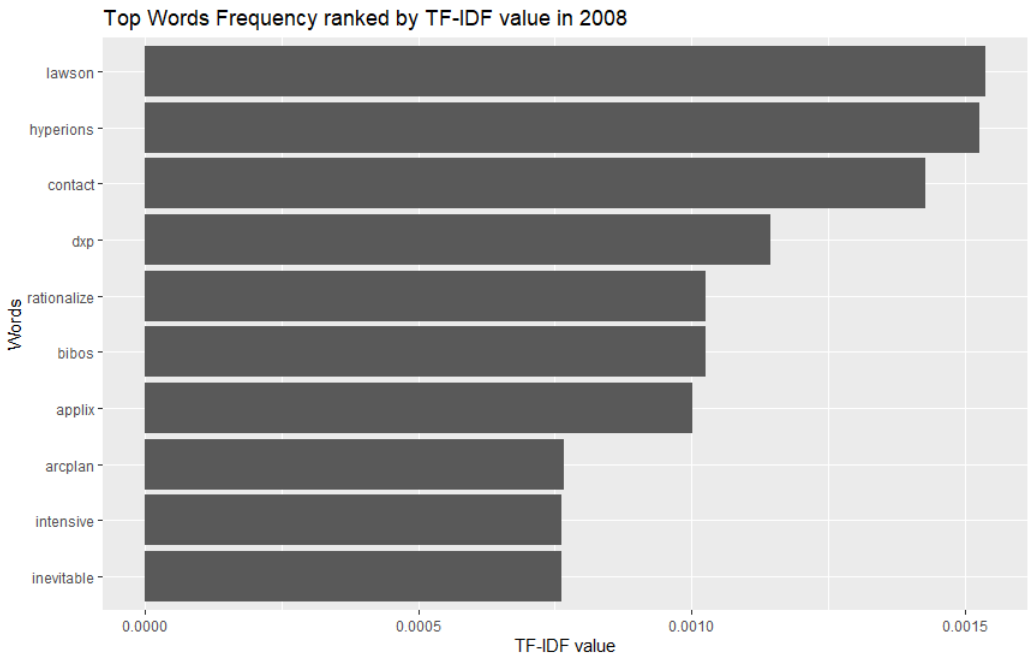


Figure 7.5: Top Words Frequency ranked by TF-IDF value in 2008

The output differs completely from the simple analysis based on the frequency. We see here that the top words using this methodology are different and don't reflect the whole purpose of the reports. These values are useful for the Topic Modeling as it gives more weight to the underlying subjects, producing in the end better results.

7.2 General Results

We now have an overview of the data we are working on. The terms related to our analysis have evolved across time and the providers that cannot be found in the descriptive analysis as well. For this part, we base ourselves on the evolution of the word frequency to assess the popularity of the different notions. The next two sections discuss those topics.

7.2.1 Evolution of Tools and Technologies

The evolution of the business intelligence tools and technologies is based on the frequency of appearance, evaluating the "popularity" of a particular element in the documents. This analysis allows us to confirm the capacity for Text Mining to accurately extract the trends in the Gartner reports. We begin by presenting the capabilities of the process by showing only one keyword and then comparing it with other terms related somewhat to each other. For instance, if we analyze the term "car", we go further by including "sports car", "autonomous car", etc.. Then, we analyze different groups of words to analyze their evolution.

First, to show the extent of possibilities of such analysis, let's take a look at the evolution of the keyword "data" shown in Figure 7.6. We can clearly see the evolution of the term, merely present at the beginning of the reports, gradually increasing to reach its peak in 2015. It shows us that the concept of data per say was probably not the most important aspect of business intelligence when the magic quadrant reports on BI and analytics started to be published. It however evolved to become a major element of the industry.

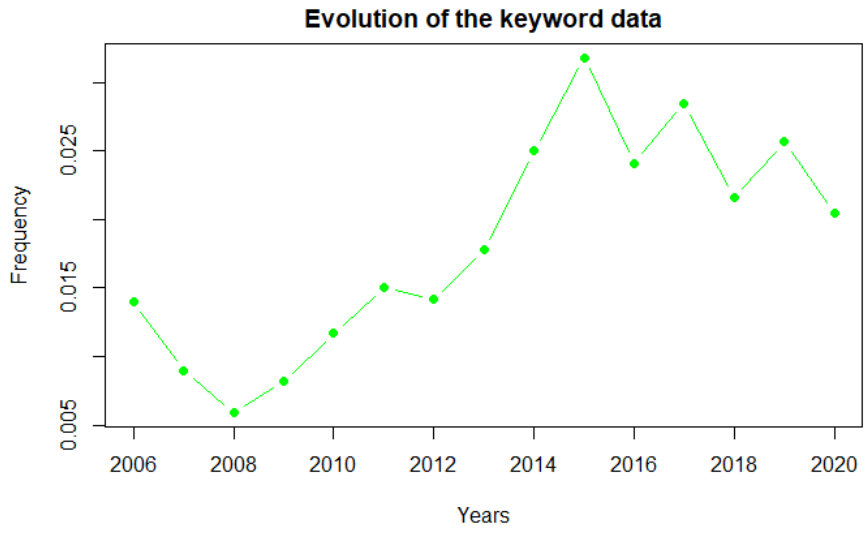


Figure 7.6: Evolution of frequency of the keyword "data"

Then, we can even go further in the analysis by merging multiple keywords in the same chart. In this case, the x-axis corresponds to the id of the document, starting with 1 for 2006. The graph in Figure 7.7 shows us how the words semi-structured and unstructured formats of data have evolved compared to each other. We see that "unstructured data" is more present than "semi structured" one in the corpus, it relates well with our literature on the subject as unstructured data is a type of information that became in the spotlight later than common structured ones (Larson & Chang, 2016).

Thanks to these two graphs, we have an overview of the capabilities of our process. In the remaining part of the results, we discuss four groups of technologies and interpret their evolution based on the theoretical aspects developed in the literature review.

Figure 7.7: Comparison of Evolution of Different Types of Data

The first graph discusses the results of an analysis in terms of insights that are given, it includes *descriptive, predictive, prescriptive, and augmented analytics*. The evolution of these elements is shown in Figure 7.8. The graph shows that from 2006, predictive analytics was a relatively popular term compared to descriptive and prescriptive ones. However, around 2013, it lost some attraction towards prescriptive analytics, which didn't take long to be replaced by augmented analytics. This one exploded from 2018 to nowadays, as opposed to descriptive analytics, already a trend of the past in the time period of our study. It relates well to the work of Andoh-Baidoo et al. (2022) stating that prescriptive analysis was typical of the 2010s years and augmented analysis of the 2020s. It is driven by the increase of computational capacity at a low cost for companies and is no longer limited to a few areas in the firm (Rodriguez, 2017).

The second one shows different elements related to the visualization of data. We included *scorecards, dashboards, data visualization, real-time, and KPIs*. Their evolution can be seen in Figure 7.9. Initial methods to report information consisted of scorecards but visualization techniques started to become more popular to the detriment of others (Bataweel, 2015). This trend is confirmed by the growing use of the words *visualization, dashboard, and real-time*. Since 2009, the association of visualization techniques composed of dashboards, delivered in real-time, has gained popularity in the reports. However, from 2014-2015 to now, it decreased, which can be caused by the reduction in the size of the reports (they do not necessarily talk in detail about tools that they have already announced in previous years), the changes in the guidelines of Gartner or a decrease of activity in the industry (Marjamäki, 2017). For the KPIs, the use of the term remained quite stable over time but relatively low, showing that such indicators are used but do not make an entire BI project. It, however, surpassed the number of occurrences of the scorecard term

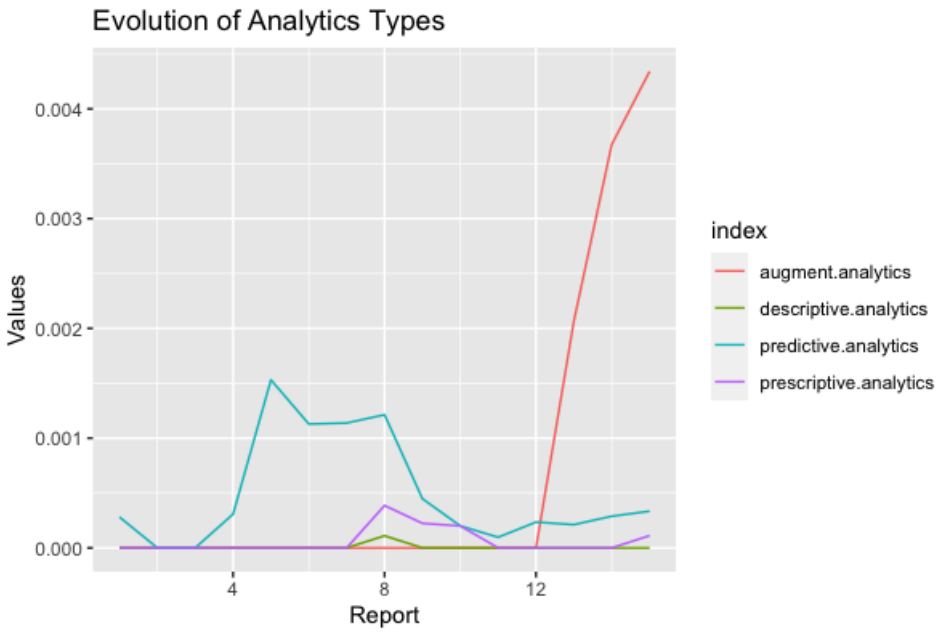


Figure 7.8: Comparison of Evolution of Different Types of Analysis

in 2015.

The third is about multiple technologies developed for BI such as *cloud, big data, self-service, data warehouse, OLAP, IoT, and ETL*. Their evolution is shown in Figure 7.10. The first trend we can observe is the decreasing use of the term OLAP tools for the benefit of cloud technologies and to a lesser extent self-service BI (Romero et al., 2021). In the same sense, IoT devices and big data started to be popular around 2014 because of their falling cost (Ebner, 2015). Data warehouses and ETL remained quite stable over time, representing the importance of the subject but as a secondary technology, just like the KPIs.

The final comparison includes elements such as *data mining, text mining, machine learning, and data science*. Their evolution is shown in Figure 7.11. At the beginning of the reports, we see that data mining was more present than any other term and that text mining is present but in smaller scale, which is confirmed by the literature as data mining represents the early time of BI processes (Panian, 2012). Then, technologies related to artificial intelligence such as data science and machine learning were increasingly used by companies to solve data-related problems (Larson & Chang, 2016) in terms of processing. The increasing use of those terms in the Gartner reports confirms this trend.

Overall, based on the evolution of the frequency of the words in the documents, we can already spot trends in the industry and check how they evolved compared to the others. In the beginning, data mining, OLAP tools, and predictive analytics were the top subjects discussed in the BI and analytics reports. The evolution of the field and its technologies slowly switched towards subjects related to artificial intelligence, big data, cloud and self-service technologies, dashboards and real-time analysis, and augmented analytics. Those

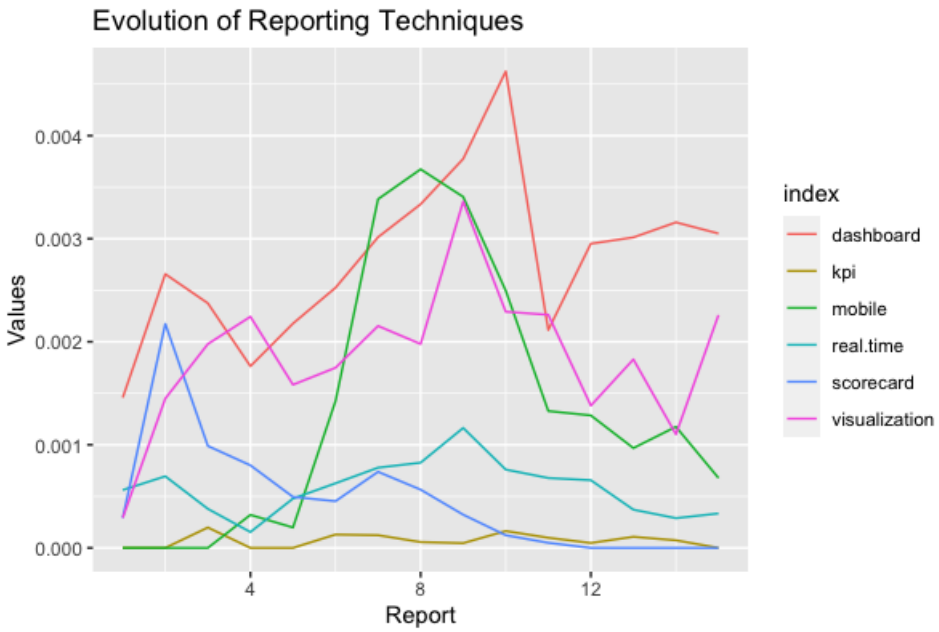


Figure 7.9: Comparison of Evolution of Reporting Techniques

results are confirmed by the work of Panian (2012), Kateeb et al. (2014) and Andoh-Baidoo et al. (2022) as the evolution of the terms used corresponds to the emerging trends in the BI industry. It gives a meaningful overview of the field of study and allows decision-makers to save time in their analysis of the market. Moreover, it justifies the use of our methodology to analyze the content of documents on a particular subject.

