

**ANALYSIS OF THE EVOLUTION OF SCIENTIFIC TOPICS USING TEXT
MINING TECHNIQUES**



Buse AÇIKALIN

AĞUSTOS 2022

**ANALYSIS OF THE EVOLUTION OF SCIENTIFIC TOPICS USING TEXT
MINING TECHNIQUES**

**BAHÇEŞEHİR ÜNİVERSİTESİ
LİSANSÜSTÜ FEN BİLİMLERİ ENSTİTÜSÜ
YÜKSEK LİSANS TEZİ**

BUSE AÇIKALIN

BÜYÜK VERİ ANALİTİĞİ VE YÖNETİMİ DALINDA

**YÜKSEK LİSANS DERECEİ İÇİN GEREKLİ ÇALIŞMALAR
YERİNE GETİRİLMİŞTİR**

AĞUSTOS 2022



T.C.
BAHCESEHIR UNIVERSITY
GRADUATE SCHOOL

MASTER THESIS APPROVAL FORM

Program Name:	Big Data Analytics And Management
Student's Name and Surname:	Buse Açıklın
Name Of The Thesis:	Analysis Of The Evolution Of Scientific Topics Using Text Mining Techniques
Thesis Defense Date:	15.09.2022

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

Prof. Dr. Ahmet ÖNCÜ
Institute Director

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

	Title/Name	Signature
Thesis Advisor's	Assoc. Prof. Dr. Tefik Aytekin	
Member's	Assist. Prof. Dr. Nilgün Güler Bayazıt	
Member's	Assist. Prof. Dr. Tarkan Aydın	

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : Buse Aıkaldn

Signature :

ÖZET

BİLİMSEL KONULARIN EVRİMİNİN METİN MADENCİLİĞİ YÖNTEMLERİ İLE ANALİZİ

Buse Açıkalin

Büyük Veri Analitiği Ve Yönetimi

Tez Danışmanı: Doç. Dr. Tevfik Aytekin

Ağustos 2022, 73 pages

Yıllar içerisinde metin kümelerin sayısı ve boyutunun artması ile metinlerden manuel olarak bilgi çıkarımının zorlaşması sonucunda, metinlerin otomatik olarak analiz edilme ihtiyacı doğmuştur. Veri madenciliğinin bir alt kümesi olan metin madenciliği; sınıflandırma, kümeleme, konu modelleme gibi yöntemlerle metin kümelerini analiz etmeye çalışmaktadır. Konu modelleme yöntemleri dokümanlardaki gizli konuların çıkarılmasını sağlar. Konu modellemede konu sayısının ve model parametrelerinin belirlenmesi ve çıkarılan konuların ne hakkında olduğunun tespiti, üzerinde çalışılmaya devam edilen ana problemlerdendir. Bu tez kapsamında uzun bir zaman dönemi için toplanan bilimsel makalelerden gizli konuların tespiti, yorumlanması ve yıllar içerisinde tartışılan konuların değişimi ile ilgili genel bir sonuç elde edilmek istenmiştir. Bu amaçla iki farklı bilim tarihi dergisinden makaleler toplanarak datasetler oluşturulmuş; konu modelleme yöntemlerinden LDA ve NMF ile kümeleme yöntemlerinden K-Means kullanılarak yıllara göre en çok konuşulan konular çıkarılıp, karşılaştırılmış ve yorumlanmıştır. Dergilerin kapsamı göz önünde bulundurularak çıkarılan konu kategorilerinin tezde uygulanan modeller sonucunda yıllar içerisinde popülaritesinin nasıl değiştiği ile bu konuların yıllar içerisinde hangi kelimeler ile temsil edildikleri belirlenmeye çalışılmıştır.

Keywords: Metin Madenciliği, Konu Modelleme, Metin Kümeleme, Gizli Dirichlet Ayrımı, Negatif Olmayan Matris Ayrıştırması

ABSTRACT

ANALYSIS OF THE EVOLUTION OF SCIENTIFIC TOPICS USING TEXT MINING TECHNIQUES

Buse Açıkalın

Big Data Analytics and Management

Thesis Supervisor: Assoc. Prof. Dr. Tevfik Aytekin

August 2022, 73 pages

Over the years, with the increasing number and size of text data, extracting and interpreting information from these data has become complex and there is a need for automatic analysis of texts. Text mining, a part of data mining, tries to make sense of large text datasets with classification, clustering and topic modeling methods. Topic modeling methods aim to extract hidden topics behind documents. In topic modeling, determining the model parameters and how the topics are obtained in the model are the main issues that continue to be studied in this field. Within the scope of the thesis, it is aimed to extract and interpret the hidden topics from the scientific publications that were collected over a long period of time, as a consequence it was desired to obtain results related to the change in the topics discussed over the years. Using Latent Dirichlet Allocation and Nonnegative Matrix Factorization which are topic modeling methods and K-Means which is clustering method the most mentioned topics by years have been extracted, compared and interpreted. It has been analyzed how the topics have changed in these journals over the years. On the other hand, it has been showed to how the words of topics have changed over the years and which words represent the related topics.

Keywords: Text Mining, Topic Modeling, Text Clustering, Latent Dirichlet Allocation, Nonnegative Matrix Factorization

To My Family



ACKNOWLEDGEMENT

I would like to thank my thesis advisor Assist. Prof. Dr. Tevfik Aytekin for his guidance, advice, criticism and support throughout my master thesis.



TABLE OF CONTENTS

ETHICAL CONDUCT	iii
ÖZET.....	iv
ABSTRACT	v
DEDICATION	vi
ACKNOWLEDGEMENT	vii
TABLE OF CONTENTS	viii
LIST OF TABLES	ix
LIST OF FIGURES.....	xii
ABBREVIATIONS.....	xiii
Chapter 1: Introduction	1
Chapter 2: Literature Review	4
Chapter 3: Methodology	9
3.1 Text Representation Methods	9
3.1.1 TF-IDF	9
3.1.2 Bag of Words	9
3.2 Clustering Method.....	10
3.2.1 K-Means	10
3.3 Topic Modeling Methods	10
3.3.1 Latent Dirichlet Allocation	11
3.3.2 Nonnegative Matrix Factorization	14
Chapter 4: Datasets and Experimental Study	15
4.1 Datasets.....	15
4.1.1 Science Dataset	16
4.1.2 British Journal For The History of Science Dataset	17
4.2 Experimental Study.....	17
4.2.1 Science Dataset Results.....	17
4.2.2 British Journal of The History Of Science Dataset Results..	60
Chapter 5: Results And Discussion.....	66
Chapter 6: Conclusion.....	72
REFERENCES.....	74

LIST OF TABLES

TABLES

Table 1 Science Dataset Categories.....	21
Table 2 Topics And Labels Which Are Extracted Using LDA Between 1880-1900 For Science Dataset.....	22
Table 3 Hot Topics Between 1880-1900 For Science Dataset.....	24
Table 4 Clusters Between 1880-1900 For Science Dataset,,,,,,.....	25
Table 5 Topics Which Are Extracted Using NMF Between 1880-1900 For Science Dataset.....	26
Table 6 Top Terms Of Electricity Topic Between 1880-1900.....	27
Table 7 Top Terms Of Geology Topic Between 1880-1900.....	28
Table 8 Topics And Labels Which Are Extracted Using LDA Between 1900-1920 For Science Dataset.....	28
Table 9 Hot Topics Between 1900-1920 For Science Dataset.....	30
Table 10 Clusters Between 1900-1920 For Science Dataset.....	30
Table 11 Topics Which Are Extracted Using NMF Between 1900-1920 For Science Dataset.....	31
Table 12 Top Terms Of Education Topic Between 1900-1920	33
Table 13 Topics And Labels Which Are Extracted Using LDA Between 1920-1940 For Science Dataset.....	33
Table 14 Hot Topics Between 1920-1940 For Science Dataset	35
Table 15 Clusters Between 1920-1940 For Science Dataset.....	35
Table 16 Topics Which Are Extracted Using NMF Between 1920-1940 For Science Dataset.....	37
Table 17 Top Terms Of Medicine Topic Between 1920-1940	38
Table 18 Top Terms Of Analytical Chemistry Between 1920-1940.....	38
Table 19 Topics And Labels Which Are Extracted Using LDA Between 1940-1960 For Science Dataset.....	39
Table 20 Hot Topics Between 1940-1960 for Science Dataset.....	41
Table 21 Topics Which Are Extracted Using NMF Between 1940-1960 For Science Dataset.....	41

