NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

Exploring patterns and trends of master dissertations

A text mining application

Bruno Filipe Brito Guerreiro

Project Work proposal presented as partial requirement for obtaining the Master's degree in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Exploring patterns and trends of master dissertations

A text mining application

Bruno Filipe Brito Guerreiro

Project Work proposal presented as partial requirement for obtaining the Master's degree in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Exploring patterns and trends of master dissertations

A text mining application

Bruno Filipe Brito Guerreiro

MEGI

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

EXPLORING PATTERNS AND TRENDS OF MASTER DISSERTATIONS
A TEXT MINING APPLICATION

by

Bruno Guerreiro

Project Work as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Information Analysis and Management

Advisor: Professor Doutor Roberto Henriques

November 2018

ABSTRACT

Data mining is a technique that is used to discover trends and patterns in a dataset (Feldman et al., 1998). However, the majority of datasets are unstructured and contain textual data (Sulova & Nacheva, 2017). Therefore, in the last decade, different text mining techniques are developed to gain competitive advantage by extracting information from these datasets. However, text mining techniques in the educational context is still scarce (Nie & Sun, 2017). Therefore, this study applied text mining techniques on master dissertations to explore trends and patterns of these studies over the years. Dissertations that were published at NOVA IMS from March 2004 until May 2018 were collected from its repository. Thereafter, different techniques were applied to pre-process the data. More specific, unevaluable characters and stop-words were removed from the dataset and stemming was applied to reduce noise and the amount of unique words. Thereafter, two clustering techniques were applied. First, Topic modelling was executed with an LDA algorithm which suggested four topics. After inferring the topics, the following definitions were indicated: Geodata Information, Behavioural Studies, Information and Decision Systems, and Implementing Systems. Thereafter, clustering k-means was executed which resulted in three clusters. The clusters indicated Geodata Analysis, Online Behaviour, and Business Analysis. Furthermore, trends and patterns were analysed for each study area (i.e. Advanced Analytics, Statistics and Information Management, Information Management, Geospatial Technologies and Geographic Information Systems and Science). The results of this study create opportunities for NOVA IMS among others to reallocate resources based on these trends (Sullivan, 2001) and to allocate financial resources based on current patterns (Simoudis, 1996) and to encourage students to examine new interesting trends (Feldman et al., 1998).

KEYWORDS

Dissertations; Text Mining; Knowledge; Trends; Patterns; Cluster; Topic Modelling;

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AA	Advanced Analytics
DM	Data mining
GISS	Geographic Information Systems and Science
GT	Geospatial Technologies
IDF	Inverse Document Frequency
IM	Information Management
LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
NOVA IMS	NOVA Information Management School
PC	Principal Component
SIM	Statistics and Information Management
TC matrix	Term Count Matrix
TF	Term Frequency
TF-IDF	Term Frequency - Inverse Document Frequency
TF matrix	Term Frequency Matrix
TM	Text mining

1. INTRODUCTION

1.1 Context

Nowadays, the aggressive and continuous growth of data volume is increasing, characterized by speed and ease of access (Sulova & Nacheva, 2017). Moreover, the majority of the documents are in unstructured text formats, such as articles, webpages, comments, and books (Sulova & Nacheva, 2017) (Talib, Hanif, Ayesha, & Fatima, 2016). These documents contain valuable information that, for many organizations, can lead to significant opportunities (Padhy, Mishra, & Panigrahi, 2012). Hence, knowledge and information extracted from these text files are key factors to survive in highly competitive markets (Padhy et al., 2012).

New technologies and information systems enable organisations to extract knowledge from those large amounts of data since traditional methodologies cannot keep up with this aggressive data evolution due to the technical and resource scarcity (time, human and monetary) (Talib et al., 2016). Therefore, in the pursuit of competitive advantage, a technological and methodological expansion occurred into a great focus in the textual analysis (Simoudis, 1996) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). The aim textual analysis is to identify and extract relevant information from unstructured data, with the purpose of organizing and interpreting them (Feldman et al., 1998) (Fan, Wallace, Rich, & Zhang, 2006). Therefore, techniques for identifying textual relations and patterns within and between documents are essential (KMWorld Magazine and SAS, 2013).

1.2 Problem Definition

Contrary to numerical data, texts have always had barriers to analyse (Fayyad et al., 1996). The difficulty of its exploration stems from the inability to apply the traditional methods of analysis and statistics to textual data due to the absence of a generic structure (Fayyad et al., 1996). In text analysis there is an intrinsic difficulty in identifying and extracting the key elements due to the consequence of the high levels of noise; large number of unevaluable information and irrelevant features (Tang, Alelyani, & Liu, 2014).

Text mining is an optimized process that allows to carry out all techniques to extract knowledge of unstructured data (Feldman et al., 1998). It starts by extracting and exploring data and then transforming this data into information, with the aim of examining patterns, relations, opportunities, threats, trends, etc. (Feldman & Sanger, 2006).

In recent years, a growing interest in data mining in the educational context (Feldman et al., 1998). Nationally, there are thousands of master dissertations published annually in the universities' repositories, however, evidence of any analysis about their evolution, patterns and trends is still lacking (Nie & Sun, 2017). The way data is nowadays presented requires a large investment of time and resources in order to gain knowledge (Feldman & Sanger, 2006). Currently, the very exposure of these contents is limited to the title, author and area of the master's degree, which makes research difficult for anyone interested in the subject. Hence, much information that create various opportunities for universities (students, teachers and managers) is undiscovered and unexplored which diminishes the advantages for all of them (Feldman et al., 1998).

1.3 Scope

This study is conducted together in collaboration with NOVA Information Management school (NOVA IMS). The university was founded in 1989 and is located in the city centre of Lisbon. The main objective of the university is to grow in projects in the area of information management by applying and developing new information technologies (NOVA IMS, 2018). Every year, students finish their Master program with either a dissertation, project or internship. Currently, NOVA IMS contains a repository with a large quantity of dissertations published by their students. Until today, knowledge about the informational deliveries of the students is lacking. To extract more information about the knowledge the students of this university reproduced, text mining techniques are applied on their dissertations.

1.4 Research Objectives

The aim of the study is to examine current patterns and trends in the dissertations published by NOVA IMS students. In order to achieve the main objective proposed, an intense and structured exploratory analysis on the dissertations is conducted. Therefore, a deep examination on how text mining could be applied in the educational context was required. Furthermore, outcomes on the text mining on the dissertations had to be analysed to gain insights about patterns and possible trends.

1.5 Research Question

The following research question was formulated to reach the objectives of this study: *What are the current patterns and trends in the dissertations published by NOVA-IMS students?*

To answer this research question, the following research sub-questions were formulated;

- How can text mining be applied in the educational context?

- What are the outcomes from applying text mining on the dissertations of students in NOVA IMS?

1.6 Study Relevance and Importance

Currently, in the present aggressiveness and competitiveness on any market, it is crucial to produce information and knowledge that translates data into information that create a competitive advantage. This advantage can provide any organization with a better quality of service, identification of opportunities, anticipation in the market, reduction of costs, and better allocation of resources (Sullivan, 2001) (Simoudis, 1996).

In particular, obtaining information of the evidenced objects of study and later identification of patterns and tendencies is justified by:

- Anticipation and investment in content covered by identified trend (Simoudis, 1996);

- Improved resource management - reallocation based on identified needs (Sullivan, 2001) (Simoudis, 1996);

- Improved availability of content (Feldman et al., 1998);

- Encouraged and created conditions for students to explore areas of influence (Feldman et al., 1998);

Therefore, this study pretends to encourage the analysis of the published dissertations that are enriched with hidden knowledge that create value to the university by anticipating on trends that past behaviours warn (Feldman et al., 1998). Hence, the quality of teaching is enhanced, and the college stands out in the teaching market and research field.

1.7 Structure of the Thesis

Chapter 2 describes the literature review which begins with a brief history and the concept of text mining. Thereafter, the process of text mining is explained gradually. The chapter ends with text mining in the educational context.

Chapter 3 is dedicated to the methodology that is used to meet the objectives of this study. First it presents a research design and thereafter it explains how data is collected and how pre-processing the data is executed. Furthermore, this chapter explains the process of Topic Modelling and K-means clustering. The chapter explains how to infer the clusters and topics and how to find trends and patterns.

Chapter 4 shows the results of the text mining techniques over the dissertations. First the data pre-processing part is displayed. Thereafter the topics from Topic Modelling and K-means clustering are presented and inferred. Finally, the trends and patterns are visualized.

Chapter 5 is dedicated to the discussion that aims to answer the research questions. Chapter 6 draws the conclusion of this study.

Thereafter, the Limitations and Recommendations for Future Works, Bibliography and Appendix are respectively presented.

2. LITERATURE REVIEW

As referred in the introduction, the purpose of this study was to explore the general knowledge and discover trends in the publications of students at NOVA IMS. To accomplish this objective, text mining techniques were applied on master dissertations published in previous years. This chapter will clarify the concept of text mining where the process is presented and explained gradually. Furthermore, it will expand on how text mining has been applied in various markets and organizations and which best practices can be applied on this study. Moreover, it will focus on other studies that applied text mining in the educational context.

2.1 A Brief History of Text Mining

Text data mining emerged at the end of the last century. Before that, text mining existed of counting the words and subsequently label the files by topics. The analyst was not able to add semantics, or to understand the meaning, to the text files.

Since 2000, new techniques were developed in the field of text mining (Miner et al., 2012). Data mining, statistical analysis and statistical learning were merged in new text mining techniques. This, resulted in the use of different techniques, such as tokenization, lexical analysis and semantic analysis (Miner et al., 2012). Furthermore, text mining took a step further in the analytical process, namely, it enabled to explore relations and complex patterns in the many unstructured text files (Miner et al., 2012).

From the last two decades data mining and knowledge discovery applications have got a rich focus due to its significance in decision making and it has become an essential component in various organizations (Reddy & Venkatadri, 2011). The field of data mining have been prospered and posed into new areas of human life with various integrations and advancements in the fields of Statistics, Databases, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation capabilities (Reddy & Venkatadri, 2011).

Followed by the technological and scientific evolution, data mining started not only to be applied to numerical and structured data but also extended to semi-structured and unstructured data. This change meant a significant evolution on this science due to the quality and quantity of hidden knowledge existent on those data sources. For instance, Hearst (1999) suggested that unstructured data express a wide amount of valuable information.

As consequence of this change, the applications fields of knowledge discovery techniques expanded significantly. Fayyad, Piatetsky-Shapiro and Smyth (1996) stated that previously the field business was the main application of text mining due to data characteristics whereas in the end of last century it extended to marketing, finance, investment, fraud detection, manufacturing, telecommunications, astronomy, education and others.

For instance, in the field of government and security, text mining is used to detect links between people and criminal organizations through digital media (suspected web sites, blogs, emails, chat lines, instant messages) (Zanasi, 2009).

On business and marketing applications, Coussement and Van den Poel (2008) applied tekst mining techniques to improve predictive analytics models for customer churn. Furthermore, Gálvez and Gravano (2017) built and validated a series of predictive models using state-of-the-art machine learning and topic discovery techniques in stock market prediction. Sentimental analysis in social media have been applied increasingly since businesses are interested in online behaviour, such as peoples likes and dislikes (Pang & Lee, 2009).

On the current days, daily user of internet face text mining algorithms without even noticing. A simple search on Google is a typical example. Moreover, the latest text mining feature introduced by google is the 'Gmail - Smart Compose' that tries to understand what is being typed in the corpus of an e-mail by the user so that artificial intelligence can suggest words and phrases to finish the user sentences (Google, 2018).

2.2 Text Mining Process

The definition of text mining is not yet clearly established, the authors divergences refer to the range of coverage of the process. Feldman and Sanger (2006) stated that text mining is an exhaustive process where the analyst applies various tools and techniques on text documents with the aim to extract knowledge. Furthermore, the core intention of text mining is defined as a process to produce information through processing textual data. To achieve that, a complex knowledge process must be executed (Feldman & Sanger, 2006). This process begins with the automation of data extraction routines and thereafter all the conditions have to be met to initialize the discovery of patterns among the subject under study.

More in detail, Feldman and Sanger (2006) argue that text mining systems rely on algorithmic and heuristic approaches to consider distributions, frequent sets, and various associations of concepts at an inter-document level in an effort to enable a user to discover the nature and relationships of concepts as reflected in the collection as a whole.

Certainly, data mining and text mining systems rely on a parallel architectural process. Computer scientists and information system specialists concentrated on the discovery of knowledge from structured, numerical databases and data warehouses. However, a large amount of available business data is captured in text files that are not overtly structured which causes a major challenge in extracting knowledge (Kroeze, Matthee, & Bothma, 2003).

Knowledge Discovery in Databases (KDD)

According to Fayyad, Piatetsky-Shapiro and Smyth (1996) Knowledge Discovery in Databases denotes to the overall process of discovering useful knowledge from data. The KDD process organizes the database content to apply data mining methods. Data mining methods consist of the application of data analysis combined with discovery algorithms. Those algorithms require that the selected data is pre-processed, subsampled and transformed (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

The authors explain that the KDD process begins with clarifying the application domain, with the aim to retain the known knowledge and establish the pretended goal. Thereafter, the target data set is defined, it requires to be cleaned and pre-processed. This operation includes noise removing, handling missing data, and accounting for time sequence information. Subsequently, dimensionality reduction and transformation methods are executed on the selected data, taking into consideration the established goal.

After executing the referred tasks, a data-mining method must be cautiously chosen to satisfy the necessities implied by the defined goal. Broadly the methods used are summarization, classification, regression and clustering. Afterwards, the exploratory analysis will be conducted together with the selection and parameterization of the model. This will initialize the exploration of patterns of interest. The mined patterns must be interpreted by consulting the visualization of the data given the extracted models. Moreover, all the conditions are met to report the extracted knowledge.

2.3 Text Mining System Architecture

The text mining typical architecture can be described by three sequential main processes; (1) the data collection, (2) the pre-processing and (3) the mining processing. In order to enrich value, a post-processing can be used to filter redundant information and cluster closely related (Feldman & Sanger, 2006).

2.3.1. Data Collection

The first step in any text mining research project is to collect the textual data for the analysis (Feldman & Sanger, 2006). The data can be collected from different types of formats, for example, documents, webpages, user comments, reviews, books. When parts of documents are examined, it is called document features (Feldman & Sanger, 2006). This indicates that a part of the document could be representative of the entire document. The use of the entire available information on each document usually produces ambiguities and it becomes impracticable due the huge volumes of data contained in text files (Feldman & Sanger, 2006).

In this research study, the document features are the abstracts from the documents under study. The collections are considered semi-structured documents because all of them respect specific layout rules.

2.3.2. Pre-Processing

At this stage, the collected documents will be converted in a format that suits the succeeding core mining process (Feldman & Sanger, 2006). It contains all the necessary procedures to progressively enrich the structure of the documents until applying a representation of the features that are the input of the core mining algorithms (Feldman & Sanger, 2006).

Natural Language Processing (NLP) technique is used to allow the referred data manipulations. NLP is defined as the automatic processing of human language (Ananiadou, Kell, & Tsujii, 2006). NLP appeals a computer through an algorithm to convert free form text into structured understandable data. This algorithm works according to Artificial Intelligence (AI) and Machine Learning techniques (ML).

This process usually combines linguistic concepts such as part-of-speech (noun, verb, adjective, etc.) and grammatical structure, taking into consideration grammatical anaphora and ambiguities. To perform such a sophisticated task, NLP makes use of diverse knowledge representations (lexicon of words and their meaning and grammatical proprieties) and a set of grammar rules. The explained method can be combined with other resources such as ontology of entities and actions, or a thesaurus of synonyms or abbreviations (Gerard Salton & McGill, 1983).

Preprocessing initializes with the task of tokenization, here each sentence is separated in words or tokens. In this research the process of tokenization converts the sentences at word level, the idea behind keeping more than a word is to retrieve knowledge from the combination of two or more words which could lead to a different concept rather the words separated, although the usage of more than a word per token increase the dimension of the data (Webster & Kit, 1992).

At this stage of the preprocessing, the collected data is characterized by its big dimensionality, since each collected document is now represented by its respective tokens. This set of data carries not only the content that can provide value but also the so-called noise, content that does not add any value to achieve the intended goal (Tang et al., 2014). Therefore, emerges the concept of 'curse of dimensionality' introduced by Richard Bellman where he concludes that after certain point, adding more features (dimensions) would only decrease the performance of the models applied in further stages of text mining. The cause of this effect underlies in the sparsity of the data, where the density of the vectors is not well represented which imply a bad effect on the entire system (Bellman, 1961).

In this context arises the need of a dimensional reduction, in summary is the reconstruction of the data into a lower dimensional space in such way that irrelevant variance in the data is discarded (Burges, 2009). After the execution of such techniques the noise and redundant features are removed, remaining the key low-dimensional uncorrelated features (Brázdil, 2016).

The main benefits of performing a dimensional reduction are: decrease of required storage, improvement of the learning performance, creation of better generalizable models, diminish of the computational complexity, and facilitates the data visualization (Brázdil, 2016) (Yan et al., 2005).

Finally, the last task of pre-processing is the document representation, where the text is transformed in order to be 'understandable' to the forthcoming algorithms. It initializes by building a text representation model, the most frequent method is Vector Space Model due to its universality and simplicity (Boulis & Ostendorf, 2005). Salton and McGill (1983) explained that a vector comprised of the keywords contained within the document can represent the document itself. Thus, in this method, each document is represented by a vector of term weights (Sebastiani, 2002), where the existing terms are contained in the resultant dictionary of the previous task (dimensional reduction).

In Vector Space Model the terms are weighted to indicate their importance for document representation (Sebastiani, 2002). Typically, weighting designs assume that a term relevance is

proportional to the number of documents that contain the respective term. The simplest measure to weight the terms in a document is the Term Frequency method, where each term importance is given by the number of times it occurs in a document, therefore the weight of a term in a document is calculated by:

$$tf(t_k, d_j) = \begin{cases} 1 + \log \#(t_k, d_j) & \text{if } \#(t_k, d_j) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Where $\#(t_k,d_j)$ represents the number of times the term k (t_k) appears in the document j (d_j) (Salton & Buckley, 1988).

Nevertheless, term frequency is limited to the term occurrence within a document which despises the term occurrence among a collection of documents. For that, Inverse Document Frequency (IDF) measure assumes that the importance of a term is inversely proportional to the number of documents that contain such term. IDF factor of a term is calculated by:

$$\log \frac{|Tr|}{\#_{Tr}(t_k)}$$

Where Tr is the total number of documents understudy and $\#Tr(t_k)$ denotes for the number of documents that contain the respective term (Salton & Buckley, 1988).

Consequently, Salton, Wong and Yang (1975) suggested that the combination of the two enunciated term weighting methods would have a significant improvement on the performance of the further algorithms. This approach is named Term Frequency Inverse Document Frequency (TF-IDF) and is given by:

$$tfidf(t_k, d_j) = tf(t_k, d_j) \cdot \log \frac{|Tr|}{\#_{Tr}(t_k)}$$

The output of this formula is a matrix where the rare terms in the collection are displayed as well as the quantity that a term occurs in a document. Both results are important in determining which terms are relevant and which are less relevant for further analysis (G. Salton, Wong, & Yang, 1975).

2.3.3. Core Mining Process

The core mining uses the pre-processed data to generate knowledge discovery, such as pattern discovery, trend analysis, and incremental knowledge discovery algorithms (Feldman & Sanger, 2006). According to the intended goal and the characteristics of the collected data, two different approaches can be used; supervised learning and unsupervised learning (Allahyari, Trippe, & Gutierrez, 2017). Both approaches require the combination of an iterative input (queries, browsing, added or subtracted constraints) and iterative output (new result-sets and subsets) (Feldman & Sanger, 2006).

