Performance evaluation of research publications through text mining applications.

Kyriakoulea Sofia

Professor: Mamakou Xenia
Program Director: Kouretas Georgios

8/29/21

Full Time MBA – Thesis
Athens University of Economics & Business

ΒΕΒΑΙΩΣΗ ΕΚΠΟΝΗΣΗΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

«Δηλώνω υπεύθυνα ότι η συγκεκριμένη μεταπτυχιακή διπλωματική εργασία για τη λήψη του μεταπτυχιακού τίτλου σπουδών του Διατμηματικού ΜΠΣ Πλήρους Φοίτησης των Τμημάτων Οργάνωσης και Διοίκησης Επιχειρήσεων και Μάρκετινγκ και Επικοινωνίας του Οικονομικού Πανεπιστημίου Αθηνών στη Διοίκηση Επιχειρήσεων: MBA (Master in Business Administration) έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει εγκριθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό. Η εργασία αυτή έχοντας εκπονηθεί από εμένα, αντιπροσωπεύει τις προσωπικές μου απόψεις επί του θέματος. Οι πηγές στις οποίες ανέτρεξα για την εκπόνηση της συγκεκριμένης διπλωματικής αναφέρονται στο σύνολό τους, δίνοντας πλήρεις αναφορές στους συγγραφείς, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο».

Η δηλούσα

Σοφία Κυριακουλέα

Verba volant, scripta manent.[1]

- Latin proverb

[1] Literal translation: “Spoken words fly away, written words remain”.

Acknowledgments

It is always confusing if this section goes before or after the table of contents, but it has been placed intentionally before them, because the present work would not be possible if I hadn't had some specific people in my life. As a nerd, these people are not a lot so I will mention them here shortly.

First and foremost, comes my dear mother, Anastasia. Words cannot describe her contribution to my life in general. If it were not for her being the person that she is, I would have drifted away so many times. I understand that half of the reasons I am who I am is because of the way she raised me, and the fact that to this day she has been by my side no questions asked. I also understand the finite game life plays with us and the constraints it puts in our existence. I could not forget to mention my grandfather, my mother's father, who I did not get to meet so well, but to this day is financing my life and education with his sweat and blood from the several years he spent as an immigrant in the United States.

Then I have to mention my professor Dr. Mamakou Xenia. There is no other academic out there with her willingness to work, help students and foster them to develop their competences and capabilities. She is unique in so many levels. She has never questioned my judgment and she always assisted me all the way through this thesis. There have been several occasions when she would answer my messages late at night or on a Sunday. Most importantly she let me pick to topic of my interest and take my own initiatives. She prompted me to take the hard way and I learned so much. I appreciate her in so many levels and I have a request. Ms. Mamakou please do not let the academic world change who you are and remain a model professor for everyone, forever.

I want to also mention Vasilis, for his always available shoulder to cry on, for his endless patience, for his putting up with my nonsense, for his willingness to take care of me, for his ability to keeping me on track, for his kindness, for the late-night snacks he fed me when I was losing it and most importantly for his love. Without him, I am not whole.

Last but not least, I would like to mention that I should thank and appreciate myself, for putting the hard work, the time and the energy into this work. I could have more night outs with friends (or even more friends maybe), I could rest a bit more, I could take care of me occasionally, but I did not. I chose to be this person, with these flaws and imperfections. Every choice shapes our future self, and I am proud of me today.

Brief Executive Summary

Research, as an activity of academia, has been growing over the past decades and that results to an increased number of scientific publications hosted in the numerous Journals that are available to publish today. It is true that society should appreciate the contribution of research in every aspect of life, not all research is conducted, written, or published equally. The quality and the societal benefit of today's published research is heavily criticized by academics, organizations, institutions, and the government. The competition between researchers, universities and institutes is fierce. That raises concerns whether the published literature should be viewed as a product. Several scientometric and bibliometric methods have been developed to characterize literature and some of them are heavily criticized like the Journal Impact Factor and the number of citations the article will receive.

To keep up with the ever-growing literature collection new age data science comes in to assist. Recent progress in computational systems and data mining could provide fundamental shifts in the way academic literature is approached. Text Mining is one of the available technologies that could aid in the analysis of millions of texts, in their categorization and even assist the already existing heuristic algorithms to increase visibility of articles that otherwise would be left in the dark. Text mining cannot be considered a novelty; however, it could make a difference in the academic world if applied properly.

Could text mining become a tool for the evaluation of the performance of published literature? Practicing text mining consists of several steps, requires teams with experience and has computational limitations. In the present thesis we analyzed a total of 724 articles and reviews published in the Journal of Management, a top tier management, business, and psychology journal with an impact factor of 11.06 in 2020. All the collected documents had the exact same format and were analyzed on the KNIME Analytics Platform, an open-source data science software available online.

Without having to manually read any document in the corpus it was possible to extract information on the most common topics by defining the keywords in the collection. Term Frequency (TF) and the product of TF and Inverse Document Frequency (IDF) were used as ways to identify the keywords. In addition, other data were collected for the corpus such as the publication year, the number of authors, the issue number of the Journal, the length of the article in pages, the number of cited refences and the term count. All these variables functioned as predictors of lifetime and yearly average citations of the collection.

There were two types of analysis preformed, besides some descriptive statistics of the collection. The first was two linear regressions with the dependent variables being the lifetime and yearly

average citations. The aim of the analysis was to determine which of the collected variables could predict the efficiency of an article in terms of citations. Significant predictors for lifetime article citations were the age of the article, the type of the article either review or research article and the number of cited references. On the other hand, for the yearly average citations the significant predictors were the type of the article, the number of cited references per record, the issue they were published, the volume of the Journal, the count of terms and the number of authors. The value of R squared is 0,171, adjusted R squared is 0,165 which does not differ from R squared and the p value of the model equals to 0,001. The value of R squared is 0,246, adjusted R squared is 0,242 which does not differ from R squared and the p value of the model equals to 0,001, therefore the model is considered significant. The second analysis was a binary logistic regression that predicts whether a record will be either highly or low cited and an article or review according to the given collection. The overall accuracy of the first model was 75% and for the second 91.7%.

In conclusion, it is worth mentioning that besides the research question being answered, the present document identifies potential future research on the topic that could shed light and explore the wealth of academia. It also raises awareness in the availability of literature for other researchers to analyze. Most importantly, this document could function as a reminder that human nature, sense of speech, brain abilities and senses are not yet substituted by computers programs, and it is questionable whether we, as a species, should let this happen.

Table of Contents

ΒΕΒΑΙΩΣΗ ΕΚΠΟΝΗΣΗΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ 1
Acknowledgments 3
Brief Executive Summary 4
Table of Figures & Tables 8
Introduction 9
Aim of the thesis 9
Scope of the thesis 9
The research marketplace today 9
Literature Review 11
Defining research 11
Where is research conducted? 11
Quality and ethical issues concerning research 11
Evaluation of research articles 12
Journal Impact Factor 12
Natural Language Processing 14
Text Mining 15
The concept of knowledge discovery & management 17
Research Questions 17
Methodology 18
Data Selection 18
Text Indexing 19
Tokenization 20
Stop-word Removal 20
Filtering 20
Stemming 20
References Removal 21
Data preprocessing process in KNIME 22
Text Encoding 23
Word Frequencies 23
Keyword Extraction 26
Vector Space Modeling 26
Dimensionality reduction methods 27

Document Vector Creation in KNIME 28
Clustering & Classification 28
System Requirements 28
Performance Evaluation 29
Performance Prediction 29
Statistical Analysis Platform 30
Results 31
Master Table Creation 31
Constraints 31
Descriptive Statistics 32
Highly or low cited 32
Term Count 33
Total Citations 34
Per year Publications 34
Cited References and number of authors 35
Word Clouds 37
Reviews 37
Articles 38
Keywords Word Clouds 39
Linear Regression 40
Lifetime Article Citations 40
Average Citations per year 40
Binomial Logistic Regression 41
Predict whether a record will be highly or low cited 41
Predict whether a record is a review or a research article 41
Discussion 43
Answering Research Questions 43
What are the most frequent keywords in the collected articles? 43
How are the articles distributed in terms of years? 43
Could text mining applications be of assistance in predicting article citations? 43
Key takeaways 44
For further research 44
For Institutions 44

Epilogue 44
Author's Notes & Questions 46
Appendix 50
References 52

Table of Figures & Tables

Figure 1: Text Mining Framework 18
Figure 2: Hard vs. Soft stemming 21
Figure 3: The three steps of the 1st phase of text indexing 21
Figure 4: KNIME window layout 23
Figure 5: Distribution of Articles and Reviews 31
Figure 6: Number of records marked as highly or low cited per type of Article 32
Figure 7: Layout of term Count 33
Figure 8: Total Citations 34
Figure 9: Publication distribution over the decade 34
Figure 10: Total number of records per type & issue number 35
Figure 11: Cited references per type 35
Figure 12: Number of authors per type of article 36
Figure 13: Authors per Article 36
Figure 14: Word Cloud for reviews - TF 37
Figure 15: Word Cloud for Reviews TF*IDF 37
Figure 16: Word Cloud for Articles TF 38
Figure 17: Word Cloud for Articles TF*IDF 38
Figure 18: Word Cloud of keywords for Reviews 39
Figure 19: Word Cloud of Keywords for Articles 39

Table 1: EXAPMLE OF TF-IDF FUNCTIONALITY 25
Table 2: Term Count and pages 33
Table 4: Coefficients Table for Total Citations As Dependent Variable 40
Table 5: Coefficients table for Average Citations per Year as Dependent Variable 40
Table 6: Variables in the equation of binary logistic predicting highly or low cited 41
Table 7: Classification table of binary logistic Predicting Highly or Low Cited 41
Table 8: Variables in the equation of binary logistic predicting Review or Article 42
Table 9: Classification Table of Binary Logistic Predicting Review or Article 42
Table 3: Distribution of records over the issues per year 50
Table 10: Top 10 words From BoW for both Types of Records 51
Table 11: Keywords word cloud metrics 51

Introduction

Aim of the thesis

The aim of the present thesis is to extract valuable information from research articles published in elite journals of management, through text mining. This information can be utilized to discover patterns in exceptional papers to assess their quality and evaluate their performance. The applications of text mining in published literature could provide insight for emerging research trends and topics of interest. The discovered knowledge from this application can be of great value to universities, funding organizations, PhD committees, hiring managers and governments who are searching for unbiased criteria for performance evaluations of individuals – scholars and students – and organizations.

Scope of the thesis

The scope of the present thesis is to identify whether text mining operations can provide assistance to deal with the ever-growing number of scientific publications int terms of their quality. In addition, the application of text mining should result in the creation of variables that can be used for analysis in statistical models. Additionally, text mining is evaluated as a way to complement what is already employed for the evaluation of papers, that being bibliometrics and scientometrics. The present thesis could become the basis for the creation of models that will predict the performance of unpublished manuscripts.

The research marketplace today

Research is viewed as a product rather than a contribution to the evolution of humankind and this is acknowledged by academics. On one hand this fact may not necessarily influence society negatively. One the other hand, as mentioned above, when the quality of research is impaired, it is difficult to accept the positive impact of published research for society.

It is significant for research performers, managers, evaluators, policy makers, sponsors, and the Government to identify the impact of research for each field and the user audience. The latter is possible with citation mining. This process can prove to be rather complex and time consuming due to the high volume of publications (Kostoff *et al.*, 2001). In addition to the above, the growing volume of literature has created classification issues that also concern storage, organization add effective access. The information provided by publications can only be of use if it is easily accessible for its users (Sulova *et al.*, 2017).

According to Rynes et. al (Rynes, Giluk and Brown, 2007) consultants or even journalists are outperforming academic management researchers, for some time now, especially as sources of ideas and advice for practice and for policy makers. It is unfortunate that academics are, now more than ever, in need for resources and insights, to produce knowledge for both more effective

and sustainable organizations. This declining academic influence in the world of policy and practice raises several questions with the first one being; *Are our major research findings truly unimportant to policy and practice?* This science-practice link is of significant importance in the discipline of management because they have real life practical implications that can impact societal-, firm- and individual-level outcomes (Antonakis *et al.*, 2014).

It is apparent now more than ever that the competition between higher education institutions has increased and this has a setback for knowledge production. The severe need of institutions to outperform others results in a rapid transformation of the higher education systems into a competitive marketplace where the focus is on profit instead of knowledge and development (Hazelkorn, 2004). The all-time high of publications is, unfortunately, not due to increased productivity. It is linked with the pressure to publish and the trend of *Publish or Perish* (van Wesel, 2016).

Literature Review

Defining research

The definition of research activity is considered a debatable topic especially in higher education. University research, otherwise known as “basic research”, is associated with discovery or the search of something new. The results of sustained enquire are subject to critical questioning of others such as peer-reviewed publications. Academia has made the concept broader by including the ability “to glean information” and respond critically to what has already been done in the field”. It is noted that these definitions differ from discipline to discipline. In sciences there is emphasis on discovery of new facts, while social sciences may implement a range of methodologies and or comments on already existing knowledge through e.g. surveys and observations. (Hazelkorn, 2004)

Where is research conducted?

Research centers, individuals, academic faculties, centers of excellence are business parks are the main loci where research is conducted. Although research is still practiced by individuals it is advised that the focus should be shifted to multidisciplinary teams and in the department level. That should be adequate to develop a community that supports a culture of grant-awarding reputations and timely outcomes while tackling the ever-growing challenges for academic institutions. (Hazelkorn, 2002)

Quality and ethical issues concerning research

Due to the fierce competition of academic institutions, the growing number of researchers around the globe and the numerous journals available, the volume of publications has increased over the years. We are now receiving millions of new publications each year and the numbers are constantly on the rise. This crisis is academic publishing puts a lot of pressure on top journals and scholars. This may impact the quality of these articles which is sometimes, marginal, or even low (Altbach, 2018). Those credibility concerns arise due to questionable research or reporting practices and the presentation of biased evidence in favor of a hypothesis (Banks *et al.*, 2016). This phenomenon is no secret in the academic world that terms such as “The seven deadly sins of research” are employed. Those include p-hacking, cherry pricking data, not publishing negative results, “salami slicing” etc. To mitigate these phenomena institutions attempt research integrity training, although these practices might prove ineffective for tenured researchers (Conroy, 2019).