7.2.2 Evolution of Providers

Each year, based Gartner's criteria, different vendors come and go in the reports. A summary of the evolution of each year's members and their magic quadrant category is shown in Annex B made upstream to verify the results of the analysis. In total, during the period between 2006 and 2020, 46 BI solution providers were on the market, some emerged, and some disappeared. Microsoft, IBM or QlikTech, for instance, are there from the beginning and are established as the major leaders. We can observe the number of vendors increasing till 2014, then decreasing. In an industry, it generally means that it was maturing (Hayes & Brown, 2022). In 2016, we can see a switch of the leaders of the market into the category of visionaries meaning that the market experienced some changes, probably due to the improvements in the data industry and the need for companies to reinvent themselves. As technologies are improving, the needs of users follow the same trend and firms have to keep up the pace of those changes in the BI market.

In the same way as for technologies, each provider was analyzed based on their frequency of appearance in the reports. The graphs can be seen in Annex A. The assumption made with those results is that if the frequency of the term is higher one year, then this particular BI vendor is considered more important than others this particular year. We consider the importance of the providers as their ability to understand the market and act accord-

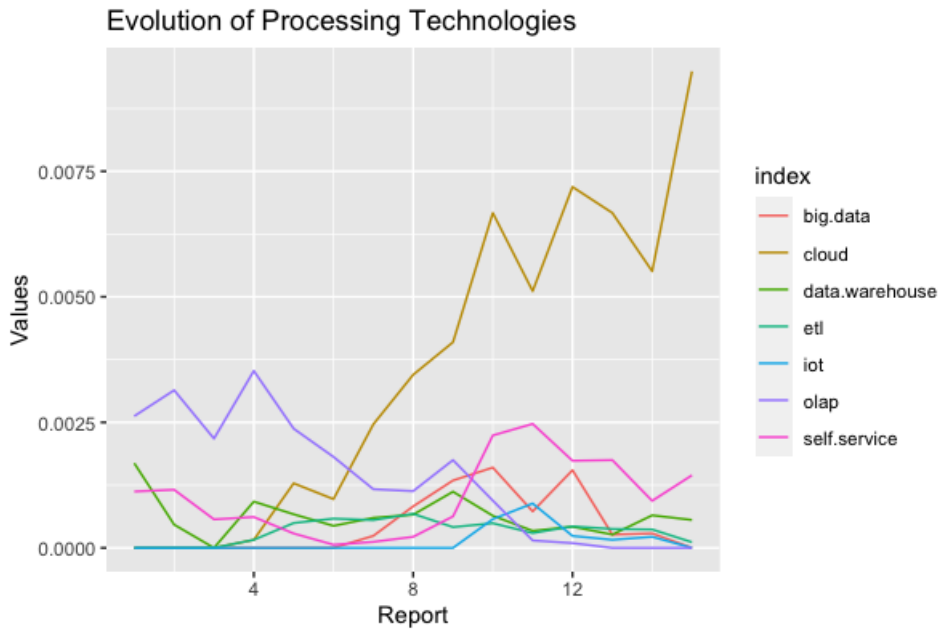


Figure 7.10: Comparison of Evolution of Processing Technologies

ingly in order to stay competitive, just like Gartner measures the place of each of them in the Magic Quadrant.

Based on that assumption, we can spot trending vendors that peaked at some time during those 15 years. We can see, for instance, that Information Builders peaked in 2009, that Microsoft was very popular in the beginning and decreased with the years, that Oracle kept a steady presence but experienced huge gaps in 2011 and 2016, and so on. All those variations are potential movements in the industry that could represent external or internal changes in terms of strategy.

A lot of providers were short-lived during the time of the analysis. It shows how difficult it is for a provider to make it in a market due to the competition, and it is no surprise knowing that a lot of companies go bankrupt in the first years of activity. Of the 16 companies that started in 2006, only half of them are still there in 2020. In the same way, of all the companies that integrated the market at some point, less than half are still on the market and all of them changed position in the magic quadrant. Microsoft is one of the only exceptions that remained a leader up to now. This evolutionary analysis made us aware of the changes in the industry in terms of providers. However, based only on this information it is not possible to interpret the situation of each company. The following part discussing the Topic Modeling analysis attempts to close this gap of knowledge.

7.3 Topic Modeling

We are now using Text Mining to extract the most talked topics in the documents according to statistical measures, selected automatically by the process. Across all the

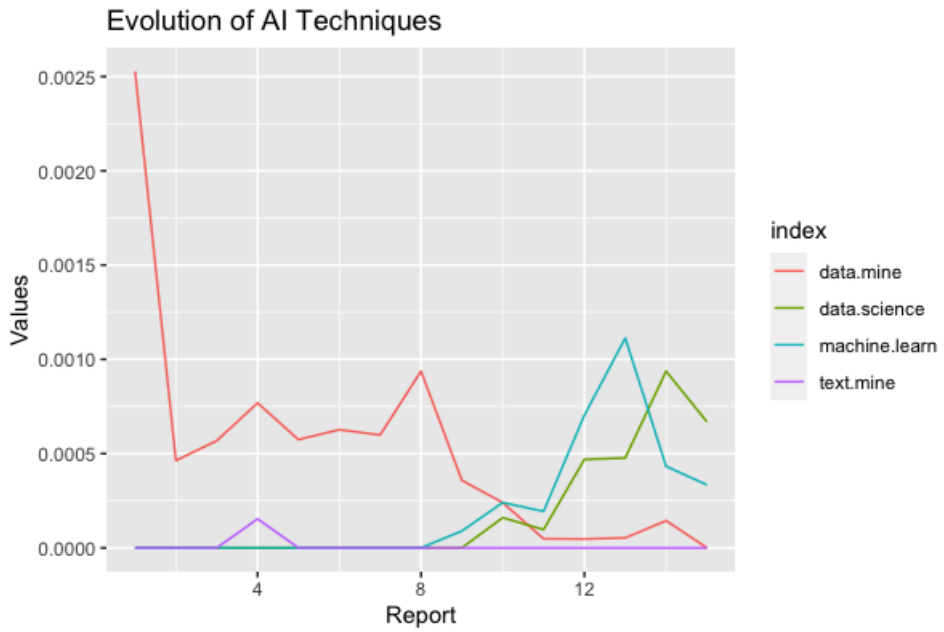


Figure 7.11: Comparison of Evolution of AI Techniques

documents analyzed in this work, some topics discuss different subjects related to the field of business intelligence. First, we figure out the optimal number of topics in the documents. Then, we analyze the topics extracted from the 15 Gartner reports, trying to give context in order to interpret them based on the theoretical information we collected in the literature review. Then, we analyze their evolution between 2006 and 2020 based on a heatmap and the related work part, and conclude on the evolution of the BI industry and its providers.

The optimal number of topics in the documents is determined with the harmonic mean methodology. We see a peak towards 11 topics in Figure 7.12 and a higher one towards 31 topics. However, there is no significant improvement between the two of them. So, we decided to go for 11 topics in order to avoid over-complexify the analysis by following the "Occam's Razor Principle" which states that if multiple solutions are possible, the least complex one should be chosen (Walsh, 1979).

The extraction of the 11 topics is shown in Table 7.1 with the keywords associated with them. We interpret each of them based on the related work and the literature review on business intelligence.

- **Topic 1** refers to BI solutions providers trying to propose smart data discovery tools but that couldn't meet the needs (cloud capabilities or modern platform) of users as they disappeared quickly. Infor, Panorama Software, and Targit all disappeared in 2015.
- **Topic 2** represents the leaders of BI solutions like IBM or Tableau that dominate the market in terms of vision and effectiveness.

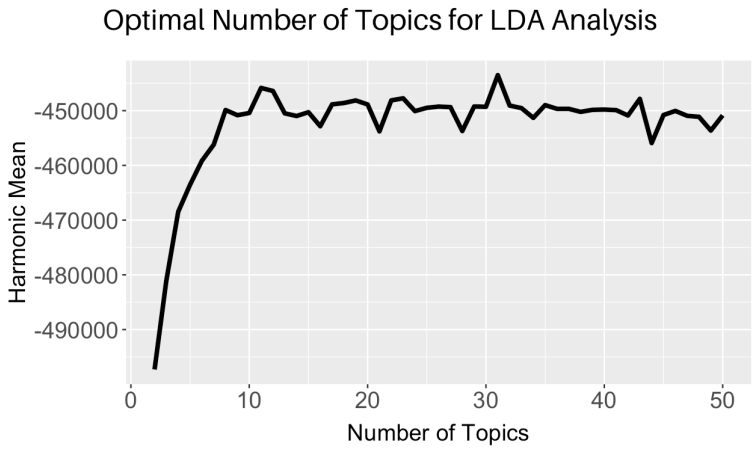


Figure 7.12: Optimal Number of Topics for LDA Analysis

Topic 1	discovery, identify, hadoop, social, infor, collaborative, smart, panorama, targit, olap
Topic 2	ibm, tableau, average, board, caution, search, ease, discovery, version, tibco
Topic 3	average, mobile, salient, percentage, discovery, cloud, alteryx, quartile, gooddata, ease
Topic 4	augment, modern, roadmap, einstein, mode, looker, thougthspot, identify, agile, salesforce
Topic 5	olap, arcplan, panorama, actuate, qliktech, netweaver, research, mine, pure, megavendors
Topic 6	cloud, qlik, preparation, exploration, logi, pyramid, premise, critical, govern, sisense
Topic 7	abi, dundas, alibaba, quick, augmentation, tableau, openness, automation, apis, prediction
Topic 8	yearmagic, infor, economic, mq, id, aris, bex, nature, loop, scheer
Topic 9	targit, tibco, logixml, jaspersoft, response, earn, cite, social, pentaho, commercial
Topic 10	hyperion, demonstrate, applix, siebel, proclarity, etc, underlie, pervasive, prove, able
Topic 11	quartile, modern, bottom, smart, wave, average, ease, discovery, cite, domo

Table 7.1: Topics generated through the Topic Modeling using Text Mining

- **Topic 3** represents all that is related to the trend of mobile analytics and the beginning of the data discovery principle as well as the beginnings of the cloud. Represented vendors being Salient Management Company, Alteryx, and GoodData.
- **Topic 4** represents an increased interest towards modern and augmented analytics with agile methodologies. Related companies include Looker, SalesForce, and ThoughtSpot. Einstein refers to AI-centric capabilities (bought by SalesForce).
- **Topic 5** represents traditional pure-play BI vendors providing OLAP tools capabilities to research information with data mining processes. Vendors are arcplan, Panorama Software, Actuate, QlikTech, and SAP with NetWeaver.
- **Topic 6** refers to the trend of cloud capabilities provided by BI vendors with exploration, preparation and governance. BI vendors are Qlik, Logi Analytics, Pyramid Analytics, and Sisense.
- **Topic 7** represents the use of analytics business intelligence platforms as a replacement for traditional BI, with also an increase in automation and augmentation capabilities and the use of APIs and prediction. Vendors related to this topic are Alibaba Cloud, Tableau, and Dundas.

- **Topic 8** refers to the events related to the economic crisis that happened in 2008 and had an impact on the BI industry the following years.
- **Topic 9** represents the companies towards which BI customers turned in response to the needs unmet by other vendors.
- **Topic 10** represents small, innovative niche players in the BI industry that could not meet users' needs and quickly disappeared from the industry.
- **Topic 11** refers to modern platforms, with easy-to-use functionalities allowing casual business users to indulge in data discovery processes.

All these topics evolved over time, some even disappeared during the 15 years of analysis. This evolution can be seen in Figure 7.13 representing the proportion of each topic by year. For the sake of summarization, each topic is represented as the five first keywords of the topic modeling analysis. We go through each year trying to interpret the distribution of topics while basing ourselves on the extended literature on business intelligence developed in Chapter 3 and the summary available at the beginning of the Gartner reports. It allowed us to base our interpretation on and verify the reliability of our analysis.