2.3.3.1. Supervised Learning

Classification is an activity of supervised learning, since the process is conducted by the knowledge within the known categories on the training instances (Sebastiani, 2002). There is a wide range of possible methods such as nearest neighbour classification, decision trees, rule-based or probabilistic classifiers (Mitchell, 1997).

The aim of text classification is to assign a predefined class to text documents (Mitchell, 1997). The process is divided in two sequential stages. Initially an automated learning algorithm is performed where a model is developed in order to receive an input and assign a category based on association knowledge driven by the training examples. There are four main models to perform such classification; Naive Bayes Classifier, Nearest Neighbour Classifier, Decision Tree classifiers, and Support Vector Machines (Allahyari et al., 2017).

Therefore, the performance of the resultant model is analysed where the categorization effectiveness is measured. Since the training sample was used to build the model, the performance results applied to this sample would bring biased results. For that reason, a group of labelled samples are not used in the training phase, so they can be used to measure the quality of the model (Feldman & Sanger, 2006).

The intention of the research is to retrieve knowledge from the object understudy without any previous categorization, therefore the focus of the study rely on unsupervised learning. For that reason, the main focus is on unsupervised learning.

2.3.3.2. Unsupervised Learning

Clustering, opposed to classification, is an unsupervised learning approach (Wagstaf, Cardie, Rogers, & Schroedl, 2001). Applying this technique without any pre-defined category, each document will be labelled to a meaningful group denoted by cluster. This grouping process is only based on the content of each sample and how the sample is related to the dataset (Feldman & Sanger, 2006). The documents that belong to the same cluster share similar characteristics and are different from the remaining clusters.

There are several types of text clustering algorithms, for instance partitioning algorithms, agglomerative clustering algorithms and probabilistic clustering algorithms (Allahyari et al., 2017). In the development of this research the algorithms k-means clustering and Latent Dirichlet Allocation were used to group the documents under study.

2.3.3.2.1 Latent Dirichlet Allocation

LDA is the most popular technique of topic modelling which was established by Blei et al. (Blei, Ng, & Jordan, 2003). Topic models can be designated as the idea that documents naturally contain a mix of topics. A topic can be defined as a distribution over the terms in the corpus of documents where each document is expressed as a probability distribution over the topics (Steyvers & Griffiths, 2006) (Aggarwal & Zhai, 2013).

LDA is a technique where a probabilistic generative model estimates the properties of a multinomial observations. The technique performs latent semantic analysis (LSA) to find the latent structure of the topics in a text corpus (Heinrich, 2008).

Deerwester et al. (1990) proved that the co-occurrence structure of terms in text documents can provide the latent topic structure, without the usage of prior knowledge. Therefore, information can be retrieved and represented in an appropriate form for matching user needs with content items on a meaning level, rather than by lexical congruence (Heinrich, 2008).

The purpose of this technique is to discover short descriptions of the contents in a corpus without interfering with the statistical relationships in the corpus (Blei et al., 2003). To achieve this discovering process, a huge amount of textual data for each document is required. Therefore, the achieved descriptions, called topics, can be used to identify similarities among the dataset. Topics are used to produce visualizations that allow the comparison and representation of each document. Additionally, it provides searching and indexing methods (Hu, 2009).

LDA premise is that the entire set of documents share the same set of topics, but each document presents a different distribution of the topics. Figure 2.1 shows an example of a scientific paper where the concept of 'hidden topics' is displayed. The highlighted terms are assigned to topics where they belong to, and the thematic structure of the collection is represented in the histogram (topics distribution) (Christou, 2016).

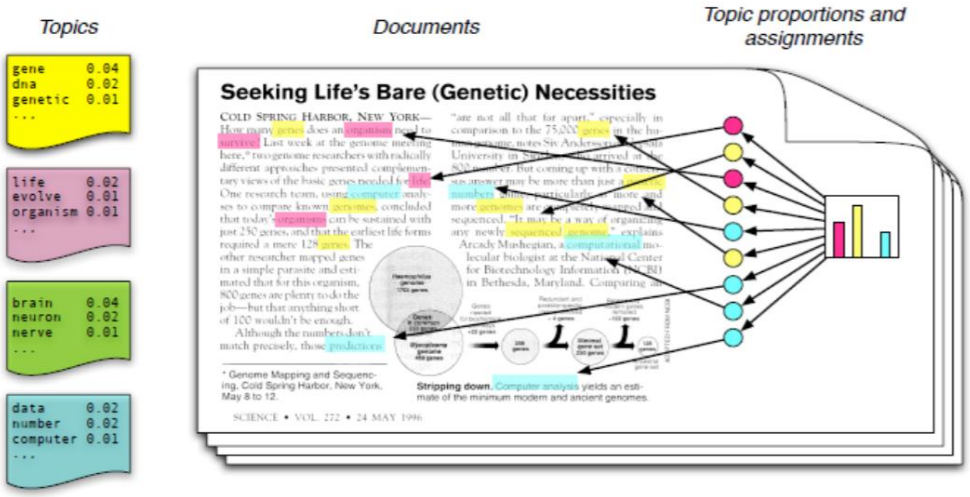

Figure 2.1 - The intuition behind LDA (Blei et al., 2003)

2.3.3.2.2 K-means Clustering

The K-means and Expectation-Maximization are the two methods mostly used in cluster analysis (Feldman & Sanger, 2006). Both are spatial clustering techniques. What distinguish them is that K-means is hard, flat, and shuffling while Expectation-Maximization is soft, flat, and probabilistic (Fayyad, Reina, & Bradley, 1998). The K-means is the method chosen to perform the clustering due to its simplicity and efficiency. Expectation-Maximization method was discarded in this study because it is usually used when data is incomplete (McLachlan & Krishnan, 2007).

K-means clustering, in the context of text mining, makes automatically a partition of the collected documents into a amount of k clusters (Wagstaf et al., 2001). The algorithm initiates with importing a vectorized representation of the dataset. Thereafter, a pre-defined number of 'seeds' are either externally selected or randomly spread among the vectors (Feldman & Sanger, 2006). Afterwards, the seeds that represents the centre of each cluster are iteratively reallocated until it has reached the minimum distance between the elements (intra-cluster similarity). At the same time, it maximizes the distance between the centres of the clusters (inter-cluster dissimilarity) (Baeza-Yates & Ribeiro-Neto, 1999).

A measure of similarity and the number of clusters has to be defined to compute the technique. By default, k-means is performed with the Euclidean Distance which is the standard metric for geometrical problems (Huang, 2008). Since k-means requires the number of clusters, usually the algorithm is computed with different values of k, where the one with best performance is chosen (Feldman & Sanger, 2006). The following figure describes the simplest form of the k-means algorithm.

Input : Document set $\mathcal{D}$, similarity measure $\mathcal{S}$, number k of cluster
Output: Set of k clusters
initialization
Select randomly k data points as starting centroids.
while *not converged* **do**
  Assign documents to the centroids based on the closest similarity.
  Calculate the the cluster centroids for all the clusters.
end
return *k clusters*

Figure 2.2 - Simplest form of the k-means algorithm, (Allahyari et al., 2017)

The classic k-means algorithm chooses randomly where to place the seed given the origin of the initial centroids. According to the importance of this initial stage and the fact that is a random process, it sometimes leads to bad clustering results. A strategy emerged to counter this ambiguity, where the initial centroids are placed far away from each other, leading to better and more consistent results (Arthur & Vassilvitskii, 2007).

2.4 Text Mining Application in Educational Context

Data mining techniques already have an extensive array of applications for the education sector, such as student marketing, selection revenue analysis, planning of courses and results analysis (Romero & Ventura, 2010). Some educational organizations use DM to learn students' performance and behaviour, in order to design course curriculum which impacts on their motivation (Goyal & Vohra, 2012).

However, there is much more valuable content hidden in unstructured data which is of no exception in the educational context. Examining the published academic papers is notably a lack of investment in this area due to lack of resources (Feldman et al., 1998).

Although, as stated by Romero, Ventura, Baker and Pechenizkiy (2010), the application of text mining in the educational field began to increase. The main goals of this change are communicating to stakeholders, maintaining and improving courses, generating recommendations, predicting student grades and learning outcomes, student modelling, and domain structure analysis (Romero, Ventura, Baker, & Pechenizkiy, 2010). Likewise, the aim of this study is to communicate to stakeholders. The core purpose is addressed to help administrators and educators with exploratory data analysis through the published dissertations in NOVA-IMS university.

Text mining applied to scientific literature is not a new approach. Krassmann, Herpich, Bercht and Cazella (2017) used text mining techniques on a corpus composed of 10 academic papers, where they found research opportunities through identifying and analysing the main trends in terms of development and applications in educations fields. Also, Nie and Sun (2017) combined clustering and bibliometric analysis applied to more than 20.000 publications to detect research trends in design research, the analysis led to shaping four academic branches and summarizing each academic branch.

Likewise, text mining applied to scientific literature was also explored in supervised learning, where the intention was to classify such literature. For instance, Sulova and Nacheva (2017) used text mining techniques to classify 242 abstracts of research papers published in an academic scientific journal in Bulgaria. The approach was successfully applied in the classification of the papers in three predefined categories (Social Sciences, Engineering and Technology, Other). Also, Baba and Kumar (2016) published a research that uses the same approach, here 40 research papers were classified in four categories (Java, Operating System, DBMS, Data Structure).

A real application of a text mining process with a similar data source (research papers) is PubMed. This is the national library of medicines online repository of medical journals (Feldman & Sanger, 2006). Likewise, PubGene is a publicly accessible search engine that combines biomedical text mining with network visualization, fed by biomedical literature.

Text mining in education have also been applied in a non-scientific context. Yadav, Bharadwaj, and Pal (2012) used these techniques to predict the retention purpose of incoming students according to their records. Moreover, the model identified whether students need special attention to reduce the drop-out rate. Furthermore, Xu and Reynolds (2012) classified students open answers to a certain leadership dilemma where they found a significant accuracy on the ratings.

3. METHODOLOGY

The aim of this study was to apply text mining techniques in the educational context. Furthermore, trends and current patterns were examined to extract knowledge from the dissertations of students of NOVA IMS. Therefore, the following research question was formulated: *What are the current patterns and trends in the dissertations published by NOVA-IMS students?*

The sub-research questions were examined to reach the initial objective of this study. Figure 3.1 summarizes the research process. First, an extensive literature research was conducted to understand the entire process of text mining. This chapter shows step by step how text mining is applied in the context of master dissertations of a university. Initially, the process of collecting data is described. Thereafter, pre-processing the dataset is explained according to three steps. Furthermore, the core mining process is explained where two different techniques are applied; Topic modelling and cluster analysis. Finally, the discovery of trends and patterns is explained.

3.1 Research Design

The initial purpose of this study is to extract knowledge from master dissertations of students of Nova IMS in the faculty of Statistics and Information Management. The exploratory part of this study aims to develop initial insights and new perspectives for future research purposes. Given a collection of Master dissertations, text mining techniques were applied to identify and discover knowledge from those informative documents. Text mining concerns discovering and extracting knowledge from unstructured data and is useful tool that combines exploration and discovery with confirmatory analysis which allows to seek for new information. In this study, Master dissertations of students are transformed into classified documents to examine explicit trends and patterns from previous years.

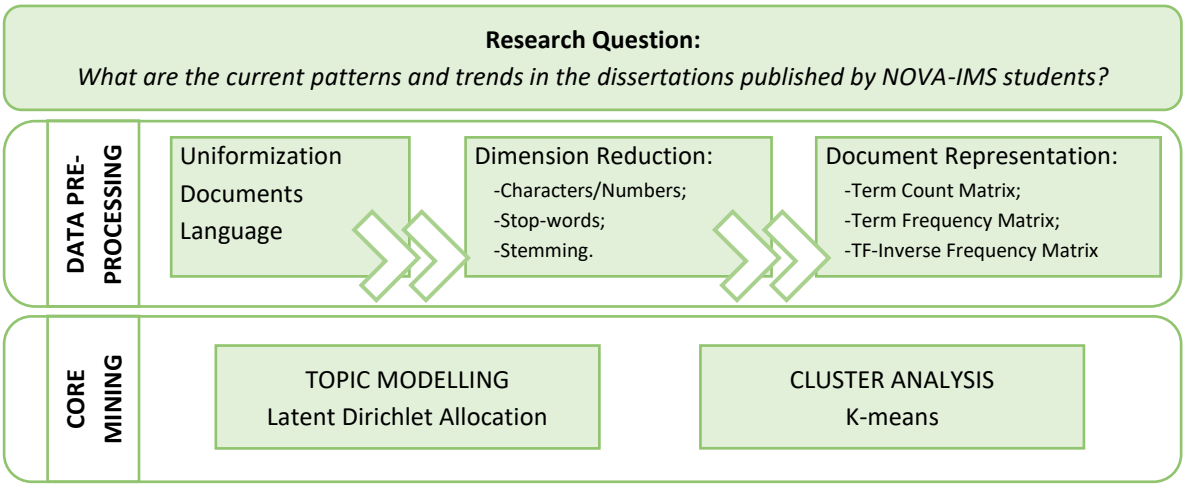

Figure 3.1 - Research Model

3.2 Data Collection

This study examines the dissertations of students from NOVA IMS that were published between March 2004 until May 2018. The dissertations were available online at the repository of documents of the NOVA IMS (run.unl.pt). To collect the document features of the all the different dissertations, Webscrapping was used. Webscrapping automatically downloads content from websites which is more efficient than performing it manually. The aim of this data collection was to extract the following document features of the dissertations; Master program, Title, Author, Date, Abstract, and Keywords.

3.3 Pre-Processing

After the data collection, data needed to be prepared for further analysis. Figure 3.2 describes the different steps that Chakraborty, Pagolu and Garla (2013) published about the text mining process. This process flow was a guidance for the performing text mining in this study.

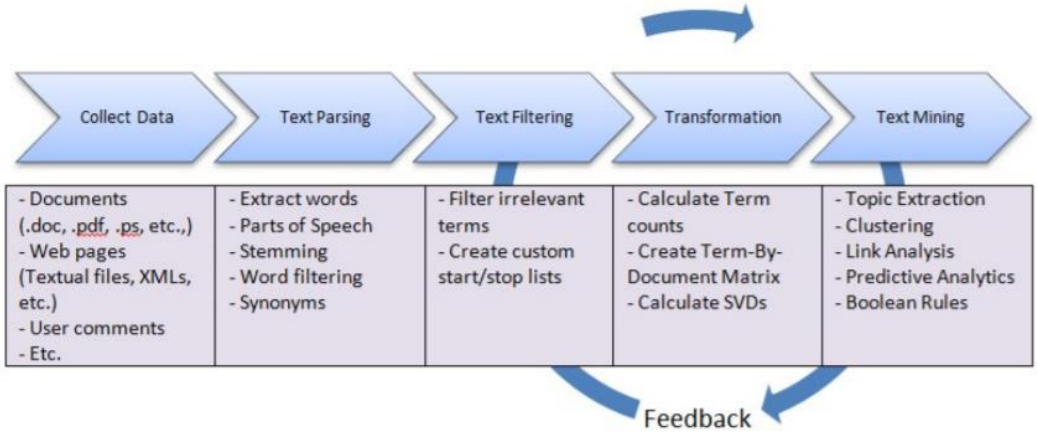

Figure 3.2 - Text mining process flow(Chakraborty, Pagolu, & Garla, 2013)

3.3.1 Uniformization of Data Language

Since this study was conducted in Portugal, dissertations were written in Portuguese and English. However, the purpose was to analyse all the dissertations together. Because the majority of textual algorithms are more developed in English and because this study is conducted in English, the Portuguese dissertations had to be translated in English first. This translating procedure was executed with the googletranslator python library. With this library the Portuguese contents were translated to English.

3.3.2 Dimension Reduction

Dimension reduction is a vital step in pre-processing the data before modelling such contents. The important information in the dissertations is accompanied by many invaluable words, structures and tokens which cause noise. The presence of too much noise disturbs the exploration of patterns and relations among the different documents. Therefore, three different techniques were applied to reduce the impact of noise.

1. **Remove non-alphabetic characters**

Different tokens that have no meaning to the context of the dissertations were removed. Also, all numeric digits were removed from the content. Furthermore, all capital letters were transformed into lower case letters in order to create equality throughout the dataset. Examples of these non-alphabetic characters are displayed below:

!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~ = ––—’‘“”€©®°·

2. **Remove stop-words**

Stop-words do not have a meaning to the context which causes a disturbance in text mining. Therefore, the stop-words were removed. Examples of stop words are listed below.

And, or, many, this, those, a, of, to, at, for, what

3. **Stemming**

Stemming is the process of transforming a word into its root form. The objective to apply this technique is to reduce the total number of unique words in the dictionary. As a result, the number of columns in the document-word matrix will be denser with less columns. To execute this procedure the python function used was SnowballStemmer from NLTK library. In the following example, nine unique words are transformed in one single unique word:

Word Before	Outcome
comparability	compar
comparable	
comparation	
comparative	
comparatively	
compare	
compared	
compares	
comparing	

Figure 3.3 - Stemming example

In this example, a part of the end of the word is removed. Therefore, the quantity of unique words is reduced as well as the amount of characters.

3.3.3 Document Representation

Document representation is the last step of pre-processing the data (Boulis & Ostendorf, 2005). To execute the following step, the python library sklearn was imported in order to execute the functions CountVectorizer and TfidfTransformer. First, a matrix was created where the rows represent the individual documents and the columns represent all the terms that are used in each document. The matrix displays how many times a word appears in an individual document. Thereafter, each word was labelled with the respective frequency weight on the document. Basically, is the count of each term divided by the total terms on the respective document. Finally, the weights were adjusted according to the times they appear in other documents. This means that the weights of terms that appear in many documents were lowered.

3.4 Core Mining Process

The aim of the core mining process is to discover patterns, analyse trends and extract knowledge from the dissertations (Feldman & Sanger, 2006). In this study, two different techniques were applied to the dataset; Topic modelling and Cluster sampling. For all the methods, to understand the outcomes of these algorithm, visualisation tools were successfully applied.

3.4.1 Topic Modelling

Topic modelling was executed with the algorithm LDA. A python machine learning algorithm was computed in order to evaluate the quality of the LDA model within different values of the required parameters: 'number of topics' and 'learning decay'. The outcome of the algorithm depicts how many topics should be chosen according to the log-likelihood score. Thereafter, the distribution of the documents over the topics was calculated. Furthermore, the key terms represented in each topic were displayed to already gain an insight of the topics. Thereafter, the topics were displayed in a multidimensional scaled matrix represented by the first two Principal Components (PC). This plot will show how the topics relate with each other's, which is valuable information when interpreting the topics.

Inferring the topics

The topics had to be interpreted to continue the analysis of the dissertations. It is the user knowledge of the collected dataset that will suggest about the interpretation of each topic. To strive for optimal objectivity, all the outcomes and visualizing tools were consulted. Therefore, the previous steps and results were deeply examined. Moreover, exclusive key terms and key terms that appeared in the remaining topics were examined to gain a deep understanding of the meaning of the topic. This was an iterative process where the python visualization tools were fundamental to explore the outcome. In this way, different perspectives were combined which was fundamental for interpreting the topics.

3.4.2 K-means Clustering

The cluster algorithm used was k-means. The python library sklearn was used to execute the cluster algorithm. The algorithm requires a pre-specified parameter from the user which is the number of clusters. Therefore, the *elbow method* was used to identify which number of clusters would be ideal for the dataset used.

Inferring the clusters

The cluster had to be interpreted to continue the analysis of the dissertations. It is the user knowledge of the collected dataset that will suggest about the interpretation of each cluster. Therefore, the previous steps and results were deeply examined. Moreover, exclusive key terms and key terms that appeared in the remaining clusters were examined to gain a deep understanding of the meaning of the cluster. Additionally, the distribution of each cluster per topic of dissertation was analysed. In this way, different perspectives were combined which was fundamental for interpreting the clusters.