The quality of a research article is interconnected with its author. It is common nowadays to measure the quality of a researcher through publication and citation ranking. Those indicators are decisive of scientific worth and play a significant role to the career of individual academics. Although, it has been proven that such quantity rankings are not objective, even geographically

biased (Pasterkamp *et al.*, 2007) and they do not effectively measure research quality (Frey and Rost, 2010) (Lindsey, 1989) these metrics are still commonly used and accepted among academics. At the same time, besides the common used metrics, quality can be measured through readability (Roberts, Fletcher and Fletcher, 1994), the chosen method of research e.g. qualitative vs. quantitative (Rolfe, 2006) and many more.

The discipline of management can be characterized as a "weak paradigm" compared to other social sciences. That means that it lacks well-established theories, methods and research questions. Due to this ambiguity a clear and successful career path is difficult to define (Glick, Miller and Cardinal, 2007).

Evaluation of research articles

To assess the importance of scientific publications, the academic world is using bibliometric and scientometric indices and peer reviews as the main source of reference for the evaluation (Ancaiani *et al.*, 2015). According to Clarivate's Web of Science report of Highly Cited Researchers for 2020 (Clarivate, 2020), there is no unique or universally agreed concept of what constitutes extraordinary research performance. Web of Science employs its tool "InCites" into the well-known database Web of Science Core Collection to identify researchers with cross-field impact, highly cited papers, and citation counts. The latter is widely accepted among academics and organizations and participates in the most important metric, the Journal Impact Factor (JIF), which in turn, plays a decisive role for the performance of an article and therefore the advancement of a scholar. It has also been widely accepted that citation count is an indicator of the impact of an article (Chen and Ho, 2015).

Besides scientometric indices, articles are also subject to peer evaluation. Peer evaluation is flawed in its own way, that is the subjectivity of the reviewers themselves. A reviewer's judgment can only by subjective, because to be objective it had to be tested – and repeated – as well as understood. The values used for the so-called objective evaluation are to their core subjective. Scientometrics are objective facts, however the idea of substituting peer evaluators completely with an objective algorithm is like removing judges from court houses and let some computer do their job (Ricker, 2017).

Journal Impact Factor

The journal impact factor (JIF) of an academic journal is a scientometric index calculated by Clarivate in the *Journal Citation Reports* (JCR) that reflects the yearly average number of citations of articles published in the last two years in each journal. It is frequently used as a proxy for the relative importance of a journal within its field; journals with higher impact factor values are often deemed to be more important, or carry more intrinsic prestige in their respective fields,

than those with lower values (Guz, Rushchitsky and Chernyshenko, 2005). The examined data base is the Web of Science.

In addition to the above, the SCIMAGO Journal and Country Rank (SJR) is a free-of-charge alternative to the one from Clarivate. It analyses publications indexed in the Scopus database, which is provided by the publisher Elsevier, dating from 1997 to the present. SJR enables users to run online searches using the Scopus platform, which is a paid-for tool, or using the SJR.

Both the JCR and the SJR separate journals in subject areas. Each subject group of magazines is divided into four quartiles: Q1, Q2, Q3, Q4. Q1 is occupied by the top 25% of journals in the list; Q2 is occupied by journals in the 25 to 50% group; Q3 is occupied by journals in the 50 to 75% group and Q4 is occupied by journals in the 75 to 100% group. The most prestigious journals within a subject area are those occupying the first quartile, Q1.

Besides the JIF the Association of Business Schools (ABS) assesses the quality of thousands of business and management publications worldwide, based on citation scores as well as judgements of leading researchers, and it is translated in a scoring system of numbers from 1 to 4. A journal ranked in 4 is considered to be one of the best in a specified field. There is also a fifth rank 4* where the Journals of Distinction are found. Usually the 4* Journals fall into the Q1.

The metrics above, although important for the stakeholders, they do not measure the quality of an article but rather the quality of the journal in which the article is published. Assessing the quality of the individual articles and not the journals, is probably a more valid metric for the evaluation of a scholar's performance. This metrics could also provide insight on what makes scientific publications successful, and potentially could be applied even before the article is published to estimate or forecast its success. These methods are applied in the entertainment industry, where it is possible, to estimate the performance of a movie based just on the plot summary (storyline), before the producers invest millions on production (Delen, Sharda and Kumar, 2007).

Criticism on the JIF and the role of article type

The JIF is basically a mean and as a metric has received a lot of criticism. It has been characterized by Nature as crude and misleading. This metric undervalues research in the not so trendy fields and as a mean – a statistical figure that stand-alone describes at best only half of the story – gives disproportionate significance to a few star-performing papers, while it implies that the not so highly cited articles are unimportant (Editorial, 2016). Thomson Reuters, the analytics firm who publishes the JIF, says that this is a broad metric that describes the output of a Journal and not the quality itself. Sadly it has been heavily misused by scientists, journals, universities, hiring committees and funding agencies (Callaway, 2016). In addition to the above the data used for

the calculation of the metric are not public. One can access the specifics of a journal only after purchase. This ambiguity alone raises awareness on the matter. When Thomson Scientific was asked whether they would consider providing a median along with the mean for the journals they replied, "*It's an interesting suggestion…The median…would typically be much lower than the mean. There are other statistical measures to describe the nature of the citation frequency distribution skewness, but the median is probably not the right choice*." (Rossner, Van Epps and Hill, 2007).

The number of original research articles is increasing and so is the number of review articles that serve the overwhelmed scientists in any field. The notion that the absolute count of reviews available in the research marketplace may bely their function is also noticeable. It is fair to assume that the type of an article will significantly affect the citation count and some studies find that this is the case (Lei and Sun, 2020). Reviews attract more attention in comparison to other article types and it is recommended that when the goal is to bring in citations a trusted method is to write substantial and comprehensive reviews (Vanclay, 2013). On the other hand, there are studies that discard this hypothesis and support "article genre" and especially reviews is not a variable that describes with significance the number of citations each article will attract (Onodera and Yoshikane, 2015) (Ketcham and Crawford, 2007).

Natural Language Processing

What is considered as a Natural Language?

Natural Language is the kind of language that evolved over time and humans write or speak this language to communicate[2] with each other. The known languages today are not developed following a plan, but instead they are the product of continuous evolution through their usage. Languages are native to certain groups of people (e.g. Greek, English, etc.) and they are the most natural way of communication. It is worth mentioning that natural languages first occur in the form of speech and afterwards writing or script develops for the most of them (Kumar, 2017).

What is Natural Language Processing?

Natural Language Processing (NLP) has many definitions, all of which revolve around the same ideas (Liddy, 2001; Collobert *et al.*, 2011; Verspoor and Cohen, 2013; Kumar, 2017; Rebala, Ravi and Churiwala, 2019). "*NLP is a theoretical motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.*"

[2] More on communication at the notes of the present work.

The phrase "*range of computational linguistics*" refers to the multiple methods and techniques available that could accomplish specific types of language analysis. By "*naturally occurring texts*" it is implied that the text being analyzed should come from human communication oral or written and should not be constructed solely for the purpose of analysis. Humans utilize a sum of levels of language to communicate which include, phonology, morphology, lexical, syntactic, semantic, discourse, and pragmatic, hence the notion of "*levels of linguistic analysis*". "*Human-like language processing*" takes into consideration that NLP attempts to perform human-like operations, therefore can be considered a discipline within Artificial Intelligence, although it still depends on a number of other disciplines (Liddy, 2001).

The goal of Natural Language Understanding (NLU)

As mentioned in the definition above, the goal is "*to accomplish human language processing.*" The word 'processing' is not randomly chosen and is completely different form understanding. Natural Language Understanding was the first termed associated with AI but it has not been accomplished yet (Collobert *et al.*, 2011). A complete NLU system should be able to:

- Paraphrase an input text.
- Translate from one language to another.
- Grasp the context and answer questions about it.
- Come to conclusions from the text (Liddy, 2001).

NLP is key to knowledge management. Cognitive technologies are evolving in a rapid rate and so is knowledge. In numerous cases computers are outperforming humans, such as chess championships and knowledge games (e.g. the case of IBM's Watson in Jeopardy). In areas like healthcare computers are utilized to discover knowledge in databases to help oncology doctors identify possible treatments for their patients. However, technology is not meant to replace the human mind, experience and intuition. The goal is to enhance and multiply human judgment abilities.

Text Mining

What is considered as text?

Text is defined as the kind of data that are characterized as unstructured. They differ from numbers on tables or records in databases. Text consists of strings which are called words (Boyce, 1990). To make the text, meaningful combination of individual strings is required. This is achieved by rules called grammars. The strings, which are written in natural language, are combined into sentences. The organized group of sentences is a paragraph, and the ordered set of paragraphs is the actual text. Sentences and paragraph can vary in the length of the strings they contain. For

the scope of this thesis artificial language such as source code or mathematical equations are excluded from the definition of text, however they are considered words, therefore strings.

What is data mining?

Data mining is the process of collecting, cleaning, processing, analyzing, and gaining useful insights from data. This term is considered a broad umbrella that describes all the different types of data processing that occur from the wide variation of applications, problem domains, formulations, and representations that the modern world faces in real applications. The society today produces an unimaginable amount of data in various formats which are called raw data. These can be ready-to-use, arbitrary or unstructured and that is the case of text. To transform those data into a standardized format, data analysts go through a pipeline process, where the raw data are collected from one ore multiple sources, cleaned, and prepared according to the desired analysis. The reality is that the vast majority of the time of the data analysts is spent in the preprocessing, rather than the actual analysis. This pipeline of processing is a concept similar to the actual ore mining to the refined end product, so the term "mining" derives from this analogy (Aggarwal, 2015).

What is text mining?

Text mining is one of the many subcategories of data mining. In a broad sense it can be described as the process of analyzing text to extract information (Witten, 2004). Several scholars have attempted to define the term. Sebastiani describes "text mining" as a system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information (Sebastiani, 2002). Others have written that text mining is defined as a process of extracting the implicit knowledge from textual data (Feldman and Sanger, 2007). Others described it as a tool used to manage textual information for the discovery of knowledge in textual databases (Tan, 1999). The aim of text mining is to employ technology to analyze more detailed information in the content of each document and to extract interesting information that can be provided only by multiple documents viewed as whole, such as trends and significant features that may be a trigger to useful actions and decision making (Nasukawa and Nagano, 2001). All these definitions are basically describing the same thing, which, in simple terms, is that text mining is the procedure of retrieving some text data, process them and analyze them to extract information and eventually knowledge.

Due to the significant number of publications in the field of Management (Table 1), text mining is a requirement. The application of this tool will prove valuable for the evaluation of the performance of scientific publications.

The concept of knowledge discovery & management

The concept of knowledge discovery is not something new. It appeared in the late '80s and in the beginning of the '90s it was defined as the nontrivial extraction of implicit, previously unknown and potentially useful information from given data (Frawley, Piatetsky-Shapiro and Matheus, 1992). It seems not far from what data mining is and indeed can be applied on structured data (Fayyad *et al.*). However, to apply the established techniques of knowledge discovery on unstructured or semi-unstructured data, some kind of structure needs to be imposed on the text. That needs to be rich enough so that the knowledge discovery algorithmic operations would bear fruit (Feldman and Dagan, 1995). There are models that apply these techniques in textual database but they had limitations (Feldman and Dagan, 1995). However, with today's technological advancement the computational power and available algorithms have the ability to process large amount of textual data.

It is important to note that, in the sense of Information Systems there is a continuum of data being analyzed to information, which is studied to provide knowledge and sometimes even wisdom within any entity, whether it is an enterprise, or a university or even a single person. The more academic view of this continuum says that data consists of facts, images, and sounds. When data are combined with meaning and interpretation, information emerges. Information is formatted, filtered, and summarized in a specific way that, when combined with action and application it becomes knowledge. Knowledge exists in forms such as ideas, rules, procedures, and instincts that guide actions and decisions.

Research Questions

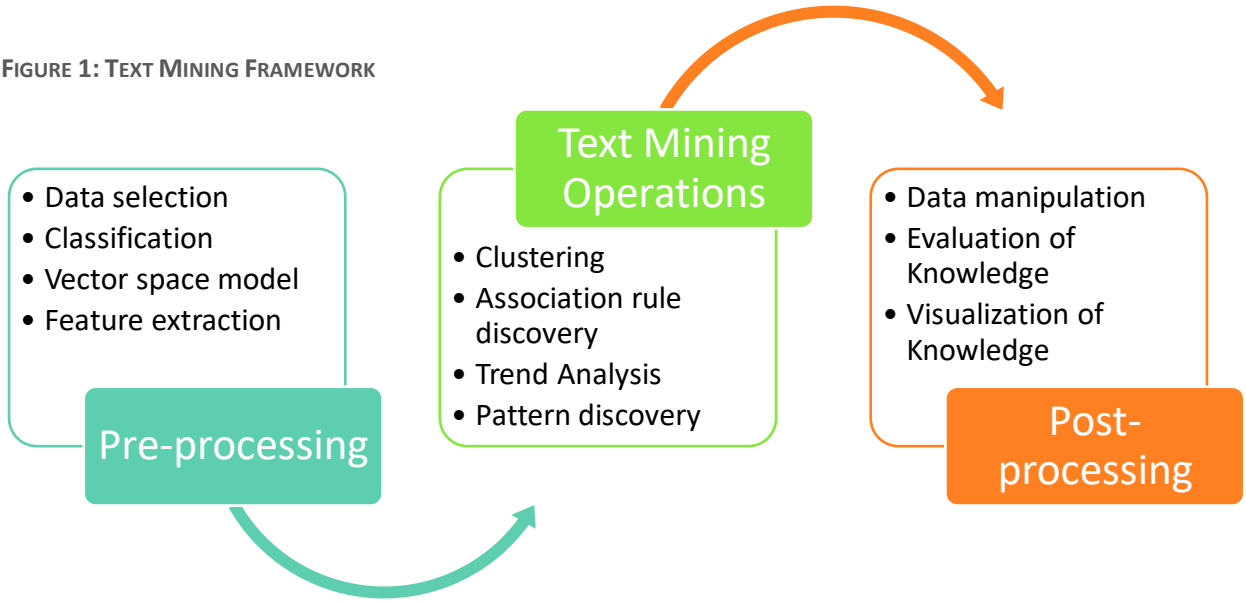
To evaluate the performance, we will try to discover if similar text patterns exist in the articles among the journals and what these patterns look like. Do these patterns describe the success of these publications in terms of citation numbers? In addition, it may be interesting to respond to the following questions (Salloum *et al.*, 2018).