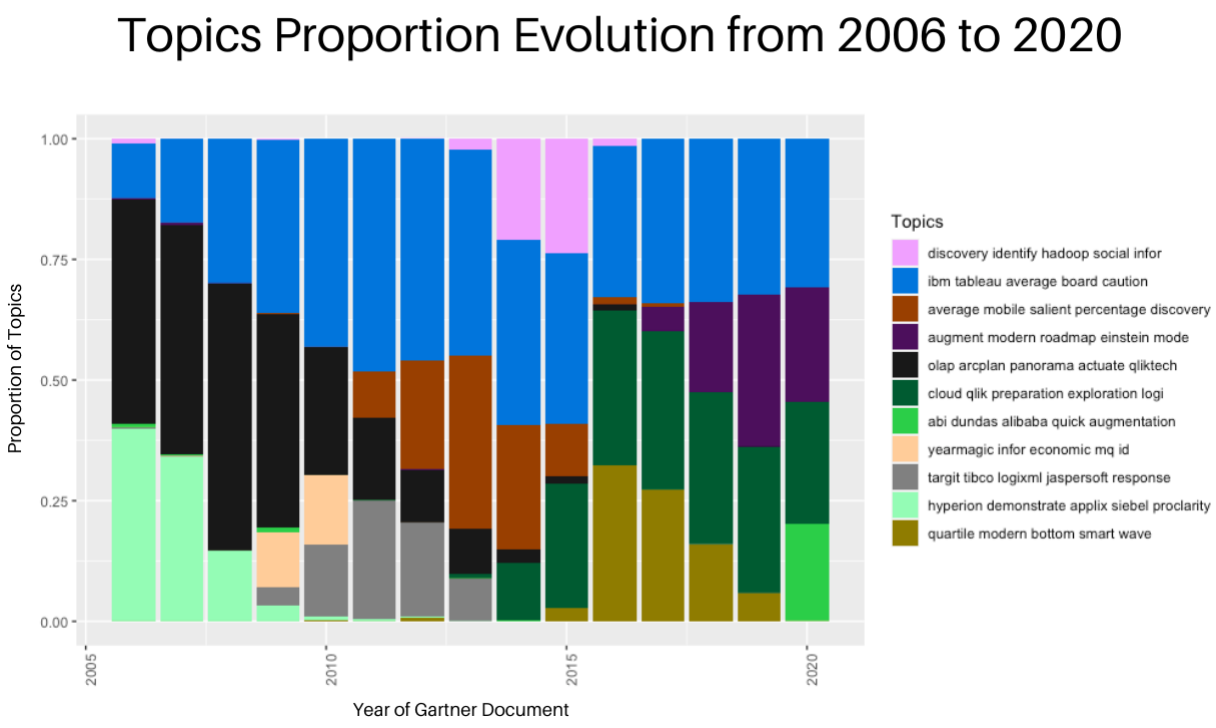


Figure 7.13: Topics Proportion Evolution from 2006 to 2020

From this evolutionary graph, we can make some inferences about the state of the BI industry between 2006 and 2020. One should keep in mind that all these interpretations are subjective because our analysis is based on a Text Mining analysis that does not provide

any context.

The initial landscape of the field was composed of three types of companies. Large providers referenced under the name of mega vendors in the Gartner reports are identified under the blue color such as IBM, Tableau, and so on. Traditional BI providers with OLAP tools like arcplan, Panorama Software, Actuate, and others with the black color. They dominated the market over the other players, probably because OLAP tools have been the norm for a decade (Bataweel, 2015). The remaining actors are smaller ones that didn't take long before disappearing from the industry and are shown with a light green color. Based on the criteria of Gartner, they lacked vision and ability to execute. Till 2008, that landscape remained stable, however, the topic about large providers got more space as well as the one about traditional BI players to the detriment of smaller ones. It can be explained by the various industries using BI (Chee et al., 2009), companies, therefore, preferred larger providers covering more functionalities over more specialized actors.

In 2008, the economic crisis, a topic that was not expected in those reports (at first glance unrelated to our analysis) and shown in beige, appears to have changed the landscape probably due to the use of analytics in the financial sector (Davenport, 2008). These moments of crisis have pushed companies to have to justify the interest of certain analyzes or make sure the data was accurate, notably via the use of data discovery processes upstream of the project (Thomas, 2010). The topic about initial BI players with OLAP capabilities lost a lot of importance to larger ones and new players, shown in gray, entered the market as an alternative. However, 2011 brought its new players, displayed in orange, oriented towards new technologies such as data discovery, cloud, and mobile BI, and confirmed by the work of Kateeb et al. (2014). The topic of data discovery kept getting more popular and even became a dominant one in 2013. Furthermore, it developed towards smart data discovery in 2014, which can be seen with the emergence of the pink color.

The cloud topic, in dark green, started to be important in 2014 with companies who switched names like QlikTech to Qlik. It can be associated with a switch of strategy towards new technologies, more appealing to those companies and their customers thanks to their better capacities to meet the growing needs of users in terms of data storage (Bataweel, 2015). That topic remained amongst the major ones up to 2020.

Modern platforms have their own topic under the gold color. They gained popularity and the investments in 2016 were mostly directed toward those platforms, as they provide fully integrated functionalities made usable by ordinary business users. They are characterized by software as a service, smart data discovery, and ease of use. This was discussed by Romero et al. (2021). Almost at the same time, around 2017, a topic related to modern and augmented platforms, in purple, got more space. Those platforms have also proven their capabilities with AI-driven functionalities, automated tools, and agile methodologies (Andoh-Baidoo et al., 2022).

Finally, we see the emergence in 2020 of a new name to qualify the BI platforms, "Analytics and BI platforms" under the abbreviation “abi”. It ends up in its own topic and is

associated to sew software such as Alibaba Cloud, Dundas, or Looker from Google, with augmentation capabilities.

Globally, initial popular topics included a few large vendors, traditional BI solutions and some small, innovative players. As time passed, those topics evolved and even disappeared for traditional players and niche players. Today, the landscape of the BI industry is composed of those large vendors who now dominate the market with augmented and modern platforms providing AI functionalities and agile methodologies, a cloud architecture with complete governance, and some emerging capabilities such as augmented and automated analytics. This analysis allowed us to have a glance at the full spectrum of the BI industry between 2006 and 2020 by showing the proportion of each major trend year by year as well as the vendors that are related to them. With some interpretation and a good sense of observation, the whole industry is summarized into one graph.

Conclusion

Throughout this work, we indulged in answering the following research question: **How has the BI industry evolved between 2006 and 2020, using Text Mining on the Gartner reports?**

This question has been answered by comparing the results of the text mining analysis with the literature developed in the background part. We confirmed with the Gartner reports that the initial state of the industry was composed of big vendors such as IBM or Tableau, traditional providers of BI solutions with OLAP tools and small niche players trying to differentiate themselves from the rest of the market. The economic crisis in 2008 has played a major role in the industry by forcing companies to innovate in terms of BI capabilities. It resulted in the emergence of new technologies such as mobile analytics or new processes like data discovery, used to get an overview of the data to check any problem upstream of the BI projects. The increasing number of sources of data, quantity of data and types of data drove new companies in the market to compete with established vendors. Their goal was to fill specific needs of firms not satisfied with current capabilities, thus traditional players disappeared to give their place to these new competitors. Nowadays, mainly cloud technologies and augmented analytics capabilities are dominating. It provides corporations with analytics possibilities accessible to any business users without the need of deep technological knowledge. Overall, the market is running towards quicker analysis, with a huge amount of data to process for any employee to make data-driven decisions for the benefit of the business.

These results were obtained thanks to the Text Mining, which proved to be useful in that kind of project. Analyzing the frequency of apparition of words in the documents allowed us to grasp the popularity of the technologies and providers, which were put in parallel with the literature to make sure it was corresponding. After the methodology was certified, the topic modeling of the whole corpus gave us a heatmap showing the proportion of the subjects addressed throughout the years. This gave us a clear overview of the market and a more visual way to understand how they evolved and how they fitted with each other.

Overall, each part of the work brought some information allowing us to answer the research question announced at the beginning of the project. On the one hand, the background part helped us understand the current state of the BI industry without the influence of the Gartner reports, allowing us to have a solid ground to compare the results of our analysis. On the other hand, the results part brought some new insights in accordance with the existing literature and new ways of visualizing the information thanks to Text Mining.

7.4 Limits of the Work

This work includes its limits due to various factors influencing the results and their interpretation. Those factors described thereafter are related to the Text Mining process and the use of the Gartner reports.

Using the text mining process brings some limitations to the results we can obtain. The major element is that the output lacks context therefore the interpretation is subjective, left to the discretion of the analyst, who could create some bias. In our analysis, discussing a provider more than another is a sign that this one is more popular whereas this is not necessarily the case. The language used is also a limitation for the extent of results we could obtain. Indeed, text mining is mainly used with the English language and cross language analysis is more complex to implement (Cancedda & Renders, 2011), limiting our research on the documents written by English-speaking analysts. The views of other nationalities on BI solution providers are thus omitted. Moreover, the final result is largely influenced by the preprocessing stage of the method. Different users could have different results, making the analysis less robust.

The Gartner reports are also a source of limitation in our work. They give an overview of the field of study but are limited to big companies such as Microsoft, Oracle, etc., omitting the smaller players that don't have the requirements to be incorporated in those reports. Moreover, the time period studied by Gartner does not reflect the complete timeline of business intelligence. Indeed, it has been around for decades whereas Gartner started publishing its reports in 2006. Some insights on these companies are then left aside for previous periods. Finally, alternative research using text mining is applied to more documents, hundreds of them in some cases, whereas Gartner only provides one report each year. The amount of the data processed is, therefore, smaller than other research, making the results harder to interpret.

In conclusion, the work gives us insightful results on the evolution of business intelligence providers and technologies but some limitations make those results less meaningful. Further research taking into account those limitations could be applied and is presented in the next section.

7.5 Future Work

Our work only scratches the surface in terms of business intelligence evolution. Based on the limitations of the work as well as personal critical thinking, different leads of research can be deepened.

The Text Mining analysis is able to create alternative result such as sentiment analysis or word graphs, it is, therefore, possible to analyze how the interactions have evolved through time by analyzing those graphs on a yearly basis. There exists plenty of measures giving information based on graphs such as centrality, degree, and connectivity (Cogis & Schwartz, 2018). Hence, the relationship between the providers and the technologies could

be deepened and highlighted with new insights to complement a text mining analysis.

We based ourselves on the Gartner reports only to apply the text mining process but the scientific literature also includes research on business intelligence. It would be interesting to apply the Text Mining process to those articles and therefore compare both of them on the topic of the technologies as the providers are not thoroughly developed in the literature as such.

Finally, one major novelty in the data industry is the current regulations such as the GDPR in Europe that reduces the potential of companies to collect information on their users. Back in time, companies could only rely on a few information such as the number of sales but as technologies improved, they were able to gather more and more pieces of information up to a point where they would know almost everything about their customers like their interests, activities, family situations, and so on. Those policies preventing them to indulge in such activities bring new challenges in order to get the same results without as much information, thus it could be interesting to analyze this evolution regarding the field of business intelligence. In the same sense, companies providing BI solutions are also subject to those changes, parallel studies with those corporations could be included as well in order to have the full spectrum of firms operating in the field.

References

Ahmed, E., Yaqoob, I., Hashem, I. A. T., Khan, I., Ahmed, A. I. A., Imran, M., & Vasilakos, A. V. (2017, December 24). The role of big data analytics in Internet of Things. Computer Networks, 129, 459 - 471. https://doi.org/10.1016/j.comnet.2017.06.013

Alghamdi, R., & Alfalqi, K. (2015, January). A Survey of Topic Modeling in Text Mining. International Journal of Advanced Computer Science and Applications, 6(1), 147 - 153. 10.14569/IJACSA.2015.060121

Andoh-Baidoo, F. K., Chavarria, J. A., Jones, M. C., & Wang, Y. (2022). Examining the state of empirical business intelligence and analytics research: A poly-theoretic approach. Information & Management.

Antioco, J. (2011, April). How I Did It: Blockbuster's Former CEO on Sparring with an Activist Shareholder. Harvard Business Review. https://hbr.org/2011/04/how-i-did-it-blockbusters-former-ceo-on-sparring-with-an-activist-shareholder

Bååth, R. (2014, December 17). Peter Norvig's Spell Checker in Two Lines of Base R. Publishable Stuff. Retrieved July 8, 2022, from https://www.sumsar.net/blog/2014/12/peter-norvigs-spell-checker-in-two-lines-of-r/

Bataweel, D. S. (2015). Business Intelligence: Evolution And Future Trends. BCG. (2022). Emerging Business Technologies Consulting | BCG. Boston Consulting Group. Retrieved July 4, 2022, from https://www.bcg.com/capabilities/digital-technology-data/emerging-technologies

Bessette, K. (2020, April 15). What is the Gartner Magic Quadrant and why is it relevant to you. Veeam. Retrieved February 07, 2022, from https://www.veeam.com/blog/gartner-magic-quadrant-for-it-pros.html

Borden, V. M. H., & Bottrill, K. V. (1994). Performance Indicators: History, Definitions, and Methods. New Directions for Institutional Research, 1994(82), 5 - 21. 10.1002/ir.37019948203

Buchanan, L., & O'Connell, A. (2006, January). A Brief History of Decision Making. Harvard Business Review.