Table 22 Topics Which Are Extracted Using NMF Between 1940-1960 For Science Dataset.....	42
Table 23 Top Terms Of Astrophysics Topic Between 1940-1960.....	44
Table 24 Top Terms Of Biochemistry Topic Between 1940-1960.....	44
Table 25 Topics and Labels Which Are Extracted Using LDA Between 1960-1980 For Science Dataset.....	44
Table 26 Hot Topics Between 1960-1980 For Science Dataset.....	46
Table 27 Clusters Between 1960-1980 For Science Dataset.....	47
Table 28 Topics Which Are Extracted Using NMF Between 1960-1980 For Science Dataset.....	48
Table 29 Top Terms Of Technology Topic Between 1960-1980	49
Table 30 Topics And Labels Which Are Extracted Using LDA Between 1980-2000 For Science Dataset.....	50
Table 31 Hot Topics Between 1980-2000 For Science Dataset.....	52
Table 32 Clusters Between 1980-2000 For Science Dataset.....	52
Table 33 Topics Which Are Extracted Using NMF Between 1980-2000 For Science Dataset.....	53
Table 34 Top Terms of Disease Topic between 1980-2000.....	55
Table 35 Topics and Labels Which Are Extracted Using LDA Between 2000-2020 For Science Dataset.....	55
Table 36 Hot Topics Between 2000-2020 For Science Dataset.....	57
Table 37 Clusters Between 2000-2020 For Science Dataset.....	57
Table 38 Topics Which Are Extracted Using NMF Between 2000-2020 For Science Dataset.....	58
Table 39 BJHS Dataset Categories.....	61
Table 40 Topics and Labels Which Are Extracted Using LDA between 1960-1980 for BJHS Dataset.....	61
Table 41 Hot Topics Between 1960-1980 For BJHS Dataset.....	62
Table 42 Topics and Labels Which Extracted Using LDA Between 1980-2000 For BJHS Dataset.....	63
Table 43 Hot Topics Between 1980-2000 For BJHS Dataset.....	64
Table 44 Topics and Labels Which Are Extracted Using LDA Between 2000-2020 For BJHS Dataset.....	64

Table 45 Hot Topics Between 2000-2020 For BJHS Dataset.....	65
Table 46 The Change Of Social Science Topic's Words Over The Years.....	66
Table 47 The Change Of Disease Topic's Words Over The Years.....	67
Table 48 Change Of Chemistry Topic's Words Over The Years.....	67
Table 49 Change Of Geology Topic's Words Over The Years.....	68
Table 50 Change Of Plant Biology Topic's Words Over The Years.....	68
Table 51 Change Of Medicine Topic's Words Over The Years.....	70
Table 52 Change Of Technology Topic's Words Over The Years.....	70
Table 53 Change Of Physics Topic's Words Over The Years.....	71



LIST OF FIGURES

FIGURES

Figure 1 The intuitions behind LDA.....	11
Figure 2 Graphical model representation of LDA.....	12
Figure 3 Metadata scheme of JSTOR data.....	15
Figure 4 Overview of the applied steps	16
Figure 5 Probabilities of plant biology, disease, paleontology over the years.....	69
Figure 6 Probabilities of astrophysics, chemistry and geology over the years.....	69
Figure 7 Probabilities of BJHS dataset's topics over the years.....	71



ABBREVIATIONS

BJHS	British Journal For The History of Science
DTM	Dynamic Topic Models
GS	Gibbs Sampling
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
NMF	Nonnegative Matrix Factorization
pLSA	Probabilistic Latent Semantic Analysis



Chapter 1

Introduction

In recent years, there has been a rapid increase in the number and size of text datasets. Texts on the internet, library databases, texts on social media, scientific publications, newspaper & magazine archives, e-mails and user comments on online platforms are the main text data sources. With the increase in internet usage and the digitization of written texts, the text datasets have become more accessible. Although text data is easy and preferable to be read and interpreted by humans when it is small in size but, when it is large in size, it is complicated to read and interpret the documents. It requires a lot of time and resources. For this reason, text mining methods have emerged.

There are many text mining methods in the literature. One of the most well-known is topic modeling. Topic models are algorithms for discovering the main themes that pervade a large understanding and otherwise unstructured collection of documents (Blei, 2012). Latent Dirichlet Allocation (LDA) is one of the best-known and most successful topic modeling methods. LDA is an unsupervised method, topics have a probability distribution over words, and text documents have a probability distribution over topics.

Topic modeling is a method that analyzes the frequency of occurrence of words in a text data set and synthesizes the corpus as consisting of one or more topics. It helps to organize, summarize and search text sets. Topic modeling can also be used if content-based matching is to be made between texts or a content-based recommendation system is to be developed. It is a valuable method for analyzing text datasets. Topic modeling brings together the keywords that are frequently used in the texts and facilitates the analysis of the texts. Topic modeling accepts the words in the documents as a bag of words. According to the bag of words method, the order of the words does not matter. The topic modeling method is a generative method and can be easily applied to new data.

In topic modeling, topics are represented as a probability distribution over words. Each document is a probability distribution over the topics. Topics are determined according to the frequency of the words in the documents and the frequency of their occurrence. In topic modeling, the probabilities of the words in each

topic can be calculated, as well as the probability of each topic being found in each document.

There are three main aims of this thesis:

1- It is to create a comprehensive dataset that consists of scientific publications collected from reputable journals and where we can see scientific knowledge change over the years.

2- It is to analyze, compare and interpret a large scientific dataset with text mining methods and thus reveal a study that can contribute to the history of science with topic modeling methods.

3- It proposes a method to see and understand the change of topics between years and to reveal which words represented the same topic over the years.

Within the scope of this thesis, the focus is on the analysis of scientific documents, the selected datasets are the articles of important journals in the history of science and collected for many years. By this way, it has been interpreted by examining which topics have been focused on over the years in this journals, and which words are mostly involved in the emerging topics.

Scientific publications can be considered where technological innovations in world history, and therefore many changes in the world, started. Therefore, essential conclusions can be drawn when examined. "Scientists come up with new approaches, terminologies, and ways of looking at the universe. These ideas necessitate new terminology, which is expressed in their scientific publications. Unlike other areas, where documents merely report on how the environment changes, scientific documents are forces that can change the world themselves." (Boyd-Graber, Hu and Minmo, 2017).

Within the scope of this thesis, more than 150,000 publications from 2 different journals were collected and analyzed by the years and the results were interpreted. After the datasets were collected it passed through various pre-processing steps. And then topics and clusters were extracted using topic models and clustering methods.

Text mining and topic model applications are mentioned in detail in the Literature Review chapter. The usage areas and the studies that have been done so far have been examined in detail. The technical details of the text mining methods used in

the Methodology section are explained. In Chapter 4, the collected datasets and the results of the analyzes are shared. It has been observed which topics are discussed in the relevant journals over the years, how science topics have changed and words of the topics have evolved. And finally in Chapters 5 and 6, the findings and conclusions are shared.