1. What are the most frequent keywords in the collected articles?
2. How are the articles distributed in terms of years?
3. Could text mining applications be of assistance in predicting article citations?

Methodology

The application of text mining consists of a few tasks, done in three steps. Each individual or group of researchers uses a customized, although similar framework according to the hypothesis and research questions. There are three major steps that are included in text mining: text pre-processing, text mining operations, and post processing. The first step is text pre-processing. It includes tasks such as data selection, classification, and feature extraction. Generally, the documents are converted into intermediate forms, which should be suitable for different mining purpose. Text mining operations are the central part of a text mining system is the second step. These include clustering, association rule discovery, trend analysis, pattern discovery and other knowledge discovery algorithms. In the third step post-processing tasks manipulate data or knowledge coming from text mining operations, such as evaluation and selection of knowledge, interpretation and visualization of knowledge (Jo, 2019) (Zhang, Chen and Liu, 2015) (Kobayashi *et al.*, 2018).

FIGURE 1: TEXT MINING FRAMEWORK



Data Selection

Data selection, as part of the text preprocessing, results in a collection of text, which is called corpus (Jo, 2019). It is a mandatory task for the research to proceed further. The corpus will be established by the exploitation of the Web of Science Core Collection from Clarivate. There is a significant number of publications, as seen in Table 1 (να βάλω όλα τα έτη). The search will be limited to ten years, from 2009 to 2019. For the purpose of this thesis the main focus will be on articles in the field of Management published in Q1 or 4* journals such as *Academy of Management Journal*, *Academy of Management Reviews, Journal of Management*. It is preferred

to study articles and reviews form renowned journals because the publishing protocols are widely accepted. The reviewers and editors are hard to please and therefore the process of publishing is strict. Those journals contain other categories of articles besides management: business and applied psychology (Table 2). It is important to note that each discipline and field differs from the others, therefore interdisciplinary comparisons would have little to no importance.

In addition to the above those journals were also chosen, since full-text articles are accessible. The Athens University of Economics and Business has a subscription plan for the timespan of choice. According to literature full-text articles and abstracts differ in structure (Cohen *et al.*, 2010). Studies generally report that valuable information can only be found in the full-text of an article and one study noted that the majority of the claims contained in an article are not reported in the abstract (Westergaard *et al.*, 2018). Although mining abstracts can be time efficient due to their smaller size and availability, favoring full-text articles greatly enhances the results with a limited impact on precision (Martin *et al.*, 2004).

The query will be refined to the specific journals, by searching all fields with the respective ISSN. It will also be limited to document type which will be Articles or Review Articles, to the category that will be Management and to the years of publication (from 2009 to 2019). Editorials, letters, conference proceedings etc., that might be published and available on these journals will be excluded from this research (Antonakis *et al.*, 2014). It is worth mentioning that no specific keywords are included for the present research. The corpus will not be limited to papers that are highly cited, therefore while refining the query in the Web of Science platform this option will not be selected. Such limitation would not be necessary for two reasons. The first reason is that the list of articles in this range will be shorted by highly cited and the second being that from each Journal the first and maximum of 350 records will be collected.

The implementation of the text mining framework will take place on the KNIME Analytics Platform where specific nodes for each operation exist to make the process time efficient and effective (Thiel, 2009).

Text Indexing

Text consists of long strings which in their raw form cannot be processed by a computer program. To transform the text into a format that can be operated by software it is necessary to index it. Text indexing is the process of converting texts into a list of words (Kowalski and Maybury, 2000). In other words, it can be viewed as the segmentation of sentences into words, that are short strings and therefore they can be encoded to numerical values. This process has three steps; Tokenization, stemming, and stop word removal, that are briefly described below.

Tokenization

After the corpus is established tokenization or enrichment of the documents is due. Tokenization, as the initial part of indexing, is the process of breaking a stream of text up into meaningful elements such as words, phrases, symbols etc. called tokens (Verma, Renu and Gaur, 2014). This step is perquisite for the other two steps which order does not affect the text mining operations.

Example of tokenization

After the corpus is established tokenization of the documents is due. Tokenization, as the initial part of indexing, is the process of breaking a stream of text up into meaningful elements such as words, phrases, symbols etc. called tokens…

Tokenization

Tokens
after
the
corpus
is
established
tokenization
of
…

Stop-word Removal

Stop words are irrelevant to the text content and function only grammatically. These words can be articles “a”, “an”, and “the”, prepositions, such as “in”, “on”, “to”, and conjunctions, like “however”, “moreover”, “and”, “or”, “but”, etc. It is wise to remove those words to improve the efficiency in text processing. Some conjunctions could prove useful for sentimental analysis since they can indicate opinion, although the sentiment of research publications does not fall into the scope of this thesis.

Filtering

Stop word removal falls into an umbrella category of filtering, which includes punctuation erasure, number filtering and case conversion. These operations may also be utilized and implemented on the corpus to increase efficiency. During this process the elimination of N-character words can be utilized if it is necessary for the analysis to follow (Verma, Renu and Gaur, 2014). Indeed, removing those terms reduces the dimensionality of the term space and makes the documents heavier with the valuable information the whole process looks after.

Stemming

The tokens created from the previous step are now converted into their own root forms using grammar rules. This refers to parts of speech such as nouns, verbs, and adverbs (Kowalski and Maybury, 2000). The two versions of stemming are Illustrated in the example below. In the hard stemming, words are converted into their respective verbs or adjectives. In the soft stemming, words are the root forms by themselves. Here the part of speech (POS) tagging is an optional step.

Varied Form	Root Form	Varied Form	Root Form
better	good	assignment	assign
best	good	complexity	complex
simpler	simple	analysis	analyze
simplest	simple	categorization	categorize
categorizing	categorize	assigning	assign
categorizes	categorize	assigned	assign

FIGURE 3: HARD VS. SOFT STEMMING

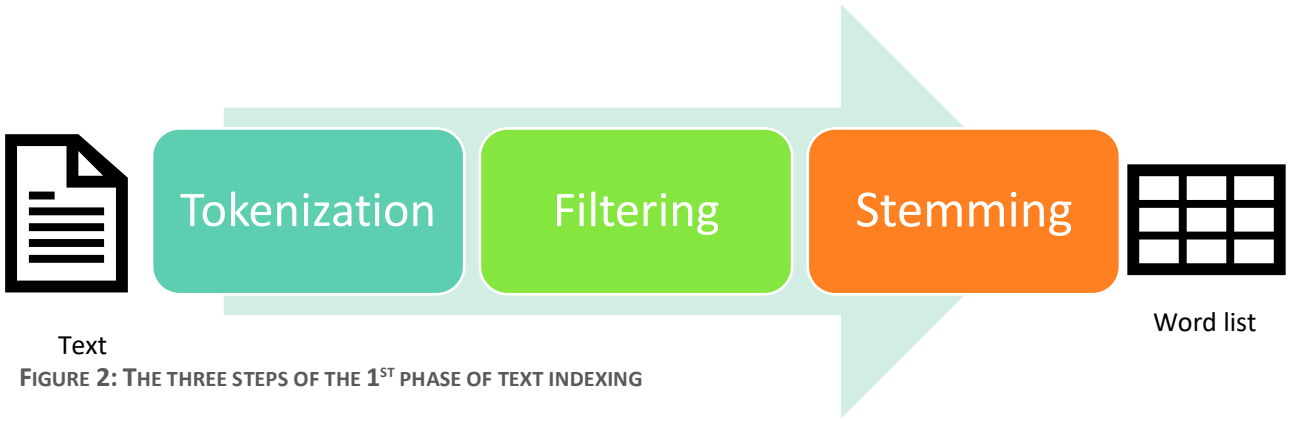


FIGURE 2: THE THREE STEPS OF THE 1ST PHASE OF TEXT INDEXING

References Removal

The articles consist of the Title, the authors, the abstract, the keywords, the main text of the article in sections and the references at the end. The latter contain text information that are not relevant to the text mining analysis of this thesis, therefore they had to be removed. It is advised to do this process before proceeding with the preprocessing steps that are analyzed in this section of the present document. To do this, the first stem is parsing the PDFs that comprise the corpus of the study. This is the input of the data to the KNIME platform. For this purpose, the first node that reads the text from locally stored files in PDF format, is the *PDF Parser*. The output of this node is a table that contains a list of documents that are ready to be processed by other nodes. The table contains a "Document" column with the respective title of each file in the description. To make sure that the document has been parsed correctly a Document Viewer node is connected to the output port. This node will assist the preprocessing phase by letting the analyst see the output document after each stage of preprocessing to determine whether each node has served its purpose. After the *PDF Parser* node, the *Document Data Extractor* node is utilized. The documents are transformed into strings contained in cells. When reviewing the output table, it is found that after the word References (with capital R) exist the strings that represent the irrelevant text.

Moving on to the *Cell Splitter* node the cells mentioned above, are separated into two columns. In the configuration dialogue the entered delimiter is the word References. It is assumed that the possibility of this word existing in the exact format within the documents is rather low. Therefore,

using this delimiter will result in the desired outcome. The first column contains the text that will be further processed and the second contains everything after References. In the next step the *Strings to Document* node is applied. This step is a necessity for the KNIME platform because the strigs cannot be analyzed in this format. They had to be transformed into the native Document format KNIME exploits for further analysis. By viewing the output documents of this process and by cross referencing with the original PDF files in Adobe Acrobat, the argument in *Cell Splitter* node is proved to have functioned properly. For the sake of simplicity, the *Column Filter* node removes the unwanted columns form the output table and the result is the document that will be preprocessed according to steps mentioned and explained above. In the following section the detailed configuration of the platform is presented.

Data preprocessing process in KNIME

After the *Column Splitter* node, the data in Document format are ready to move to the preprocessing phase in the KNIME Analytics Platform. The documents contain a significant number of names that can be tagged at this stage, in order to be removed at a further point in the analysis if they cause any problems. The tagging of names is done with the *OpenNLP NE tagger* node. This node recognizes named entities based on OpenNLP Name Finder models and assigns the corresponding tags to them. The version of the underlying OpenNLP framework is 1.8.4. The built-in models, pre-trained models are from OpenNLP version 1.5.

Right after the name tagging the next step is the removal of the tagged Named Entities in the tagged document with a *Tag Filter* node. Moving on the next step is the part of speech tagging of the document. The *POS Tagger* node is configured to tokenize the documents with the OpenNLP English WordTokenizer and executed. This node assigns to each term of a document, a part of speech tag, e.g., verb, noun, adverb, adjective etc. Although it is not an obligatory step, it is advised to precede the aggressive document cleanup that comes afterwards. The next node is the *Case Converter* node which is configured to transform all upper-case characters to lower-case ones and creates a new “Preprocessed Document” column. Then comes the *Number Filter* node that removes numbers from the preprocessed documents. In the filter options/filtering mode “Filter terms containing numbers” is selected. Moving on to the *N Chars Filter* node, which filters all term containing less than a specified number of characters. In this case number 4 is selected in the configuration dialog. This step is helpful because it removes from the preprocessed documents small words like conjunctions and articles that carry little to no meaning. The Punctuation Erasure node follows the N Chars Filter, and it removes punctuation marks like, periods, semi colons, commas, etc. The *Stop Word Filter* node filters all terms of the input preprocessed documents, which are contained in the specified stop word list. The node provides built-in stop word lists for the English language (Porter Stemmer), and it is configured accordingly.

The penultimate step in the preprocessing/cleanup of the corpus is the stemming. The *Snowball Stemmer* node stems all the terms in the preprocessed document with the Snowball stemming library. The final step in this stage is the creation of a Bag of Words (BoW). The *Bag of Words Creator* node is selected and configured to create the BoW from the preprocessed document. It is important to exclude the Document column in the node dialog, to make sure that the preprocessing and cleanup serve their purpose. The output is a table that contains a list (at least one column) of the terms/words that are included in the preprocessed documents.