Burstein, F. (2008). Handbook on Decision Support Systems 2: Variations (F. Burstein & C. W. Holsapple, Eds.). Springer. https://doi.org/10.1007/978-3-540-48716-6_9

Cancedda, N., & Renders, J.-M. (2011). Cross-Lingual Text Mining. Encyclopedia of Machine Learning, 243 - 249. https://doi.org/10.1007/978-0-387-30164-8_189

Capron, L. (2015, September 8). How Should Companies Evolve? | Yale Insights. Yale Insights. https://insights.som.yale.edu/insights/how-should-companies-evolve

Carter, R. (2017, July 3). How Trustworthy is the Gartner Magic Quadrant? UC Today. Retrieved April 12, 2022, from https://www.uctoday.com/unified-communications/gartner-magic-quadrant/

Cebotarean, E. (2011). Business Intelligence. Journal of Knowledge Management, Economics and Information Technology.

Chee, T., Chan, L.-K., Chuah, M.-H., Tan, C.-S., Wong, S.-F., & Yeoh, W. (2009). Business Intelligence Systems: State-of-the-art Review and Contemporary Applications. Symposium on Progress in Information & Communication Technology, 96 - 101.

Cogis, O., & Schwartz, C. (2018). Théorie des graphes. Cassini.

Collins English Dictionary. (2012). Evolution Definition & Meaning. Dictionary.com. Retrieved April 17, 2022, from https://www.dictionary.com/browse/evolution

Davenport, T. H. (2008, September 25). Is This an Analytics-Driven Financial Crisis? Harvard Business Review.

Denter, N., Caferoglu, H., & Moehrle, M. G. (2019). Applying Dynamic Topic Modeling for Understanding the Evolution of the RFID Technology. Technology Management in the World of Intelligent Systems.

Dittrich, J., & Quiané-Ruiz, J.-A. (2015, August 01). Efficient big data processing in Hadoop MapReduce. Proceedings of the VLDB Endowment, 5(12), 2014 - 2015. 10.14778/2367502.2367562

Ebner, J. (2015, January 27). How Sensors Will Shape Big Data and the Changing Economy. Dataconomy. Retrieved June 6, 2022, from https://dataconomy.com/2015/01/how-sensors-will-shape-big-data-and-the-changing-economy/

Feinerer, I. (2020, November 18). Introduction to the tm Package Text Mining in R. 1 - 8.

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014, January). Text Mining Methods and Techniques. International Journal of Computer Applications, 85.

Gartner. (2022). Magic Quadrant Research Methodology. Gartner. Retrieved July 1, 2022, from https://www.gartner.com/en/research/methodologies/magic-quadrants-research

Gartner. (2022). What We Do and How We Got Here. Gartner. Retrieved July 1, 2022, from https://www.gartner.com/en/about

Griffiths, T. L., & Steyvers, M. (2004, April 06). Finding scientific topics. Proceedings Of The National Academy Of Sciences, 101, 5228-5235. https://doi.org/10.1073/pnas.0307752101

Grigorescu, A., Baiasu, D., & Chitescu, R. I. (2020, August). Business Intelligence, the New Managerial Tool: Opportunities and Limits. Ovidius University Annals, Economic Sciences Series, 651 - 657.

Han, J., E, H., Le, G., & Du, J. (2011). Survey on NoSQL database. 2011 6th International Conference on Pervasive Computing and Applications, 363 - 366. 10.1109/ICPCA.2011.6106531

Hasan, S. A. (2017, August). The Impact of Business Intelligence on Strategic Decision-Making. European Journal of Business and Management, 9.

Hayes, A., & Brown, J. R. (2022, March 28). Mature Industry Definition. Investopedia. Retrieved August 3, 2022, from https://www.investopedia.com/terms/m/matureindustry.asp

Hiter, S. (2021, May 4). What is Raw Data? Datamation. Retrieved July 6, 2022, from https://www.datamation.com/big-data/raw-data/

Inmon, W. H. (2013). Evolution of Business Intelligence. Business Intelligence and Performance Management, Springer, 263 - 269.

Jourdan, Z., Rainer, K., & Marshall, T. (2008, March). Business Intelligence: An Analysis of the Litterature. Information Systems Management, 25, 121 - 131.

Kaplan, R. S., & Norton, D. P. (1996). The Balanced Scorecard: Translating Strategy into Action. Harvard Business Review Press.

Kateeb, I., Humayun, S., & Bataweel, D. (2014). The Evolution of Business Intelligence.

Khare, A., Burgartz, T., Mack, O., & Krämer, A. (Eds.). (2015). Managing in a VUCA World. Springer International Publishing. https://doi.org/10.1007/978-3-319-16889-0_1

Kinsey, A., & Seidel, M. (2019, June 1). Importance of Decision Making in Management. Bizfluent. Retrieved June 4, 2022, from https://bizfluent.com/about-6618914-importance-decision-making-management.html

Kontostathis, A., & Pottenger, W. M. (2002). Detecting Patterns in the LSI Term-Term Matrix.

Kumar, A., & Paul, A. (2016). Mastering Text Mining with R (A. Kumar, Ed.). Packt Publishing.

Larson, D., & Chang, V. (2016, April). A review and future direction of agile, business intelligence, analytics and data science. International Journal of Information Management,

700 - 710.

Lebied, M. (2017, September 27). The History of Business Intelligence: A Dive Into The Roots of BI. Datapine. Retrieved May 16, 2022, from https://www.datapine.com/blog/history-of-business-intelligence/

Limp, P. (2020). A Complete History of Business Intelligence. Toptal. Retrieved May 16, 2022, from https://www.toptal.com/project-managers/it/history-of-business-intelligence

Marjamäki, P. (2017). Evolution and Trends of Business Intelligence Systems: A Systematic Mapping Study.

Murzintcev, N. (2020, April 20). Select number of topics for LDA model. The Comprehensive R Archive Network. Retrieved June 3, 2022, from https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html

Obeidat, M., North, M., Richardson, R., & Rattanak, V. (2015). Business Intelligence Technology, Applications, and Trends. In International Management Review Journal (Vol. 11, pp. 47-56).

O'Grady, S. (2021, March 1). The RedMonk Programming Language Rankings: January 2021 – tecosystems. RedMonk. https://redmonk.com/sogrady/2021/03/01/language-rankings-1-21/

Ozgul, F. (2013). Incorporating Data and Methodologies for Knowledge Discovery for Crime. Intelligent Systems for Security Informatics, 181 - 197. https://doi.org/10.1016/B978-0-12-404702-0.00009-4

Pan, W., & Wei, H. (2012). Research on Key Performance Indicator (KPI) of Business Process. In Second International Conference on Business Computing and Global Informatization. 10.1109/BCGIN.2012.46

Panian, Z. (2012). The Evolution of Business Intelligence: From Historical Data Mining to Mobile and Location-based Intelligence.

Park, J., Fables, W., Parker, K. R., & Nitse, P. S. (2010, July-September). The role of culture in business intelligence. In International Journal of Business Intelligence Research.

Patel, N. (2021, August 20). The Evolution of Business Intelligence and Analytical Reporting. BMC Software. Retrieved May 16, 2022, from https://www.bmc.com/blogs/analytical-reporting/

Petrini, M., & Pozzebon, M. (2009). Managing Sustainability with the Support of Business Intelligence Methods and Tools. Information Systems, Technology and Management, 88 - 99.

Raj, R., Wong, S. H., & Beaumont, A. J. (2016, November 11). Business Intelligence Solution for an SME: A Case Study. Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016), 41 - 50. https://doi.org/10.5220/0006049500410050

Rekatsinas, T., Dong, X., Getoor, L., & Srivastava, D. (2015). Finding Quality in Quantity: The Challenge of Discovering Valuable Sources for Integration. CIDR.

Rodriguez, E. (2017). Evolution of Analytics Concept. In Data Analytics Applications in Latin America and Emerging Economies (pp. 3 - 19).

Romero, C. A. T., Ortiz, J. H., Khalaf, O. I., & Prado, A. R. (2021). Business Intelligence: Business Evolution after Industry 4.0. Sustainability.

Schweinberger, M. (2022, May 21). Topic Modeling with R. Language Technology and Data Analysis Laboratory (LADAL). Retrieved July 11, 2022, from https://slcladal.github.io/topicmodels.html

senhasegura. (2021, September 30). How important is the Gartner report? senhasegura. Retrieved April 12, 2022, from https://senhasegura.com/how-important-is-the-gartner-report/

Sullivan, O. (2020, December 3). Does Anyone Really Care About Gartner's Magic Quadrant? Finance Monthly. Retrieved April 12, 2022, from https://www.finance-monthly.com/2020/12/does-anyone-really-care-about-gartners-magic-quadrant-reports-anymore/

Syed, S., & Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 165-174.

Talaoui, Y., & Kohtamäki, M. (2020, September). 35 years of research on business intelligence process: a synthesis of a fragmented literature. Management Research Review, 44, 677 - 717.

Thomas, P. J. (2010, March 28). Limitations of Business Intelligence. Peter James Thomas. Retrieved July 12, 2022, from https://peterjamesthomas.com/2010/03/28/limitations-of-bi/

Tidyverse packages. (2022). Tidyverse. Retrieved August 7, 2022, from https://www.tidyverse.org/packages/

Vaisman, A., & Zimányi, E. (2014). Data Warehouse Systems: Design and Implementation. Springer Berlin Heidelberg.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009, June 14). Evaluation methods for topic models. ICML '09: Proceedings of the 26th Annual International

Conference on Machine Learning, 1105 - 1112.

Walsh, D. (1979, July). Occam's Razor: A Principle of Intellectual Elegance. American Philosophical Quarterly, 241 - 244.

Watson, H., & Wixom, B. (2007, October). The Current State of Business Intelligence. Computer, 96 - 99. 10.1109/MC.2007.331

Wickham, H. (2022). stringr: Simple, Consistent Wrappers for Common String Operations. Simple, Consistent Wrappers for Common String Operations. Retrieved May 5, 2022, from http://stringr.tidyverse.org

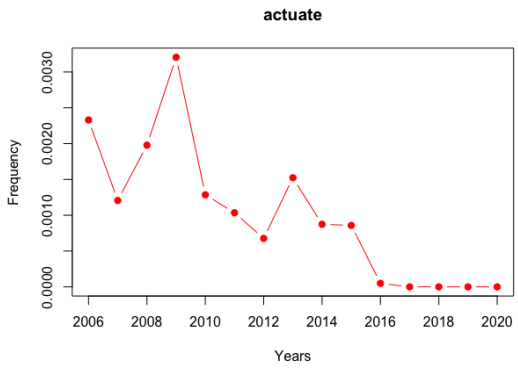
Wilkins, J. (2001, January-April). Defining Evolution. Reports of the National Center for Science Education, 21.