Chapter 2

Literature Review

“Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources.” (Breiman, 2003). Information retrieval, text categorization, text clustering, text summarization, topic model, sentiment analysis are mainly methods of text mining. Topic modeling provides easy, fast, and human-free extraction of key themes among large document sets. It draws a general framework about the topics that the given large set of documents talks about.

Topic modeling is a method of discovering word patterns in a corpus. There are many topic modeling methods in the literature. The first method used in topic modeling is the latent semantic analysis (LSA) method, which uses the linear algebra approach (Deerwester, Dumais, Landauer, Furnas and Harshman, 1990). On the basis of latent semantic structures, LSA was able to categorize documents collectively under the presumption that words with related semantic ranges frequently appear together. Users no longer needed to use the same terms in their searches that exist in linked documents since similarity could be found via the greater semantic range of the term, which made this method superior to previous techniques of document retrieval (Deerwester et al., 1990).

After LSA, 2 probabilistic topic model methods were developed which are called Latent Dirichlet Allocation (Blei, Ng, Jordan, 2003) and Probabilistic Latent Semantic Analysis (Hofmann, 2013). PLSA can be considered historically and Technically as the basis of LDA. LDA measures the distribution of a given topic across a set of documents as well as the distribution of a specific word over a set of topics. By analyzing word co-occurrence patterns throughout a corpus, or collection of documents, LDA discovers the themes. Each topic in the model will be assigned a probability distribution across each document. Each document is a distribution of topics, and each topic is identified by a certain collection of words, according to the logic of LDA. In such situation, LDA looks for words that often appear together and treats them as topics. The observable terms that actually exist in the documents may be used to evaluate how frequently these topics occur in various documents. The

mathematical technical details of LDA to be used in the analyzes in this thesis will be explained in Chapter 3.

In topic modeling, it is important to prepare the corpus and how to set the parameters. A corpus of documents must first be prepared in a way that ensures that only the most informative features of those documents are used for the topic model. While preparing the corpus, some common preprocessing steps are performed first. Make all letters lowercase, eliminate any special characters and numbers and remove stop words. In addition to these, there may be situations where different approaches are adopted. For example, in some studies, the corpus was prepared to consist of only one language, while in some studies it was preferred to establish a multilingual model. Mimno (2012) makes the conscious choice to retain other languages in his corpus since it enables him to highlight variations between traditional research carried out in Germany and English-speaking nations.

Discarding stop words is a common practice for topic modeling methods. Stop words are both frequently used in the document and do not provide information about the content. Suominen and Toivanen (2016) removed the topic models of scientific publications, in addition to stop words, as well as acknowledgments that they believe are less informative. Different methods have also been tried to eliminate words. Székely and Vom Brocke (2017) removed words appearing in less than two documents, Suominen and Toivanen remove words that occur only once in the corpus (Suominen, Toivanen, 2016). Martin and Johnson (2015) offer another alternative for finding the most informative features of a corpus: it eliminates all words except nouns.

Another important issue for topic modeling is parameter selection. There are different studies in the literature on how the parameters should be decided. Suominen and Toivanen conclude that the number of selected topics can only be evaluated in light of the model's real world performance. Therefore, researchers take a trial-and-error approach when choosing the number of topics that the model should exclude from the corpus (Suominen, Toivanen, 2016). Carron-Arthur and their friends try K different models ranging from 10 to 100 in increments of 10. Researchers qualitatively evaluate the results by looking at the top words in each topic. They observed that models with more than 30 topics tended to repeat, but models under 20 topics tended to combine what the researchers believed should be different topics. Therefore, 25 topics were selected for a corpus of 131,004 documents (Carron-Arthur, Reynolds, K. Bennett, A. Bennett, Griffiths, 2016). Székely and vom Brocke test the number of

topics as 3, 5, 10, 20, 50, 70, and 100 for a corpus of 9,514 documents and decide $K = 70$ as the best balance between very general topics (Székely, Vom Brocke, 2017). Carter and his friend tests models with $K = 10, 15, 20, 50$ and 100. They report results for models with $K = 10$ and 100 for their corpus of 7476 documents (Carron-Arthur et al., 2016). Suominen and Toivanen choose $K = 60$ for their corpus of 144,081 documents. As it can be seen, there is no certain way to select the number of topic and it is generally decided by trying it according to the relevant corpus (Suominen, Toivanen, 2016).

Topic modeling has many applications. Information extraction, interpretation of historical documents, interpretation of scientific publications, multilingual data and machine translation, suggestion systems may be among the main application areas. Topical modeling has become a popular technique among social scientists as it makes large sets of documents much easier to interpret.

In various studies, LDA has been used for ad-hoc information retrieval and successful results have been obtained (Wei, Croft, 2006). “The study shows that improvements over retrieval using cluster-based models can be obtained with reasonable efficiency. Topic models are a basic indexing tool that can be used. Users should look for topics that have a high probability of matching a question word, and then look for documents that have a high probability of matching these topics.” (Boyd-Graber et al., 2017)

Topic modeling has also been frequently used in the analysis of historical documents. Historical books, newspapers, and scholarly literature can contain a large number of documents because their dates can go back a long time. That's why they need to be analyzed automatically. In one study, Matt Erlin examined German and English historical books with LDA (Erlin, 2017).

Topical models were extracted and interpreted with the LDA on 8,000 articles referring to state support for the arts in the USA, published in five newspapers between 1986 and 1997 (DiMaggio, Nag, Blei, 2013).

80,000 articles collected between the years 1728-1800 were applied to LDA, and how the topics changed over time, and historical examples of the issues that emerged with LDA were given (Blei, Lafferty, 2006). Historic text datasets have a customized LDA method that captures topics that have changed over the years which is called Dynamic Topic Model (DTM). It shows how a topic has changed over the years and how the topics related to that topic have changed. In the above article, Blei

and his friends collected nearly 30,000 articles from the JSTOR database and analyzed them with DTM.

Topic modeling has also been frequently used in the analysis of scientific documents. Scientific publications have the potential to transform the world by incorporating new technologies. And this is why they are so important. Accurately analyzing datasets of major scientific publications can also contribute to the study of the history of science. Because scientific publications are not a report or a commentary, they are an innovation. There is a scientific article behind the great technological breakthroughs that changed the world (Boyd-Graber et al., 2017). Since the histories of scientific publications are very old and they are very large in number, they are very important in terms of the information they will give. However, it is very difficult to analyze and trend with the human eye. For this reason, topic modeling is one of the most used methods in extracting the topics and following the changes over the years or seeing how the words belonging to the same topic have changed over the years.

There is a very important difference that distinguishes scientific publications from general documents. Words used in Scientific documents are precisely and carefully measured. Some common words may appear frequently in each topic and gain different meanings depending on where they are used. Topic modeling has the ability to capture which topic these common words in the documents are related to and which topic should be included.

One of the first studies to use topic modeling in scientific publications was the work by Griffiths and Steyvers (2004). They collected the abstracts of the articles in PNAS and applied the LDA algorithm on them. With this study, besides extracting the topics in the document set and the semantic significance of the words in these topics, they revealed some relationships between scientific articles in different disciplines. Griffiths and Steyvers turned to documents in different disciplines in their studies, Talley et al. he focused on a certain specific field in his work in 2011. He collected studies in the field of biological and health research from the American National Institutes of Health database (Talley et al., 2011). Talley et al. discovered in this study that many institutes are working on similar issues. They also observed that mapping multi-words to single tokens with preprocessing steps contributed significantly to performance.