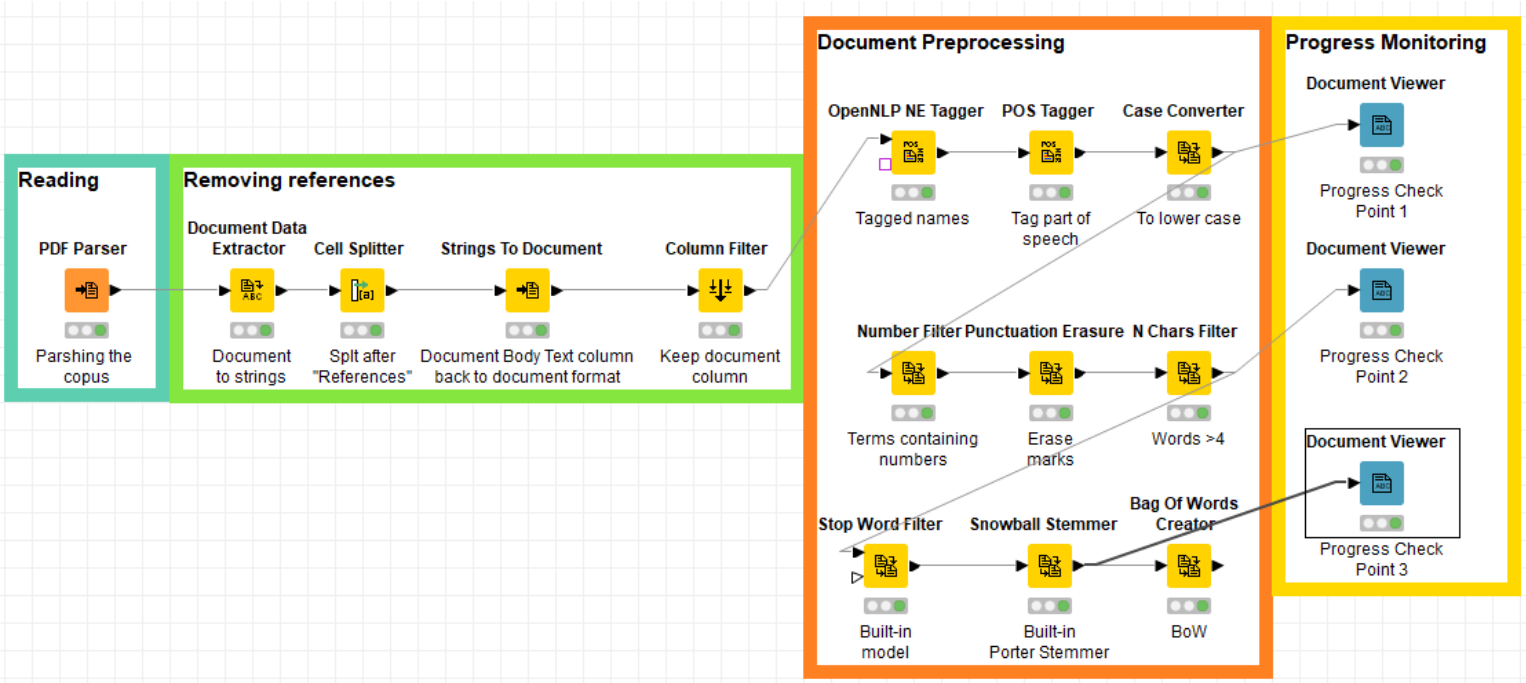


FIGURE 4: KNIME WINDOW LAYOUT

Text Encoding

The output of the process described in detail above, will become the input of the following phase of the preprocessing stage which is text encoding. The goal of text encoding is to move from a word-based representation of the data included in the corpus to a number-based representation. The latter type of representation is especially important if the ultimate steps in the Text Mining process are classification, clustering, prediction, or the application of machine learning algorithms in general.

Word Frequencies

The result of the first phase of preprocessing is the bag of words for each document, which is essentially a series-list of words. These words can also be called features. The idea here is that if a word occurs more often in one document that it does in others, then it should receive a higher

frequency score. This score is the absolute number - count of occurrences of a word in document represents the **absolute frequency (AF)**. This metric would be applicable if the documents had the same lengths, which in this case they do not. For the corpus of the present thesis, where each document varies in length, **relative frequency** would be a better metric. Relative frequency is the normalized absolute frequency, where the term AF is divided by the total number of words in the text. In addition, document length normalization here may not be necessary, however in information retrieval, when documents vary in lengths, it is an advised step (Singhal *et al.*, 1996).

However, sometimes just counting is a mere crude measure, and may not be enough to proceed with analysis. In 1972 another, more sophisticated, statistic was introduced to help researchers understand the term relevance of a single document in a group of similar documents, in this case research publications, which is the **Inverse Document Frequency (IDF)**. It is also known as the weighting function (Sparck Jones, 1972).

$$\mathrm{IDF}(t_1) = \log \frac{N}{f_i} \quad (1)$$

IDF assigns a weight on a given term by counting the number of occurrences f_i of the given term in the whole document collection N according to formula (1). Originally the base of the logarithm was 2, but it has been simplified to the current form over the years (Robertson, 2004). This formula assigns a higher score, meaning a stronger weight, to words occurring less frequently across the corpus, because of a smaller denominator and vice versa. The core of the function lays in the idea that less frequent terms are better identifiers of the document they belong to, by standing out from the rest of the word list. It is heuristic and although sometimes problematic, there are good theoretical justifications in the traditional probabilistic model of information retrieval.

To make the analysis more robust, yet another frequency measure is introduced, the **TF-IDF** or **term frequency – inverse document frequency**. This is the normalized word frequency, the most popular weighting schema, with many applications in information retrieval. The "-" represents a hyphen and not a minus sign. This measure computes the *product* of TF and IDF following formula (2) (Salton and Yang, 1973).

$$w_{t,d} = \log\left(1 + tf_{t,d}\right) \times \log_{10}\left(\frac{N}{df_t}\right) \quad (2)$$

- $w_{t,d}$: weight of the given term in the document, the inverse document frequency
- $tf_{t,d}$: normalized term frequency – how many times the word occurs in the document
- N: number of documents in the corpus
- df_t: document frequency – how many documents contain the word.

According to the formula (2) results can vary from high to low scores/values depending on the times term *t* is found and how many documents comprise the corpus.

- The highest value: when term *t* occurs often, and the N number of documents is small.
- Lower values: when term *t* less often, and the N number of documents is big.
- The lowest value (0): when term *t* occurs in all the N documents. To mitigate that effect, instead of the normalized IDF Figure (1), the probabilistic IDF could be used[3].

Considering the example illustrated below it is possible to understand better how this formula works. A random sample of 5 articles were selected from the Journal of Management. The manually searched term was "*innovation*". In one document the word occurred only in the references section of the document. This section will be trimmed from our documents, hence TF=0.

DOI	**Term Frequency**	**TF-IDF**	**Excel Function**
10.1177/0149206311406265	65 (E4)	0,176	=LOG10(1+E4)*LOG10(E9/4)
10.1177/0149206314527128	**366**	**0,249**	
10.1177/0149206314527130	4	0,068	
10.1177/0149206308321554	28	0,142	
10.1177/0149206311415280	**0**	**0**	
N =	5 (E9)		
Term =	innovation		

TABLE 1: EXAPMLE OF TF-IDF FUNCTIONALITY

The formulas above were developed by testing relatively short texts such as article abstracts or short catalogue records. Here the documents vary significantly in length, therefore the good document length normalization is of critical importance (Robertson, 2004). It is preferred to use the relative term frequency, to avoid over- or underestimation by the text lengths.

To achieve that it is possible to modify function (2) by replacing the absolute TF with the relative frequency. The equation (3) is as follows:

$$w_{t,d} = \log\left(1 + \frac{tf_{t,d}}{tf_{max}}\right) \times \log_{10}\left(\frac{N}{df_t}\right) \quad (3).$$

Assigning frequencies to words is essential to the next step which is the document vector creation. After all, the ultimate goal of text mining is to transform words to numbers who can be further analyzed with traditional statistics and machine learning algorithms. This TF-IDF method can also be utilizes to remove common words from the corpus that do not carry significant

[3]Probabilistic $IDF(t_i) = \log\frac{N-f_i}{f_i}$

meaning and have not been filtered out in previous steps (Sebestyén, Domokos and Abonyi, 2020).

Keyword Extraction

To extract the keywords of each document the *Keygraph Keyword Extractor* node was employed. This node analyses documents and extracts relevant keywords using the graph-based approach. The method is described in length in the "KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor" by Yukio Ohsawa (Ohsawa, Benson and Yachida, 1998). On the configuration pane it is set to return 10 keywords per document. Afterwards a Word Cloud operation is performed to view the most common keywords.

Vector Space Modeling

After the computation of TF-IDF it is time to represent each document in the collection as a vector of dimension $W, w_d = (w_{1d}, \dots, w_{Wd})$ (Salton and McGill, 1983). To do so, features must be selected. The feature candidates are the words contained in the BoW created in previous steps. These features are associated with the dimensions. Each entry w_{id} is the TF-IDF of the term i in the document d and the computations of TF-IDF is thoroughly explained in the previous section. Moreover, binary values could be assigned to the features. The process is called hot encoding however since TF-IDF is an important metric it is preferred to it. The result is a matrix where the words/ features are the columns, and the documents are the rows.

The vector space provides an approachable method to similarity computations among the documents in the collection (Maas *et al.*). It is also necessary for the next steps of mining such as classification and clustering. However, the vector space model has two distinct flaws. The first one is huge dimensionality. It is causes by the several hundred thousand of features in the matrix. Several theories have been proposed for the reduction of dimensionality. This is a problem especially for representation, understanding and visualization of the data but can also lead to high demand of computational power to perform mining tasks. The second issue is that zero values are more dominant within the matrix than non-zero ones. This sparse distribution of each feature vector can affect the expected performance of classification and clustering, because of the lack of discrimination among them (Jo, 2006).

To simply illustrate why dimensionality is important an example, outside of data science, could be considered. Let us imagine our palm, a three-dimensional object, in a dark room in front of a wall. Now we illuminate our palm with a light source from the front, and we examine the shadow it casted on the wall behind it. The shadow is the two-dimensional image of our three-dimensional palm. If we can see our five fingers and the general shape of our palm, we can still understand that the object that casted this shadow is a palm with certainty, without having to

verify that it was a palm indeed in the three-dimensional space. In this example we were able to reduce the dimensions by one without misinterpreting the initial object. Could this be the case for the corpus of the present thesis?

Dimensionality reduction methods

To reduce the dimensions of the space created, improve accuracy, and eliminate irrelevant data, it is possible to consider four schemes of selecting features, rather than keeping all the words as features in the BoW. These are the:

- Wrapper approach
- Principal Component Analysis (PCA)
- Independent Component Analysis (ICA)
- Singular Value Decomposition (SVD)

Wrapper Approach

In this scheme the selection or elimination of a subset features is based on the observed performance of the model. In a way, the algorithm evaluates the optimization of the accuracy rate (Sebban and Nock, 2002).

Principal Component Analysis

PCA is not a novel transform method and has applications in many different fields from genetics to computer science and data analysis. For the issue of feature selection, the eigenvector is exploited to evaluate the contribution to the result of the feature extraction of each feature component. It has shown encouraging results when implemented in dimensionality reduction (Song, Guo and Mei, 2010).

Independent Component Analysis

Like PCA, ICA is also an eigenspace method. ICA generally performs better when the data do not follow the Gaussian distribution (also known as normal distribution). In this method the original feature vectors are transformed into statistically independent directions by maximizing the degree of non-Gaussianity of the new directions (Prasad, Sowmya and Koch, 2004).

Singular Value Decomposition

SVD is another way to manipulate a matrix by factorizing it into singular vectors and singular values. It is the process through which the $n \times N$ matrix is decomposed in three components: the orthogonal matrix $n \times n$, the diagonal matrix $n \times N$ and the orthogonal $N \times N$ matrix (Poole, 2014).

Document Vector Creation in KNIME

For the encoding operations in the KNIME Analytics Platform there are some predecessor nodes to the *Document Vector* node that will represent the documents in the vector space. The first one being the *TF* node. This step calculates the relative term frequency of the words in each document. Then comes the *IDF* node which computes the smooth inverse document frequency of each term / word in the collection of documents. Afterwards, by utilizing the *Math Formula* node the product of TF*IDF is computed. To take a first peek at the results the *Tag Cloud* node is applied at the TF and at the TF*IDF point. At the configuration window, the number of displayed rows is set to a 1.000 instead of the 2.500 default number. For visualization purposes this number can be further reduced. By making the word cloud, it is possible to have a visualization of the most frequent terms and get an idea of what is the collections most talked about terms without having to read thousands of documents.

Clustering & Classification

By creating a vector space, the clustering and classification processes are becoming feasible. To perform those applications a data mining model has to be trained by partitioning the dataset in a training set and a test set. These operations fall outside of the scope of the present work; however, they could be considered as the following steps. It is important to note that the creation of such model could predict whether a future scientific publication would be accepted and published in renowned academic journals. To create this supervised machine learning model the team of people would need to analyze, a significant amount of already published articles – a lot more than this thesis -, a number of rejected manuscripts, and a well-informed team of peer reviewers to supervise the outcomes and the training of the model.

System Requirements

To analyze this volume of data in the KNIME Analytics Platform the following System Requirements for hardware and software had to be met:

- Operating System: Windows 10 (latest version)
- KNIME Analytics Platform, downloaded from the respective website, kept updated on the latest version
- Processor: Intel Core i5-7400 @ 3.00GHz
- Memory: 16 GB (DDR4)

Additional steps to make the analysis possible:

- Set RAM priority in task manager in high every time the program is initiated, to avoid software crashing during heavy computational operations like Bag of Words creation or relative term frequency calculation.

- The platform will still run out of memory. An additional step that seemed to resolve the memory issue was to change the default memory. The settings are located in the file knime.ini, which is in the installation in the directory knime-vversion/bin/arch/knime/. The file should contain the following settings **-Xmx1024m**. In order to increase the amount of heap memory the -Xmx setting should be modified. For example, to allocate 2048 MB of heap space to the JVM change -Xmx1024m to -Xmx2048m. In this case it was changed to -Xmx8g.

Performance Evaluation

To assess the factors that contribute to a high number of citations the following dependent and independent variables are selected. Linear and logistic regression models are going to be utilized to determine the relationships between the variables if any (Antonakis *et al.*, 2014). The dependent and independent variables for the linear regression model are listed and explained below.

Dependent Variables

1. Lifetime article citations
2. Average article citations per year, considering publication year

Independent Variables

1. Article type (Dummy variable 0; Research article and 1; Review article)
2. Number of authors
3. Age of article (Number ranging from 2 to 12; result of the subtraction of 2021 – 2009 = 12 if the article was published in 2009)
4. The Issue number of the Journal
5. The length of the article as the number of pages
6. Number of references (absolute number of the cited literature in each article)
7. Sum of the TF*IDF result from the text mining analysis
8. Term count in each article. This number represents only words with 5 or more characters and excludes words that carry little significant meaning such as articles, pronouns etc. and it was calculated from the text mining operations mentioned above.

Performance Prediction

To predict weather a manuscript will be highly cited or not the variables above can be utilized if the corpus is divided in two groups. To do that the average of total citations is computed separately for Reviews and Articles. If the record has a total number of citations above this average, then it's marked as highly cited, or low cited otherwise. This rough classification will function as the basis of a binary logistic regression that will predict using the variables above

(excluding the dependent variables) whether a manuscript will be highly cited before being published.

Statistical Analysis Platform

All statistical analysis took place on IBM SPSS software, while the preparation of tables and visualization of descriptive statistics was performed using Microsoft Excel pivot tables (latest version).