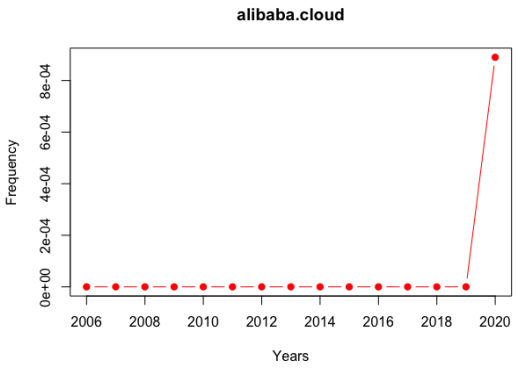
Part III

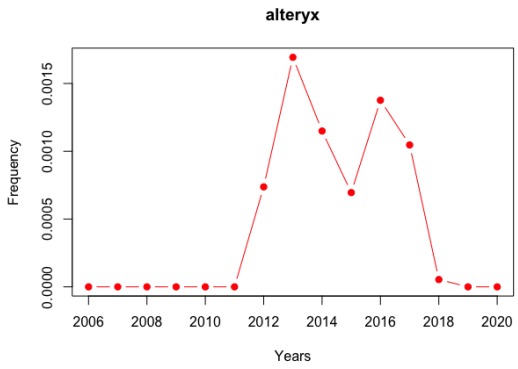
Appendices

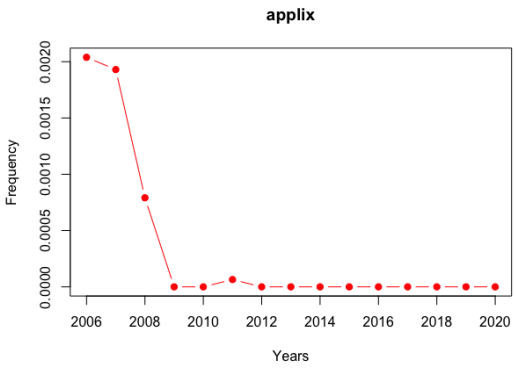
Appendix A

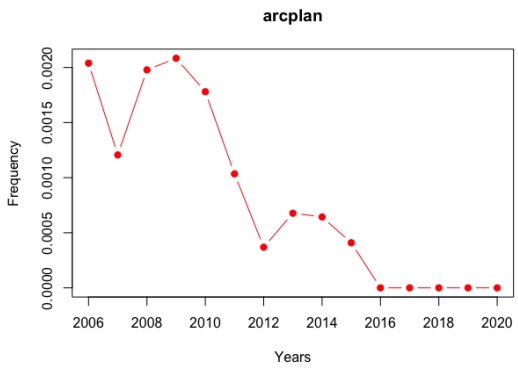
Frequency Evolution of Providers

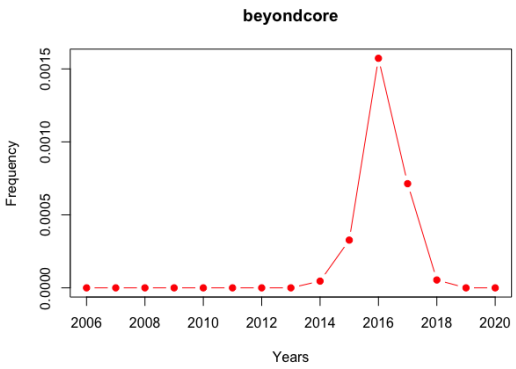


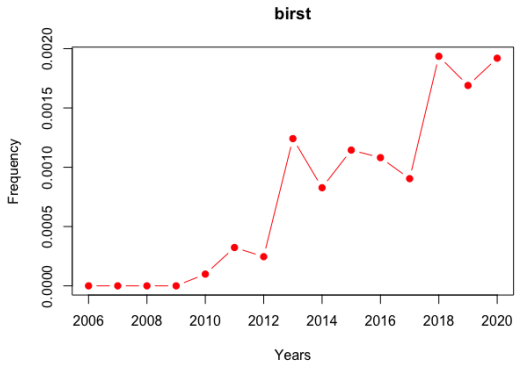


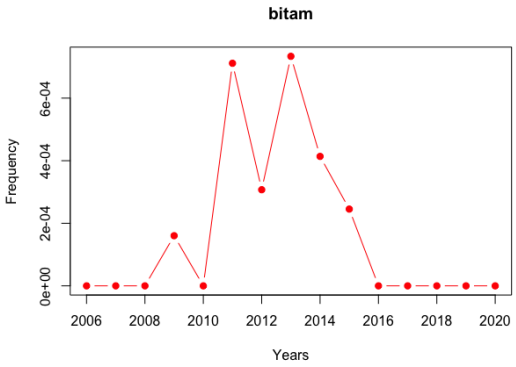


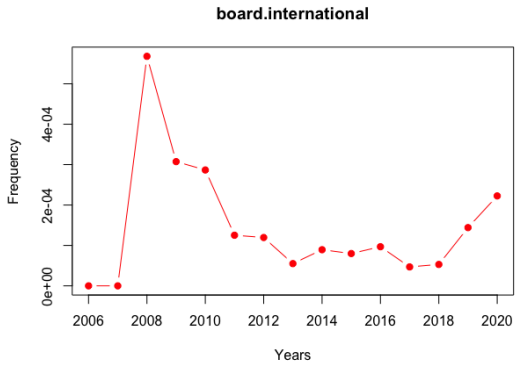


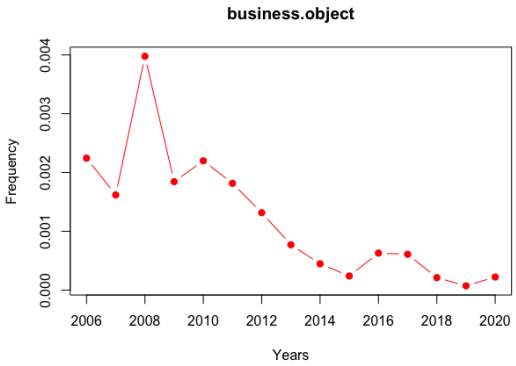


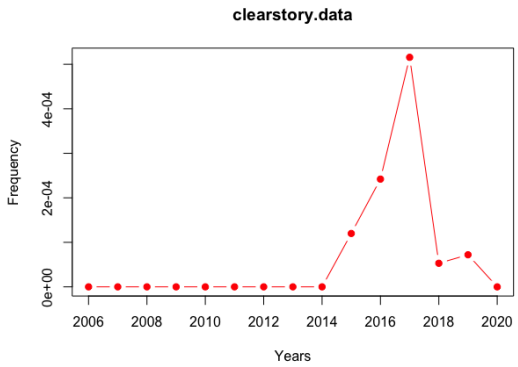


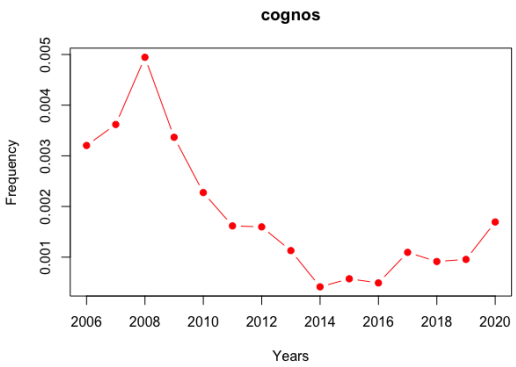

birst
Frequency
0.0000
0.0005
0.0010
0.0015
0.0020
2006
2008
2010
2012
2014
2016
2018
2020
Years

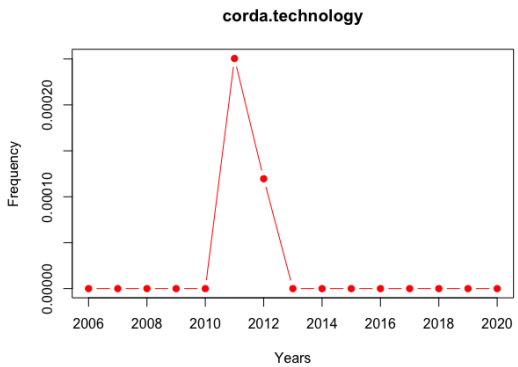

bitam
Frequency
0e+00
2e-04
4e-04
6e-04
2006
2008
2010
2012
2014
2016
2018
2020
Years

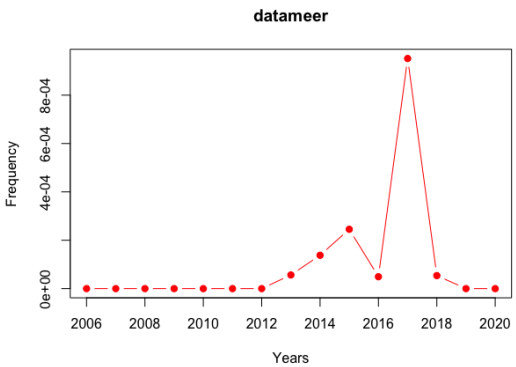

board.international
Frequency
0e+00
2e-04
4e-04
2006
2008
2010
2012
2014
2016
2018
2020
Years

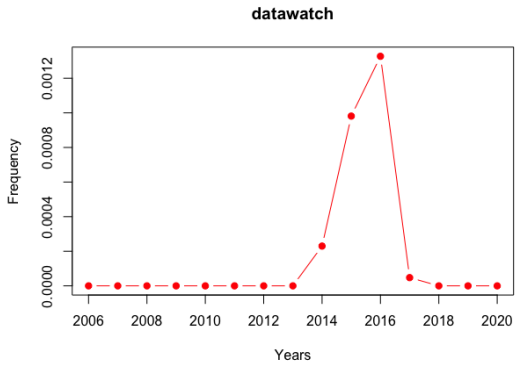

business.object
Frequency
0.000
0.001
0.002
0.003
0.004
2006
2008
2010
2012
2014
2016
2018
2020
Years

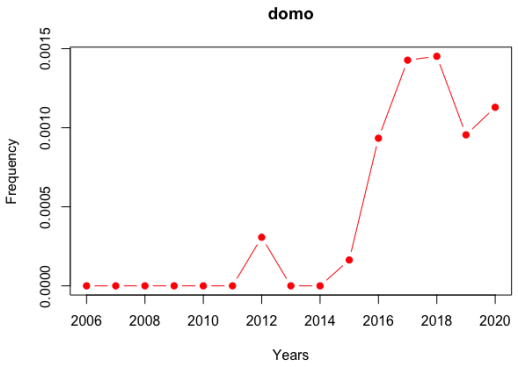

clearstory.data
Frequency
0e+00
2e-04
4e-04
2006
2008
2010
2012
2014
2016
2018
2020
Years

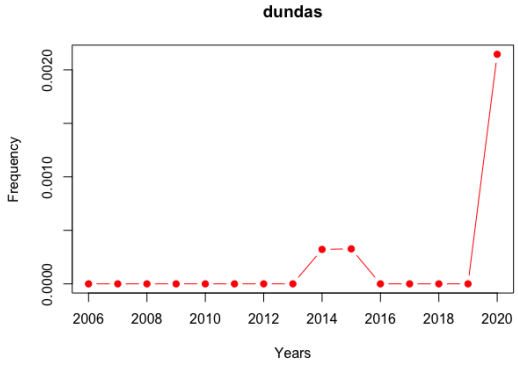

cognos
Frequency
0.001
0.002
0.003
0.004
0.005
2006
2008
2010
2012
2014
2016
2018
2020
Years

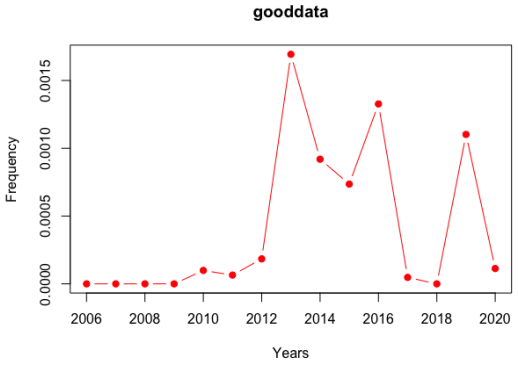

corda.technology
Frequency
0.00000
0.00010
0.00020
2006
2008
2010
2012
2014
2016
2018
2020
Years

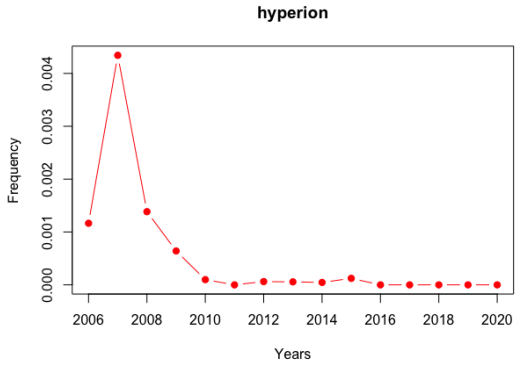

datameer
Frequency
0e+00
4e-04
6e-04
8e-04
2006
2008
2010
2012
2014
2016
2018
2020
Years

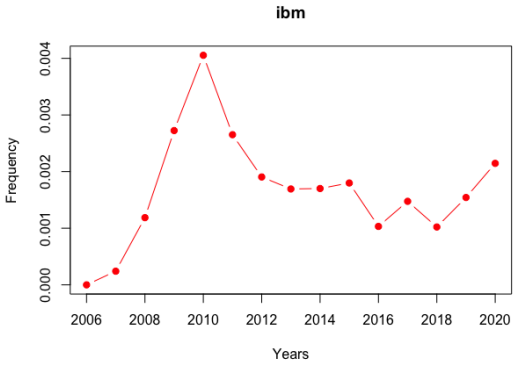

datawatch
Frequency
0.0000
0.0004
0.0008
0.0012
2006
2008
2010
2012
2014
2016
2018
2020
Years

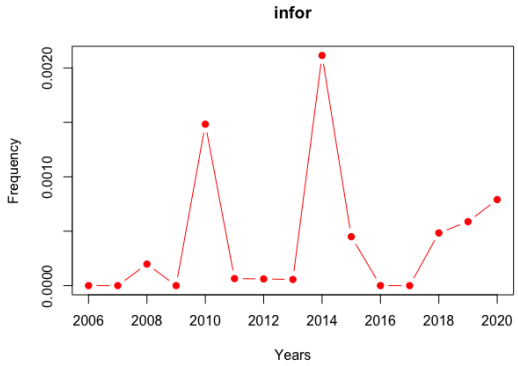

domo
Frequency
0.0000
0.0005
0.0010
0.0015
2006
2008
2010
2012
2014
2016
2018
2020
Years

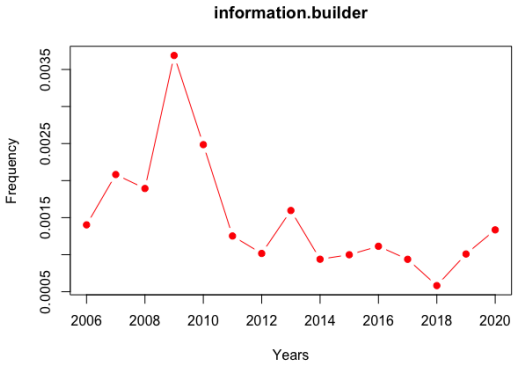

dundas
Frequency
0.0000
0.0010
0.0020
2006
2008
2010
2012
2014
2016
2018
2020
Years