Mimno (2012) analyzed the articles he collected from 24 journals from JStor with LDA. His purpose is to detect the change in scientific publications in the 20th

century. In this study, the polylingual topic model was used. It has been seen that there are more studies in the field of law and oratory in articles in German. Cultural and economic studies are more common in English-language studies.

Yau, Porter, Newman and Suominen (2014) use a topic model to organize their investigation of a corpus that over a century. They drew very important conclusions and made comments from the model they established with 21,000 documents and 150 topics.

LDA was applied on scientific publications and they wanted to create a series of documents containing academic papers from a variety of fields in order to see whether papers from the same field are grouped together. (Ranaei, Suominen, Dedehayir, 2017)

Topic modeling was used to show the relationship between science and technology in a specific field (Thomas, Adams, Hassan, Blostein, 2012). Latent Dirichlet Allocation, to detect the semantic relationship between patent and scientific publication corpus is examined in this paper. According to study, the topic clustering algorithm is a useful approach in detecting patent and publication linkage, according to a qualitative study of the clusters. In this study, hypothesized that using topics to explain the evolution of software repositories may help with maintenance and comprehension. They discovered that the vast majority of topic evolutions (87–89%) lead to real code change activities by developers. And they are encouraged to use topic models as methods for researching software system evolution.

Topical modeling has also been frequently used by social scientists. LDA was used in the analysis of both historical and sociological texts and in extracting their topics. In the literature, there are many studies conducted using LDA in the field of Economics (Wehrheim, 2019). LDA also has been used in the literature to explain large sets of texts in the field of psychology (Bittermann, Fischer, 2018).

There are also many studies in the literature comparing different topic model methods. NMF based methods have been successfully applied to topic modeling by Choo, Lee, Reddy and Park (2013), Choo, Lee, Reddy and Park (2015), H. Kim, Choo, J. Kim, Reddy and Park (2015). Topics were extracted with NMF and LDA, and it was seen that similar words were removed in the topic labeled as the same topic (MacMillan, Wilson, 2017).

Chapter 3

Methodology

3.1 Text Representation Methods

3.1.1 TF-IDF. TF-IDF is a weighting factor calculated by statistical methods that shows how important a term is in the document. TF is the frequency of a term in the document.

$$TF(w, d) = \frac{\text{occurrences of } w \text{ in document } d}{\text{total number of words in document } d}$$

IDF is a measure of the significance of the word. The logarithmically scaled version is calculated as the logarithm of the total number of documents divided by the number of documents containing the term. The more documents a term includes, the less information it contains, for example, frequently used terms such as conjunctions can appear in any document. If this term is mentioned in every document, it can be said that it does not contain information.

$$IDF(w, D) = \ln \left(\frac{\text{Total number of documents } (N) \text{ in corpus } D}{\text{number of documents containing } w} \right)$$

TF-IDF weighting is calculated by multiplying the TF and IDF values.

$$TFIDF(w, d, D) = TF(w, d) * IDF(w, D)$$

3.1.2 Bag of words. A bag-of-words model is a manner of extracting options from text to be used in modeling. The approach is very simple and flexible, and can be used in a myriad of ways for extracting features from documents. In this method, there is no information about the order or structure of words in the document. The model is only concerned with whether known words occur in the document, not where in the document.

3.2 Clustering Method

3.2.1 K-Means. K-means algorithm is one of the oldest and most commonly used clustering algorithms (Yang, Sinaga, 2019). With the input parameter k , the K-Means algorithm divides a set of n objects into K -clusters with a high intra-cluster similarity and a low inter-cluster similarity. The mean value of the items in the cluster, which may be thought of as the cluster's center of gravity, is used to quantify cluster similarity. Using Euclidean distance, the phrases' syntactical similarity is determined.

Steps of K-Means clustering are below.

- 1- Choose k observations as initial cluster centroids.
- 2- Assign each observation to the cluster with the closest centroids.
- 3- Recalculate the positions of the k centroids after all of the observations have been assigned.
- 4- Repeat until the cluster centroids stop changing.

This method's benefit is that it is simple to use and may be used to extract unknown groupings of data from large data sets. Another benefit is that trying to minimize the total SSE across a number of repetitions as a criteria function, it ensures convergence.

3.3 Topic Modeling Methods

Topic modeling is a statistical data analysis method for extracting meaningful information from large data sets. Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents (Blei, 2012).

Topical modeling allows the discovery of hidden thematic structures in large document sets. Topic modeling automatically extracts topics via computer. It is a method that analyzes the frequency of occurrence of words in a text data set and synthesizes the corpus as consisting of one or more topics. It helps to organize, summarize and search text sets. Topic modeling can also be used if content-based

matching is to be made between texts or a content-based recommendation system is to be developed.

Topic modeling accepts the words in the document as a bag of words. According to the word bag method, the order of the words does not matter. The topic modeling method is a generative method and can be easily applied to new data.

In topic modeling, topics are represented as a probability distribution over words. Each document is a probability distribution over the topics. Topics are calculated according to the frequency of occurrence of the words in the documents and the frequency of their occurrence together. In topic modeling, the probabilities of the words in each topic can be calculated, as well as the probability of each topic being found in each document.

3.3.1 Latent dirichlet allocation. LDA is a topic modeling method used to reveal the issues that make up the document. It is important that the words are found together, not their order and place in the document. The idea behind LDA is that all documents consist of multiple topics. LDA is a generative model used to reveal the issues that make up the document. K-Means assigned each document to a topic while grouping documents. LDA, on the other hand, calculates the probability of each document being found in each topic. For example, the document shown in Figure 1 is about genetics, life, brain and computer science at the same time.

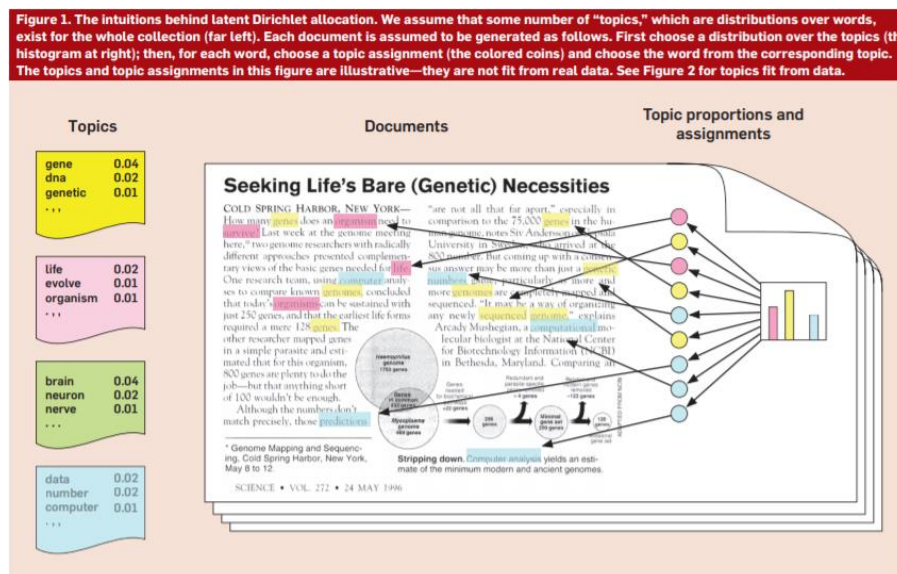


Figure 1. The intuitions behind LDA

LDA is a generative model. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words (Blei, 2012).

Figure 2 illustrates the graphical model representation of LDA. White circles are unobserved variables that will be estimated in the model; the shaded circle represents the observed variable. In the figure, D denotes the set of data resources that selected by a user, while N represents the keywords set extracted from a resource $d \in D$ respectively.