Results

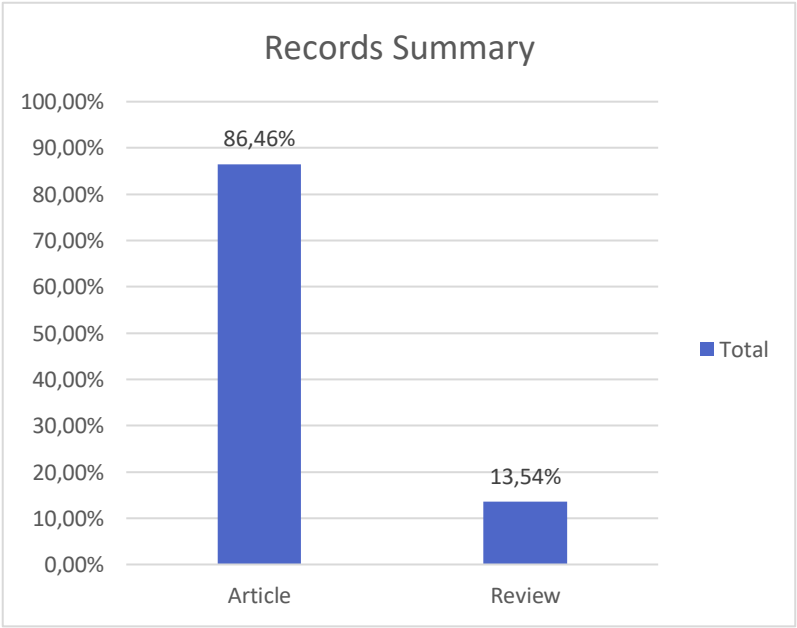
Master Table Creation[4]

The results of the analysis that took place in the KNIME Analytics platform were extracted in Excel tables so that the rest of the variables could be included to be further analyzed. The exported tables were pivoted by article title, sum of TF*IDF and term count and the respected values were matched with the master table that contained all other variables. The values of the other variables were exported from the Web of Science search report and were matched accordingly. To finalize the master data table the average of lifetime article citations was calculated and was used as a threshold to characterize two distinct groups in the collection: highly cited and not so highly cited. Articles that had more lifetime citations than this average were considered as highly cited and vice versa.

Constraints

The pdf files from Journal of Management were consistent in format and the algorithm was able to extract all relevant information from the documents therefore the pivots worked as they should, and it was possible to match the text mining variables with the rest of the data. This was not the case for the other two Journals. The files that contained the documents were not all created equal like the ones form Journal of management. As a result, the algorithm was not able to read them properly and that led to a misinterpretation of TF*IDF and term count. That led to the decision to exclude them from further analysis and representation in this Thesis.

The KNIME algorithm was accurate with the pdfs from Journal of management so all articles from the studied period were collected. In total our viable dataset consists of one Journal and 724 records, 626 of which are articles and 98 are reviews, 86.5% and 13.5% respectively. For this dataset further analysis was conducted and will be presented below both visually frow word clouds that were created on the KNIME Analytics Platform along with logistic



[4] The Master data table consists of 725 of rows and 13 columns. That makes it impossible to include in the Appendix. The reader can view it here.

regression statistical analysis. In addition, some descriptive statistics are depicted below.

Descriptive Statistics

Highly or low cited

The studied groups would not consist of different journals due to the algorithm limitation mentioned above. To create two groups in our collection form Journal of Management the average of lifetime article citations was computed. This number was used as a threshold to characterize two distinct groups in the collection: highly cited and not so highly cited. Articles and reviews that had more lifetime citations than this average was considered as highly cited and vice versa. In the bar chart below, low cited and highly cited articles are depicted in terms of their type, ether Article or Review.

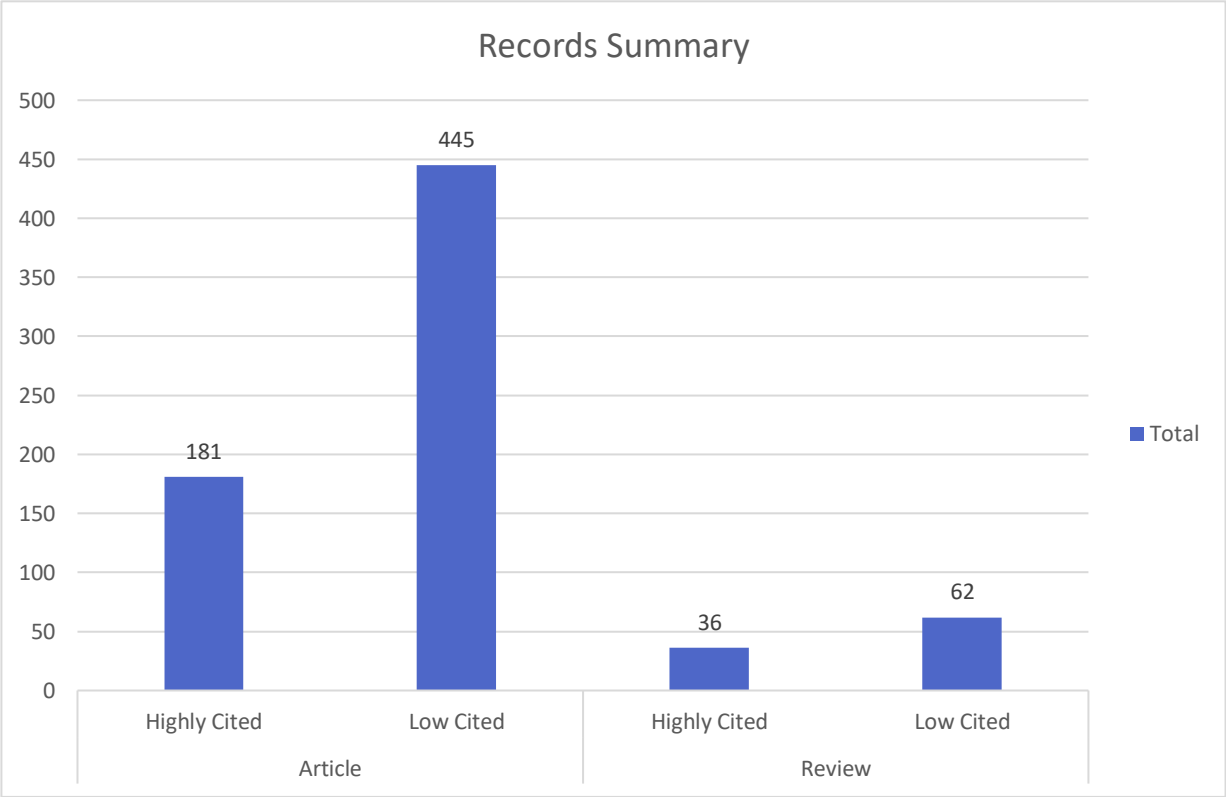


FIGURE 6: NUMBER OF RECORDS MARKED AS HIGHLY OR LOW CITED PER TYPE OF ARTICLE

Term Count

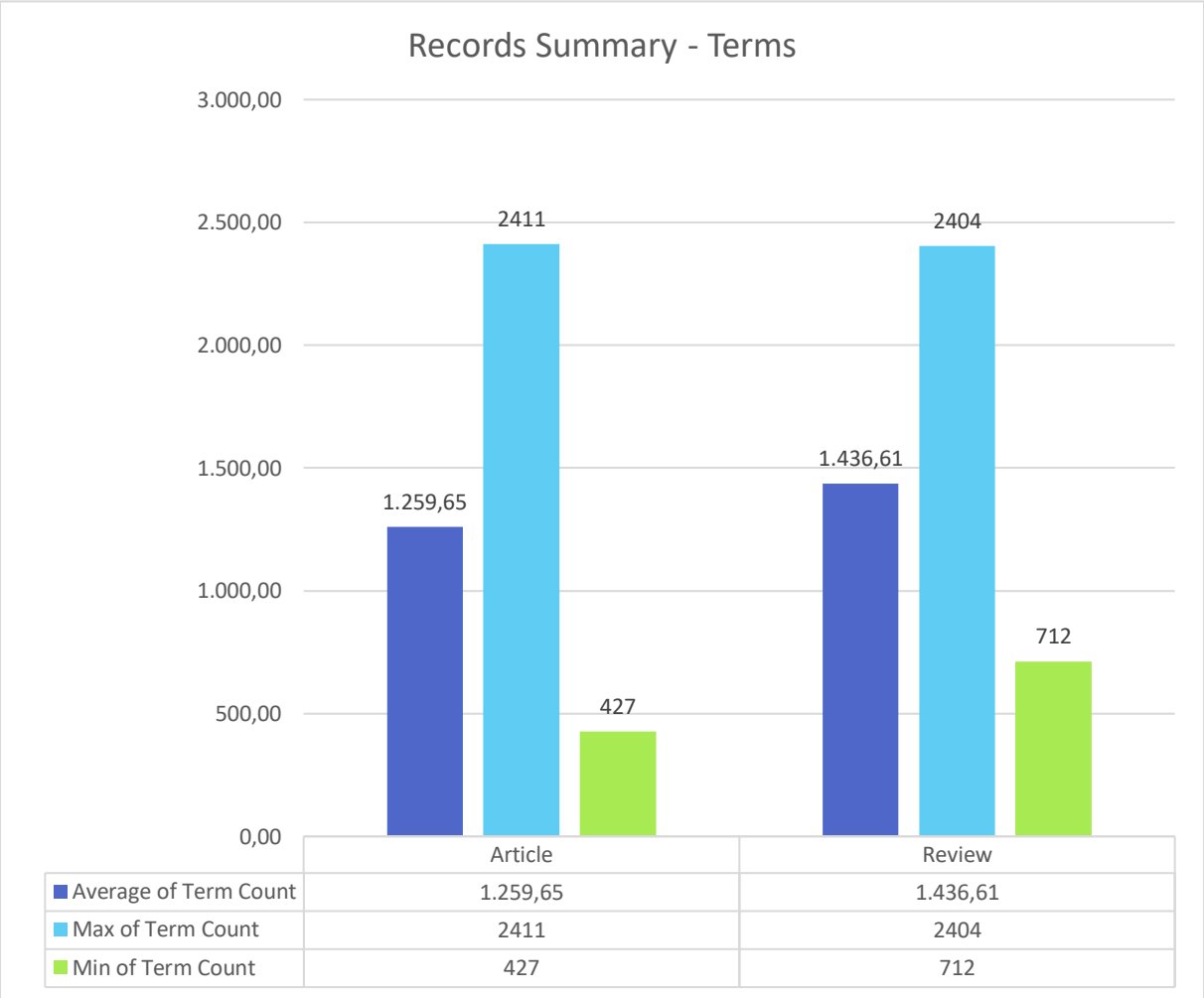


FIGURE 7: LAYOUT OF TERM COUNT

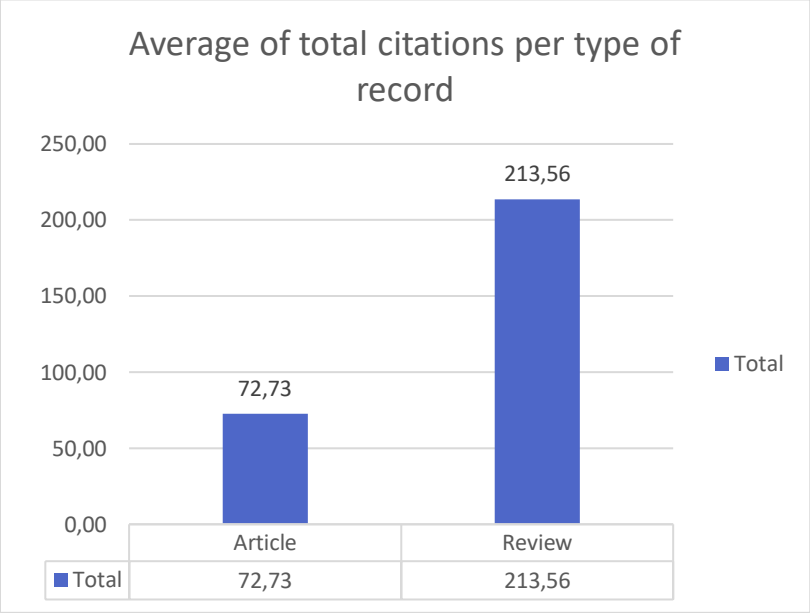
On average it is apparent that reviews have more words than research articles which is represents the reality since reviews have on average more pages than articles (see Table 2). The minimum term count for articles is 427 words and for reviews is 712. The maximum is approximately 2400 words for both types of records.

Row Labels	Average of Term Count	Average of Pages
Reviews	1436,6	**33,26**
Articles	1258,9	**27,72**
Grand Total	1282,9	28,47

TABLE 2: TERM COUNT AND PAGES

Although, as seen on the table above, Reviews have on average 6 more pages than research articles it is interesting to note that the maximum word count does not differ dramatically, only by 7 words. This can be due to the presence of tables or even pictures on the pages of a review article.

Total Citations



It is worth mentioning the average of total citations in Reviews and in Articles. Although reviews are only a fraction of articles on average, they get more cited than pure research articles. This me be since they function as extended summaries of other research articles.