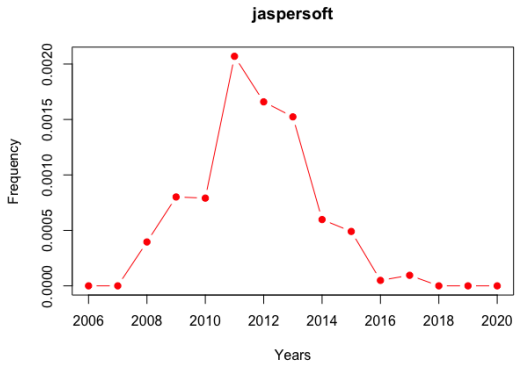

gooddata
Frequency
0.0000
0.0005
0.0010
0.0015
2006
2008
2010
2012
2014
2016
2018
2020
Years

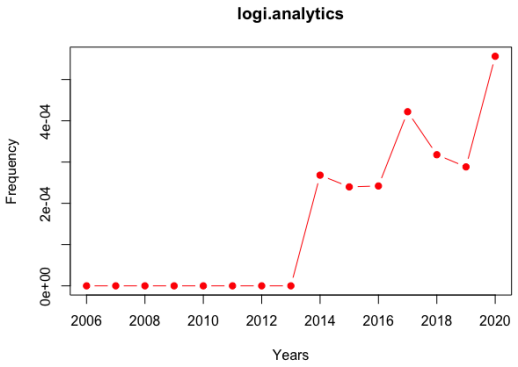

hyperion
Frequency
0.000
0.001
0.002
0.003
0.004
2006
2008
2010
2012
2014
2016
2018
2020
Years

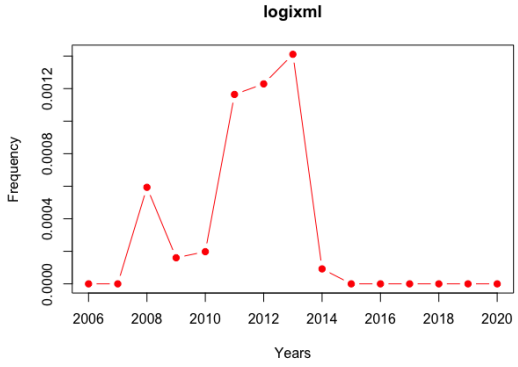

ibm
Frequency
0.000
0.001
0.002
0.003
0.004
2006
2008
2010
2012
2014
2016
2018
2020
Years

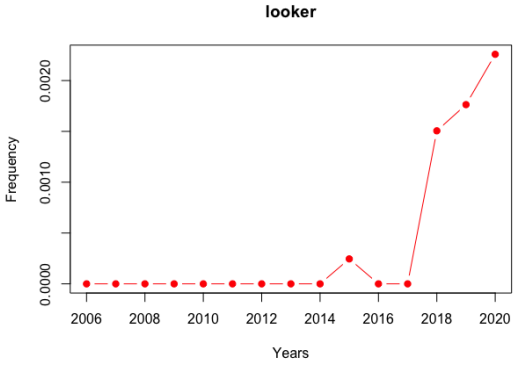

infor
Frequency
0.0000
0.0010
0.0020
2006
2008
2010
2012
2014
2016
2018
2020
Years

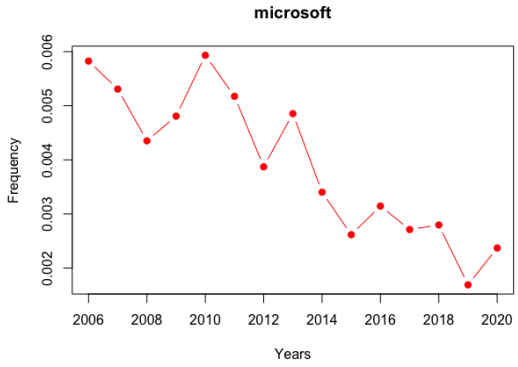

information.builder
Frequency
0.0005
0.0015
0.0025
0.0035
2006
2008
2010
2012
2014
2016
2018
2020
Years

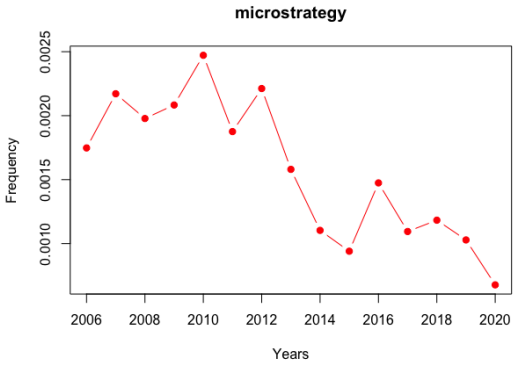

jaspersoft
Frequency
0.0000
0.0005
0.0010
0.0015
0.0020
2006
2008
2010
2012
2014
2016
2018
2020
Years

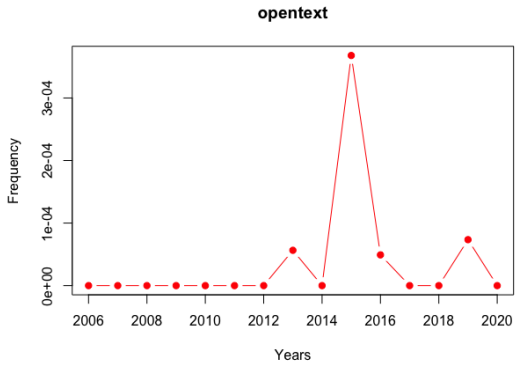

logi.analytics
Frequency
0e+00
2e-04
4e-04
2006
2008
2010
2012
2014
2016
2018
2020
Years

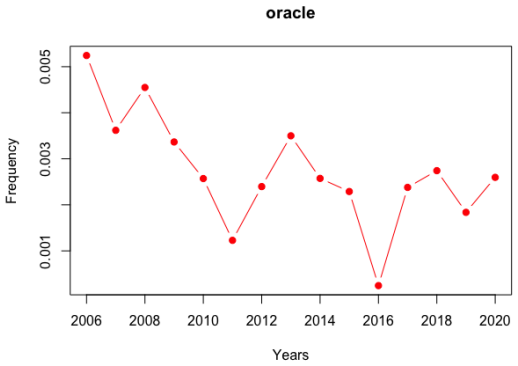

logixml
Frequency
0.0000
0.0004
0.0008
0.0012
2006
2008
2010
2012
2014
2016
2018
2020
Years

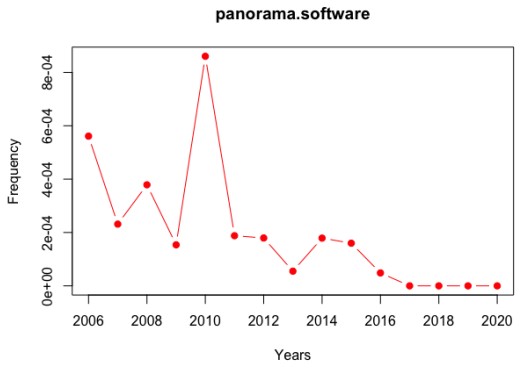

looker
Frequency
0.0000
0.0010
0.0020
2006
2008
2010
2012
2014
2016
2018
2020
Years

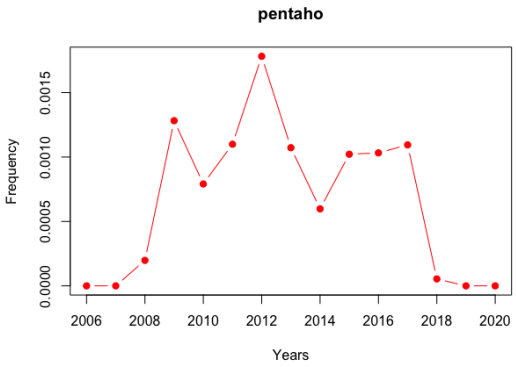

microsoft
Frequency
0.002
0.003
0.004
0.005
0.006
2006
2008
2010
2012
2014
2016
2018
2020
Years

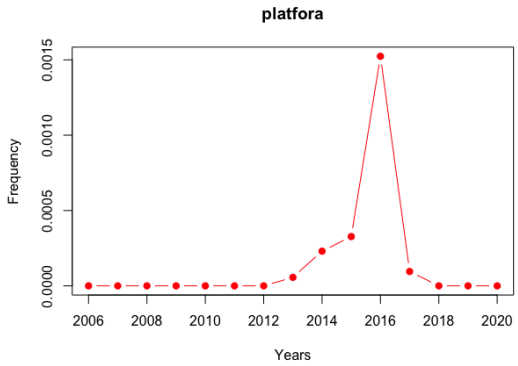

microstrategy
Frequency
0.0010
0.0015
0.0020
0.0025
2006
2008
2010
2012
2014
2016
2018
2020
Years

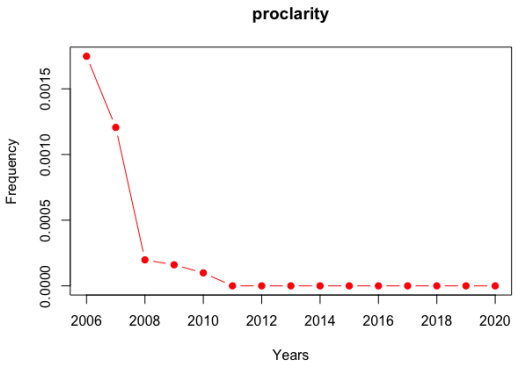

opentext
Frequency
0e+00
1e-04
2e-04
3e-04
2006
2008
2010
2012
2014
2016
2018
2020
Years

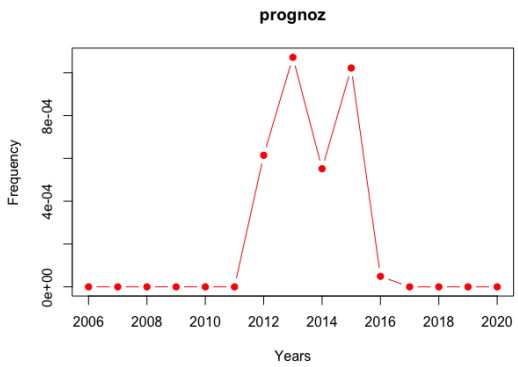

oracle
Frequency
0.001
0.003
0.005
2006
2008
2010
2012
2014
2016
2018
2020
Years

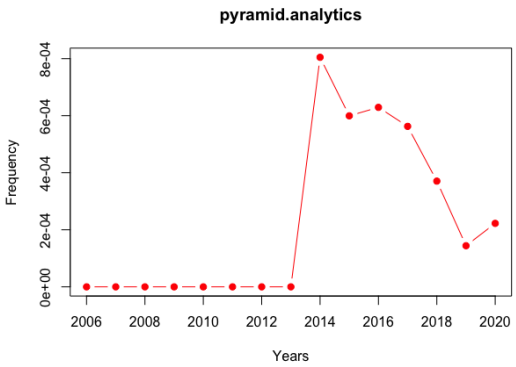

panorama.software
Frequency
0e+00
2e-04
4e-04
6e-04
8e-04
2006
2008
2010
2012
2014
2016
2018
2020
Years

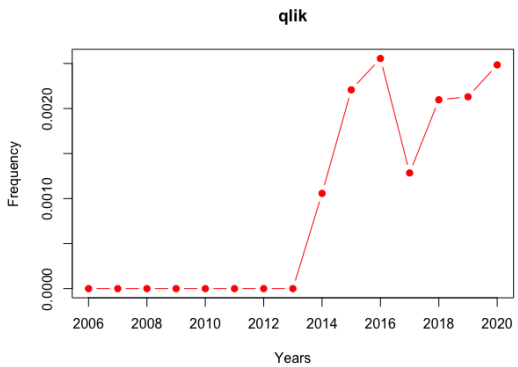

pentaho
Frequency
0.0000
0.0005
0.0010
0.0015
2006
2008
2010
2012
2014
2016
2018
2020
Years

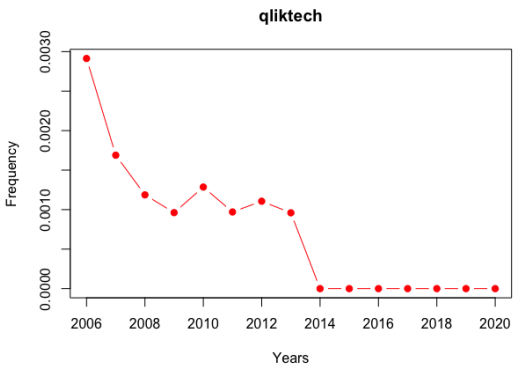

platfora
Frequency
0.0000
0.0005
0.0010
0.0015
2006
2008
2010
2012
2014
2016
2018
2020
Years

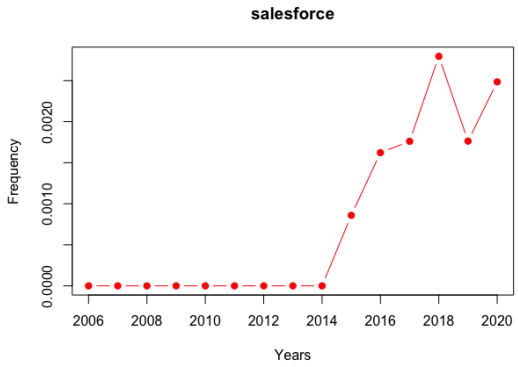

proclarity
Frequency
0.0000
0.0005
0.0010
0.0015
2006
2008
2010
2012
2014
2016
2018
2020
Years