3.4.3 Finding trends and patterns

At this stage, the outcomes of the previous tasks were merged in order to achieve knowledge from the dataset. Initially, the resultant topics and clusters were analysed independently. Moreover, the outcomes of all the text mining techniques together were used in the examination each dissertation field separately, in order to explore trends and patterns.

LDA topics and K-means clusters analysis

The analysis of trends and patterns for the topics and clusters started by displaying the frequency of the documents per topic/cluster. Thereafter, the weight of each study field per topic/cluster was observed to identify relations. Moreover, a plot was developed where each year displayed how many dissertations were published in each specific topic/cluster in order to understand how the topics/clusters were trending.

Dissertations fields analysis

The intended goal of this stage was to understand how each dissertation field was related to the dataset, the founded topics and clusters (patterns), also how these relations have been changing over the years (trends).

Therefore, the analysis started by observing the quantity of documents published over the years. Moreover, a pie chart presented the weight of the dissertation field compared to the remaining dataset. Afterwards, the most frequent words and respective counts over the years were displayed in a table. The examination of this data allowed to understand the consistency of the keywords over the years, the emerging concepts and the vanished ones.

Thereafter, the weight of each topic was plotted over the years (of the respective dissertation field). Insights of trends and patterns were reached through the examination of the produced data. Finally, the previous approach was applied to the clusters to understand how they were trending in the respective dissertation field over the years.

4. RESULTS

4.1 Data Collection

The original Master dissertations were gathered electronically available at the repository of documents of NOVA IMS. In total, 594 dissertations were collected. The dissertations were written in Portuguese and English in the period of March 2004 until May of 2018 (see Figure 4.1). Since 2014 the number of published dissertations have a clear positive trend, the highest record is in 2017 with a total of 82 dissertations published, followed by 2018 which in five months were published 72.

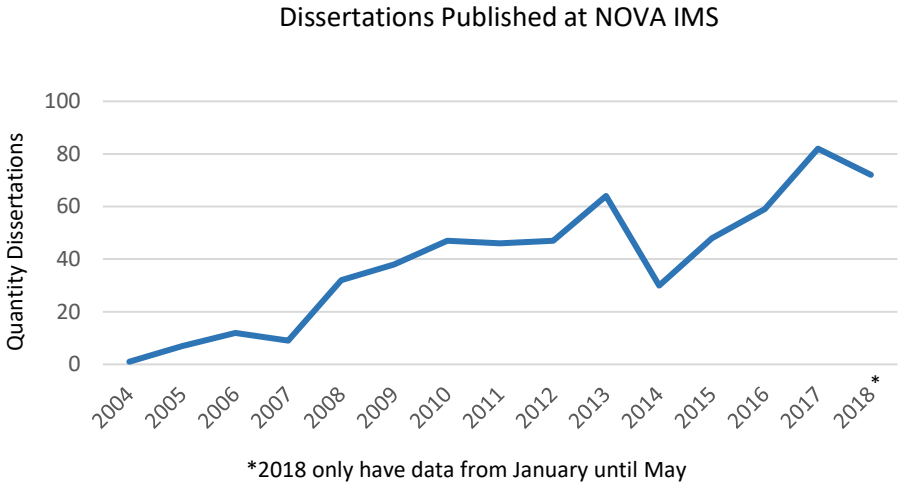

Figure 4.1 - Published dissertations at NOVA IMS

The students were following a master's degree at Nova IMS in the field of Geospatial Technologies, Statistics and Information Management, Information Management, Advanced Analytics, and Geographic Information Science and Systems. As demonstrated in Figure 4.2, most of the master dissertations were written in the field of Statistics and Information Management (31%). Advanced Analytics is with 2% the field where the least dissertations are published.

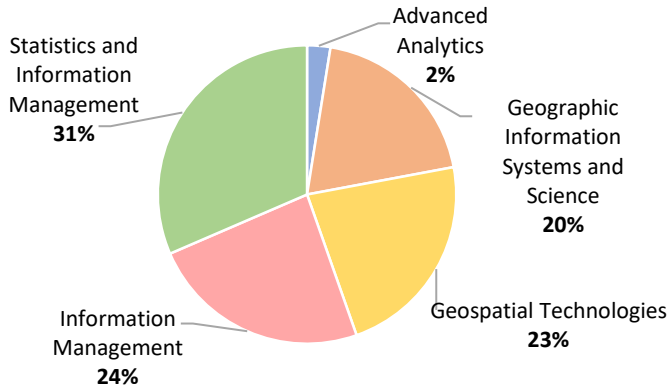

Figure 4.2 - Master dissertations published in the different fields of NOVA IMS

4.2 Pre-Processing

4.2.1 Uniformization of Data Language

In total, 340 dissertations were written in Portuguese. Therefore, these dissertations were successfully translated into English with googletranslator python library.

4.2.2 Dimension Reduction

In order to decrease noise in the dataset three techniques were applied to the collected dataset: remove non-alphabetic characters, remove stop-words, and stemming. Below, the results are presented respectively. The first tables show the top 10 removed symbols from the dataset. The Comma was the symbol that was removed the most, namely 6.888 times. The second table displays the top 10 numbers that are removed and the quantity. It demonstrates that '2010' is removed 19 times from the dataset. The third table displays the top ten stop-words, words that are un-valuable for interpreting the data, that were removed from the dataset. It is clear from the table that the words 'the', 'in', and 'of' are removed the most times. The last table shows the stemming where a part of the word termination is removed so that words in different form converge to the same. Most of the times, the '-s' and '-e' at the end of a word were removed to create the same word. For instance; 'competitors' became 'competitor' and 'culture' becomes 'cultur'.

Table 4.1 - Removed Non-alphabetic Characters; Symbols

Symbol	Quantity Removed
,	6.888
.	4.717
-	878
)	614
(	614
:	156
/	69
’	56
%	34
&	30

Table 4.2 - Removed Numbers

Numbers	Quantity Removed
2010	19
2001	14
2000	13
2013	12
2003	12
2007	11
2009	11
2005	10
2012	10
2006	9

Table 4.3 - Removed Stop-words

Stop-Words	Quantity Removed
the	9.377
in	8.149
of	4.686
and	4.622
to	3.134
at	2.178
o	1.702
is	1.667
for	1.665
what	1.429

Table 4.4 - Stemming

Numbers	Quantity Removed
s	8.728
e	6.487
ed	3.933
es	3.174
ation	2.579
ing	2.525
ion	2.100
al	1.892
ic	855
ions	608

The following table presents the amount reduced of characters, words and unique words among the three different techniques applied in the dimension reduction. The first technique achieved a reduction of 5.374 unique words that represents 37% of the total unique words on the dataset. The removing of Stop-Words provided a reduction of 43 of the total words. Followed by stemming technique that allowed a reduction of 117.992 characters, that represents 13% of the characters in the dataset. In summary, after applying the referred techniques the resultant dataset had a reduction of 61% unique words, a noteworthy noise reduction that will improve the quality of the core mining outcomes. Also, the average character per word was reduced from 8,2 to 6,8 a noteworthy increase in performance and efficiency on the algorithms required in the further tasks.

Table 4.5 - Impact of dimension reduction on the dataset

	Characters and numbers		Stop-Words		Stemming		TOTAL	
Characters	21.207	2%	222.878	25%	117.992	13%	362.077	40%
Words	492	0%	59.637	43%	0	0%	60.129	43%
Unique Words	5.374	37%	124	1%	3.232	22%	8.730	61%

4.2.3 Document Representation

After finishing the dimension reduction, the documents were ready to be represented in 3 sequential matrices. All of them had a shape of 594 lines (documents) and 5.340 columns (unique terms). Since the non-zero values on the matrix were 29.805, thus the matrices have a sparsity of 0.94%.

First, the term count matrix was developed followed by the term frequency matrix (TF) and finally the inverse document frequency matrix (TF-IDF). The following three tables represent parts of the matrices. For instance, the tables demonstrate that the TF matrix compared with the TF-IDF loses weight on the term ‘algorithm’ among the documents that contain this word. The reason of this loss is due to the frequency that the word appears in the entire dataset, the more ‘exclusive’ a word is the more importance gets.

Table 4.6 - Term Count Matrix

	Term 1 ababa	Term 2 abandon	Term 3 abbrevi	...	Term 154 algorithm	...	Term 5340 µgm³
Document 1	0	0	0	...	1	...	0
Document 2	0	0	0	...	0	...	0
Document 3	0	0	0	...	0	...	0
Document 4	0	0	0	...	2	...	0
Document 5	0	0	0	...	0	...	0
Document 6	0	0	0	...	0	...	0
...	...	...	...	...	...	...	...
Document 594	0	0	0	...	0	...	0

Table 4.7 - Term Frequency Matrix

	Term 1 ababa	Term 2 abandon	Term 3 abbrevi	...	Term 154 algorithm	...	Term 5340 µgm³
Document 1	0%	0%	0%	...	1.0%	...	0%
Document 2	0%	0%	0%	...	0%	...	0%
Document 3	0%	0%	0%	...	0%	...	0%
Document 4	0%	0%	0%	...	2.9%	...	0%
Document 5	0%	0%	0%	...	0%	...	0%
Document 6	0%	0%	0%	...	0%	...	0%
...	...	...	...	...	...	...	...
Document 594	0%	0%	0%	...	0%	...	0%

Table 4.8 - Term Frequency - Inverse Document Frequency Matrix

	Term 1 ababa	Term 2 abandon	Term 3 abbrevi	...	Term 154 algorithm	...	Term 5340 µgm³
Document 1	0%	0%	0%	...	0.8%	...	0%
Document 2	0%	0%	0%	...	0%	...	0%
Document 3	0%	0%	0%	...	0%	...	0%
Document 4	0%	0%	0%	...	2.3%	...	0%
Document 5	0%	0%	0%	...	0%	...	0%
Document 6	0%	0%	0%	...	0%	...	0%
...	...	...	...	...	...	...	...
Document 594	0%	0%	0%	...	0%	...	0%

*The original values of the TF-IDF were weighted to allow the comparison of TF and TF-IDF

4.3 Core Mining Process

4.3.1 Latent Dirichlet Allocation

4.3.1.1 Model parameters

The technique used to perform topic modelling was LDA. A python machine learning algorithm was computed in order to evaluate the quality of the LDA model within different values of the required parameters: 'number of topics' and 'learning decay'. The following figure demonstrate the Log Likelihood Score (measure of quality) for the different combinations of parameters. The LDA model shows that a model with four topics has the highest Log Likelihood Score, namely -129.488. Furthermore, at four number of topics, the red line has the highest value which indicates a learning decay of 0,7. Therefore a model with four topics and a learning decay of 0,7 is suggested.

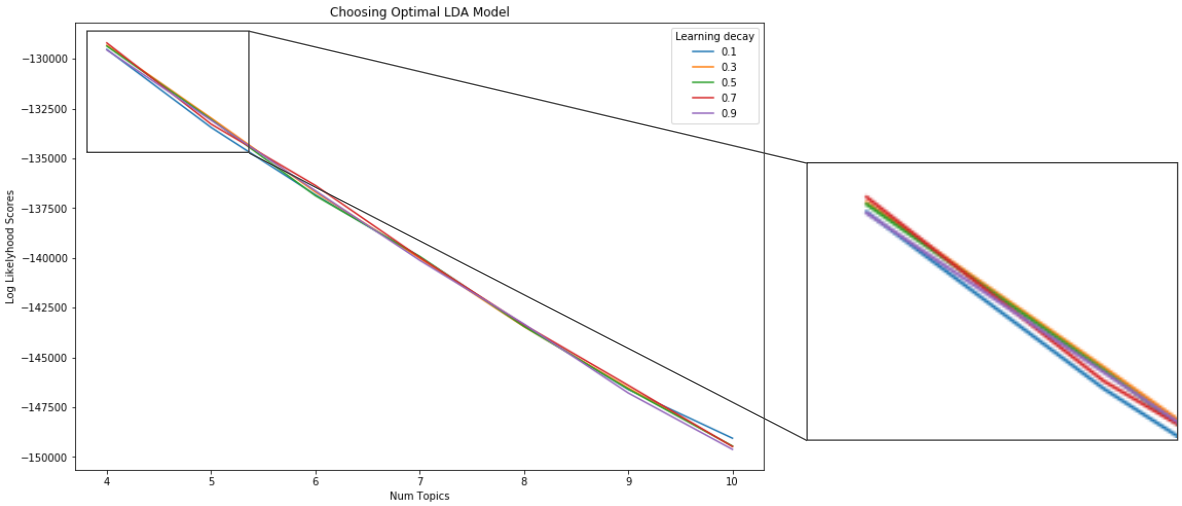

Figure 4.3 - Topics distribution according to LDA Model

4.3.1.2 Model outputs

The dissertations were distributed over the four topics with each topic containing 119, 227, 103 and 145 topics respectively. As an example, Table 4.9 presents part of the attribution of the dominant topic to each document. It can be observed that some documents are fully representative of a topic, however, some documents have ambiguous dominant topic since they could be represented in two or more topics (e.g. document 4). This outcome is useful for future interpretations since documents can be detected to gain more insight of the topics.

Table 4.9 - Dominant topic per document

	Topic 1	Topic 2	Topic 3	Topic 4	Dominant Topic
Document 1	0.99	0	0	0	1
Document 2	0	0	0.99	0	3
Document 3	0	0	0	0.99	4
Document 4	0.56	0	0.43	0	1
Document 5	0	0	0	0.99	4
Document 6	0.01	0.98	0.01	0.01	2
…	…	…	…	…	…
Document 594	0	0	0.99	0	3

Table 4.10 shows the top ten terms that represent a topic. Figures 9.1, 9.2, 9.3 and 9.4 in the appendix show the top 30 most relevant terms for the topics. They are divided into overall term frequency and estimated term frequency within the selected topic (Sievert & Shirley, 2014).

Table 4.10 - Top ten key terms representing the topic

Topic 1	Topic 2	Topic 3	Topic 4
land	satisfaction	insurance	clustering
urban	students	web	tourism
cover	consumers	algorithm	experience
forest	sample	internal	Infrastructure
classification	index	framework	Scale
images	client	city	Tourist
detection	loyalty	adoption	Group
accuracy	dimensions	participation	Emergency
classes	transport	competitive	survey
algorithm	teaching	report	report

Via multidimensional scaling, the topics were plotted on two axes. The figure below shows the output of the PC plot. In PC1, topic 2 and topic 4 contain highly controversial characteristics, therefore they are far apart from each other's in the horizontal axis. However, topic 1 and topic 3 are not correlated with those characteristics. On the other hand, in PC2, topic 1 and topic 3 are controversial and therefore at the end of the vertical axis. And here counts as well; topic 2 and topic 4 are not related with these characteristics. Those opposite characteristics should be considered for when inferring the topics

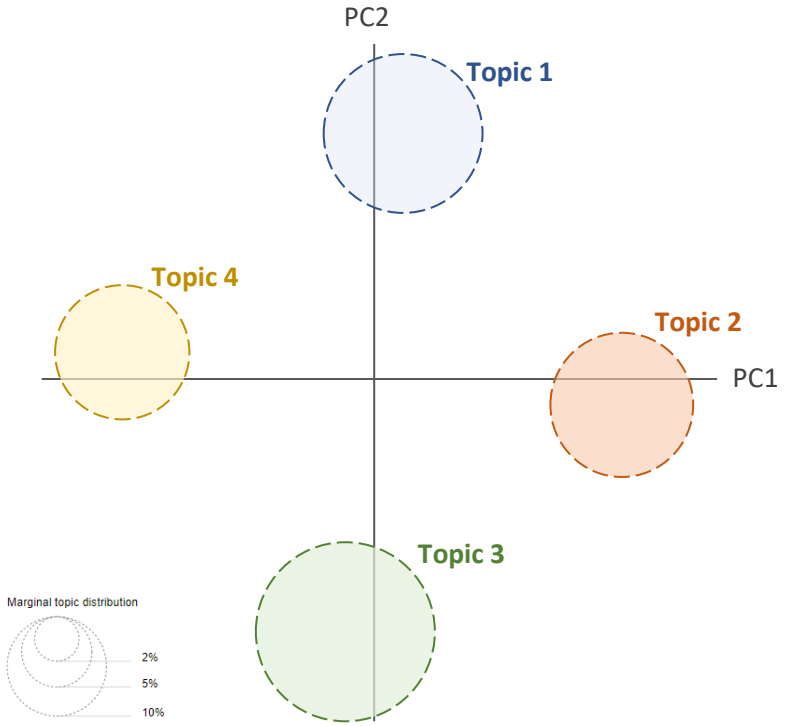

Figure 4.4 - Topics represented in distance map via multidimensional scaling

4.3.1.3 Inferring the topics

Different results were used to infer the topics. First, the key terms in each topic from Table 4.10 were deeply examined. The more complete figures in the appendix provide more information whether the term is represented in other topics as well. Moreover, Table 4.9 played a central role in this examination to understand which documents are represented in the topics. Furthermore, the PCs enabled to explore how each significant term was represented on the remaining topics in order to understand the differences. For example, the term 'tourist' in topic 3 and topic 4 do not have any frequency in topic 1 and topic 2. This indicates that 'tourist' is an exclusive term for the topic 3 and topic 4 which is highly relevant for interpretation. As mentioned above, the axes on the PCs are a great asset in inferring the topics. The matrix shows that topic 2 and topic 4 have opposite characteristics, as well as topic 1 and topic 3.

Therefore, topic 1 is considered as Geodata Information. Topic 2 is mostly concerned with students that examined behavioural topics in organisations, consequently called Behavioural Studies. Furthermore, topic 3 contained dissertations about Information and Decision Systems. Lastly, topic 4 indicated a topic about Implementing Systems. See the figure below for a visual representation.

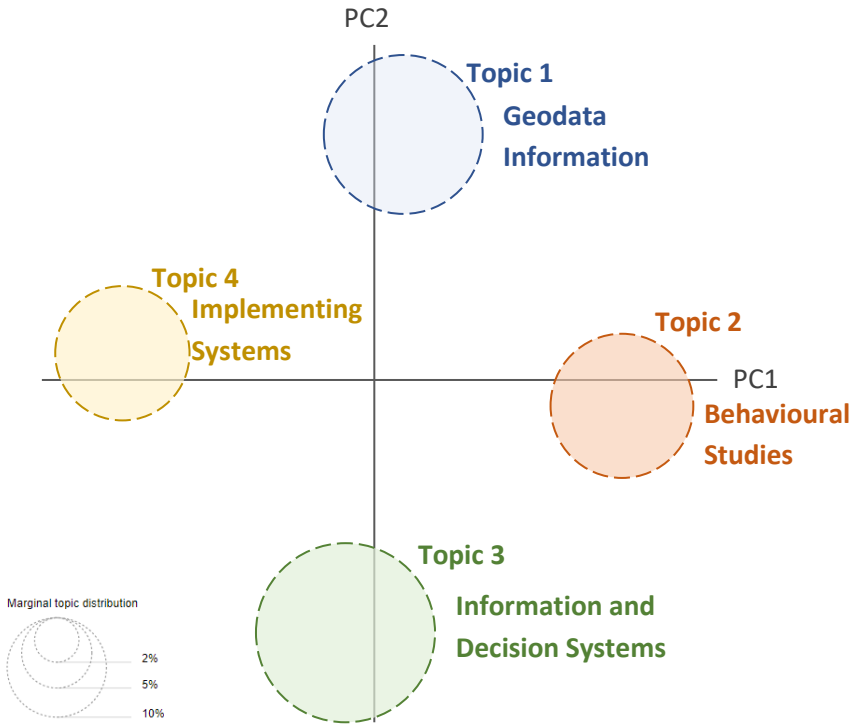

Figure 4.5 - Inferred topics represented in distance map via multidimensional scaling

Figure 4.5 shows the results of the inferred topics. First, it clearly shows controversies between topic 2 and 4 which are not important features in topic 1 and topic 3. Implementing systems included master dissertations that were concerned optimizing processes by implementing a new system in institutions. On the other hand, behavioural studies were more concerned on the personal level where different surveys were conducted to employees, tourists or clients to draw conclusions about satisfaction or loyalty in organisations. As PC1 already indicates, the left side of the axis is more concerned on the institutional level whereas the right side is more on the personal level. Furthermore, topic 1 and topic 3 are not related with these levels and are therefore in the middle of the axis.