FIGURE 8: TOTAL CITATIONS

Per year Publications

Moving on to see how the publications are growing per year, as shown on the trend below, it is apparent that more and more articles are published every year that goes by. In 2009 Journal of Management published only 54 articles in its 8 issues per year. By 2019 that number reached 89 and in 2020 the total published articles in the Journal were 98. Furthermore, it could be interesting to look at how the articles are distributed over the issues over the years as a total count per issue. Titles that are published in the 7 the first seven

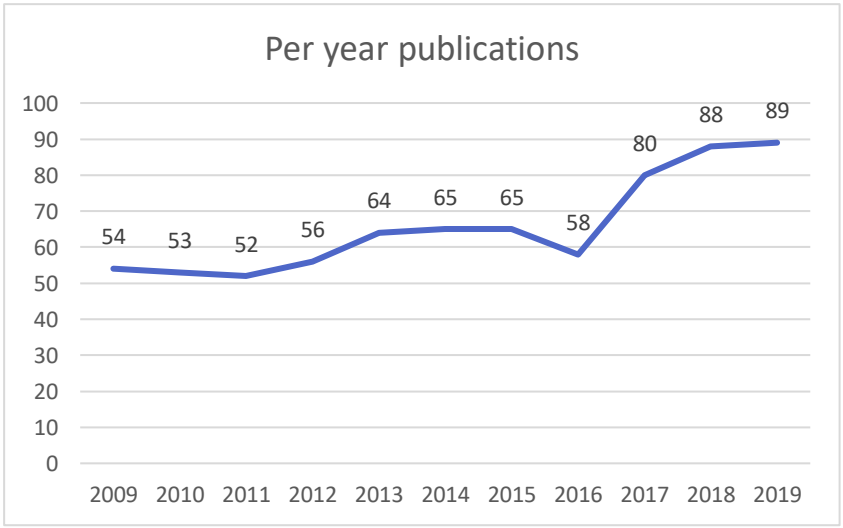


FIGURE 9: PUBLICATION DISTRIBUTION OVER THE DECADE

issues are almost even. The exception is the 8th issue of the year which has the lowest count of published articles. That is because the 8th issue appeared later in the decade, in 2017. Please see Table 3 in the Appendix for each year analysis.

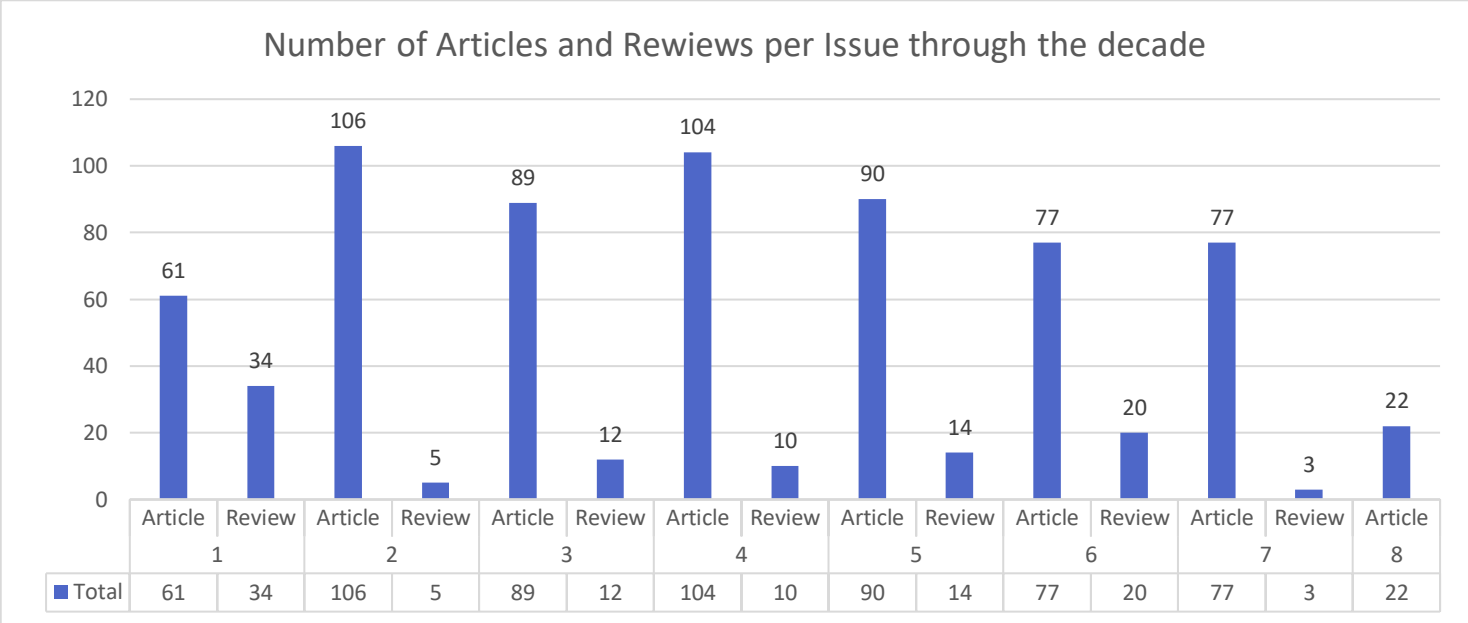


FIGURE 10: TOTAL NUMBER OF RECORDS PER TYPE & ISSUE NUMBER

Cited References and number of authors

In the bar chart on the right, it is apparent that on average reviews cite 42% more articles in comparison to pure research articles and this is expected due to their nature. In the following chart it is worth mentioning that the average of the number of authors is around three across the collection. There is one article with 23 authors which is an absurd number of authors, although there are some cases of articles that are composed by whole research departments and therefore all members of the department are considered as authors. There are 46 records in total that are composed by one

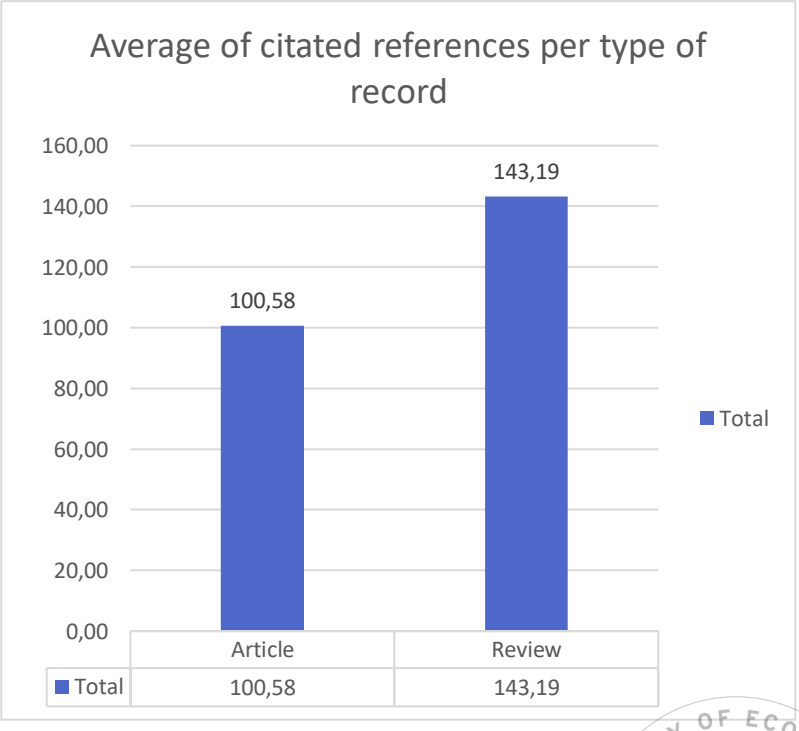


FIGURE 11: CITED REFERENCES PER TYPE

author and five records that have seven authors. Only 4 records have more than 7 authors as seen on Figure 13.

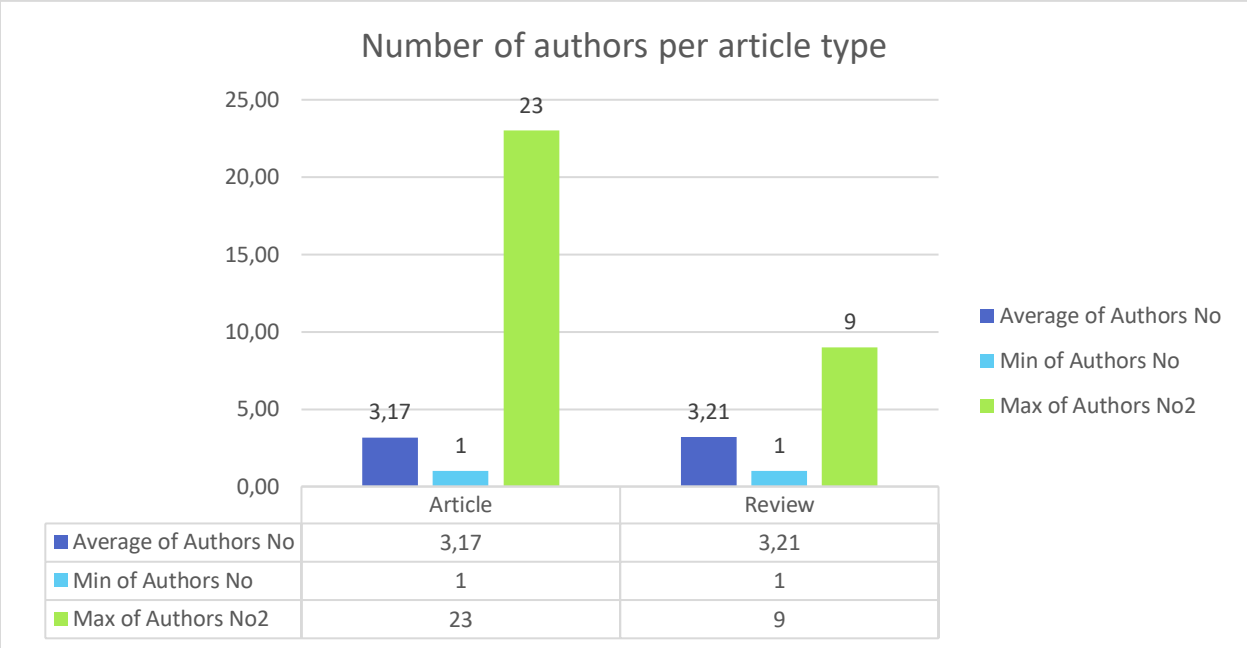


FIGURE 12: NUMBER OF AUTHORS PER TYPE OF ARTICLE

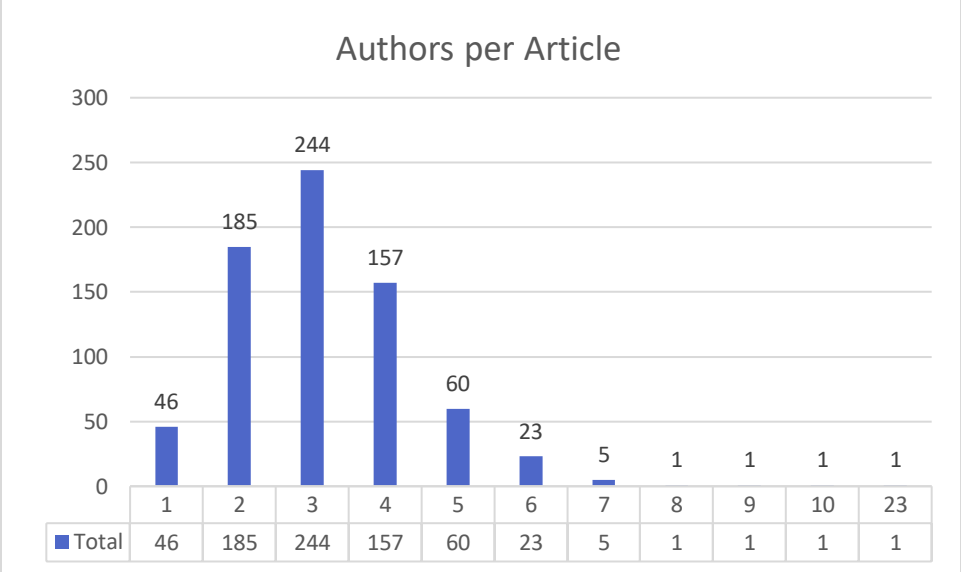


FIGURE 13: AUTHORS PER ARTICLE

Word Clouds

Reviews

The KNIME Analytics Platform has the option to create word clouds after the creation of BoW and the calculation of TF and TF*IDF for each word. Word clouds are a way of visualizing results in terms of word frequencies and can function as an overview of what the given collection is about. Below the word clouds for Reviews.

FIGURE 14: WORD CLOUD FOR REVIEWS - TF

FIGURE 15: WORD CLOUD FOR REVIEWS TF*IDF

Articles

Below the word clouds for Articles.

FIGURE 16: WORD CLOUD FOR ARTICLES TF

FIGURE 17: WORD CLOUD FOR ARTICLES TF*IDF

From the created word clouds and without physically reading any of the downloaded pdfs it is safe to assume that reviews are mostly about stress, stress factors and business phycology, while articles talk about employees' commitment and attitude, growth, and organizational change. The

word research is omnipresent and that is an indicator of good algorithm performance. Some metrics are available in Table 10 in the Appendix.

Keywords Word Clouds

Below the word cloud of the keywords for Reviews.

FIGURE 18: WORD CLOUD OF KEYWORDS FOR REVIEWS

FIGURE 19: WORD CLOUD OF KEYWORDS FOR ARTICLES

In Table 11 in Appendix are the top 10 key words for Reviews and Articles.

Linear Regression

Lifetime Article Citations

The first linear regression that was performed had as the dependent variable the total citations of each record and the method was forward. This could give the best model considering all the predictors. The result was three models and here the third model is analyzed. The predictors for the best model as a result of the forward method were the age of the article, the type of the article either review or research article and the number of cited references. The value of R squared is 0,246, adjusted R squared is 0,242 which does not differ from R squared and the p value of the model equals to 0,001, therefore the model is considered significant. The model assumptions were met. Please see Table 3 below for respective coefficient metrics.

Coefficients		
Model 3	**B**	**Significance**
Constant	11,74	0,614
Age	14,198	***
Review/Article	-85,877	***
Cited References	0,591	***

Table 3: Coefficients Table for Total Citations As Dependent Variable

Average Citations per year

The second linear regression performed had as the dependent variable the average citations per year of each record. It is worth mentioning that this average is calculated separately for each record taking into consideration the year of publication. The result was six models and here the sixth model is analyzed. The predictors for the best model as a result of the forward method were the type of the article, the number of cited references per record, the issue they were published, the volume of the Journal, the count of terms and the number of authors. The value of R squared is 0,171, adjusted R squared is 0,165 which does not differ from R squared and the p value of the model equals to 0,001, therefore the model is considered significant. The model assumptions were met. Please see Table 4 below for respective coefficient metrics.