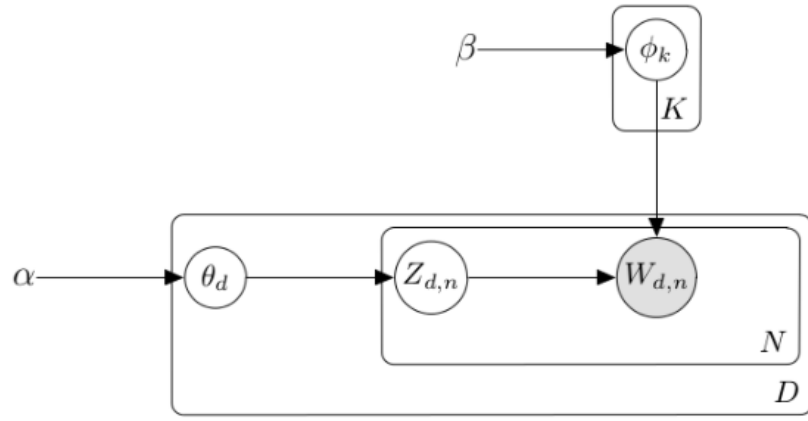


Figure 2. Graphical model representation of LDA

The keyword sets from the data resources chosen by each user correspond to the set of documents to be categorized in the LDA model, and each set of keywords retrieved from a data resource in our dataset is considered as a document in the LDA model. In the LDA model, θ_d and ϕ_k are assumed to be multinomial distributions, which are drawn from two separated Dirichlet distributions with priors α and β (Blei et al., 2003). The multinomial distribution is a distribution over discrete outcomes and the summation for the probabilities of all outcomes is 1. Dirichlet distribution is a conjugate prior to multinomial distribution. Given a number Z of latent topics, our generative model of LDA can be explained as follows:

1. For each data resource $d \in D$
 - Choose $\theta_d \sim \text{Dir}(\alpha)$
 - For each location d in n
 - * Choose a topic $z_{dn} \sim \text{Multinomial}(\theta_r)$

* Choose a word $w_{dn} \sim \text{Multinomial}(\phi z)$

2. For each topic $z \in Z$

• Choose $\phi_k \sim \text{Dir}(\beta)$

N represents the total number of documents in the corpus. w_{dn} represents word in that position n in document d , z_{dn} represents word's topic in that position n in document d . K is the total number of topics. θ indicates the probability of the topics being found in the document, and ϕ shows the distribution of the words in the topics. α and β are Dirichlet parameters.

The statistical model is based on the assumption that documents consist of a mixture of more than one topic. In the first step of the generative model, the topics show the probability distribution according to the Dirichlet distribution over the words in V , which is a fixed dictionary. This fixed dictionary consists of words in the document corpus. In the second step, the probability of each topic being included in the relevant document for each document is again determined according to the Dirichlet distribution. For each word in the document, the topics are sampled according to the multinomial distribution. Finally, the word for the relevant topic is sampled according to the polynomial distribution.

The combined distribution of latent and observed random variables in LDA $p(\phi_{1:K}, \theta_{1:M}, z_{1:M}, w_{1:M})$ is given in equation.

$$\left(\prod_{k=1}^K p(\phi_k | \beta) \right) \left(\prod_{m=1}^M \rho(Q_m | \alpha) \right) \left(\prod_{n=1}^N \rho(z_{m,n} | \theta_m) (w_{m,n} | z_{m,n}, \phi_k) \right)$$

The main goal with LDA is to obtain hidden variables, that is, model parameters. For this purpose, the posterior distribution in below equation is used.

$$P(\phi_{1:K}, \theta_{1:M}, z_{1:M} | w_{1:M}) = \frac{p(\phi_{1:K}, \theta_{1:M}, z_{1:M}, w_{1:M})}{p(w_{1:M})}$$

In order to calculate the marginal probability, which gives the probability of the document set under any topic model, it is necessary to collect the common distribution over all samples of the hidden topic structure. For this, sampling

algorithms are used. The Mallet package used for LDA in this thesis uses Gibbs sampling.

3.3.2 Nonnegative matrix factorization. Non-Negative Matrix factorization is a method to scale back the dimension of the input corpora. It uses factor analysis techniques to supply comparatively less weightage to words with less coherence.

Nonnegative matrix factorization (NMF) methods reduce a matrix into two low-rank matrices with the feature that none of the three matrices contain negative components.

$$A \approx WH.$$

The nonnegative basis vectors, which are represented as the columns of W , may be combined additively to approximate each data point, which is represented as the column of A . NMF especially performs well as a document clustering and topic modeling method (Kuang, Choo, Park, 2014).

NMF fundamentally produces the same output kinds as traditional topic modeling techniques like latent Dirichlet allocation (LDA): a topicwise document representation and a keyword-wise topic representation (the columns of W) (the columns of H). The main distinction, as opposed to the LDA outputs, is that the columns of W and H lack a unit L1-norm. However, this distinction is insignificant because NMF may be utilized as a topic modeling alternative and can be controlled using diagonal scaling matrices.

Chapter 4

Datasets And Experimental Study

4.1 Datasets

Within the scope and delimitations of the thesis, scientific publications from 2 different journals were collected from the JSTOR database over a wide period. These journals are Science Magazine, British Journal of the History of Science Journal. In Figure 3 metadata scheme of JSTOR data is shown. JSTOR shares two types of data. The first is the data in XML format and which is shown in Figure 3. XML data keep information such as the year of the article, when, where and by whom it was published. The second group of data provides the words in all documents in N-Gram format. In this thesis, two groups of data were combined and the required information was obtained from both formats. Year information is obtained from XML data while corpus is taken from data in N-Gram format.

The `journal-meta` element contains elements that describe the journal, significantly:

- `journal-id` `journal-id-type="jstor"` or `"publisher-id"`, a unique system identifier for that journal.
- `issn` elements, for the ISSN values that apply to the journal
- `journal-title`, as it was included in the source, not necessarily what the currently-publishing title is.
- `publisher-name`, as it was included in the source, not necessarily the publisher of the currently-publishing title.

The `article-meta` element can contain a variety of data. Elements that are reliably present:

- `article-id`
 - When `pub-id-type="doi"`, this is a registered DOI
 - When `pub-id-type="jstor"`, this is NOT a registered DOI
 - IDs that are not labeled as DOI have not been registered with CrossRef by JSTOR and should not be used with DOI resolution services.
- `article-categories` may be present, describing or grouping this article with others.
- `article-title`, if there was a title given in the source.
- `contrib-group`, for authors and possibly other creator roles identified in the source, as `contrib` elements within
- At least one `pub-date` element, with as much date information as the print source or electronic source contained. Some journals at times have multiple dates on a given issue, and these are all listed.
- An `fpage` ("first page") where the source provided it.
- Where the article is a review article, there may be one or more `product` elements describing works reviewed. A permissions element containing whatever copyright-statement or license statement the source provided.
- A `self-uri` element with an `xlink:href` containing a URL to the landing page for that publication at JSTOR.
- A `kwd-group` is included if the source supplied one.
- A `custom-meta-group` containing a custom-meta representing the languages present in the article. JSTOR's digitized backfile metadata tracks the important languages present, allowing for there to be more than one, and they are listed here in roughly order of quantity.

Figure 3. Metadata scheme of JSTOR data

After data is collected, firstly the data is passed through various pre-processing steps. HTML tags, punctuations, special characters and numbers are removed. All texts are converted to lowercase. Stemming algorithm was applied to words and stop words were removed from the dataset. The following figure shows the steps applied in the thesis.