Second, the figure shows controversies between topic 1 and 3 which are not important features in topic 2 and topic 4. Namely, Geodata Information, is a topic that is concerned with networks in space, such as geography, urban planning and maps. On the other hand, topic 3, decision systems, is focused on the network within a system. Therefore, the PC2 is considered as special networks versus system networks. In addition, Topic 2 and topic 4 are not related with those characteristics.

4.3.2 K-means Clustering

4.3.2.1 Model parameters

The following analysis was executed using k-means. Figure 4.7 presents the distortion values among the number of clusters. Three clusters are chosen since the elbow occurs when the number of clusters is three and since the variation is highest at this number (see Table 4.13).

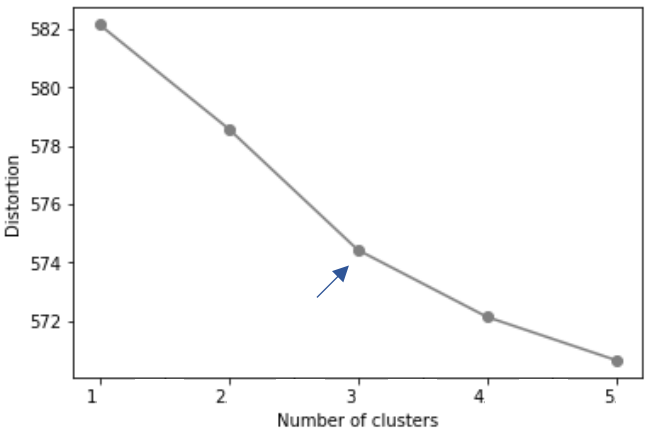

Figure 4.6 - Distortion per number of clusters

Table 4.11 - Distortion per number of clusters

Nº Clusters	Distortion	Variation
1	582.2	
2	578.6	3.6
3	**574.4**	**4.2**
4	572.1	2.3
5	570.6	1.5
6	567.8	2.9
7	564.8	3.0
8	563.4	1.3
9	561.6	1.9

4.3.2.2 Model outputs

Table 4.12 shows the top ten terms that represent each cluster. Figures 9.5, 9.6 and 9.7 in the appendix show a word cloud per each cluster.

Table 4.12 - Top ten key terms representing the cluster

Cluster 1	Cluster 2	Cluster 3
land	web	insurance
forest	consumers	bank
urban	digital	financial
cover	participation	marketplace
detection	city	client
classification	community	competitive
images	cheers	power
city	students	satisfaction
growth	sig	report
pattern	content	segmentation
lulc	human	algorithm
vegetation	open	intelligence
satellite	intention	internship
Landsat	online	internal
accuracy	sample	plug

4.3.2.3 Inferring the clusters

Below, Figure 4.7 the words of Cluster 1 is presented and examined. The words are mainly concerned with Geodata (e.g. land, forest, urban) and Analysis (e.g. classification, growth, pattern). Therefore, combining the two identified groups, Cluster 1 is referred as 'Geodata Analysis'.

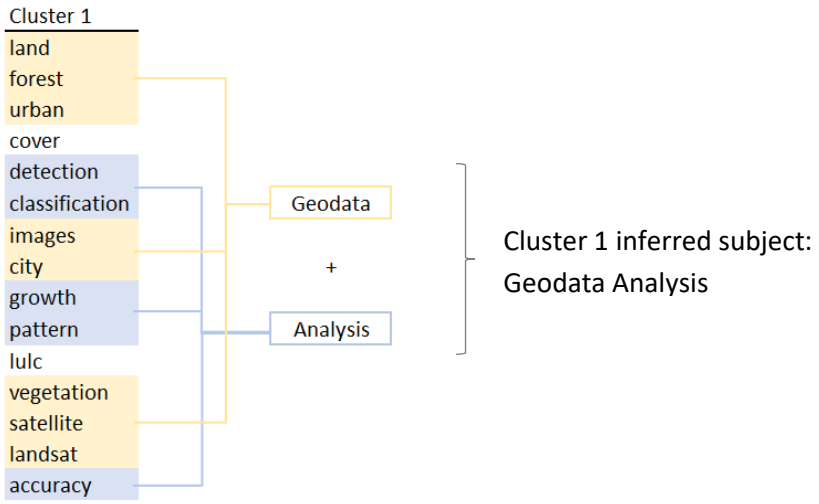

Figure 4.7 - Cluster 1 keywords and inferred subject

In Cluster 2, three subjects were engaged from the cluster keywords. Figure 4.8 supports the association made to inferring the final subject of the cluster. Firstly, the subject Online was the outcome of three keywords of the cluster: web, digital, and online. The second subject is People related to four keywords: consumers, community, student and human. Finally, the words participation, content, open, intention, and sample were summarized in the subject Understand. Therefore, combining the three identified subjects the Cluster 2 is referred as 'Online Behaviour'.

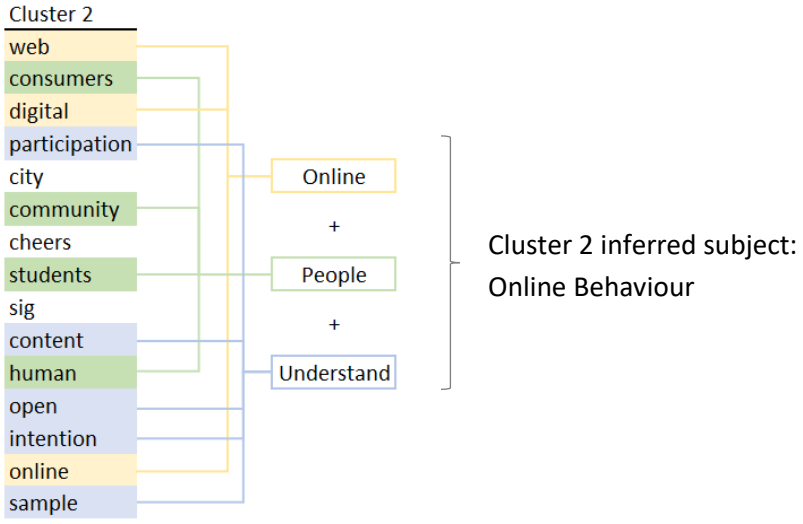

Figure 4.8 - Cluster 2 keywords and inferred subject

Lastly, the Figure 4.9 shows the four groups identified for the Cluster 3. First group was referred as Finance since it was composed by the words insurance, bank, and financial. Business took place of the second group, composed by the words client, competitive, and satisfaction. The third identified group was based on the words marketplace and segmentation denoted by Marketing. Finally, with the words report, algorithm and intelligence the last group was called Analysis. Therefore, combining the four identified groups the Cluster 3 is referred as 'Business Analysis'.

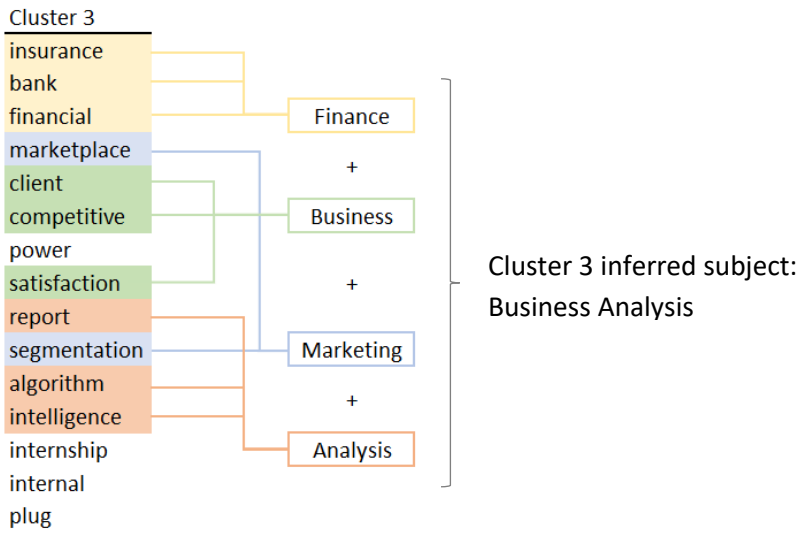

Figure 4.9 - Cluster 3 keywords and inferred subject

4.3.3 Finding trends and patterns

4.3.3.1 LDA topics analysis

The table below shows how many dissertations were published in each topic. 'Information and Decision Systems' was a topic where most dissertations were focused on (38%). Contrasting, 'Implementing systems' was, with 17%, the topic where the least dissertations were published in.

Table 4.13 - Percentage of dissertations per topic

Topic	Topic Name	Quantity Dissertations	%
1	Geodata Information	145	24%
2	Behavioural Studies	119	20%
3	Information and Decision Systems	227	**38%**
4	Implementing systems	103	17%

The different topics were examined per field of the dissertations published by NOVA IMS. Table 4.14 shows the weight of each dissertation field per topic. Interestingly, the topic 'Geodata Information' is mainly related to dissertations in the field of GISS and GT, where both fields have the same weight of 33%. Moreover, the dissertation field of SIM represent 47% of the topic 'Behavioural Studies'.

Furthermore, 'Information and Decision System' topic is mostly represented by the fields IM and SIM (i.e. 34% and 39% respectively). Finally, the topic 'Implementing Systems' is represented by the fields GT and GISS with 28% and 33% respectively.

Table 4.14 - Weight of each dissertation field per topic

Dissertation Field	Geodata Information	Behavioural Studies	Information and Decision Systems	Implementing Systems
AA	2%	0%	5%	0%
GISS	**33%**	22%	6%	**28%**
GT	**33%**	14%	15%	**33%**
IM	16%	17%	**34%**	20%
SIM	16%	**47%**	**39%**	18%

Figure 4.10 shows the dissertations topic distribution in different time periods. In the period of 2004 – 2008 the weights were quite similar. 'Geodata Information' was slightly more frequent than the others. In the period of 2009 – 2013 the distribution is still identic, however the weight in 'Geodata Information' decreased and the weight of 'Information and Decision Systems' increased. This trend continued in 2014 – 2018 where this topic is by far the most published topic with 46%. Among time, the topics about 'Implementing Systems' and 'Behavioural Studies' lost weight.

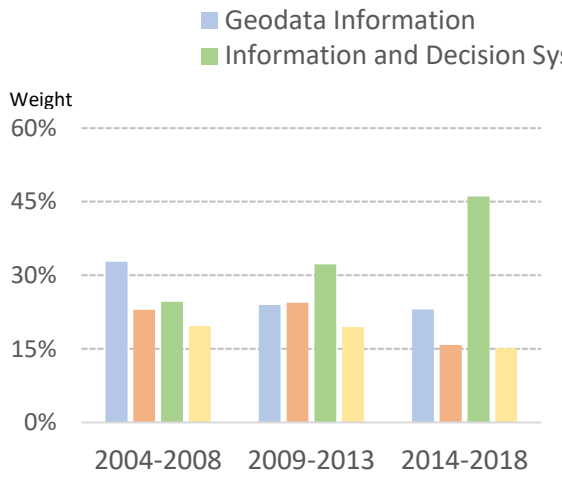

Figure 4.10 - Weight of dissertation published per topic over the years

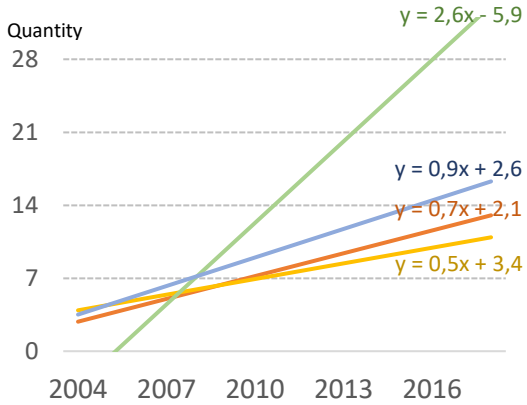

Figure 4.11 - Linear trendline of dissertations published per topic over the years

The previous Figure 4.11 displays a linear trendline of the quantity of thesis published per topic. All topics present a positive, nevertheless 'Information and Decision Systems' reveal a much higher positive trend compared to the remaining topics. Therefore, each slope of the linear equations reveals how much the quantity of dissertation published increase on a variation of one year. For instance, the topic 'Geodata Information' in next year would increase approximately one more dissertation published, according to the linear trend analysis.

4.3.3.2 K-means clusters analysis

The table below shows how many dissertations were published in each cluster. Online Behaviour was the cluster where most dissertations were focused on (51%), followed by Business Analysis with 40%, the remaining cluster, Geodata Analysis, had the lowest quantity of dissertations published (8%).

Table 4.15 - Percentage of dissertations per cluster

Cluster	Cluster Name	Quantity Dissertations	%
1	Geodata Analysis	50	8%
2	Online Behaviour	304	51%
3	Business Analysis	240	40%

The different clusters were examined per field of the dissertations published by NOVA IMS. Table 4.16 shows the weight of each dissertation field per cluster. Interestingly, the cluster 'Geodata Analysis' is mainly related to dissertations in the field of GISS and GT (i.e. 26% and 72% respectively). Moreover, the cluster 'Online Behaviour' presents to be generic since the weights among the dissertation fields are similar (expect for AA). Finally, the cluster 'Business Analysis' is mainly represented by the fields IM and SIM with 35% and 48% respectively.

Table 4.16 - Weight of each dissertation field per cluster

Dissertation Field	Geodata Analysis	Online Behaviour	Business Analysis
Advanced Analytics	0%	1%	5%
Geographic Information Systems & Science	26%	26%	10%
Geospatial Technologies	**72%**	30%	3%
Information Management	0%	19%	**35%**
Statistics and Information Management	2%	23%	**48%**

Figure 4.12 shows the dissertations cluster distribution in different time periods. In the period of 2004 – 2008 the weight of 'Online Behaviour' had the highest weigh (over 60%). In the period of 2009 – 2013 the distribution is still identic, although 'Online Behaviour' cluster lost weight contrary to 'Business Analyst'. Moreover, in the period 2014 – 2018 these two clusters share a similar weight. Finally, the cluster 'Geodata Analysis' during the entire period record the lowest weight.

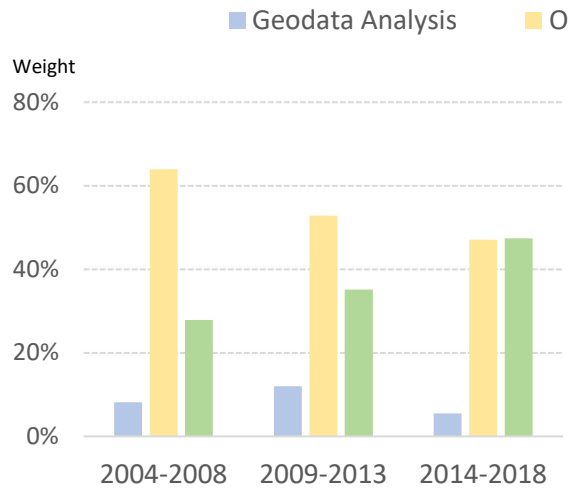

Figure 4.12 - Weight of dissertation published per cluster over the years

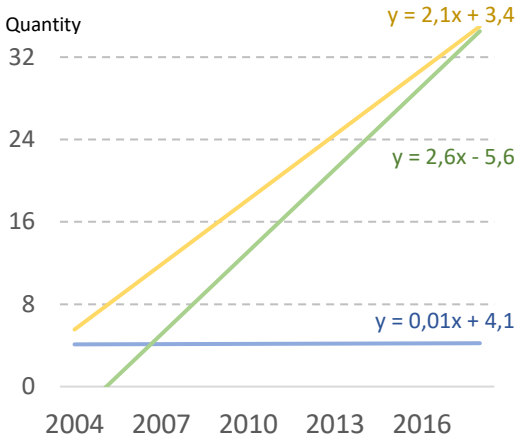

Figure 4.13 - Linear trendline of dissertations published per cluster over the years

The previous Figure 4.13 displays a linear trendline of the quantity of thesis published per cluster. The clusters 'Online Behaviour' and 'Business Analysis' shows a similar positive trend. Contrary, the cluster 'Geodata Analysis' revealed a static trend, which means that there is no projection of an increase, neither decrease, of the amount of thesis published yearly. Therefore, each slope of the linear equations reveals how much the quantity of dissertation published increase on a variation of one year. For instance, the cluster 'Online Behaviour' in next year would increase approximately two more dissertations published, according to the linear trend analysis.

4.3.3.3 Dissertations fields analysis

The collected data is grouped in five dissertation fields: Advanced Analytics, Geographic Information Systems and Science, Geospatial Technologies, Information Management, and Statistics and Information Management. In this section each field will be analysed according to the quantity of dissertations delivered over the years and their weight compared to the remaining fields. Moreover, the top frequent words over the years are presented. Finally, per year is presented how the contained dissertations have been grouped by the founded topics and clusters.

1. Advanced Analytics

The table below shows that the field of AA had the first published dissertation in the last three years of the period, contrary to the remaining fields that published the first dissertation during the first five years. Moreover, AA have the most dissertations published on the last two year.

Table 4.17 - AA and remaining field dissertations published over the years

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
Advanced Analytics	0	0	0	0	0	0	0	0	0	0	0	0	3	6	6	15
Remaining thesis	1	7	12	9	32	38	47	46	47	64	30	48	56	76	66	579
Total	1	7	12	9	32	38	47	46	47	64	30	48	59	82	72	594

*The green highlighted cells refer to the highest record of published dissertations on the field of AA in the period in analysis.

The Table 4.18 indicates that the three the most frequent word are 'algorithm', 'programming', and 'genetic' with the respective frequency 24, 20, and 18. Moreover, AA is the smallest field under study, counts for 3% of the overall dissertations published under study (see Figure 4.14).

Table 4.18 - AA 10 most frequent words and respective count per grouped years

Word Rank	2004-2006	2007-2009	2010-2012	2013-2015	2016-2018
#1	No dissertations Published	No dissertations Published	No dissertations Published	No dissertations Published	algorithm (24)
#2					programming (20)
#3					genetic (18)
#4					model (17)
#5					solutions (17)
#6					different (14)
#7					classification (14)
#8					proposed (14)
#9					results (13)
#10					techniques (13)

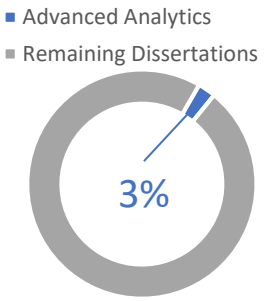

Figure 4.14 - Weight of published dissertations on the field AA.

In the Figure 4.15 is displayed the weights of each topic per year on the field of AA. The dissertations published in this field are only presented in the topics 'Geodata Information' and 'Information and Decision Systems'. In the first two years there is no published dissertations assigned to the topic 'Geodata Information'. Furthermore, in 2018 the two topics that AA contains share the same percentage (i.e. 50%).