Coefficients		
Model 6	**B**	**Significance**
Constant	33,992	***
Review/Article	-8,782	***
Cited References	0,088	***
Issue	-0,761	**
Volume	-0,403	*
Term Count	-0,006	*
Authors	0,700	*

Table 4: Coefficients table for Average Citations per Year as Dependent Variable

Binomial Logistic Regression

Predict whether a record will be highly or low cited

By grouping the dataset of published records in two classes, highly or low cited a binomial logistic regression was performed. The selected method was forward LR and it resulted in five models. Here the fifth model is analyzed. The variables in the equation are the Issue, the volume, the number of cited references, the sum of TF*IDF for each record and the type of the article. The total citations and the average citations were not included in the model because they are used to calculate the classification variable and their inclusion could give false results and overfitting. Please see Table 5 below for the metrics of variables in the equation. The overall percentage of the best model given the variables was 75%. Low cited articles are better predicted by the model in an accuracy of 88,8%, while the highly cited are predicted with an accuracy of 42,9%. It is worth considering at this point that highly cited records are less that low cited ones. The model assumptions were met. Please see Table 6 below for the classification table.

Variables in the Equation		
Step 5	**B**	**Significance**
Issue	-0,234	***
Volume	-0,276	***
Cited References	0,013	***
Sum TF*IDF	-3,182	*
Review/Article	1,133	***
Constant	10,71	***

Table 5: Variables in the equation of binary logistic predicting highly or low cited

Classification Table			
Step 5	**Predicted**		
Observed	**Low**	**High**	**Percentage Correct**
Low	450	57	88,8
High	124	93	42,9
Overall Percentage			75,0

Table 6: Classification table of binary logistic Predicting Highly or Low Cited

Predict whether a record is a review or a research article

Given that the dataset is grouped in reviews and research articles the second binomial logistic that was performed consisted of a model that attempted to predict the type of a record. The selected method was forward LR and it resulted in six models. Here the sixth model is analyzed. The variables in the equation are the volume, the number of cited references, the length of the record in pages, the term count, the sum of TF*IDF and the total citations. Here the average citations variable was not included in the model. Please see Table 7 below for the metrics of variables in the equation. The overall percentage of the best model given the variables was 91,7%. Reviews are predicted with an accuracy of 51% while articles are predicted with an

accuracy of 98,1%. Here it is taken into consideration that articles are a lot more than reviews. The model assumptions were met. Please see Table 8 below for the classification table.

Variables in the Equation		
Step 6	**B**	**Significance**
Volume	0,109	*
Cited References	-0,012	*
Pages	0,071	*
Term Count	-0,006	***
Sum TF*IDF	30,002	***
Total Citations	-0,03	**
Constant	-13,56	***

TABLE 7: VARIABLES IN THE EQUATION OF BINARY LOGISTIC PREDICTING REVIEW OR ARTICLE

Classification Table			
Step 5	**Predicted**		
Observed	**Review**	**Article**	**Percentage Correct**
Review	50	48	51,0
Article	12	614	98,1
Overall Percentage			91,7

Table 8: Classification Table of Binary Logistic Predicting Review or Article

Discussion

Answering Research Questions

All three research questions have been addressed and answered in the present thesis, through the methodology steps mentioned in the cited references, along with some innovative approaches for the topic, such as the binomial logistic regression.

What are the most frequent keywords in the collected articles?

Keywords have been extracted from articles where tables and word clouds are available and presented. With the *Keygraph Keyword Extractor* node it was made possible to identify the top ten keywords for each collection of reviews and articles according to a score that has been assigned to them by the algorithm. This can be a valuable tool not only for authors who want to improve their keyword appearance in the respective section of their articles, but also for database architects and programmers to improve search engines and heuristic algorithms. It is important alongside the word clouds to have a sense of what the corpus is about, what are the common topics and whether there is academic interest to further invest in PhDs and research. Additionally, it would be interesting to be able to compare keywords among journals, or even among decades to determine trends in topics and other possible similarities and differences.

How are the articles distributed in terms of years?

The distribution of the articles over the years is uneven and rising. This corresponds to what was written in the beginning of the thesis, that published research is ever-growing. In one decade alone the number in the journal of study has doubled, and to facilitate this increase two more issues per year were added after 2014. It is expected that more literature would mean a better world. Although this would be significantly hard to measure, it is safe to say that in the world of research it would be better to err in the side of quality rather than quantity.

Could text mining applications be of assistance in predicting article citations?

For the third research question that considered whether text mining applications can be of assistance in predicting article citations, the results show that TF*IDF among other variables can be utilized in prediction models. The utilization of variables that occur form text mining operations in prediction models is encouraging for further research on the topic. For the present thesis it would be of great importance to have the technical ability to include at least the two more Journals that were in the initial plan. Additionally, as a further step Journals of lower impact factor could be included in the models to compute true similarities and differences of the published literature.

Key takeaways

For further research

With the number of published literature growing, it is important to utilize all available techniques for classification and categorization of articles in the databases. Besides heuristic algorithms have still a long way to go to become even better in returning relevant results. Text mining could contribute to this field significantly. In addition to the model presented here, the next steps of Vector Space Modeling, clustering, and classification, briefly presented in the methodology could be a way to create mining models that could predict if an unpublished manuscript would be successful in terms of citations or even what is the best journal to publish the specific article. That could save a significant amount of time and resources both for the researchers and for the editor teams.

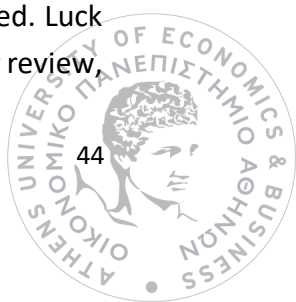
Moreover, it is necessary for data in terms of published papers to be readily available for text mining analysis. Each Journal has its own pdf format which makes it difficult for off-the-shelf algorithms to adapt and analyze them correctly, as was the case in the present thesis. If the texts were available in text (.txt) format, the mining operations would be much more efficient. Furthermore, it would be an important training measure for the models to include rejected manuscripts from peer reviewers. It goes without saying that peer reviewers would be of great value in the development, training, and evaluation of these models and that any model should ever replace the trained human brain.

For Institutions

To begin with, research instructions such as laboratories, universities and knowledge centers could review text mining results to guide their research programs and funds to research trends in each field. That does not mean that topics that are out of trend should be abandoned, however these indicators would function as a focus point for future research. Afterall the interest of the researcher on a topic is a factor that contributes to their devotion on research, therefore it results to a better outcome. Committees and Funds should not expect each and every published work to be groundbreaking or life changing. On the contrary, each paper should have a small contribution to the scientific world and that is something extremely hard to measure especially with all the publications noise that exists today. By ensuring small and steady contribution of each published paper, researchers, journals, and evaluators can reassure society that academia is doing its best to move the world to a better place, in good faith, one small step at a time.

Epilogue

The characteristics of research articles are both qualitative and quantitative. They can be numerus and hard to measure and there are still a lot of things that cannot be measured. Luck must be considered as a factor in publishing, because when a manuscript is submitted for review,



peer reviewers will evaluate it among others and comparisons are hard to avoid. As a result, a good manuscript could be rejected if others in the same batch were considered as better. To truly evaluate an article many factors should be taken into consideration and our mind or computational systems are not able to comprehend them all, yet.

Author's Notes & Questions

And a little background story[5]

At the beginning of the present thesis, I was extremely enthusiastic about the concept that I came up with. To elaborate, prior to my MBA I was fortunate enough to occupy a PhD position at the Agricultural University of Athens. A Professor of mine suggested I did it, and I just said yes. I was my first and last (I hope) chance to see the "dark side of the moon" or in other words, Academia. It more like a glimpse of the scholar sector because I was a PhD student for something less than two years.

Doing a PhD was at the time, and this may sound oxymoron, the easy choice for me. I was a good student, devoted to studying, passionate about the topic, young, ambitious to become a professor since I also enjoyed teaching, and most importantly I was unemployed. Although that at the time seemed like a straight line, it was more like a roller coaster with the chance of becoming a "*Professor*" having a big question mark at the end. Thankfully, I didn't spend much time on the rather arduous task of a PhD. I decided to walk away, despite the escalation of commitment was already high. That difficult decision was a powerful lesson for me.

During those years I had the chance to become even more familiar with scientific writing and publications. PhDs are literature heavy documents, in a way that in my opinion, it defies their purpose. But to become a "Doctor" one must be ready to comply. Since I started, I read something around 2.500 thousand papers and several books around my topic and the result was, what I call scientific fatigue. All this scientific contribution and yet humans are dying from malaria and cancer, the environment is degrading, our dignity as a species is fading and I could go on, but I will not. What are we doing wrong?

As an individual plagued by my own thoughts, I couldn't let this go. I joined the MBA at the one and only "Harvard" of Greece, Athens University of Economics and Business, which is a remarkable institution indeed, however the thought that the abundance of knowledge and scientific contribution was not enough to pull humanity from the depressing track it has been for some years, was still troubling me. So, I started exploring my options. I was somewhat familiar with data analysis and data mining. I figured, why not do some text mining on the readily available scientific publications to explore automated ways to utilize all this knowledge. This is how I came up with the topic of the present work.

At the same time, as a person who is still unemployed, I do some side-jobs to boost my income. I have created a network of high school students whom I teach them Algebra and Geometry and

[5] Who doesn't love a good story, without references?

I have the unique opportunity to see what the younger generation is actually doing. More on that later.

Being in the position to study NLP and TM, I have come to some conclusions and a few rather philosophical questions. So, technology is in fact advancing. However, I am not entirely sure if we as a species will ever be able to make computers able to fully understand our language. At least not in the future that I will be around. And by understanding, I mean the ability to draw inferences from unstructured text, as it is written by humans for ages now. Algorithms will keep getting better, computers will keep getting faster and human communication will keep getting more complex or ambiguous.

On the problem of our own lack of understanding.

My limited life experience has taught me that people have a hard time understanding each other even if they speak the same language. Problems of communication occur in every kind of relationship and especially in the business world where information flow is crucial to success, we still experience bottlenecks and failures. To overcome these barriers, companies have come up with all sorts of forms and reports to keep the flow of information on track, however this have only made matters worse for the people involved.

This situation, and I will not be calling it a problem, concerns society and in a broader concept, culture itself. Cultural legacies are so deeply rooted in our behavior that is almost impossible to overcome them. To be honest with you, dear reader, I am concerned with why we would want to put them aside in the first place. Like in nature, biodiversity plays the most significant role in our future survival, the same goes for variability among humans. It is from the lack of understanding that people get creative and bare ideas that will take the whole species one step further. After all it is our uniqueness that makes us who we are, and by trying to conform with a standard erases our identity. Language is not a one-size-fits-all element that we all use the same way. And it should not by any means be!

Communication, in my absolutely personal opinion is getting worse by the minute. People chose to talk less and write more, while their texts are getting shorter. When was the last time you received an email with an attachment and a long text? For the majority of people, the first thing to do is open the attachment and never read the actual email. Information is being lost and communication is somewhat broken. As a talkative person, I hate it when I am with my peers and there are long periods of silence. So many things are going on in people's lives and yet they have nothing to talk about. And when they do, they feel that no one can truly understand them. Why?

From my experience with teenagers on the other hand, things look even more depressing. Their ability to communicate with each other is limited to emojis and their willingness to communicate with adults, unless cultivated from a young age, is nonexistent. They speak the same language and yet it is so different. As a teacher, I used to relay on non-verbal ques for most of my students to understand when they are not following what I'm saying. Only a few can put into words what they do not understand and ask for an explanation. With online synchronous classes even that has become a distant memory. Their emotions have become stagnant and monotonous. That has nothing to do with the operations of text mining but has a lot to do with our approach on AI and therefore NLP as part of it.

All the above make it even harder to put together a computer system that can effectively, understand and mimic the way humans attempt to communicate.

Should we let computers understand our language?

I believe that with NLP we are trying to deal with one problem by creating another one. How can we possibly let a machine understand human nature so well to a level that it can make decisions for us. Why should we teach the computer of the future our language? It will only make matters of communication worse. And why should we trust an unconscious object with the first and probably the only thing that separates us from other animals on this poor and abused planet?

Data analysis has proven valuable for countless of reasons and of course we should keep practicing it. We have a conscious mind. We should, by all means, be data informed but not data dictated. There is this gut feeling that we are so willing to give up and let computers be in charge. Is it our incapability of making the decisions or is it the fear of the consequences? Either way I am afraid that we are gradually becoming the audience instead of the actors, and this is no ancient tragedy where we will experience katharsis at the end of the play. The so-called play is our here and now, the future of the society and our personal take in it. We are letting computers do the heavy lifting and instead of taking ownership we incrementally slip away from the lead.

To answer the question, and let you reader go about your life, I will bluntly say no. No because I have trust issues, because I still feel superior in comparison to a machine, because I am not alone on the planet (and I include other species), because if there is another life after this one, I will have to deal with the consequences of my current choices and because I am proud enough to own these choices. You can always disagree.

Closing this epilogue/ rant/ food for thought I would like to say that everything is easy until it is difficult. The tipping point is the sweet spot where it gets interesting. Lack of understanding keeps things spicy and kills boredom, in my opinion.

I have three last questions for you, dear reader before you go.

1. Why have we lost connection to meaningful things?
2. Why did the easy way, became the only way?
3. Why are we letting go of power?