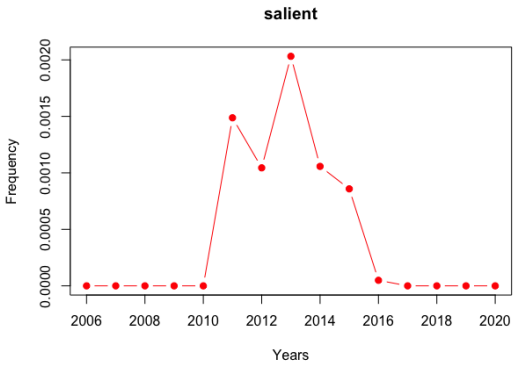

prognoz
Frequency
0e+00
4e-04
8e-04
2006
2008
2010
2012
2014
2016
2018
2020
Years

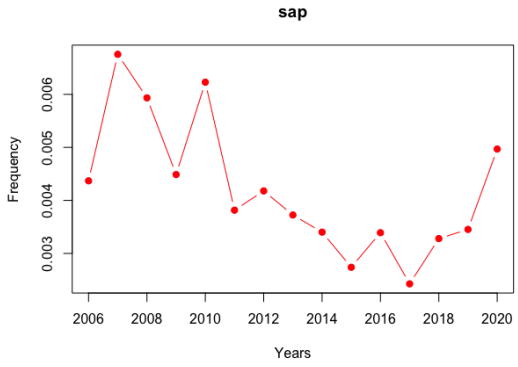

pyramid.analytics
Frequency
0e+00
2e-04
4e-04
6e-04
8e-04
2006
2008
2010
2012
2014
2016
2018
2020
Years

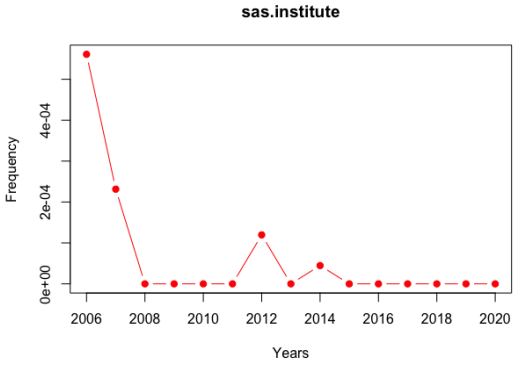

qlik
Frequency
0.0000
0.0010
0.0020
2006
2008
2010
2012
2014
2016
2018
2020
Years

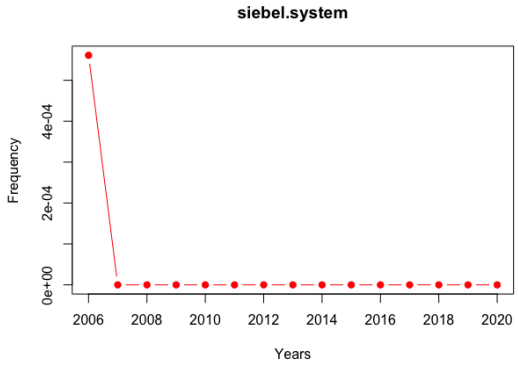

qliktech
Frequency
0.0000
0.0010
0.0020
0.0030
2006
2008
2010
2012
2014
2016
2018
2020
Years

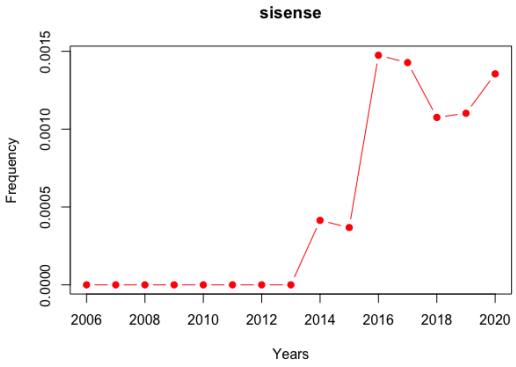

salesforce
Frequency
0.0000
0.0010
0.0020
2006
2008
2010
2012
2014
2016
2018
2020
Years

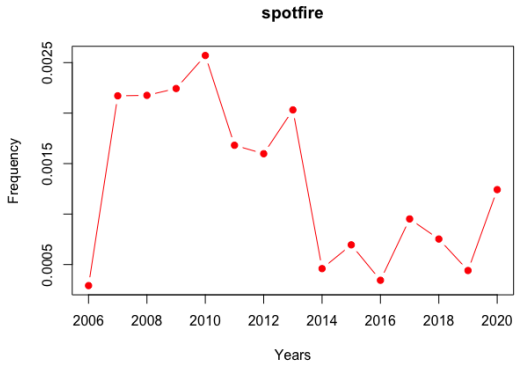

salient
Frequency
0.0000
0.0005
0.0010
0.0015
0.0020
2006
2008
2010
2012
2014
2016
2018
2020
Years


sap
Frequency
0.003
0.004
0.005
0.006
2006
2008
2010
2012
2014
2016
2018
2020
Years


sas.institute
Frequency
0e+00
2e-04
4e-04
2006
2008
2010
2012
2014
2016
2018
2020
Years


siebel.system
Frequency
0e+00
2e-04
4e-04
2006
2008
2010
2012
2014
2016
2018
2020
Years


sisense
Frequency
0.0000
0.0005
0.0010
0.0015
2006
2008
2010
2012
2014
2016
2018
2020
Years


spotfire
Frequency
0.0005
0.0015
0.0025
2006
2008
2010
2012
2014
2016
2018
2020
Years

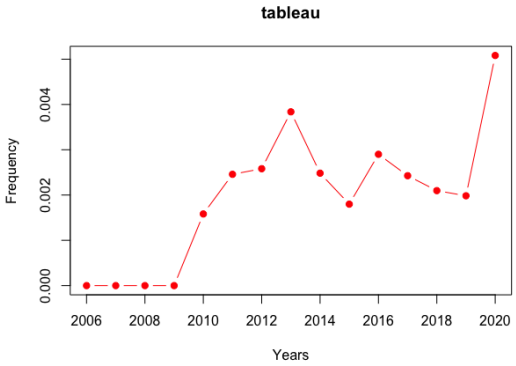
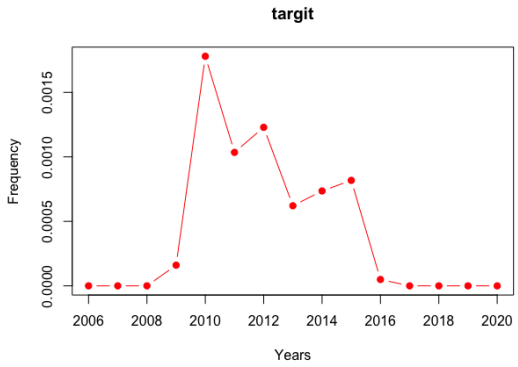
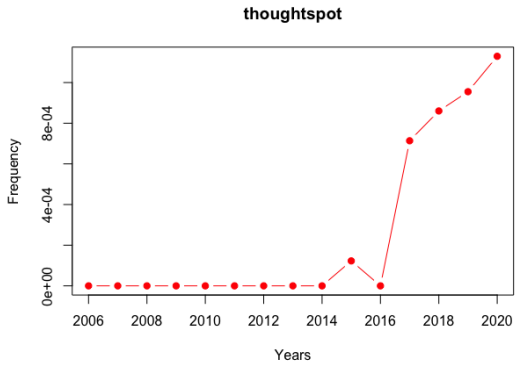
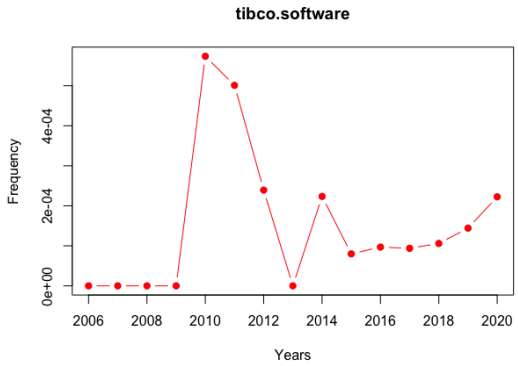
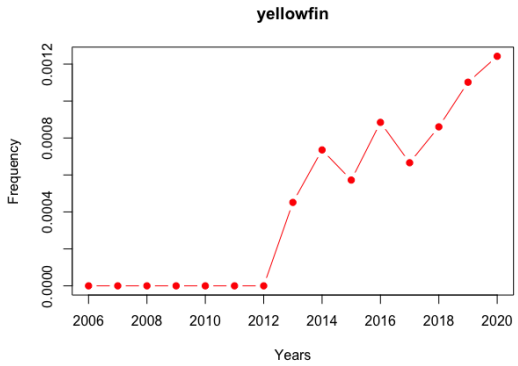


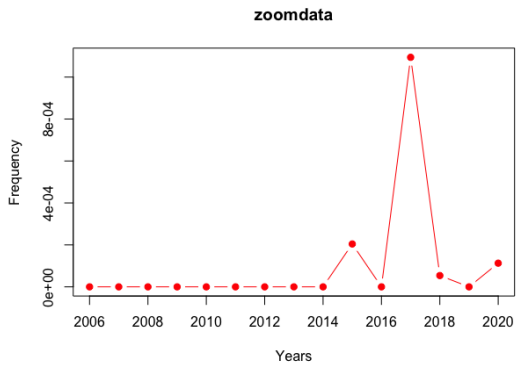
tableau
Frequency
0.000
0.002
0.004
2006
2008
2010
2012
2014
2016
2018
2020
Years


targit
Frequency
0.0000
0.0005
0.0010
0.0015
2006
2008
2010
2012
2014
2016
2018
2020
Years


thoughtspot
Frequency
0e+00
4e-04
8e-04
2006
2008
2010
2012
2014
2016
2018
2020
Years

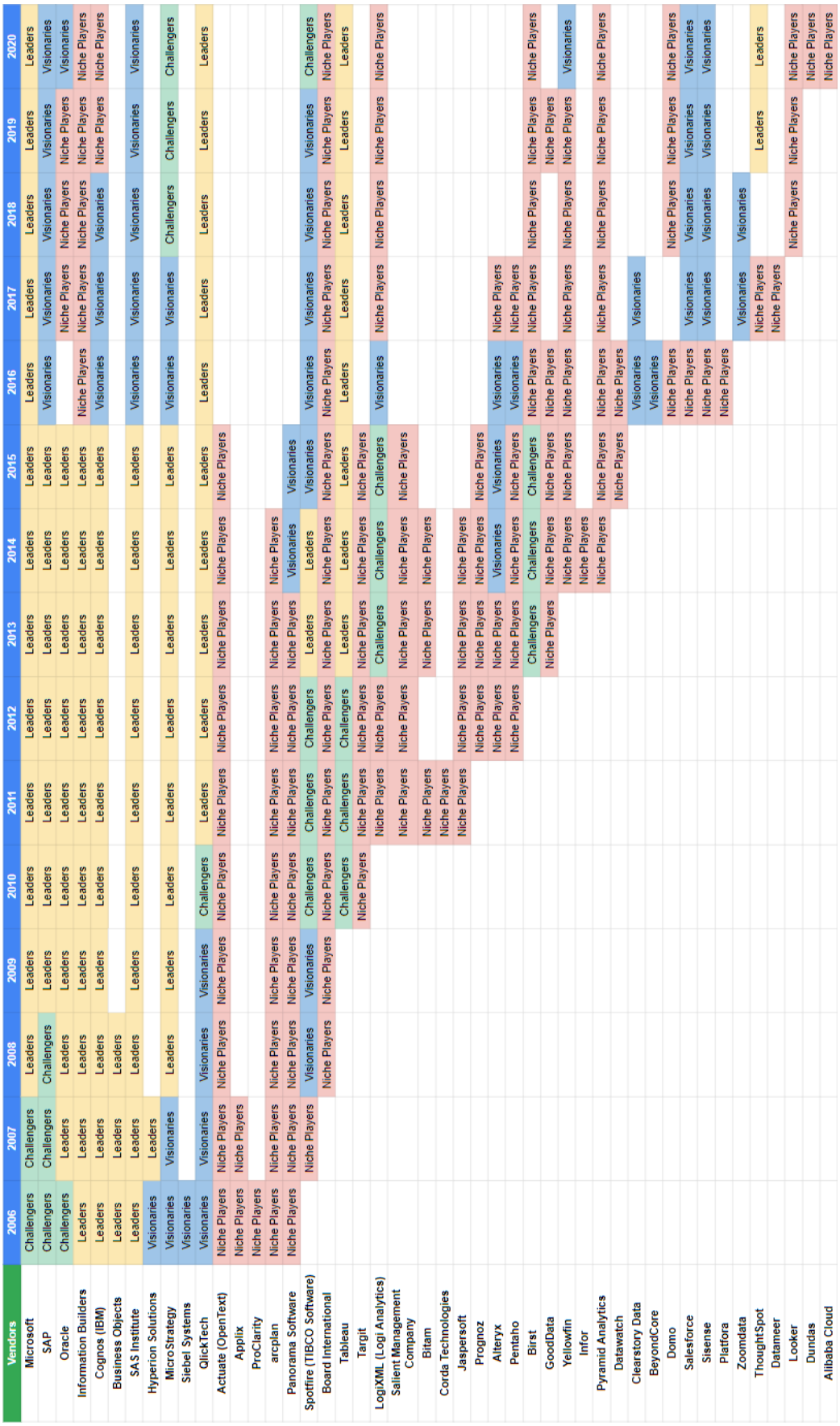

tibco.software
Frequency
0e+00
2e-04
4e-04
2006
2008
2010
2012
2014
2016
2018
2020
Years


yellowfin
Frequency
0.0000
0.0004
0.0008
0.0012
2006
2008
2010
2012
2014
2016
2018
2020
Years


zoomdata
Frequency
0e+00
4e-04
8e-04
2006
2008
2010
2012
2014
2016
2018
2020
Years