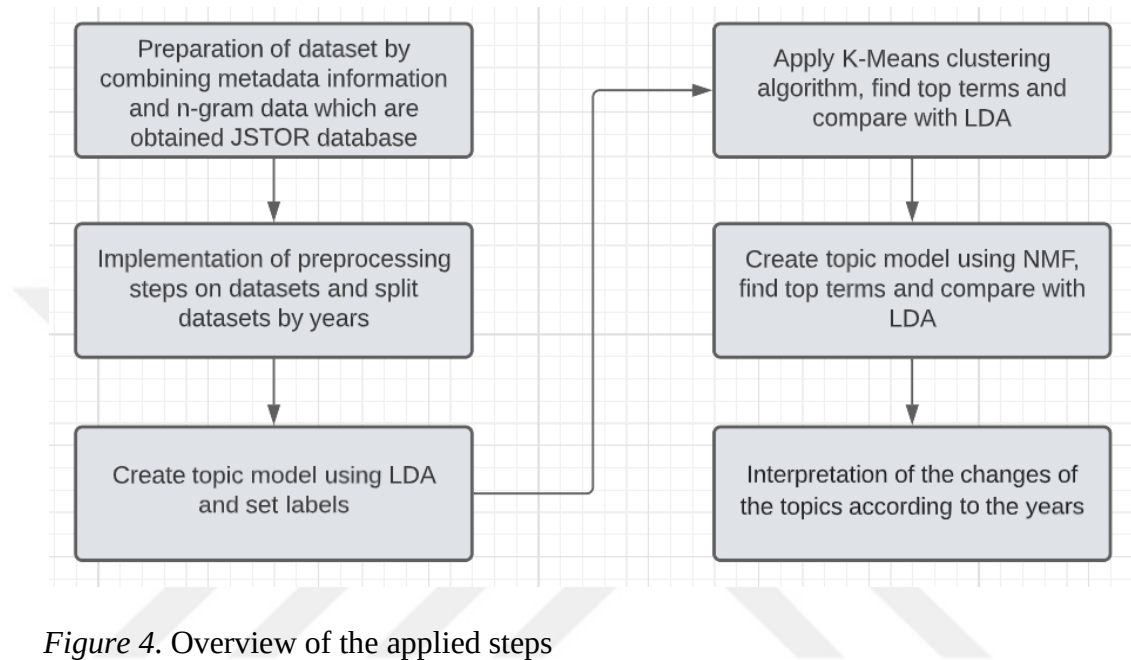


Figure 4. Overview of the applied steps

4.1.1 Science dataset. Science Magazine by the Science American Association for the Advancement of Science and is considered one of the most prestigious journals in the world. The first issue was published in 1880 and still continues. Besides scientific research and research views, it includes science news and science policies and publications. In addition to the general interest in all scientific disciplines, it has recently frequently included topics such as biology, genetics and life sciences. 15,008 scientific documents between 1880-1900, 17,741 documents between 1900-1920, 23993 documents between 1920-1940, 24,752 documents between 1940-1960, 23448 documents between 1960-1980, 24,110 documents between 1980-2000, 24,057 documents between 2000-2020, and totally 153,109 documents were obtained from JSTOR database. Because the dataset covers a wide period, the size is large and the journal is reputable, it is important to extract and interpret the topics automatically.

4.1.2 British journal for the history of science dataset. The British Journal for the History of Science is an international academic journal published quarterly by Cambridge University Press in association with the British Society for the History Of Science. It started to be published in 1962 and still continues. The journal contains articles in the fields of medicine, technology and social sciences. Approximately 12,000 scientific documents between 1960-2020 were collected. 2,732 documents between 1962-1980, 5,000 documents between 1980-2000, 4,838 documents between 2000-2020 were obtained from JSTOR database. The aim of using the dataset is automatically extract which topics were talked about more in the history of science over the years and observe the change in scientific publications topics.

4.2 Experimental Study

4.2.1 Science dataset results. Documents were collected from the Science Magazine from 1880 to 2020. The documents were divided into 10-year and 20-year intervals and the results were analyzed. The change of topics was tried to be understood with three different methods. LDA and NMF were used as topic modeling methods. K-means was used as a clustering method. The changes and trends were examined comparatively through the extracted topics and clusters.

Mallet library is used for LDA implementation. For Science Magazine, the number of topics was chosen as 30 for data divided into 20-year intervals. The number of topics were tested from 10 to 100, and it was seen that the 30 topics were best separated from the dataset. In 10-year intervals, it has been observed that the number of topics that separated the topics best was 20. The parameter α was chosen as 50/topic as suggested by Griffiths and Steyvers (2004) β is set to 0.1 by default.

The main issue here is how to follow the change of topics over the years. First of all, considering the scope of the Science Magazine, 60 sub-topics were determined manually and are shown in Table 1. These topics are Academic Research, Environmental Sciences, Evolution, Zoology, Organismal Biology, Scientific Writing, Geochemistry, Plant Biology, Physiology, Mathematics, Astrophysics, History of the USA, Paleontology, Social Science, Phonetic, Disease, Electricity, Anthropology, Scientific Law, Ecology, Developmental Biology, Optic, Nuclear Physics, Medicine, Chemical Engineering, Cell Biology, Agriculture, Astrophysics, Earth Science,

Scientific Theory, Geography, Chemistry, Geology, Research Methodology, Economic Science, Electromagnetic, Education, United States of America, World History, Medical Education, Human Science, Scientific Research, Analytical Chemistry, Anthropology, Neurophysics, Nutrition, Biochemistry, Sociology, Engineering, Hardware Technology, Metallurgy, Organic Chemistry, Technology, Neuroscience, Animal Testing, Genetic, Nuclear Energy, Electrophysics, Data Science, Healthcare, Cell Culture, Geophysics, Energy Systems, Molecular Biology, Medical Genetics.

Each of the 60 sub-topics can be grouped into 4 main categories: biological science, physiological science, social science and the scientific method. The trends of sub-topics in year intervals were examined. The rates of mentioning the main topics in the result and discussion section were also compared. There is a possibility that sub-topics dealing with similar areas in the main categories may be confused with each other. How did assign topic labels according to words are mentioned below.

Academic Research, Scientific Writing, Research Methodology, Scientific Research, Scientific Theory, Data Science, Scientific News and Social Science topics are related to the scientific method. Scientific Method topics have always been one of the most talked topics in every year ranges. Topics in this category are not a specific discipline, they are topics that can help every branch of science, and it has been seen that more general words are collected in these topics. Academic research is systematic and scientific work on any topic, and studies in this field are often associated with the academy. The most general definition of Scientific Writing is to write about scientific topics. Within the scope of this thesis, topics with writing and science-related words are labeled as Scientific Writing. Topics with words related to methods of science are labeled as Research Methodology. Scientific Research is topics that contain general words related to scientific research such as research, science scientific. Scientific Theory is the explanation of some aspects of nature, obtained through the scientific method through experiment and observation, which is constantly tested and verified. Mostly theoretical words are gathered in this topic within the scope of this thesis. In recent years, data science topics have started to be discussed and these topics have also been evaluated within the scope of this main topic.