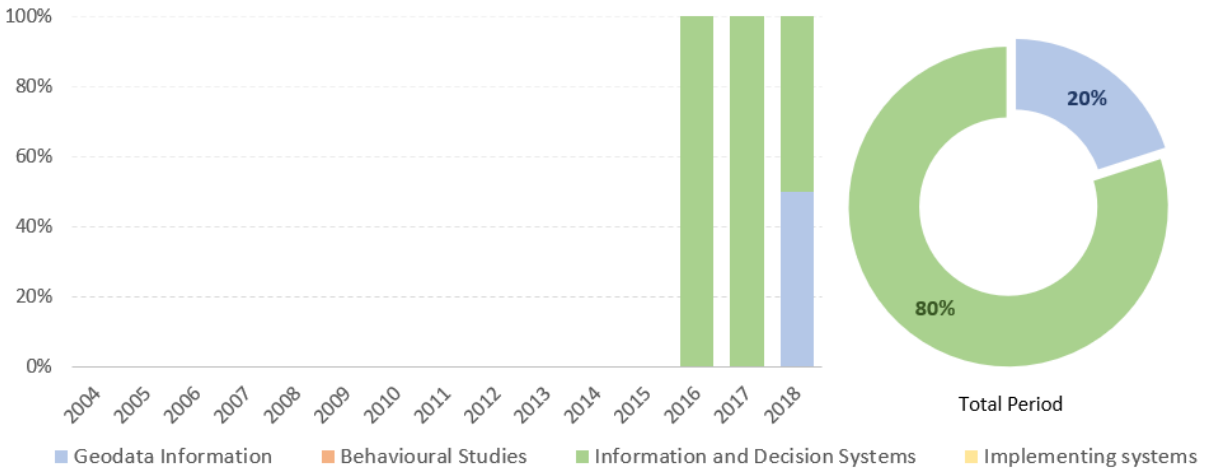

Figure 4.15 - AA weight of dissertation per topic among time

The Figure 4.16 shows the weights of each cluster per year on the field of AA. Over the entire period, there were no dissertations published in the cluster 'Geodata Analysis'. Moreover, in the year 2017 all the dissertations published belonged to the cluster 'Business Analysis'. Therefore, the cluster 'Online Behaviour' have a weight of 20% during the total period.

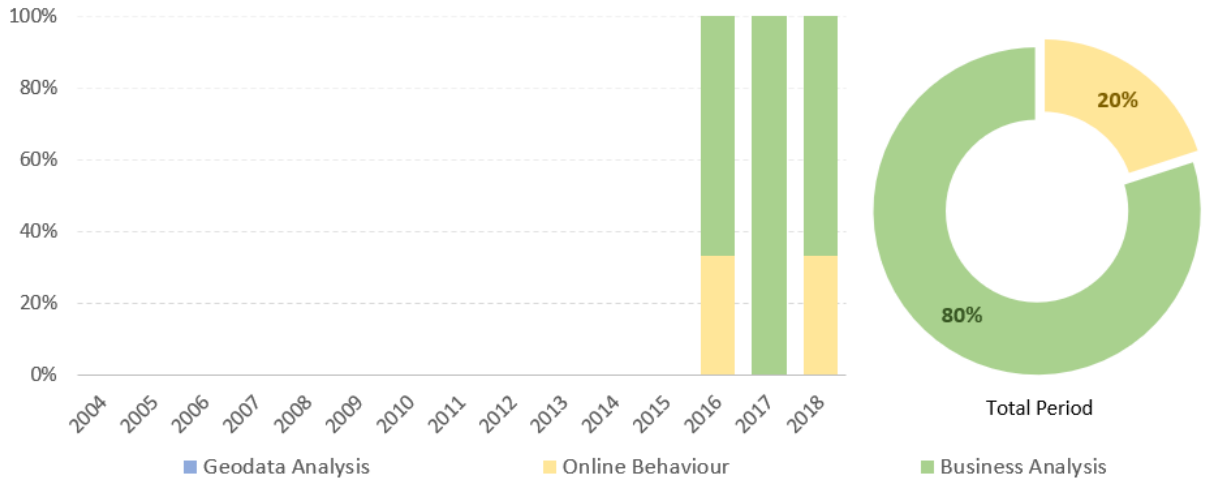

Figure 4.16 - AA weight of dissertation per cluster among time

2. Geographic Information Systems and Science

The table below shows that the field of GISS had the first published dissertation in 2004. In 2008 students published the most dissertations (i.e. 17 dissertations). Among the years, the quantity of published dissertation of this field are decreasing, inversely to the tendency of the remaining thesis.

Table 4.19 - GISS and remaining field dissertations published over the years

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
GISS	1	7	11	8	17	9	12	6	9	11	5	6	7	4	3	116
Remaining thesis	0	0	1	1	15	29	35	40	38	53	25	42	52	78	69	478
Total	1	7	12	9	32	38	47	46	47	64	30	48	59	82	72	594

*The green highlighted cells refer to the highest record of published dissertations on the field of GISS in the period in analysis.

The Table 4.20 indicates that in all periods the keyword 'information' was always the most frequent word. Furthermore, the words 'region', 'public' and 'tools' are exclusive of the period 2016-2018. Although, the general the keywords remained the same over the years, which reveals consistency over time. Lastly, GISS counts for 20% of the overall dissertations published under study (see Figure 4.17).

Table 4.20 - GISS 10 most frequent words and respective count per grouped years

Word Rank	2004-2006	2007-2009	2010-2012	2013-2015	2016-2018
#1	information (32)	information (70)	information (49)	information (47)	maps (20)
#2	analyze (32)	geographical (45)	geographical (36)	data (40)	information (20)
#3	data (26)	development (37)	data (28)	development (31)	area (19)
#4	geographical (26)	data (29)	area (26)	management (30)	analyze (17)
#5	development (26)	area (26)	management (25)	model (25)	data (16)
#6	model (24)	space (26)	model (23)	change (23)	region (15)
#7	space (23)	management (25)	maps (20)	space (23)	public (14)
#8	study (22)	model (22)	development (19)	geographical (21)	management (13)
#9	sig (17)	resources (21)	software (14)	results (20)	tools (13)
#10	area (15)	occupation (21)	based (14)	maps (19)	based (13)

*The green highlighted cells refer to words that appear in more than one grouped year.

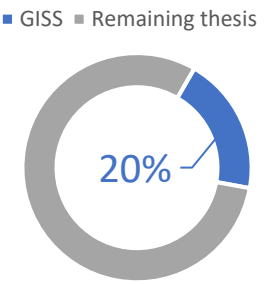

Figure 4.17 - Weight of published dissertations on the field GISS.

In the Figure 4.18 is displayed the weights of each topic per year on the field of GISS. The dissertations published in this field are present in all topics, however the weights are not consistent over time. In the last two years there is no published dissertations assigned to the topic 'Implementing Systems'. Furthermore, in 2018 there is only dissertations published on the topics 'Geodata Information' and 'Behavioural Studies'. Lastly, over the total period, the highest weight is the topic 'Geodata Information' (i.e. 41%) and smallest 'Information and Decision Systems' (i.e. 11%).

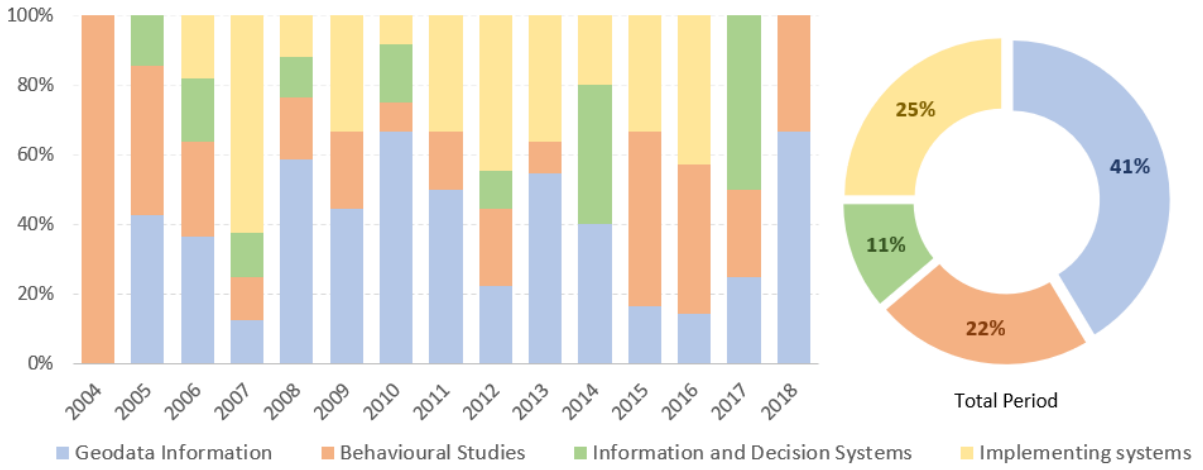

Figure 4.18 - GISS weight of dissertation per topic among time

The Figure 19 shows the weights of each cluster per year on the field of GISS. Apart from the year 2014, the cluster 'Online Behaviour' has the highest weight. Moreover, in the last four years the cluster 'Geodata Analysis' did not had any record. Lastly, the cluster with lower percentage of dissertations over the entire period is 'Geodata Analysis' (i.e. 11%) contrary to 'Online Behaviour' (i.e. 69%).

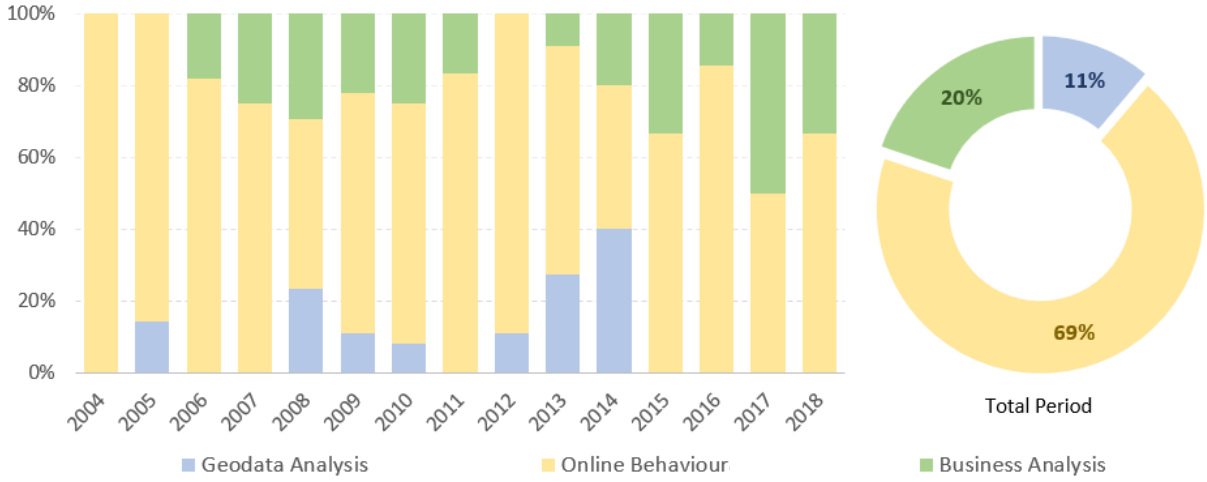

Figure 4.19 - GISS weight of dissertation per cluster among time

3. Geospatial Technologies

The table below shows that the field of GT had the first dissertation published in 2009. In 2011 students published the most dissertations (i.e. 20 dissertations).

Table 4.21 - GT and remaining field dissertations published over the years

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
GT	0	0	0	0	0	15	9	20	11	18	10	14	10	8	19	134
Remaining thesis	1	7	12	9	32	23	38	26	36	46	20	34	49	74	53	460
Total	1	7	12	9	32	38	47	46	47	64	30	48	59	82	72	594

*The green highlighted cells refer to the highest record of published dissertations on the field of GT in the period in analysis.

The Table 4.22 indicates that in all periods the keyword 'data' was always one of the most frequent words. Furthermore, the words 'based', 'performance' and 'city' are exclusive of the period 2016-2018. Although, the general the keywords remained the same over the years, which reveals consistency over time. Lastly, GT counts for 23% of the overall dissertations published under study (see Figure 4.20).

Table 4.22 - GT 10 most frequent words and respective count per grouped years

Word Rank	2004-2006	2007-2009	2010-2012	2013-2015	2016-2018
#1	No dissertations Published	data (30)	data (82)	study (79)	data (76)
#2		change (29)	model (56)	data (74)	study (58)
#3		maps (26)	land (55)	land (66)	land (46)
#4		spatial (23)	development (53)	information (63)	results (45)
#5		information (23)	study (52)	area (62)	maps (41)
#6		land (19)	spatial (50)	spatial (57)	spatial (38)
#7		urban (18)	different (49)	model (55)	based (33)
#8		model (17)	methods (46)	maps (45)	area (33)
#9		development (16)	area (44)	results (43)	performance (31)
#10		study (14)	information (44)	change (43)	city (30)

*The green highlighted cells refer to words that appear in more than one grouped year.

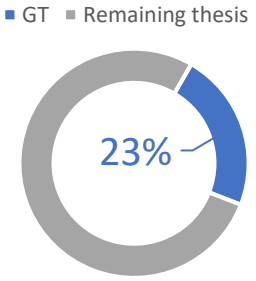

Figure 4.20 - Weight of published dissertations on the field GT.

In the Figure 4.21 is displayed the weights of each topic per year on the field of GT. The dissertations published in this field are present in all topics, however the weights are not consistent over time. The only topic that had dissertations published every year is 'Information and Decision Systems'. Furthermore, in 2017 there is only dissertations published on the topics 'Geodata Information' and 'Information and Decision Systems'. Lastly, over the total period, the highest weight is the topic 'Geodata Information' (i.e. 36%) and smallest 'Behavioural Studies' (i.e. 13%).

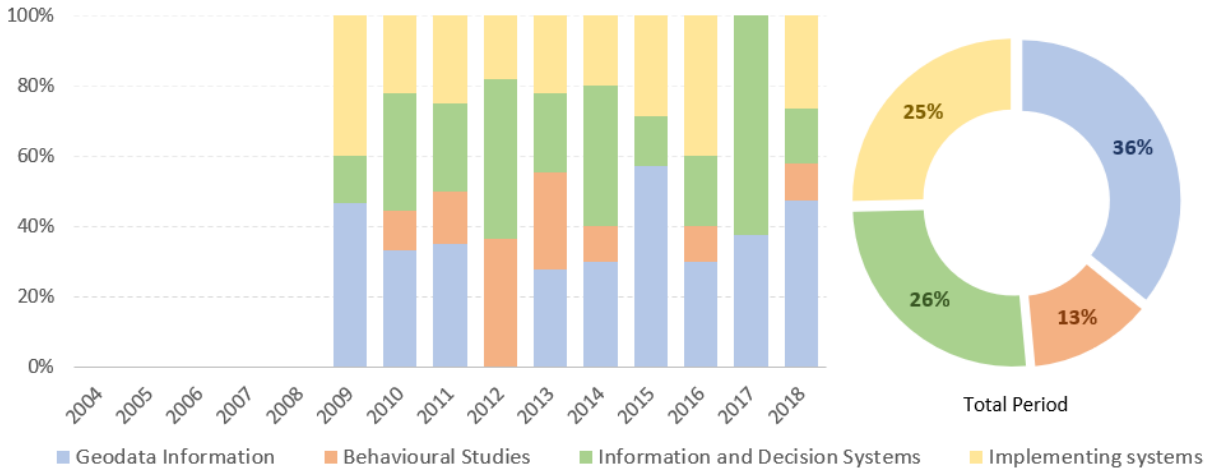

Figure 4.21 - GT weight of dissertation per topic among time

The Figure 4.22 shows the weights of each cluster per year on the field of GT. Over the entire period, the cluster 'Online Behaviour' has the highest weight. Moreover, there was four years the cluster 'Business Analysis' did not had any record. Lastly, the cluster with lower percentage of dissertations over the entire period is 'Business Analysis' (i.e. 4%) contrary to 'Online Behaviour' (i.e. 69%).

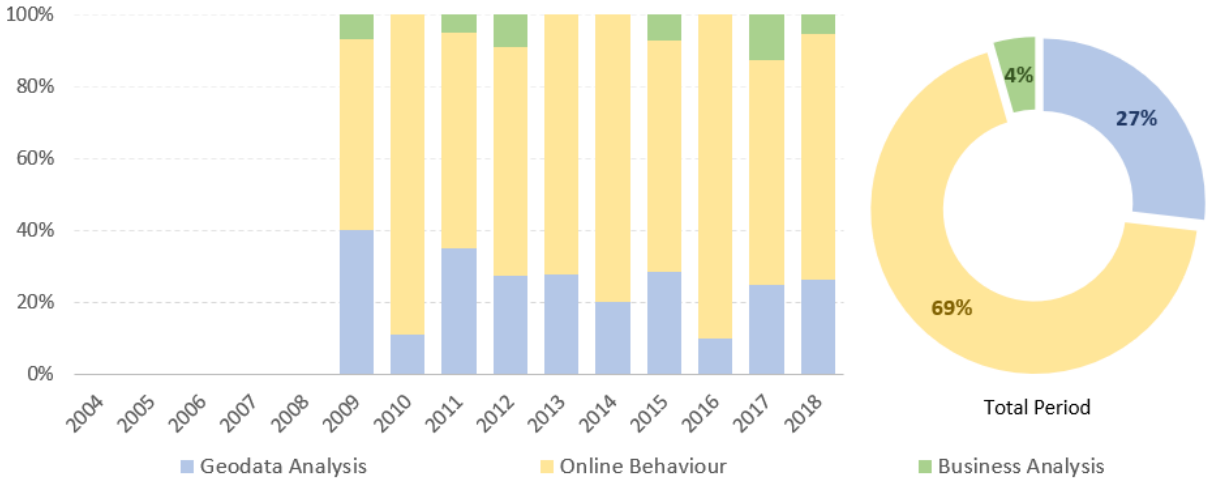

Figure 4.22 - GT weight of dissertation per cluster among time

4. Information Management

The table below shows that the field of IM had the first published dissertation in 2009. In 2017 students published the most dissertations (i.e. 43 dissertations). Among the years, the quantity of published dissertation of this field are increasing.

Table 4.23 - IM and remaining field dissertations published over the years

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
IM	0	0	0	0	0	3	3	0	1	7	10	17	27	43	31	142
Remaining thesis	1	7	12	9	32	35	44	46	46	57	20	31	32	39	41	452
Total	1	7	12	9	32	38	47	46	47	64	30	48	59	82	72	594

*The green highlighted cells refer to the highest record of published dissertations on the field of IM in the period in analysis.

The Table 4.24 indicates that the word 'study' is the only keyword repeated over the periods. Furthermore, the words 'data', 'information', 'management', 'business', 'model', 'organizations' are exclusively mentioned in the last two periods. The keywords of the first two periods seem both to be unique from the other periods. However, the quantity of dissertations published in the first two periods is much lower. Lastly, IM counts for 24% of the overall dissertations published under study (see Figure 4.23).

Table 4.24 - IM 10 most frequent words and respective count per grouped years

Word Rank	2004-2006	2007-2009	2010-2012	2013-2015	2016-2018
#1	No dissertations Published	region (8)	sector (12)	information (60)	data (161)
#2		demographic (8)	innovation (12)	model (54)	information (150)
#3		population (6)	customers (10)	study (42)	management (105)
#4		possible (5)	relationship (10)	companies (41)	technologies (99)
#5		mortality (4)	study (9)	services (37)	business (95)
#6		birth (4)	loyalty (8)	development (36)	results (95)
#7		level (4)	portugal (8)	organizations (36)	process (91)
#8		study (4)	services (7)	management (36)	study (90)
#9		description (4)	firms (7)	data (35)	model (88)
#10		investigation (3)	prone (6)	business (34)	organizations (82)

*The green highlighted cells refer to words that appear in more than one grouped year.

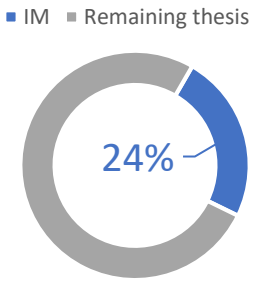

Figure 4.23 - Weight of published dissertations on the field IM.

In the Figure 4.24 is displayed the weights of each topic per year on the field of IM. The topic 'Geodata Information' only have records on the last four years. Moreover, the topic 'Information and Decision System' was the only topic on the year 2012 and from then on always kept the highest weight. Lastly, over the total period, the highest weight is the topic 'Information and Decision Systems' (i.e. 55%) and the remaining topics have a similar weight.