Appendix B

Evolution of Provider's Category in the Magic Quadrants



Vendors	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Microsoft	Challengers	Challengers	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders
SAP	Challengers	Challengers	Challengers	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Visionaries	Visionaries	Visionaries	Visionaries	Visionaries
Oracle	Challengers	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders		Niche Players	Niche Players	Niche Players	Visionaries
Information Builders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players
Cognos (IBM)	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Visionaries	Visionaries	Visionaries	Niche Players	Niche Players
Business Objects	Leaders	Leaders	Leaders												
SAS Institute	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Visionaries	Visionaries	Visionaries	Visionaries	Visionaries
Hyperion Solutions	Visionaries	Leaders													
MicroStrategy	Visionaries	Visionaries	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Visionaries	Visionaries	Challengers	Challengers	Challengers
Siebel Systems	Visionaries														
QlickTech	Visionaries	Visionaries	Visionaries	Visionaries	Challengers	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders
Actuate (OpenText)	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players					
Applix	Niche Players	Niche Players													
ProClarity	Niche Players														
arcplan	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players						
Panorama Software	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Visionaries	Visionaries					
Spotfire (TIBCO Software)		Niche Players	Visionaries	Visionaries	Challengers	Challengers	Challengers	Leaders	Leaders	Visionaries	Visionaries	Visionaries	Visionaries	Visionaries	Challengers
Board International			Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players
Tableau					Challengers	Challengers	Challengers	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders	Leaders
Targit					Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players					
LogiXML (Logi Analytics)						Niche Players	Niche Players	Challengers	Challengers	Challengers	Visionaries	Niche Players	Niche Players	Niche Players	Niche Players
Salient Management Company						Niche Players	Niche Players	Niche Players	Niche Players	Niche Players					
Bitam						Niche Players		Niche Players	Niche Players						
Corda Technologies						Niche Players									
Jaspersoft						Niche Players	Niche Players	Niche Players	Niche Players	Niche Players					
Prognoz							Niche Players	Niche Players	Niche Players	Niche Players					
Alteryx							Niche Players	Niche Players	Visionaries	Visionaries	Visionaries	Niche Players			
Pentaho							Niche Players	Niche Players	Niche Players	Niche Players	Visionaries	Niche Players			
Birst								Challengers	Challengers	Challengers	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players
GoodData								Niche Players	Niche Players	Niche Players	Niche Players		Niche Players	Niche Players	
Yellowfin									Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Visionaries
Infor									Niche Players						
Pyramid Analytics									Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players	Niche Players
Datawatch										Niche Players	Niche Players				
Clearstory Data											Visionaries	Visionaries			
BeyondCore											Visionaries				
Domo											Niche Players		Niche Players	Niche Players	Niche Players
Salesforce											Niche Players	Visionaries	Visionaries	Visionaries	Visionaries
Sisense											Niche Players	Visionaries	Visionaries	Visionaries	Visionaries
Platfora											Niche Players				
Zoomdata												Visionaries	Visionaries		
ThoughtSpot												Niche Players		Leaders	Leaders
Datameer												Niche Players			
Looker													Niche Players	Niche Players	Niche Players
Dundas															Niche Players
Alibaba Cloud															Niche Players

Figure B.1: Evolution of Provider's Category in the Magic Quadrants

Appendix C

Data Table of Gartner Documents

	id	word	n	total	tf	idf	tf_idf
1	gartner-2015.txt	data	777	24999	0.031081243	0.00000000	0.0000000000
2	gartner-2017.txt	data	597	21317	0.028005817	0.00000000	0.0000000000
3	gartner-2014.txt	data	543	22340	0.024306177	0.00000000	0.0000000000
4	gartner-2016.txt	data	490	20648	0.023731112	0.00000000	0.0000000000
5	gartner-2014.txt	bi	450	22340	0.020143241	0.00000000	0.0000000000
6	gartner-2015.txt	bi	415	24999	0.016600664	0.00000000	0.0000000000
7	gartner-2014.txt	customer	410	22340	0.018352731	0.00000000	0.0000000000
8	gartner-2018.txt	customer	402	18873	0.021300270	0.00000000	0.0000000000
9	gartner-2018.txt	data	402	18873	0.021300270	0.00000000	0.0000000000
10	gartner-2015.txt	customer	398	24999	0.015920637	0.00000000	0.0000000000
11	gartner-2011.txt	bi	386	15953	0.024196076	0.00000000	0.0000000000
12	gartner-2015.txt	platform	384	24999	0.015360614	0.00000000	0.0000000000
13	gartner-2017.txt	customer	375	21317	0.017591594	0.00000000	0.0000000000
14	gartner-2018.txt	analytics	351	18873	0.018597997	0.00000000	0.0000000000
15	gartner-2015.txt	user	350	24999	0.014000560	0.00000000	0.0000000000
16	gartner-2019.txt	data	349	13876	0.025151340	0.00000000	0.0000000000
17	gartner-2012.txt	bi	344	16701	0.020597569	0.00000000	0.0000000000
18	gartner-2013.txt	customer	335	18141	0.018466457	0.00000000	0.0000000000
19	gartner-2015.txt	use	329	24999	0.013160526	0.00000000	0.0000000000
20	gartner-2019.txt	analytics	320	13876	0.023061401	0.00000000	0.0000000000
21	gartner-2014.txt	user	318	22340	0.014234557	0.00000000	0.0000000000
22	gartner-2013.txt	data	314	18141	0.017308858	0.00000000	0.0000000000
23	gartner-2012.txt	customer	309	16701	0.018501886	0.00000000	0.0000000000
24	gartner-2013.txt	bi	307	18141	0.016922992	0.00000000	0.0000000000
25	gartner-2014.txt	use	307	22340	0.013742167	0.00000000	0.0000000000
26	gartner-2011.txt	customer	298	15953	0.018679872	0.00000000	0.0000000000
27	gartner-2015.txt	analytics	298	24999	0.011920477	0.00000000	0.0000000000
28	gartner-2014.txt	platform	296	22340	0.013249776	0.00000000	0.0000000000
29	gartner-2017.txt	use	289	21317	0.013557255	0.00000000	0.0000000000
30	gartner-2014.txt	capability	276	22340	0.012354521	0.00000000	0.0000000000
31	gartner-2010.txt	bi	275	10454	0.026305720	0.00000000	0.0000000000
32	gartner-2017.txt	analytics	275	21317	0.012900502	0.00000000	0.0000000000
33	gartner-2014.txt	product	273	22340	0.012220233	0.00000000	0.0000000000
34	gartner-2015.txt	capability	270	24999	0.010800432	0.00000000	0.0000000000
35	gartner-2011.txt	vendor	268	15953	0.016799348	0.00000000	0.0000000000
36	gartner-2015.txt	business	265	24999	0.010600424	0.00000000	0.0000000000
37	gartner-2014.txt	vendor	264	22340	0.011817368	0.00000000	0.0000000000
38	gartner-2019.txt	customer	261	13876	0.018809455	0.00000000	0.0000000000

Figure C.1: Sample of Data from the Gartner Documents

Appendix D

Data Table of Gartner Documents sorted by TF-IDF

	id	word	n	total	tf	idf	tf_idf
1	gartner-2020.txt	abi	73	8984	0.0081255565	2.70805020	0.0220044150
2	gartner-2020.txt	alibaba	13	8984	0.0014470169	2.70805020	0.0039185945
3	gartner-2020.txt	dundas	19	8984	0.0021148709	1.60943791	0.0034037534
4	gartner-2020.txt	ml	14	8984	0.0015583259	2.01490302	0.0031398756
5	gartner-2010.txt	yearmagic	12	10454	0.0011478860	2.70805020	0.0031085328
6	gartner-2020.txt	einstein	21	8984	0.0023374889	1.32175584	0.0030895896
7	gartner-2019.txt	ml	21	13876	0.0015134044	2.01490302	0.0030493632
8	gartner-2006.txt	ebis	4	3562	0.0011229646	2.70805020	0.0030410446
9	gartner-2006.txt	siebels	4	3562	0.0011229646	2.70805020	0.0030410446
10	gartner-2020.txt	looker	20	8984	0.0022261799	1.32175584	0.0029424663
11	gartner-2019.txt	reviewer	20	13876	0.0014413376	2.01490302	0.0029041554
12	gartner-2017.txt	modern	81	21317	0.0037997842	0.76214005	0.0028959677
13	gartner-2017.txt	quartile	96	21317	0.0045034480	0.62860866	0.0028309064
14	gartner-2019.txt	modern	48	13876	0.0034592101	0.76214005	0.0026364026
15	gartner-2018.txt	roadmap	45	18873	0.0023843586	1.09861229	0.0026194857
16	gartner-2006.txt	applix	7	3562	0.0019651881	1.32175584	0.0025974988
17	gartner-2020.txt	augment	57	8984	0.0063446126	0.40546511	0.0025725191
18	gartner-2018.txt	quartile	76	18873	0.0040269168	0.62860866	0.0025313547
19	gartner-2018.txt	salesforce	52	18873	0.0027552588	0.91629073	0.0025246181
20	gartner-2007.txt	applixs	4	4319	0.0009261403	2.70805020	0.0025080345
21	gartner-2009.txt	mq	8	6507	0.0012294452	2.01490302	0.0024772129
22	gartner-2007.txt	applix	8	4319	0.0018522806	1.32175584	0.0024482627
23	gartner-2018.txt	modern	60	18873	0.0031791448	0.76214005	0.0024229536
24	gartner-2016.txt	quartile	76	20648	0.0036807439	0.62860866	0.0023137475
25	gartner-2019.txt	looker	24	13876	0.0017296051	1.32175584	0.0022861156
26	gartner-2020.txt	salesforce	22	8984	0.0024487979	0.91629073	0.0022438108
27	gartner-2011.txt	corda	22	15953	0.0013790510	1.60943791	0.0022194969
28	gartner-2019.txt	roadmap	28	13876	0.0020178726	1.09861229	0.0022168596
29	gartner-2018.txt	einstein	31	18873	0.0016425582	1.32175584	0.0021710608
30	gartner-2007.txt	hyperion	18	4319	0.0041676314	0.51082562	0.0021289329
31	gartner-2016.txt	modern	57	20648	0.0027605579	0.76214005	0.0021039318
32	gartner-2020.txt	cloud	84	8984	0.0093499555	0.22314355	0.0020863823
33	gartner-2016.txt	preparation	56	20648	0.0027121271	0.76214005	0.0020670207
34	gartner-2018.txt	looker	28	18873	0.0014836009	1.32175584	0.0019609582
35	gartner-2016.txt	qlik	52	20648	0.0025184037	0.76214005	0.0019193763
36	gartner-2019.txt	augment	64	13876	0.0046122802	0.40546511	0.0018701187
37	gartner-2020.txt	preparation	22	8984	0.0024487979	0.76214005	0.0018663269
38	gartner-2020.txt	qlik	22	8984	0.0024487979	0.76214005	0.0018663269

Figure D.1: Sample of Data from the Gartner Documents sorted by TF-IDF value

Appendix E

R Packages used in the project

Package name	Utility for the project
tidyverse	Include different packages useful for data analysis in R such as ggplot2, tidyr and dplyr
tm	Handle all text mining related activities in R like managing corpus of documents, importing data or preprocessing the documents
dplyr	Make it easier to handle data
tidyr	Transform data into a format easier to handle for text mining related analysis
tidytext	Provide functions helpful for text mining related analysis
ggplot2	Create graphical visualizations of the data
readtext	Import text files in various formats into R
SnowballC	Include functionalities to stem the words into their root form to make the analysis easier to interpret
slam	Handle matrices data structures and functions to manage them
qdap	Provide useful function to go from qualitative data to quantitative analysis
topicmodels	Provide functionalities to do Topic Modeling using the LDA algorithm

Table E.1: R Packages used in the project