Appendix

Count of Title												
Row Labels	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**Total**
1	**7**	**11**	**12**	**11**	**9**	**9**	**11**	**7**	**9**	**4**	**5**	**95**
Article	5	3	6	5	8	9	8	3	6	3	5	61
Review	2	8	6	6	1		3	4	3	1		34
2	**11**	**8**	**7**	**10**	**9**	**10**	**11**	**7**	**11**	**15**	**12**	**111**
Article	8	7	6	10	9	10	11	7	11	15	12	106
Review	3	1	1									5
3	**8**	**9**	**9**	**6**	**10**	**10**	**9**	**9**	**5**	**9**	**17**	**101**
Article		9	9	6	9	8	9	9	4	9	17	89
Review	8				1	2			1			12
4	**11**	**8**	**11**	**11**	**8**	**8**	**7**	**9**	**13**	**11**	**17**	**114**
Article	9	6	8	8	8	8	7	9	13	11	17	104
Review	2	2	3	3								10
5	**7**	**8**	**5**	**10**	**8**	**9**	**10**	**8**	**13**	**13**	**13**	**104**
Article	7	8	5	10	5	6	5	5	13	13	13	90
Review					3	3	5	3				14
6	**10**	**9**	**8**	**8**	**10**	**11**	**8**	**8**	**10**	**9**	**6**	**97**
Article	3	9	8	8	9	11	8	6	4	8	3	77
Review	7				1			2	6	1	3	20
7					**10**	**8**	**9**	**10**	**14**	**15**	**14**	**80**
Article					10	8	9	7	14	15	14	77
Review								3				3
8									**5**	**12**	**5**	**22**
Article									5	12	5	22
Total	**54**	**53**	**52**	**56**	**64**	**65**	**65**	**58**	**80**	**88**	**89**	**724**

TABLE 9: DISTRIBUTION OF RECORDS OVER THE ISSUES PER YEAR

Word Cloud Metrics – Top 10					
Reviews			Articles		
Term	TF	TF*IDF	Term	TF	TF*IDF
effect[]	0,634797166	0,191092988	behavior[]	3,313614093	1,043373036
firm[]	0,817532201	0,295523618	effect[]	4,953509375	1,494518912
individu[]	0,478411744	0,144016285	employe[]	3,532250579	1,192799293
manag[]	0,823323894	0,247845188	firm[]	5,676499405	2,157820289
organiz[]	0,595364059	0,180525029	manag[]	4,71812625	1,420297525
perform[]	0,739120999	0,22742346	model[]	4,163123766	1,270390021
relationship[]	0,594748472	0,181659447	perform[]	4,911291454	1,498695692
research[]	1,108840855	0,333794358	research[]	4,919687908	1,482642191
studi[]	0,901313578	0,271322422	studi[]	4,794840972	1,448281072
team[]	0,46408115	0,175513612	team[]	2,725960519	1,139679307

TABLE 10: TOP 10 WORDS FROM BOW FOR BOTH TYPES OF RECORDS

Reviews		Articles	
Keyword	Score	Keyword	Score
employe[]	5297	behavior[]	29092
firm[]	13312	effect[]	40686
manag[]	8660	employe[]	31823
measur[]	5319	firm[]	70125
organiz[]	6515	manag[]	31899
perform[]	10218	model[]	25512
relationship[]	4846	perform[]	48929
research[]	9861	research[]	25466
studi[]	9187	studi[]	33539
team[]	5276	team[]	29672

TABLE 11: KEYWORDS WORD CLOUD METRICS

References

1. Aggarwal, C. C. (2015) *Data mining: the textbook.* Springer.
2. Altbach, P. G. (2018) 'Too much academic research is being published', *University World News*.
3. Ancaiani, A., Anfossi, A. F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., Cicero, T., Ciolfi, A., Costa, F., Colizza, G., Costantini, M., Di Cristina, F., Ferrara, A., Lacatena, R. M., Malgarini, M., Mazzotta, I., Nappi, C. A., Romagnosi, S. and Sileoni, S. (2015) 'Evaluating scientific research in Italy: The 2004–10 research evaluation exercise', *Research Evaluation,* 24(3), pp. 242-255.
4. Antonakis, J., Bastardoz, N., Liu, Y. H. and Schriesheim, C. A. (2014) 'What makes articles highly cited?', *Leadership Quarterly,* 25(1), pp. 152-179.
5. Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S. and Rupp, D. E. (2016) 'Editorial: Evidence on Questionable Research Practices: The Good, the Bad, and the Ugly', *Journal of Business and Psychology,* 31(3), pp. 323-338.
6. Boyce, B. R. (1990) 'AUTOMATIC TEXT-PROCESSING - THE TRANSFORMATION ANALYSIS, AND RETRIEVAL OF INFORMATION BY COMPUTER - SALTON,G', *Journal of the American Society for Information Science,* 41(2), pp. 150-151.
7. Callaway, E. (2016) 'BIBLIOMETRICS Publishing elite turns against impact factor', *Nature,* 535(7611), pp. 210-211.
8. Chen, H. Q. and Ho, Y. S. (2015) 'Highly cited articles in biomass research: A bibliometric analysis', *Renewable & Sustainable Energy Reviews,* 49, pp. 12-20.
9. Clarivate (2020) *Highly Cited Researchers 2020*: Web of Science.
10. Cohen, K. B., Johnson, H. L., Verspoor, K., Roeder, C. and Hunter, L. E. (2010) 'The structural and content aspects of abstracts versus bodies of full text journal articles are different', *Bmc Bioinformatics,* 11.
11. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P. (2011) 'Natural Language Processing (Almost) from Scratch', *Journal of Machine Learning Research,* 12, pp. 2493-2537.
12. Conroy, G. (2019) 'The 7 deadly sins of research', *Nature Index*.
13. Delen, D., Sharda, R. and Kumar, P. (2007) 'Movie forecast guru: A Web-based DSS for Hollywood managers', *Decision Support Systems,* 43(4), pp. 1151-1170.
14. Editorial (2016) 'Time to Remodel the Journal Impact Factor”', *Nature,* 535(466).
15. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. 'Advances in knowledge discovery and data mining'. 1996: American Association for Artificial Intelligence.
16. Feldman, R. and Dagan, I. 'Knowledge Discovery in Textual Databases (KDT)'. 1995, 112-117.
17. Feldman, R. and Sanger, J. (2007) *The text mining handbook: advanced approaches in analyzing unstructured data.* Cambridge university press.
18. Frawley, W. J., Piatetsky-Shapiro, G. and Matheus, C. J. (1992) 'Knowledge discovery in databases: An overview', *AI magazine,* 13(3), pp. 57-57.
19. Frey, B. S. and Rost, K. (2010) 'DO RANKINGS REFLECT RESEARCH QUALITY?', *Journal of Applied Economics,* 13(1), pp. 1-38.

20. Glick, W. H., Miller, C. C. and Cardinal, L. B. (2007) 'Making a life in the field of organization science', *Journal of Organizational Behavior,* 28(7), pp. 817-835.
21. Guz, A. N., Rushchitsky, J. J. and Chernyshenko, I. S. (2005) 'On a modern philosophy of evaluating scientific publications', *International Applied Mechanics,* 41(10), pp. 1085-1091.
22. Hazelkorn, E. (2002) 'Challenges of growing research at new and emerging HEIs'.
23. Hazelkorn, E. (2004) 'Growing research: challenges for latedevelopers and newcomers', *Higher Education Management and Policy,* 16(1), pp. 119-140.
24. Jo, T. (2006) *The Implementation of Dynamic Document Organization using the Integration of Text Clustering and Text Categorization.* Ph.D, University of Ottawa, Canada2021-05-13T11:53:32).
25. Jo, T. (2019) 'Text mining', *Studies in Big Data. Cham: Springer International Publishing*.
26. Ketcham, C. M. and Crawford, J. M. (2007) 'The impact of review articles', *Laboratory Investigation,* 87(12), pp. 1174-1185.
27. Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihok, G. and Den Hartog, D. N. (2018) 'Text Mining in Organizational Research', *Organizational Research Methods,* 21(3), pp. 733-765.
28. Kostoff, R. N., del Rio, J. A., Humenik, J. A., Garcia, E. O. and Ramirez, A. M. (2001) 'Citation mining: Integrating text mining and bibliometrics for research user profiling', *Journal of the American Society for Information Science and Technology,* 52(13), pp. 1148-1156.
29. Kowalski, G. J. and Maybury, M. T. (2000) *Information storage and retrieval systems: theory and implementation.* Springer Science & Business Media.
30. Kumar, R. (2017) 'Natural Language Processing', *Machine Learning and Cognition in Enterprises*: Apress, pp. 65-73.
31. Lei, L. and Sun, Y. M. (2020) 'Should highly cited items be excluded in impact factor calculation? The effect of review articles on journal impact factor', *Scientometrics,* 122(3), pp. 1697-1706.
32. Liddy, E. D. (2001) 'Natural language processing'.
33. Lindsey, D. (1989) 'Using citation counts as a measure of quality in science measuring what's measurable rather than what's valid', *Scientometrics,* 15(3-4), pp. 189-203.
34. Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y. and Potts, C. 'Learning word vectors for sentiment analysis'. 2011, 142-150.
35. Martin, E. P. G., Bremer, E. G., Guerin, M.-C., DeSesa, C. and Jouve, O. 'Analysis of protein/protein interactions through biomedical literature: Text mining of abstracts vs. text mining of full text articles'. 2004: Springer, 96-108.
36. Nasukawa, T. and Nagano, T. (2001) 'Text analysis and knowledge mining system', *IBM systems journal,* 40(4), pp. 967-984.
37. Ohsawa, Y., Benson, N. E. and Yachida, M. 'KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor'. 1998: IEEE, 12-18.
38. Onodera, N. and Yoshikane, F. (2015) 'Factors Affecting Citation Rates of Research Articles', *Journal of the Association for Information Science and Technology,* 66(4), pp. 739-764.

39. Pasterkamp, G., Rotmans, J. I., de Kleijn, D. V. P. and Borst, C. (2007) 'Citation frequency: A biased measure of research impact significantly influenced by the geographical origin of research articles', *Scientometrics,* 70(1), pp. 153-165.
40. Poole, D. (2014) *Linear algebra: A modern introduction.* Cengage Learning.
41. Prasad, M., Sowmya, A. and Koch, I. 'Efficient feature selection based on independent component analysis'. *Proceedings of the 2004 Intelligent Sensors, Sensor Networks and Information Processing Conference, 2004.*, 14-17 Dec. 2004, 427-432.
42. Rebala, G., Ravi, A. and Churiwala, S. (2019) 'Natural Language Processing', *An Introduction to Machine Learning*: Springer International Publishing, pp. 117-125.
43. Ricker, M. (2017) 'Letter to the Editor: About the quality and impact of scientific articles', *Scientometrics,* 111(3), pp. 1851-1855.
44. Roberts, J. C., Fletcher, R. H. and Fletcher, S. W. (1994) 'Effects of Peer Review and Editing on the Readability of Articles Published in Annals of Internal Medicine', *JAMA,* 272(2), pp. 119-121.
45. Robertson, S. (2004) 'Understanding inverse document frequency: on theoretical arguments for IDF', *Journal of Documentation,* 60(5), pp. 503-520.
46. Rolfe, G. (2006) 'Validity, trustworthiness and rigour: quality and the idea of qualitative research', *Journal of Advanced Nursing,* 53(3), pp. 304-310.
47. Rossner, M., Van Epps, H. and Hill, E. (2007) 'Show me the data', *Journal of Cell Biology,* 179(6), pp. 1091-1092.
48. Rynes, S. L., Giluk, T. L. and Brown, K. G. (2007) 'The very separate worlds of academic and practitioner periodicals in human resource management: Implications for evidence-based management', *Academy of Management Journal,* 50(5), pp. 987-1008.
49. Salloum, S. A., Al-Emran, M., Monem, A. A. and Shaalan, K. (2018) 'Using text mining techniques for extracting information from research articles', *Intelligent natural language processing: Trends and Applications*: Springer, pp. 373-397.
50. Salton, G. and McGill, J. M. (1983) *Introduction to Modern Information Retrieval.* United States of America: McGraw-Hill Inc.
51. Salton, G. and Yang, C.-S. (1973) 'On the specification of term values in automatic indexing', *Journal of documentation*.
52. Sebastiani, F. (2002) 'Machine learning in automated text categorization', *Acm Computing Surveys,* 34(1), pp. 1-47.
53. Sebban, M. and Nock, R. (2002) 'A hybrid filter/wrapper approach of feature selection using information theory', *Pattern Recognition,* 35(4), pp. 835-846.
54. Sebestyén, V., Domokos, E. and Abonyi, J. (2020) 'Focal points for sustainable development strategies—Text mining-based comparative analysis of voluntary national reviews', *Journal of Environmental Management,* 263, pp. 110414.
55. Singhal, A., Salton, G., Mitra, M. and Buckley, C. (1996) 'Document length normalization', *Information Processing & Management,* 32(5), pp. 619-633.
56. Song, F., Guo, Z. and Mei, D. 'Feature Selection Using Principal Component Analysis'. *2010 International Conference on System Science, Engineering Design and Manufacturing Informatization*, 2010-11-01: IEEE.
57. Sparck Jones, K. (1972) 'A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL', *Journal of Documentation,* 28(1), pp. 11-21.

58. Sulova, S., Todoranova, L., Penchev, B. and Nacheva, R. 'Using text mining to classify research papers'. 2017, 647-654.
59. Tan, A.-H. 'Text mining: The state of the art and the challenges'. 1999: Citeseer, 65-70.
60. Thiel, K. (2009) 'The KNIME text processing plugin', *Nycomed Chair for Bioinformatics and Information Mining, University of Konstanz,* 78457.
61. van Wesel, M. (2016) 'Evaluation by Citation: Trends in Publication Behavior, Evaluation Criteria, and the Strive for High Impact Publications', *Science and Engineering Ethics,* 22(1), pp. 199-225.
62. Vanclay, J. K. (2013) 'Factors affecting citation rates in environmental science', *Journal of Informetrics,* 7(2), pp. 265-271.
63. Verma, T., Renu, R. and Gaur, D. (2014) 'Tokenization and filtering process in RapidMiner', *International Journal of Applied Information Systems,* 7(2), pp. 16-18.
64. Verspoor, K. and Cohen, K. B. (2013) 'Natural Language Processing', in Dubitzky, W., Wolkenhauer, O., Cho, K.-H. and Yokota, H. (eds.) *Encyclopedia of Systems Biology*. New York, NY: Springer New York, pp. 1495-1498.
65. Westergaard, D., Staerfeldt, H. H., Tonsberg, C., Jensen, L. J. and Brunak, S. (2018) 'A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts', *Plos Computational Biology,* 14(2).
66. Witten, I. H. 2004. Text Mining.
67. Zhang, Y., Chen, M. and Liu, L. 'A review on text mining'. 2015: IEEE, 681-685.