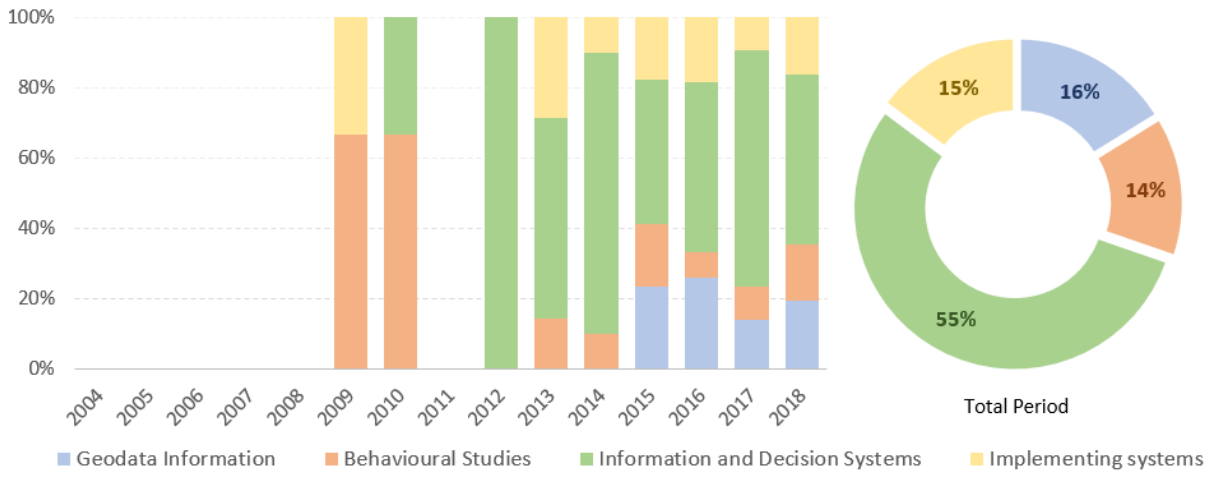

Figure 4.24 IM weight of dissertation per topic among time

The Figure 4.25 shows the weights of each cluster per year on the field of IM. Over the entire period, there were no dissertations published in the cluster 'Geodata Analysis'. Moreover, in the last two years of the period, the cluster 'Business Analysis' had higher weight compared. Lastly, the cluster 'Online Behaviour' have a weight of 41% during the total period.

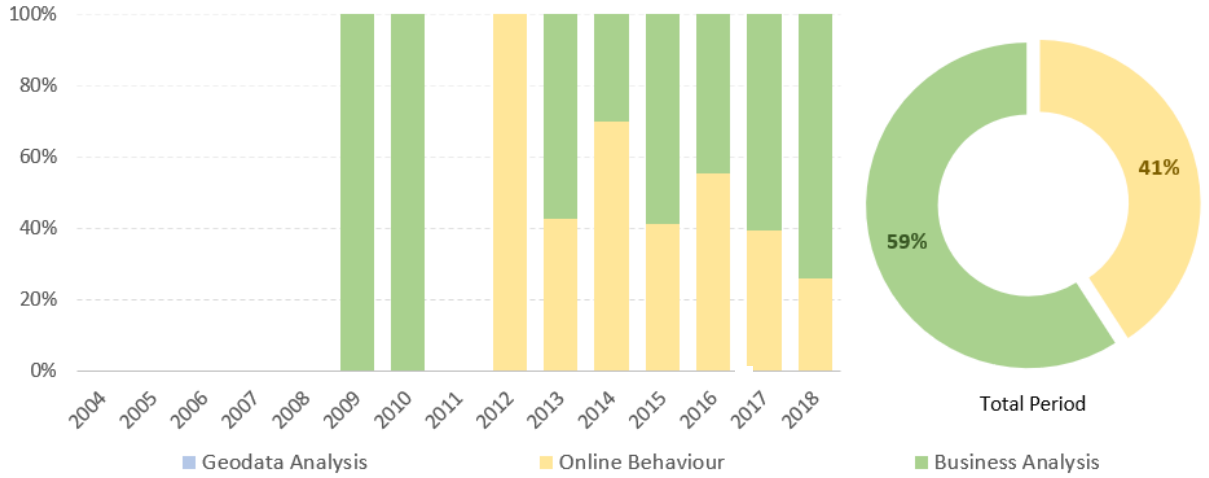

Figure 4.25 - IM weight of dissertation per cluster among time

5. Statistics and Information Management

The table below shows that the field of SIM had the first published dissertation in 2006. In 2013 students published the most dissertations (i.e. 28 dissertations).

Table 4.25 - SIM and remaining field dissertations published over the years

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
SIM	0	0	1	1	15	11	23	20	26	28	5	11	12	21	13	187
Remaining thesis	1	7	11	8	17	27	24	26	21	36	25	37	47	61	59	407
Total	1	7	12	9	32	38	47	46	47	64	30	48	59	82	72	594

*The green highlighted cells refer to the highest record of published dissertations on the field of SIM in the period in analysis.

The Table 4.26 indicates that in all periods the words 'risk', 'results', 'satisfaction' and 'impact' were exclusive of the period 2016-2018. Furthermore, the words 'model', 'analyze', 'data' and 'development' are repeated over the last three periods. The keywords of the first period seem to be unique from the other periods. However, the quantity of dissertations published in the first period is much lower. Lastly, SIM counts for 31% of the overall dissertations published under study (see Figure 4.26).

Table 4.26 - SIM 10 most frequent words and respective count per grouped years

Word Rank	2004-2006	2007-2009	2010-2012	2013-2015	2016-2018
#1	ict (6)	information (28)	information (106)	model (76)	model (75)
#2	contribute (4)	analyze (27)	data (82)	analyze (68)	analyze (68)
#3	understand (4)	data (27)	model (80)	study (50)	study (59)
#4	determinants (4)	development (23)	development (80)	insurance (45)	data (52)
#5	adoption (4)	companies (22)	management (74)	development (39)	risk (44)
#6	diffusion (4)	job (22)	analyze (57)	data (37)	results (42)
#7	consensual (2)	customers (19)	organizations (56)	information (34)	satisfaction (40)
#8	technologies (2)	planning (18)	study (56)	based (31)	development (38)
#9	information (2)	study (18)	based (51)	important (31)	impact (34)
#10	communication (2)	area (17)	process (47)	job (31)	management (33)

*The green highlighted cells refer to words that appear in more than one grouped year.

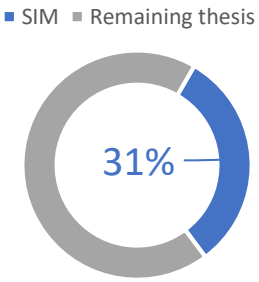

Figure 4.26 - Weight of published dissertations on the field SIM.

In the Figure 4.27 is displayed the weights of each topic per year on the field of SIM. Apart from the years 2006, 2017, 2014, and 2017, the dissertations published in this field are present in all topics, however the topic 'Information and Decisions Systems' represents most of the field followed by 'Behavioural Studies'. Moreover, the remaining topics kept a similar low weight among time. Lastly, over the total period, the topic 'Information and Decisions Systems' contains almost half of the dissertations published (i.e. 48%).

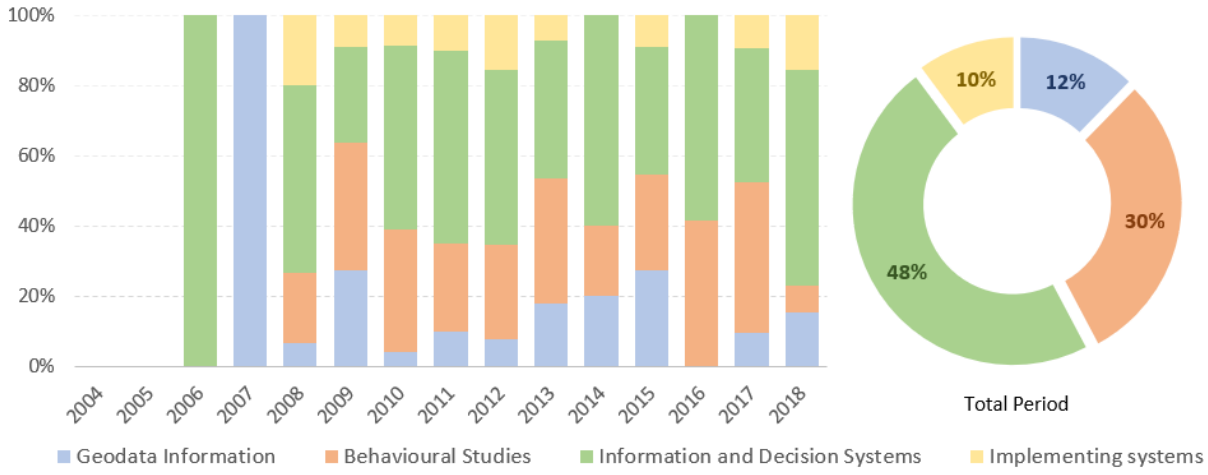

Figure 4.27 - SIM weight of dissertation per topic among time

The Figure 4.28 shows the weights of each cluster per year on the field of SIM. The cluster 'Geodata Analysis' only had a record in the year 2013. Moreover, in the last seven years the cluster 'Business Analysis' always kept a higher weight compared to 'Online Behaviour'. Lastly, the distribution of the clusters did not have any abrupt change, over the entire period 'Business Analysis' lead the field with 61% of thesis published.

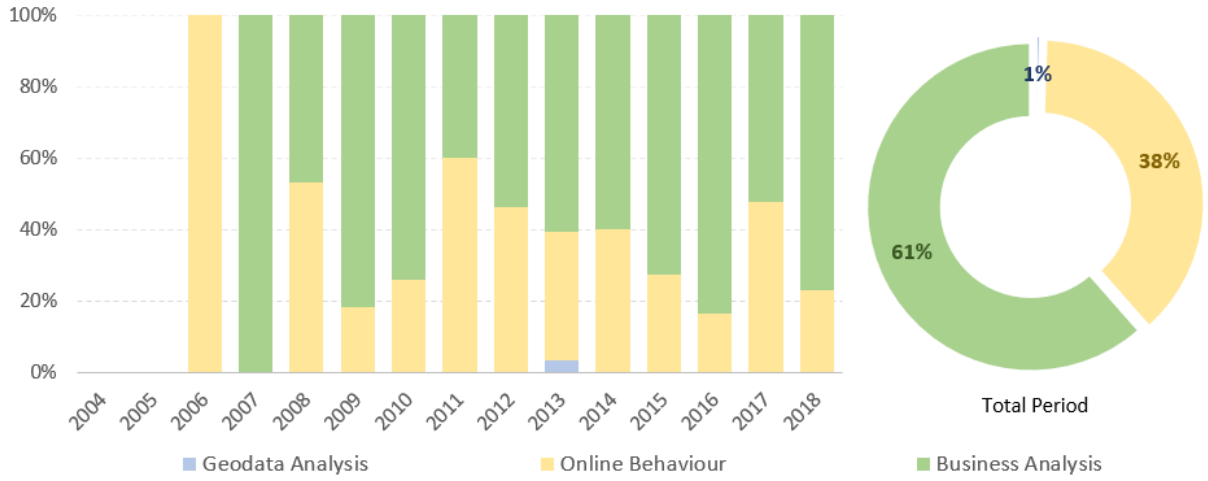

Figure 4.28 - SIM weight of dissertation per cluster among time

5. DISCUSSION

Current study applied text mining techniques on published dissertations of students of NOVA IMS. The research question was: *What are the current patterns and trends in the dissertations published by NOVA-IMS students?* To answer the main research question, first the two sub-research questions were examined.

Sub-research question 1: How can text mining be applied in the educational context?

An exhaustive literature review was executed to understand the history and the concept of text mining. Thereafter, it was examined how text mining techniques were applied in the educational context. Then, different steps were conducted to apply text mining to master dissertations of NOVA IMS.

In this study, 594 dissertations that were published between March 2004 and May 2018 were collected to examine patterns and trends. After gathering the abstracts from the different dissertations, the dataset needed to go through a text mining process flow in order to structure the dataset. First, the data language was uniformized in English since many dissertations were written in Portuguese. Thereafter, pre-processing of the data played a major role since the text formats contained noise, such as numbers, characteristics and words without a significant meaning (Chakraborty, Pagolu, & Garla, 2013). Therefore, the amount of unique words was reduced through removing non-alphabetic characters, stop-words and applying stemming technique.

Sub-research question 2: What are the outcomes from applying text mining on the dissertations of students in NOVA IMS?

After the dataset was prepared, two cluster methods were applied on the dataset. First, topic modelling, a probabilistic generative model, was applied with the LDA algorithm. The suggested optimal number of topics was four. The PCs depicted two groups of topics that showed controversial characteristics which was important for inferring the topics. Finally, the following topics were suggested; Geodata Information, Behavioural Studies, Information and Decision Systems and Implementing Systems. The two of the PCs were interpreted as personal level versus institutional level and as special networks versus system networks. Thereafter, k-means clustering was applied. The optimal number of clusters in the k-means algorithm was three. After a deep examination, the clusters were indicated as; Geodata Analysis, Online Behaviour and Business Analysis.

Main research question: What are the current patterns and trends in the dissertations published by NOVA-IMS students?

Patterns and trends of Topic Modelling

The analysis of the above-mentioned topics shows that 'Geodata Information' is mainly related to the fields of Geographic Information Systems and Science and Geospatial Technologies, which is according to the expectations. Furthermore, 'Behavioural Studies' is mostly represented by the field of Statistics and Information Management. Moreover, 'Information and Decision System' is strongly represented

in the field of Statistics and Information Management and also Information Management. Lastly, 'Implementing System' is almost equally distributed among the dissertation fields.

Furthermore, a trend is detected in the topics of the dissertations. In the beginning of the period the topics were similar distributed. Although, over the years 'Information and Decision Systems' became the most frequent topic.

Finally, the quantity of dissertations published among the year reveal a positive trend for each topic. However, 'Information and decision Systems' indicates a much stronger positive trend compared to the remaining topics that share similar growth rate.

Patterns and trends of Clustering

The analysis of the referred clusters demonstrates that 'Geodata Analysis' is mostly represented by the field of Geospatial Technologies, followed by Geographic Information Systems and Science, which is according to the expectations. Thereafter, 'Online Behaviour' is more equally distributed among the different dissertation fields. Finally, 'Business Analysis' is strongly represented by the field of Statistics and Information Management and Information Management.

Furthermore, a trend is identified in the clusters of the dissertations. In the beginning of the period, the cluster 'Online Behaviour' represented the majority of the dissertations. Although, over the years 'Business Analysis' gained weight and became balanced with this cluster. Moreover, 'Geodata Analysis' remained the less represented cluster.

Finally, the quantity of dissertations published among the year has shown a positive and similar trend for the clusters 'Online Behaviour' and 'Business Analysis'. Nevertheless, the cluster 'Geodata Analysis' revealed a static trend.

Below the patterns and trends per dissertation field are discussed.

Advanced Analytics

Advanced Analytics concerns the smallest field of the dataset. Results show that dissertations have mainly been related to the subjects of 'Business Analysis' and 'Information and Decision Systems'. Furthermore, no trends were identified due to the short period of the collected data.

Geographic Information Systems and Science

The quantity of dissertations of Geographic Information Systems and Science decreased over the years, which is contrary to the behaviour of the other fields. However, it has remained a relevant weight over all the dissertations, namely 20%. Furthermore, the general the keywords remained the same over the years which reveals consistency of the published subjects over time. Furthermore, these dissertations have been concerning the subjects of 'Geodata Information' and 'Online Behaviour'.

Geospatial Technologies

The subjects of dissertations in the field of Geospatial technologies were consistent over time. Furthermore, the distribution of the topics of the dissertations are similar, which indicates how generic

the subjects under study can be. However, the cluster analysis showed that the field is highly related to ‘Online Behaviour’ subject.

Information Management

The dissertations in the field of Information Management has increased over the years. During the last 6 years, the dissertations manifested the frequent use of the keywords data, business, model, and organizations. Moreover, the field is mainly described by the topic ‘Information and Decision Systems’ and the clusters ‘Online Behaviour’ and ‘Business Analysis’. Contrary, the subject ‘Geodata’ do not have much relation with this field.

Statistics and Information Management

Statistics and Information Management consists of approximately one third of the dataset. It is observable that in the last period new keywords emerged, namely ‘risk’, ‘results’, ‘satisfaction’ and ‘impact’. Moreover, the subject ‘Geodata’ does not appear in this field. Besides, Statistics and Information Management is mostly described by the topic ‘Information and Decision Systems’ and the cluster ‘Business Analysis’.

6. CONCLUSION

This study applied text mining techniques on dissertations published by students from NOVA IMS. The results of this study give new insights on trends and patterns in the academic field, which could solely be explored by applying text mining techniques. These trends and patterns over the last years could create opportunities for universities and play a central role in the future career of their students. Concerning the limitations, this study is a first essential step for NOVA IMS in examining published dissertations. Since this study shows which topics and clusters are trending, the university can use this knowledge as an asset.

7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Text mining, even data driven, can be very challenging since the input data is often vague and inconsistent. The textual contents have a lot of noise that obstructs the intended goal: extraction of knowledge. The noise is not only represented by the words and/or characters around the concepts of each sentence but also in the use of irony, different dialects, words with different meanings etc. For that reason, the entire process of text mining requires an exhaustive investment on the preparation of the collected data.

Nevertheless, the major limitation of the research is the subjectivity around the interpretation of the topics and clusters. The outcomes from the performed algorithms only suggest terms that are representative of each topic or cluster. It is the user that needs to explore and suggest an interpretation for that outcome. Although, as applied on this research, there are different techniques and visualization tools that help the user to frame each topic.

Moreover, a fundamental aspect on the entire process is the sample size. In this study, the quantity of samples was not sufficient which might have caused problems in the quality of grouping the documents per topic and cluster.

Furthermore, the content that represented each document was limited to the abstracts, which only represents a small portion of the entire content. Westergaard, Stærfeldt, Tønsberg, Jensen and Brunak (2018) made a comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. They conclude that text mining of full-text articles significantly transcends the quality of outcomes compared with using abstracts only.

Despite the mentioned limitations, the purpose of this research was to encourage the use of text mining techniques to gain knowledge from the dissertations. In fact, this study only presents a simple application of text mining in the educational context. Therefore, different approaches, for instance the use of more samples, more investment in the preparing data phase, trying different algorithms and respective parameters or even applying supervised learning are recommended for future research.

8. BIBLIOGRAPHY

Aggarwal, C. C., & Zhai, C. X. (2013). *Mining text data*. *Mining Text Data*. https://doi.org/10.1007/978-1-4614-3223-4

Allahyari, M., Trippe, E. D., & Gutierrez, J. B. (2017). A Brief Survey of Text Mining : Classification , Clustering and Extraction Techniques. *ArXiv:1707.02268 [Cs]*.

Ananiadou, S., Kell, D. B., & Tsujii, J. ichi. (2006). Text mining and its potential applications in systems biology. *Trends in Biotechnology*, *24*(12), 571–579. https://doi.org/10.1016/j.tibtech.2006.10.002

Arthur, D., & Vassilvitskii, S. (2007). K-Means++: the Advantages of Careful Seeding. In *Proc ACM-SIAM symposium on discrete algorithms.* https://doi.org/10.1145/1283383.1283494

Baba, S. W., & Kumar, R. S. (2016). *Data Mining : Text Classification System for Classifying Abstracts of Research Papers*.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. *New York*. https://doi.org/10.1080/14735789709366603

Bellman, R. E. (1961). Adaptive control processes: A guided tour. *Princeton University Press*. https://doi.org/69756

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation David. *Journal of Machine Learning Research2*. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Boulis, C., & Ostendorf, M. (2005). Text Classification by Augmenting the Bag-of-Words Representation with Redundancy-Compensated Bigrams *. In *Proc., Int. Workshop Feature Selection in Data Mining*.

Brázdil, J. (2016). Dimensionality reduction methods for vector spaces.