Zoology, evolution, organismal biology, Plant Biology, Paleontology, Disease, Ecology, Developmental Biology, Medicine, Cell Biology, Agriculture, Nutrition, Biochemistry, Neuroscience, Animal Testing, genetic, Healthcare, Cell Culture,

molecular biology and medical genetics can be evaluated as biological science. Zoology is the branch of science that studies animals. It studies the development, behavior, heredity and interactions of animals from past to present. Paleontology is the science that aims to write the history of life on earth by using fossils as data. It has been seen that the words of the topics Evolution, Ecology and developmental biology, organismal biology are similar to each other. Evolution is the change of genes and traits in a population over generations. It deals with the origins of species and the relationships between them. Ecology studies the relationships of living things with each other and with their environment. Developmental biology investigates the growth and development of organisms and examines the evolution and change of cells that form tissues and systems. Organismal biology studies the diversity of biological organisms. Cell biology is a branch of science that studies the physiology, structure, organelles it contains, its relationship with the environment, life cycle, division and death of cells. Genetic and molecular biology topics, which have been talked about more recently, also focus on similar words. Molecular Biology studies how molecules interact with each other to perform the functions of life in living organisms. Although genetics seems to cover the topic of molecular biology, it can also cover non-molecular fields such as evolutionary genetics. Biochemistry is the science that studies the chemical substances in the structure of all living things in the form of plants, animals, and microorganisms and the chemical processes that go on throughout the life of the living thing. Cell culture is the process of growing cells under controlled conditions. In practice, the term cell culture is used for culturing cells originating from multicellular eukaryotes, especially animal cells. Separating sub-topics that examine similar topics is also one of the main challenges. Neuroscience is the science that studies and researches the nervous system and brain. In addition, with the acceleration of the application of genetic studies in the medical field in recent years, it has been seen that the topic of medical genetics has started to be discussed since the beginning of the 2000s. Medicine topics have been one of the most talked about topics since the 1880s. Disease and healthcare topics have also been among the topics mentioned every year. Plant biology examines the physiological structure of plants and has been one of the most talked about topics in every period. The relationship of plant biology with pharmacy can be associated with the fact that it is on the agenda in every period. Animal Testing is the procedure performed for scientific purposes on non-human living animal species, and it came to the fore in 1960-1980 and was not discussed

afterward. The Agriculture topic became one of the most talked about topics in the years when agriculture was still on the agenda. The topic of nutrition was also discussed more in similar years.

Astrophysics, Electricity, Geochemistry, Environmental science, Physiology, Mathematics, Optic, Nuclear Physics, Chemical Engineering, Earth Science, Geography, Chemistry, Geology, Electromagnetic, Analytical Chemistry, Neurophysics, Engineering, Hardware Technology, Metallurgy, Organic Chemistry, Technology, Nuclear Energy, Electrophysics, Geophysics, Energy Systems topics can be evaluated under physical science. Electrophysics, electromagnetic and electricity have been observed as topics that can be confused with each other. Electrophysics represents the overlap between physics and electrical and computer engineering. Electrophysics products ultimately fit into other areas of electrical and computer engineering such as communications and signal processing, computer engineering, microelectronics and controls. Electromagnetic; it is the physical interaction of two charged particles with an electric force. If the charges are the same, they repel each other; if the charges are opposite, they attract each other; The force between them is called electromagnetic. While these two topics stand out as more specialized topics in the field of electricity, the topics with less specialized words related to electricity were evaluated under electricity. Nuclear energy is the energy found in the nucleus of the atom. The energy stored in an atom's nucleus is known as nuclear energy. Energy systems is a field of study that aims for maximum efficiency in energy production, transmission and use. Electromagnetic is the physical force that causes the interaction between electrically charged particles. The fields where this interaction takes place are defined as electromagnetic fields. Geology, geophysics, geochemistry, geography, earth science, and Environmental science are topics that contain similar words. Geology is the science that studies the earth's core and layers. Geophysics is the science that studies the study of the Earth, including the hydrosphere and atmosphere, by utilizing the basic principles of physics. Geochemistry studies the chemical composition, processes and reactions of the Earth and other planets. Astrophysics, or stellar physics, is a branch of astronomy that uses the principles of physics and chemistry to determine the positions of celestial bodies in space and their structures rather than their motions and has been one of the topics that have attracted the attention of scientists since the 1800s. Basic sciences such as mathematics, chemistry and Physics were evaluated in this main category and they were among the most talked

about topics every year range. Over the years, we see that sub-categories of these basic sciences such as analytical chemistry, organic chemistry, nuclear physics and optics have started to be discussed. With the acceleration of engineering and technology studies, it has been seen that the topics of engineering, chemical engineering, technology, and hardware technology have been frequently discussed in recent years.

Social Science, Phonetic, Economic Science, Education, Science History, Human Science, Anthropology, Philology, and Culture sub-topics were generally evaluated in the social science category.

Table 1

Science Dataset Categories

Biological Science	Physical Science	Social Science	Scientific Method
Zoology	Astrophysics	Social Science	Academic Research
Evolution	Electricity	Phonetic	Scientific Writing
Organismal Biology	Geochemistry	Economic Science	Research Methodology
Plant Biology	Environmental Science	Education	Scientific Research
Paleontology	Physiology	Science History	Scientific Theory
Disease	Mathematics	Human Science	Data Science
Ecology	Optic	Anthropology	Scientific News
Developmental Biology	Nuclear Physics	Philology	Social Science
Medicine	Chemical Engineering	Culture	
Cell Biology	Earth Science		
Agriculture	Geography		
Nutrition	Chemistry		
Biochemistry	Geology		
Neuroscience	Electromagnetic		
Animal Testing	Analytical Chemistry		
Genetic	Neurophysics		
Healthcare	Engineering		

Table 1 (continued)

Cell Culture	Hardware Technology		
Molecular Biology	Metallurgy		
Medical Genetics	Organic Chemistry		
	Technology		
	Nuclear Energy		
	Electrophysics		
	Geophysics		
	Energy Systems		

Topics which are extracted by LDA and their labels between the years of 1880-1900 are in Table 2.

Table 2

Topics And Labels Which Are Extracted Using LDA Between 1880-1900 For Science Dataset

Topic 1 Academic Research	Topic 2 Environmental Sciences	Topic 3 Evolution	Topic 4 Zoology	Topic 5 Organismal Biology
university year science laboratory museum work school scientific student institution	coast north island sea west river south expedition great east	plant species form variation group character type study evolution tion	species fish museum collection specimen form find animal number work	cell form organ body nerve egg structure find animal tissue

Table 2 (continued)

Topic 6 Scientific Writing	Topic 7 Geochemistry	Topic 8 Plant Biology	Topic 9 Physiology	Topic 10 Mathematics
science letter writer fact question case editor year opinion good	air temperature wind earth pressure surface heat great velocity storm	plant experiment year soil food cent station milk seed result	bone form tooth fig skull find brain large specimen present	die geometry science number mathematical point theory paris equation tion
Topic 11 Astrophysics	Topic 12 Economic Science	Topic 13 Paleontology	Topic 14 Social Science	Topic 15 Phonetic
observation star time observatory sun year line great solar instrument	work cent science american price free lafayette year send boston	rock geological bed formation deposit find geology coal fossil survey	state work year country report united government survey public states	word language child sound english number deaf letter form time
Topic 16 Disease	Topic 17 Electricity	Topic 18 Anthropology	Topic 19 Scientific Theory	Topic 20 Ecology
disease case health fever death water time animal medical result	electric current power electrical wire engine electricity light steam work	find man race tribe people ancient stone indians american human	theory fact law matter force body nature science tion form	bird find tree insect day time year animal large egg
Topic 21 Geography	Topic 22 Chemistry	Topic 23 Geology	Topic 24 Research Methodology	Topic 25 Economic Science
river lake foot water mountain valley region great south north	gas acid chemical solution water form substance liquid metal chemistry	foot water inch place large small end stone surface iron	work subject student method text study general present treat good	Exchange science price send work sale good free american cent

Table 2 (continued)

Topic 26 Electromagnetic	Topic 27 Education	Topic 28 Scientific News	Topic 29 Science History	Topic 30 Other
light color eye line ray plate wave experiment case red	man science great work study life school scientific time knowledge	science american section secretary academy member washington present read university	french england paris history france english united british map report	title and great and high place present the time state

Labels are assigned to the topics manually. Then, in order to determine the most talked about topics, the dominant topics in the range of every two decades were determined. The highest probability topic of each document in the corpus was considered as the dominant topic for that document. Dominant topics which the most documents were assigned were evaluated as the most talked about topic.