Burges, C. J. C. (2009). Dimension Reduction: A Guided Tour. *Foundations and Trends® in Machine Learning*. https://doi.org/10.1561/2200000002

Chakraborty, G., Pagolu, M., & Garla, S. (2013). Text Mining and Analysis. In *Text mining and Analysis. Practical method, examples and case studies using SAS*.

Christou, D. (2016). *Feature extraction using Latent Dirichlet Allocation and Neural Networks: A case study on movie synopses*. Retrieved from http://arxiv.org/abs/1604.01272

Coussement, K., & Van den Poel, D. (2008). Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems*. https://doi.org/10.1016/j.dss.2007.10.010

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*. https://doi.org/10.1145/1151030.1151032

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Fayyad, U., Reina, C., & Bradley, P. S. (1998). Initialization of Iterative Refinement Clustering Algorithms. In *KDD '98*. https://doi.org/10.1393/ncc/i2014-11719-1

Feldman, R., Fresko, M., Hirsh, H., Aumann, Y., Liphstat, O., Schler, Y., & Rajman, M. (1998). Knowledge Management: A Text Mining Approach. *Proc of the 2nd Int Conf on Practical Aspects of Knowledge Management （PAKM98*. https://doi.org/10.1016/j.aqpro.2013.07.003

Feldman, R., & Sanger, J. (2006). *The Text Mining Handbook*. https://doi.org/10.1017/CBO9780511546914

Gálvez, R. H., & Gravano, A. (2017). Assessing the usefulness of online message board mining in automatic stock prediction systems. *Journal of Computational Science*. https://doi.org/10.1016/j.jocs.2017.01.001

Google - Gmail (2018). Google Products Gmail. Retrieved from https://www.blog.google/products/gmail/subject-write-emails-faster-smart-compose-gmail/

Goyal, M., & Vohra, R. (2012). Applications of Data Mining in Higher Education, *9*(2), 113–120.

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics -*. https://doi.org/10.3115/1034678.1034679

Heinrich, G. (2008). Parameter Estimation for Text Analysis. *Introduction of Gibbs Sampling*. https://doi.org/10.2514/2.3375

Hu, D. (2009). Latent Dirichlet Allocation for Text, Images, and Music. *University of California, San Diego*.

Huang, A. (2008). Similarity measures for text document clustering. In *New Zealand Computer Science Research Student Conference (NZCSRSC)*. https://doi.org/10.1109/ICDMW.2009.61

Krassmann, A. L., Herpich, F., Bercht, M., & Cazella, S. C. (2017). Analyzing trends in academic papers about ubiquitous virtual worlds in education using text mining. *International Journal for Innovation Education and Research*.

Kroeze, J. H., Matthee, M. C., & Bothma, T. J. D. (2003). Differentiating Data- and Text-Mining Terminology. In *Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology*. https://doi.org/10.4102/sajim.v6i4.353

Manago, M., & Auriol, E. (1996). Using data mining to improve feedback from experience for equipment in the manufacturing & transport industries. In *IEE Colloquium (Digest)*.

McLachlan, G. J., & Krishnan, T. (2007). *The EM Algorithm and Extensions: Second Edition*. *The EM Algorithm and Extensions: Second Edition*. https://doi.org/10.1002/9780470191613

Miner, G., Elder, J., Hill, T., Nisbet, R., Delen, D., & Fast, A. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*.

Mitchell, T. M. (1997). Bayesian learning. In *Machine Learning*. https://doi.org/10.1016/j.rpsmen.2012.11.001

Nie, B., & Sun, S. (2017). Using Text Mining Techniques to Identify Research Trends: A Case Study of Design Research. *Applied Sciences*. https://doi.org/10.3390/app7040401

NOVA IMS (2018). Universidade NOVA de Lisboa. Retrieved from http://www.novaims.unl.pt/quem-somos/

KMWorld Magazine and SAS (2013). From Big Data to Meaningful Information.

Padhy, N., Mishra, D. P., & Panigrahi, R. (2012). The Survey of Data Mining Applications and Feature Scope. *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3,Page No-43 June 2012*. https://doi.org/10.5121/ijcseit.2012.2303

Pang, B., & Lee, L. (2009). Opinion mining and sentiment analysis. *Computational Linguistics*. https://doi.org/10.1162/coli.2009.35.2.311

Reddy, C. L., & Venkatadri, M. (2011). A Review on Data mining from Past to the Future. *International Journal of Computer Applications*. https://doi.org/10.5120/1961-2623

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*. https://doi.org/10.1109/TSMCC.2010.2053532

Romero, C., Ventura, S., Baker, Rs., & Pechenizkiy, M. (2010). *Handbook of Educational Data Mining (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*. *Lavoisierfr*. https://doi.org/10.1201/b10274

RUL (n.d.). Information Management School. Retrieved from http://www.run.unl.pt/

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*. https://doi.org/10.1016/0306-4573(88)90021-0

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval. Introduction to modern information retrieval*. https://doi.org/10.3233/978-1-61499-289-9-1124

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*. https://doi.org/10.1145/361219.361220

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization, *34*(1), 1–47.

Sievert, C., & Shirley, K. E. (2014). LDAvis : A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. https://doi.org/10.1.1.100.1089

Simoudis, E. (1996). Reality check for data mining. *IEEE Expert-Intelligent Systems and Their Applications*. https://doi.org/10.1109/GlobalSIP.2016.7906004

Steyvers, M., & Griffiths, T. L. (2006). Probabilistic topic models. In *In Landauer, T., McNamara, D., Dennis, S., and Kintsch, W., editors, Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum. Tang, Z. and MacLennan, J*. https://doi.org/10.1109/TKDE.2009.122

Sullivan, D. (2001). *Document Warehousing and Text Mining. Techniques for Improving Business Operations, Marketing, and Sales*. John Wiley & Sons, Inc. New York, NY, USA ©2001.

Sulova, S., & Nacheva, R. (2017). *Using Text Mining To Classify Research Papers*. https://doi.org/10.5593/sgem2017/21/S07.083

Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *IJACSA) International Journal of Advanced Computer Science and Applications*.

Tang, J., Alelyani, S., & Liu, H. (2014). Feature Selection for Classification: A Review. In *Data Classification: Algorithms and Applications*. https://doi.org/10.1111/j.1749-6632.2008.03634.x

Wagstaf, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). Constrained K-means Clustering with Background Knowledge. In *Eighteenth International Conference on Machine Learning*. https://doi.org/10.1109/TPAMI.2002.1017616

Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. In *Proceedings of the 14th conference on Computational linguistics* -. https://doi.org/10.3115/992424.992434

Westergaard, D., Stærfeldt, H. H., Tønsberg, C., Jensen, L. J., & Brunak, S. (2018). A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1005962

Xu, Y., & Reynolds, N. (2012). Using Text Mining Techniques to Analyze Students' Written Responses to a Teacher Leadership Dilemma. *International Journal of Computer Theory and Engineering*. https://doi.org/10.7763/IJCTE.2012.V4.535

Yadav, S., Bharadwaj, B., & Pal, S. (2012). Mining Education Data to Predict Student's Retention: A comparative Study. *International Journal of Computer Science and Information Security*.

Yan, J., Liu, N., Zhang, B., Yan, S., Chen, Z., Cheng, Q., & Ma, W. (2005). OCFS: optimal orthogonal centroid feature selection for text categorization. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*. https://doi.org/10.1145/1076034.1076058

Zanasi, A. (2009). Virtual weapons for real wars: Text mining for national security. In *Advances in Soft Computing*. https://doi.org/10.1007/978-3-540-88181-0_7

9. APPENDIX

9.1 Python Scripts

9.1.1 Data collection – Webscraping

Pulling from RUN all the links where each dissertations content is stored:

```
#import libraries
from bs4 import BeautifulSoup
import requests
import pandas as pd

#parameters
url_prefix = "https://run.unl.pt/handle/10362/2063/simple-search?query=NOVA+IMS&filter_field_1=type&filter_type_1" \
             "=equals&filter_value_1=masterThesis&sort_by=dc.date.issued_dt&order=desc&rpp=100&etal=0&start="
url_sufix = ['0', '100', '200', '300', '400', '500'] #6 pages > ~600thesis
res = []

#data_extraction
for url_num in url_sufix:
    url = url_prefix + url_num
    r = requests.get(url)
    soup = BeautifulSoup(r.content, "html.parser")
    links = soup.find_all("a")
    for link in links:
        res.append((link.get("href"), link.text))

#data_storage
my_df = pd.DataFrame(res)
my_df.to_csv('01_Web_Scraping_output.csv', index=False, header=False)
```

From each dissertations link, extract and store the title, author, keywords, date, master field, and abstract:

```
#import libraries
from bs4 import BeautifulSoup
import requests
import csv
import xlwt
from tempfile import TemporaryFile

#parameters
ids = []
titles = []
authors = []
keywordss = []
dates = []
masters = []
abstracts = []

with open('02_1_links.csv', 'r') as f:
  reader = csv.reader(f)
  your_list = list(reader)

#data_extraction
i = 0
for x in your_list:
    url = your_list[i]
    r = requests.get(url[0])
    soup = BeautifulSoup(r.content, "html.parser")
    gg_data = soup.find_all("table", {"class": "table itemDisplayTable"})

    for item in gg_data:
        title = item.contents[1].text
        author = item.contents[3].text
        keywords = item.contents[7].text
        date = item.contents[9].text
        master = item.contents[11].text
        abstract = item.contents[13].text

        ids.append(i)
        titles.append(title)
        authors.append(author)
        keywordss.append(keywords)
        dates.append(date)
        masters.append(master)
        abstracts.append(abstract)
```

```
        i+=1

#data_storage
book = xlwt.Workbook()
sheet1 = book.add_sheet('output')

for i,e in enumerate(ids):
    sheet1.write(i,0,e)

for i,e in enumerate(titles):
    sheet1.write(i,1,e)

for i, e in enumerate(authors):
    sheet1.write(i, 2, e)

for i, e in enumerate(keywordss):
    sheet1.write(i, 3, e)

for i, e in enumerate(dates):
    sheet1.write(i, 4, e)

for i, e in enumerate(masters):
    sheet1.write(i, 5, e)

for i, e in enumerate(abstracts):
    sheet1.write(i, 6, e)


name = "02_1_Web_Scraping_output.xls"
book.save(name)
book.save(TemporaryFile())
```

9.1.2 Pre-processing – Language detection

Language detection per document:

```
#import libraries
from langdetect import detect
import xlwt
from tempfile import TemporaryFile

#parameters
languages = []

with open('03_Abstracts.csv', 'r') as f:
    abstracts = f.read().splitlines()
    for abs in abstracts:
        languages.append(detect(abs))


#data_storage
book = xlwt.Workbook()
sheet1 = book.add_sheet('output')

for i,e in enumerate(languages):
    sheet1.write(i,0,e)

name = "03_Languages.xls"
book.save(name)
book.save(TemporaryFile())
```

9.1.3 Pre-processing – Language translation

Translation of the Portuguese contents to English detection per document:

```
#Import the necessary packages
from googletrans import Translator
import pandas

years_list = [2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018]

# TRANSLATOR ------------------------------------------------------------------------------
def text_translate(trans_text, trans_from, trans_to):
    translator = Translator()
    trans_text = translator.translate(trans_text, src=trans_from, dest=trans_to).text
    return trans_text


# Sampling the dataset ----------------------------------------
df_dataset = pandas.read_csv('0_dataset.csv', sep=';', header=0) #import_dataset

for year_num in years_list:
    df_sample_pt = df_dataset.loc[(df_dataset['Language'] == 'PT')&(df_dataset['Year'] == year_num)] #filter_dataset
    print(df_sample_pt)
    df_sample_pt['Abstract'] = df_sample_pt.Abstract.map(lambda x: ' '.join([text_translate(y,'pt','en') for y in x.split(' ')]))
    df_sample_pt.to_pickle('df_PT_translated_' + str(year_num) +'.pkl')



#Import the necessary packages/modules
import pandas

pkl_list = ['df_PT_translated_2004', 'df_PT_translated_2005', 'df_PT_translated_2006', 'df_PT_translated_2007',
            'df_PT_translated_2008', 'df_PT_translated_2009', 'df_PT_translated_2010', 'df_PT_translated_2011',
            'df_PT_translated_2012', 'df_PT_translated_2013', 'df_PT_translated_2014', 'df_PT_translated_2015',
            'df_PT_translated_2016', 'df_PT_translated_2017', 'df_PT_translated_2018']

df_PT_translated_2004 = pandas.read_pickle('df_PT_translated_2004.pkl')
df_PT_translated_2005 = pandas.read_pickle('df_PT_translated_2005.pkl')
df_PT_translated_2006 = pandas.read_pickle('df_PT_translated_2006.pkl')
df_PT_translated_2007 = pandas.read_pickle('df_PT_translated_2007.pkl')
df_PT_translated_2008 = pandas.read_pickle('df_PT_translated_2008.pkl')
df_PT_translated_2009 = pandas.read_pickle('df_PT_translated_2009.pkl')
df_PT_translated_2010 = pandas.read_pickle('df_PT_translated_2010.pkl')
df_PT_translated_2011 = pandas.read_pickle('df_PT_translated_2011.pkl')
df_PT_translated_2012 = pandas.read_pickle('df_PT_translated_2012.pkl')
df_PT_translated_2013 = pandas.read_pickle('df_PT_translated_2013.pkl')
df_PT_translated_2014 = pandas.read_pickle('df_PT_translated_2014.pkl')
df_PT_translated_2015 = pandas.read_pickle('df_PT_translated_2015.pkl')
df_PT_translated_2016 = pandas.read_pickle('df_PT_translated_2016.pkl')
df_PT_translated_2017 = pandas.read_pickle('df_PT_translated_2017.pkl')
df_PT_translated_2018 = pandas.read_pickle('df_PT_translated_2018.pkl')

frames = [df_PT_translated_2004, df_PT_translated_2005, df_PT_translated_2006, df_PT_translated_2007,
          df_PT_translated_2008, df_PT_translated_2009, df_PT_translated_2010, df_PT_translated_2011,
          df_PT_translated_2012, df_PT_translated_2013, df_PT_translated_2014, df_PT_translated_2015,
          df_PT_translated_2016, df_PT_translated_2017, df_PT_translated_2018]

result = pandas.concat(frames)
print(result)

# result.to_pickle('df_PT_translated_total.pkl')
```

9.1.4 Pre-processing – Dimension reduction

IMPORT LIBRARIES ----------

```
#Import the necessary packages/modules
import matplotlib.pyplot as plt
from __future__ import print_function
# from A_defining_treatment import tokenize, stemming, remove_stopwords, text_translate
import numpy as np
import pandas
from nltk.stem.snowball import SnowballStemmer
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from itertools import islice
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
from matplotlib import cm
from sklearn.metrics import silhouette_samples
import re
import string
```

PARAMETERS + LIST_CREATION ----------

```
# Parameter
language = "english"
stemmer = SnowballStemmer("english")
pandas.options.mode.chained_assignment = None
stop_words = stopwords.words('english')

# Adding_info
print(string.punctuation)
new_punctuation = '—’‘“”€©®°.'
remove_punctuation = string.punctuation + new_punctuation

stop_words.append('use')
print(stop_words)
```

IMPORT DATA ----------

```
# Sampling the dataset ----------------------------------------
df_dataset = pandas.read_csv('0_dataset.csv', sep=';', header=0) #import_dataset
df_sample = df_dataset.loc[(df_dataset['Language'] == 'EN')] #filter_dataset
```

Import transalted Portuguese > English Abstracts ----------

```
df_PT_translated_total = pandas.read_pickle('df_PT_translated_total.pkl')
frames = [df_sample, df_PT_translated_total]
df_sample = pandas.concat(frames)

#sorting
df_sample = df_sample.sort_values(by=['ID'])
```

DATA PREPARING ----------

RAW DATA ----------

```
print('Original Absctract:')
print(df_sample['Abstract'].iloc[170])
```

LowerCase + Remove:Punctuation&digits ----------

```
df_sample['Abstract_Words'] = df_sample['Abstract'].apply(lambda x: x.lower())
df_sample['Abstract_Words'] = df_sample['Abstract_Words'].apply(lambda x: x.translate(str.maketrans('','',remove_punctuation)))
df_sample['Abstract_Words'] = df_sample['Abstract_Words'].apply(lambda x: x.translate(str.maketrans('','',string.digits)))
print('Absctract only words:')
print(df_sample['Abstract_Words'].iloc[170])
```

STOPWORDS ----------

```
df_sample['Abstract_stopwords'] = df_sample['Abstract_Words'].apply(lambda x: [word for word in x.split()
                                                                               if word not in stop_words]).str.join(' ')
df_sample['Abstract_stopwords_removed'] = df_sample['Abstract_Words'].apply(lambda x: [word for word in x.split()
                                                                                      if word in stop_words]).str.join(' ')
print('Absctract after stopwords:')
print(df_sample['Abstract_stopwords'].iloc[0])
print('')
print('stopwords removed:')
print(df_sample['Abstract_stopwords_removed'].iloc[0])
```

STEMMING ----------

```
df_sample['Abstract_stemmed'] = df_sample.Abstract_stopwords.map(lambda x: ' '.join([stemmer.stem(y) for y in x.split(' ')]))
print('Absctract after stemming:')
print(df_sample['Abstract_stemmed'].iloc[0])
```

STOPWORDS (again) ----------

```
df_sample['Abstract_stemmed_stopwords'] = df_sample['Abstract_stemmed'].apply(lambda x: [word for word in x.split()
                                                                                        if word not in stop_words]).str.join(' ')
df_sample['Abstract_stemmed_stopwords_removed'] = df_sample['Abstract_stemmed'].apply(lambda x: [word for word in x.split()
                                                                                        if word in stop_words]).str.join(' ')
print('Absctract stemmmed after 2nd stopwords:')
print(df_sample['Abstract_stemmed_stopwords'].iloc[0])
print('')
print('2nd stopwords removed:')
print(df_sample['Abstract_stemmed_stopwords_removed'].iloc[0])
```

Save to Excel: before/after stop_words+stemming dataframe -----

```
writer = pandas.ExcelWriter('0_prepared_data.xlsx')
df_sample['Abstract'].to_excel(writer,'Original')
df_sample['Abstract_Words'].to_excel(writer,'Abstract_Words')
df_sample['Abstract_stopwords'].to_excel(writer,'Stop_Words')
df_sample['Abstract_stemmed'].to_excel(writer,'Stemming')
df_sample['Abstract_stemmed_stopwords'].to_excel(writer,'Stemming_Stop')
df_sample['Abstract_stopwords_removed'].to_excel(writer,'Stop_Words_Removed')
df_sample['Abstract_stemmed_stopwords_removed'].to_excel(writer,'Stemmed_Stop_Words_Removed')
writer.save()
```

Save entire df----------

```
df_sample.to_pickle('df_total.pkl')
```

Create a Text Column with [Abstract+Master+Year+Language] ----------

```
df_sample['Master'] = df_sample['Master'].apply(lambda x: "{}{}".format('Master_', x.replace(" ", "_")))
df_sample['Language'] = df_sample['Language'].apply(lambda x: "{}{}".format('Language_', x))
df_sample['Year'] = df_sample['Year'].apply(lambda x: "{}{}".format('Year_', x))

df_sample['Text'] = df_sample.Abstract_stemmed_stopwords + ' ' + df_sample.Master + ' '
                  + df_sample.Language + ' ' + df_sample.Year
del df_sample['Title']
del df_sample['Master']
del df_sample['Abstract_stemmed']
del df_sample['Abstract']
del df_sample['Language']
del df_sample['Year']
del df_sample['ID']
del df_sample['Abstract_stopwords']
del df_sample['Abstract_stopwords_removed']
del df_sample['Abstract_stemmed_stopwords']
del df_sample['Abstract_stemmed_stopwords_removed']
del df_sample['Abstract_Words']
df_sample.head()
```

Save only corpus df----------

```
df_sample.to_pickle('df_corpus.pkl')
```

```
df_sample['Text'].iloc[0]
```

9.1.5 Pre-processing – Document representation

IMPORT LIBRARIES ----------

```
#SKLEARN LIBRARIE
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer

#GENERAL LIBRARIES
import pandas
import scipy.sparse
import pickle
from itertools import islice
```

IMPORT DATA ----------

```
df = pandas.read_pickle('df_corpus.pkl')
```

```
df['Text'].iloc[0]
```

CountVectorizer ----------

```
#tc-matrix (term count matrix)
df_tolist = df['Text'].tolist()
#dataframe to bag-of-words - vector represent the ...
tc_vectorizer = CountVectorizer(max_df=0.1, min_df=1, stop_words='english', ngram_range=(1,1))
tc = tc_vectorizer.fit_transform(df['Text'].tolist())
print(list(islice( tc_vectorizer.fit(df['Text'].tolist()).vocabulary_.items(), 20)))

#Save matrix
scipy.sparse.save_npz('1_tc.npz', tc)

#Print info:
print('sparse matrix shape:', tc.shape)
print('nonzero count:', tc.nnz)
print('sparsity: %.2f%%' % (100.0 * tc.nnz / (tc.shape[0] * tc.shape[1])))

#get feature_names
tc_feature_names = tc_vectorizer.get_feature_names()

#Save feature names
with open('1_tc_feature_names.pkl', 'wb') as f:
    pickle.dump(tc_feature_names, f)

#print dataframe
tc_df = pandas.DataFrame(tc.toarray(), columns=tc_feature_names)
tc_df.head()

#Save matrix to EXCEL
# writer = pandas.ExcelWriter('1_tc.xlsx')
# tc_df.to_excel(writer,'tc')
# writer.save()
```

Inverse Document Frequency (IDF) ----------

```
#tf-idf-matrix

#dataframe to bag-of-words - vector represent the ...
tfidf_vectorizer = TfidfVectorizer(use_idf=True, max_df=0.1, min_df=1, stop_words='english', ngram_range=(1,1))
tfidf = tfidf_vectorizer.fit_transform(df['Text'].tolist())
print(list(islice( tfidf_vectorizer.fit(df['Text'].tolist()).vocabulary_.items(), 20)))

#Save matrix
scipy.sparse.save_npz('1_tfidf.npz', tfidf)

#Print info:
print('sparse matrix shape:', tfidf.shape)
print('nonzero count:', tfidf.nnz)
print('sparsity: %.2f%%' % (100.0 * tfidf.nnz / (tfidf.shape[0] * tfidf.shape[1])))

#get feature_names
tfidf_feature_names = tfidf_vectorizer.get_feature_names()

#Save feature names
with open('1_tfidf_feature_names.pkl', 'wb') as f:
    pickle.dump(tfidf_feature_names, f)

#print dataframe
idf_df = pandas.DataFrame(tfidf.toarray(), columns=tfidf_feature_names)
idf_df.head()

#Save matrix to EXCEL
# writer = pandas.ExcelWriter('1_tfidf.xlsx')
# idf_df.to_excel(writer,'tfidf')
# writer.save()
```

9.1.6 Core mining process – Topic Modelling

Import: libraries

```
#SKLEARN LIBRARIES
from sklearn.decomposition import NMF, LatentDirichletAllocation
from sklearn.model_selection import GridSearchCV

#pyLDAvis LIBRARIES
import pyLDAvis
import pyLDAvis.sklearn

#GENERAL LIBRARIES
import pandas
import scipy.sparse
import pickle
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

Load: files

```
#Load Term Count Matrix
tc = scipy.sparse.load_npz('1_tc.npz')
tc_feature_names = pickle.load(open('1_tc_feature_names.pkl', 'rb'))

#Load TF-IDF Matrix
tfidf = scipy.sparse.load_npz('1_tfidf.npz')
tfidf_feature_names = pickle.load(open('1_tfidf_feature_names.pkl', 'rb'))
```

Auxiliary: create functions

```
#this functions allows to print the top words per topic per model
def display_topics(model, feature_names, top_words_quantity):
    for topic_idx, topic in enumerate(model.components_):
        print("Topic %d:" % (topic_idx+1))
        print(" ".join([feature_names[i]
                        for i in topic.argsort()[:-top_words_quantity - 1:-1]]))
```

Model LDA: Latent Dirichlet Allocation

Matrix information

```
# Compute Sparsicity: Percentage of Non-Zero cells
tc_dense = tc.todense()
print("Sparsicity: ", ((tc_dense > 0).sum()/tc_dense.size)*100, "%")
```

Find the most suitable number of topics

```
# Define Search Param
search_params = {'n_components': [4,5,6,7,8,9,10], 'learning_decay': [.1, .3, .5, .7, .9]}

# Init the Model
lda = LatentDirichletAllocation(learning_method='batch')

# Init Grid Search Class
model = GridSearchCV(lda, param_grid=search_params)

# Do the Grid Search
model.fit(tc)

GridSearchCV(cv=None, error_score='raise',
       estimator=LatentDirichletAllocation(batch_size=128, doc_topic_prior=None,
             evaluate_every=-1, learning_decay=0.7, learning_method='batch',
             learning_offset=10.0, max_doc_update_iter=100, max_iter=10,
             mean_change_tol=0.001, n_components=10, n_jobs=1,
             n_topics=None, perp_tol=0.1, random_state=123,
             topic_word_prior=None, total_samples=1000000.0, verbose=0),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'n_components': [4,5,6,7,8,9,10], 'learning_decay': [0.1, 0.3, 0.5, 0.7, 0.9]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring=None, verbose=0)
```

```
# Get Log Likelyhoods from Grid Search Output
n_components = [4,5,6,7,8,9,10]
log_likelyhoods_1 = [round(gscore.mean_validation_score) for gscore in model.grid_scores_ if gscore.parameters['learning_decay']==0.1]
log_likelyhoods_3 = [round(gscore.mean_validation_score) for gscore in model.grid_scores_ if gscore.parameters['learning_decay']==0.3]
log_likelyhoods_5 = [round(gscore.mean_validation_score) for gscore in model.grid_scores_ if gscore.parameters['learning_decay']==0.5]
log_likelyhoods_7 = [round(gscore.mean_validation_score) for gscore in model.grid_scores_ if gscore.parameters['learning_decay']==0.7]
log_likelyhoods_9 = [round(gscore.mean_validation_score) for gscore in model.grid_scores_ if gscore.parameters['learning_decay']==0.9]

# Show graph
plt.figure(figsize=(12, 8))
plt.plot(n_components, log_likelyhoods_1, label='0.1')
plt.plot(n_components, log_likelyhoods_3, label='0.3')
plt.plot(n_components, log_likelyhoods_5, label='0.5')
plt.plot(n_components, log_likelyhoods_7, label='0.7')
plt.plot(n_components, log_likelyhoods_9, label='0.9')
plt.title("Choosing Optimal LDA Model")
plt.xlabel("Num Topics")
plt.ylabel("Log Likelyhood Scores")
plt.legend(title='Learning decay', loc='best')
# plt.axis([4,5,-133000,-129333])
plt.show()
```

Dominant Topic per each document

```
# Create Document - Topic Matrix
lda_output = best_lda_model.transform(tc)

# column names
topicnames = ["Topic" + str(i) for i in range(best_lda_model.n_components)]

# index names
docnames = ["Doc" + str(i) for i in range(len(tc.toarray()))]

# Make the pandas dataframe
df_document_topic = pandas.DataFrame(np.round(lda_output, 2), columns=topicnames, index=docnames)

# Get dominant topic for each document
dominant_topic = np.argmax(df_document_topic.values, axis=1)
df_document_topic['dominant_topic'] = dominant_topic

# Export to Excel
writer = pandas.ExcelWriter('0_dominant_topic_per_document.xlsx')
df_document_topic.to_excel(writer,'topic_per_doc')
writer.save()

# Styling
def color_green(val):
    color = 'green' if val > .1 else 'black'
    return 'color: {col}'.format(col=color)

def make_bold(val):
    weight = 700 if val > .1 else 400
    return 'font-weight: {weight}'.format(weight=weight)

# Apply Style
df_document_topics = df_document_topic.head(15).style.applymap(color_green).applymap(make_bold)
df_document_topics
```

Topics distribution acress documents

```
df_topic_distribution = df_document_topic['dominant_topic'].value_counts().reset_index(name="Num Documents")
df_topic_distribution.columns = ['Topic Num', 'Num Documents']
df_topic_distribution
```

LDA Visualization with oyLDAvis

```
from sklearn.feature_extraction.text import CountVectorizer
tc_vectorizer = CountVectorizer(max_df=0.1, min_df=1, stop_words='english', ngram_range=(1,1))

df = pandas.read_pickle('df_corpus.pkl')
tc = tc_vectorizer.fit_transform(df['Text'].tolist())

pyLDAvis.enable_notebook()
panel = pyLDAvis.sklearn.prepare(best_lda_model, tc, tc_vectorizer, mds='tsne')
panel
```

Topic's keywords

```
# Topic-Keyword Matrix
df_topic_keywords = pandas.DataFrame(best_lda_model.components_)

# Assign Column and Index
df_topic_keywords.columns = tc_feature_names
df_topic_keywords.index = topicnames


# Export to Excel
writer = pandas.ExcelWriter('0_topics_keywords.xlsx')
df_topic_keywords.T.to_excel(writer,'keywords') #transposed matrix
writer.save()

# Export df
df_topic_keywords.to_pickle('df_topics_keywords.pkl')

# View
df_topic_keywords.head()
```

top 15 keywords each topic

```
# Show top n keywords for each topic
def show_topics(vectorizer=tc_vectorizer, lda_model=best_lda_model, n_words=20):
    keywords = np.array(vectorizer.get_feature_names())
    topic_keywords = []
    for topic_weights in lda_model.components_:
        top_keyword_locs = (-topic_weights).argsort()[:n_words]
        topic_keywords.append(keywords.take(top_keyword_locs))
    return topic_keywords


topic_keywords = show_topics(vectorizer=tc_vectorizer, lda_model=best_lda_model, n_words=15)

# Topic - Keywords Dataframe
df_topic_keywords = pandas.DataFrame(topic_keywords)
df_topic_keywords.columns = ['Word '+str(i) for i in range(df_topic_keywords.shape[1])]
df_topic_keywords.index = ['Topic '+str(i) for i in range(df_topic_keywords.shape[0])]
df_topic_keywords
```

9.1.7 Core mining process – K-means clustering

Import: libraries

```
#Import the necessary packages/modules
import matplotlib.pyplot as plt
from __future__ import print_function
# from A_defining_treatment import tokenize, stemming, remove_stopwords, text_translate
import numpy as np
import pandas
from nltk.stem.snowball import SnowballStemmer
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from itertools import islice
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
from matplotlib import cm
from sklearn.metrics import silhouette_samples, silhouette_score
import re
import string


import scipy.sparse
import pickle
```

Load: files

```
#Load Term Count Matrix
tc = scipy.sparse.load_npz('1_tc.npz')
tc_feature_names = pickle.load(open('1_tc_feature_names.pkl', 'rb'))

#Load TF-IDF Matrix
tfidf = scipy.sparse.load_npz('1_tfidf.npz')
tfidf_feature_names = pickle.load(open('1_tfidf_feature_names.pkl', 'rb'))
```

Cluster Algorithm: K-means ¶

determine the optimal number of clusters

```
distortions = []
for i in range(1, 6):
    km = KMeans(n_clusters=i,
                init='k-means++',
                n_init=10,
                max_iter=300,
                tol=1e-04,
                random_state=0)
    km.fit(tfidf)
    distortions.append(km.inertia_)
plt.plot(range(1,6), distortions, marker='o', color='grey')
plt.xlabel('Number of clusters')
plt.ylabel('Distortion')

plt.show()
distortions
```

number of clusters found

```
number_of_clusters=3
```

modelling the k-means algorithm

```
km = KMeans(n_clusters=number_of_clusters,
            init='k-means++',
            n_init=10,
            max_iter=300,
            tol=1e-04,
            random_state=0)
y_km = km.fit_predict(tfidf)
```

quality of clustering [silhouett plots]

```
cluster_labels = np.unique(y_km)
n_clusters = cluster_labels.shape[0]
silhouette_vals = silhouette_samples(tfidf,y_km,metric='euclidean')
y_ax_lower, y_ax_upper = 0, 0
yticks = []

for i, c in enumerate(cluster_labels):
    c_silhouette_vals = silhouette_vals[y_km == c]
    c_silhouette_vals.sort()
    y_ax_upper += len(c_silhouette_vals)
    color = cm.jet(float(i) / n_clusters)
    plt.barh(range(y_ax_lower, y_ax_upper),
             c_silhouette_vals,
             height=1.0,
             edgecolor='none',
             color=color)
    yticks.append((y_ax_lower + y_ax_upper) / 2.)
    y_ax_lower += len(c_silhouette_vals)
silhouette_avg = np.mean(silhouette_vals)
plt.axvline(silhouette_avg,
            color="red",
            linestyle="--")
plt.yticks(yticks, cluster_labels + 1)
plt.ylabel('Cluster')
plt.xlabel('Silhouette coefficient')
plt.show()

print("The average silhouette_score is :", silhouette_avg)
```

```
df_cluster_1 = pandas.DataFrame(
    {'terms': terms,
     'Weight': km.cluster_centers_[0],
    })
df_cluster_1.to_pickle('df_cluster_1.pkl')

df_cluster_2 = pandas.DataFrame(
    {'terms': terms,
     'Weight': km.cluster_centers_[1],
    })
df_cluster_2.to_pickle('df_cluster_2.pkl')

df_cluster_3 = pandas.DataFrame(
    {'terms': terms,
     'Weight': km.cluster_centers_[2],
    })
df_cluster_3.to_pickle('df_cluster_3.pkl')
```

```
print("Top terms per cluster:")
order_centroids = km.cluster_centers_.argsort()[:, ::-1]
terms = tfidf_feature_names
for i in range(number_of_clusters):
    top_ten_words = [terms[ind] for ind in order_centroids[i, :15]]
    print("Cluster {}: {}".format(i, ' '.join(top_ten_words)))
```

```
df_sample = pandas.read_pickle('df_corpus.pkl')
print(df_sample.head())

results = pandas.DataFrame()
results['text'] = df_sample['Text'].tolist()
results['category'] = km.labels_
results.head()

# writer = pandas.ExcelWriter('clusters.xlsx')
# results.to_excel(writer,'clusters')
# writer.save()
```

9.2 LDA outcomes

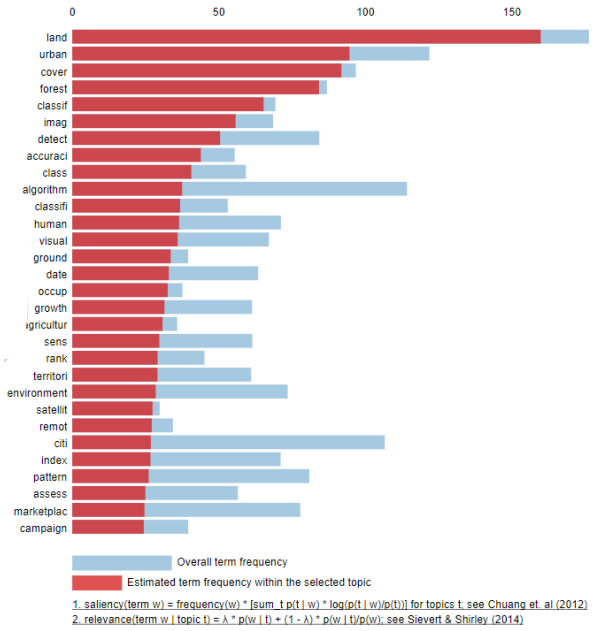

Figure 9.1 - Topic 1 Most Relevant Terms

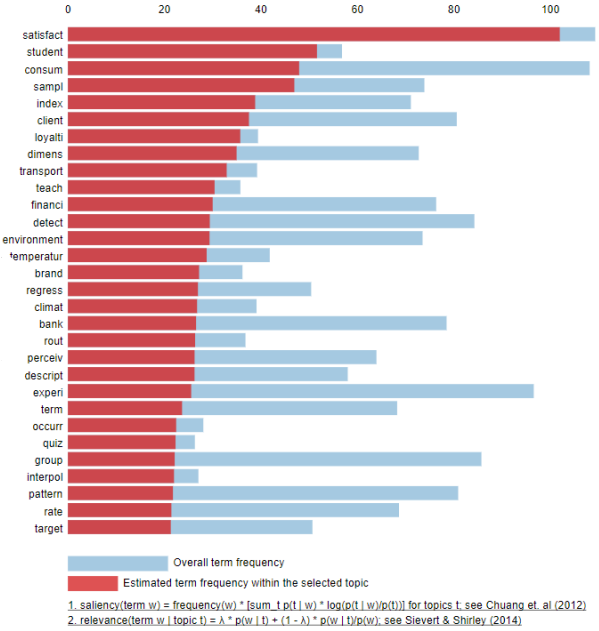

Figure 9.2 - Topic 2 Most Relevant Terms

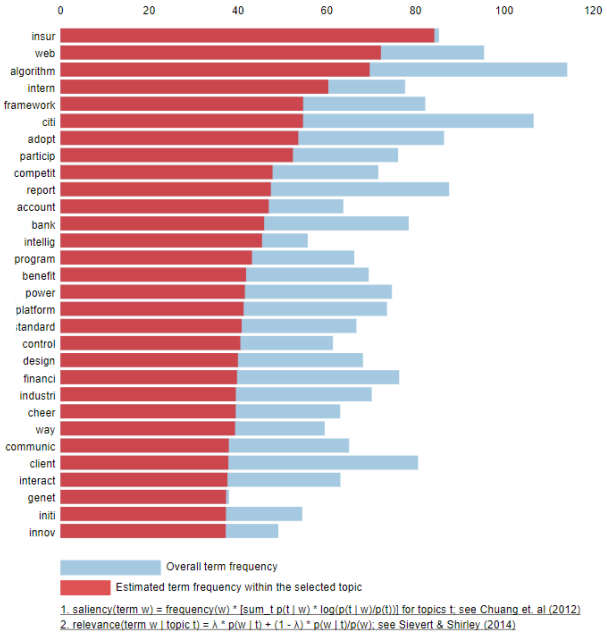

Figure 9.3 - Topic 3 Most Relevant Terms

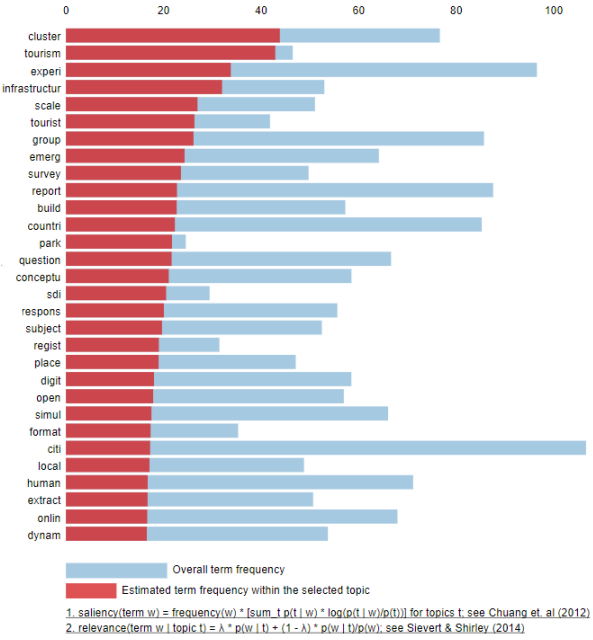

Figure 9.4 - Topic 4 Most Relevant Terms

9.3 Cluster outcomes

Figure 9.5 - Wordcloud representation of cluster 1

Figure 9.6 - Wordcloud representation of cluster 2

Figure 9.7 - Wordcloud representation of cluster 3