We see that the most mentioned topics between 1880-1900 were anthropology, electricity, geology, disease and geography, respectively.

Table 3

Hot Topics Between 1880-1900 For Science Dataset

Topic	Number of Document	Probability
Anthropology	2436	0.1488
Electricity	847	0.0518
Geology	838	0.0512
Disease	838	0.0512
Geography	726	0.0444

Clusters extracted with k-means are shown in Table 4, and topics extracted with NMF are shown in Table 5. The hot topics obtained with LDA were compared with the ones obtained with K-means and NMF, and the words of similar topics obtained in different algorithms were examined.

Table 4

Clusters Between 1880-1900 For Science Dataset

Cluster	Top Terms
Cluster 1	science, great, time, form, know, fact, work, life, use, tion
Cluster 2	shool, work, science, study, year, great, time, course, child, teach
Cluster 3	rock, form, great, foot, bed, science, new, deposit, formation, present
Cluster 4	science, new, form, time, great, work, year, use, state, present
Cluster 5	time, use, year, great, work, case, state, result, new, form
Cluster 6	new, work, science, cent, price, year, state, use, school, number
Cluster 7	result, form, work, science, time, great, theory, line, case, tion
Cluster 8	foot, great, mile, water, time, form, large, ice, high, mountain
Cluster 9	work, science, new, year, great, time, scientific, state, study, tion
Cluster 10	atom, element, chemistry, compound, weight, form, chemical, know, science, theory
Cluster 11	air, gas, heat, temperature, pressure, liquid, water, work, point, use
Cluster 12	plant, form, specie, science, year, time, great, grow, leave, study
Cluster 13	milk, fermentation, bacteria, sour, produce, cow, use, long, cause, trouble
Cluster 14	specie, form, new, science, great, work, present, time, plant, fact
Cluster 15	inch, observation, work, line, photograph, great, time, plate, observatory, instrument
Cluster 16	work, science, new, year, state, great, time, use, present, study
Cluster 17	time, great, use, work, science, form, year, point, case, water
Cluster 18	cell, live, form, matter, organism, science, structure, body, tion, work
Cluster 19	state, year, great, new, science, country, work, time, forest, use
Cluster 20	station, work, state, experiment, year, investigation, number, agricultural, water, farmer
Cluster 21	specie, insect, plant, form, case, great, time, family, know, year
Cluster 22	school, student, work, course, science, study, year, teach, subject, education
Cluster 23	seed, test, result, sprout, sample, great, soil, germination, plant, cent
Cluster 24	science, work, great, scientific, time, year, new, knowledge, tion, fact
Cluster 25	earth, time, great, year, heat, present, science, surface, form, rock

Table 4 (continued)

Cluster 26	work, science, new, year, state, great, time, present, number, study
Cluster 27	water, great, time, form, use, surface, work, year, new, large
Cluster 28	form, great, time, water, specie, new, work, present, large, region
Cluster 29	wave, component, cam, phase, case, length, particle, velocity, curve, axle
Cluster 30	battery, charge, plate, use, light, lead, effect, cell, current, experiment

Table 5

Topics Which Are Extracted Using NMF Between 1880-1900 For Science Dataset

Topic	Top Terms
Topic 1	science, great, time, form, know, fact, work, life, use, tion
Topic 2	shool, work, science, study, year, great, time, course, child, teach
Topic 3	rock, form, great, foot, bed, science, new, deposit, formation, present
Topic 4	science, new, form, time, great, work, year, use, state, present
Topic 5	time, use, year, great, work, case, state, result, new, form
Topic 6	new, work, science, cent, price, year, state, use, school, number
Topic 7	result, form, work, science, time, great, theory, line, case, tion
Topic 8	foot, great, mile, water, time, form, large, ice, high, mountain
Topic 9	work, science, new, year, great, time, scientific, state, study, tion
Topic 10	atom, element, chemistry, compound, weight, form, chemical, know, science, theory
Topic 11	air, gas, heat, temperature, pressure, liquid, water, work, point, use
Topic 12	plant, form, specie, science, year, time, great, grow, leave, study
Topic 13	milk, fermentation, bacteria, sour, produce, cow, use, long, cause, trouble
Topic 14	specie, form, new, science, great, work, present, time, plant, fact
Topic 15	inch, observation, work, line, photograph, great, time, plate, observatory, instrument
Topic 16	work, science, new, year, state, great, time, use, present, study
Topic 17	time, great, use, work, science, form, year, point, case, water
Topic 18	cell, live, form, matter, organism, science, structure, body, tion, work

Table 5 (continued)

Topic 19	state, year, great, new, science, country, work, time, forest, use
Topic 20	station, work, state, experiment, year, investigation, number, agricultural, water, farmer
Topic 21	specie, insect, plant, form, case, great, time, family, know, year
Topic 22	school, student, work, course, science, study, year, teach, subject, education
Topic 23	Seed, test, result, sprout, sample, great, soil, germination, plant, cent
Topic 24	science, work, great, scientific, time, year, new, knowledge, tion, fact
Topic 25	earth, time, great, year, heat, present, science, surface, form, rock
Topic 26	work, science, new, year, state, great, time, present, number, study
Topic 27	water, great, time, form, use, surface, work, year, new, large
Topic 28	form, great, time, water, specie, new, work, present, large, region
Topic 29	wave, component, cam, phase, case, length, particle, velocity, curve, axle
Topic 30	battery, charge, plate, use, light, lead, effect, cell, current, experiment

The similarities of the clusters in the cluster center which were extracted with K-means and the topics extracted with NMF and LDA were examined. It has been observed that electricity and geology, which are the most talked about topics and extracted with LDA, are also captured with K-means and NMF. The top terms of electricity topic for three algorithms is shown in Table 6 and the top terms of geology topic for three algorithms is shown in Table 7.

Table 6

Top Terms Of Electricity Topic Between 1880-1900

LDA	K-Means	NMF
species	battery	current
bird	charge	electrical
fish	plate	car
insect	use	wire
find	light	battery
water	lead	electricity
animal	effect	light
number	cell	company
form	current	coil
life	experiment	use

Table 7

Top Terms Of Geology Topic Between 1880-1900

LDA	K-Means	NMF
foot great mile water time form large ice high mountain	foot great mile water time form large ice high mountain	foot mile mountain level stream plain great region height land

Topics that are extracted by LDA and their labels between 1900-1920 are in Table 8.

Table 8

Topics And Labels Which Are Extracted Using LDA Between 1900-1920 For Science Dataset

Topic 1 Ecology	Topic 2 Disease	Topic 3 Mathematics	Topic 4 Human Science	Topic 5 Chemistry
species bird fish insect find water animal number form life	disease bacteria animal case blood organism culture find fever day	mathematical force theory number body tion point mathematics system equation	man great science life world day scientific thing time human	acid solution water chemical method salt cent compound substance action
Topic 6 Developmental Biology	Topic 7 Optic	Topic 8 Paleontology	Topic 9 Other	Topic 10 Plant Biology
character color female male variation produce case sex white factor	light end place small water room tube large foot glass	fossil form museum type find bone specimen species animal group	whieh work tion ean great study sueh present seienee year	plant species tree botanical find form leave flower flora grow

Table 8 (continued)

Topic 11 Education	Topic 12 Research Methodology	Topic 13 Geochemistry	Topic 14 Nuclear Physics	Topic 15 Economic Science
student school work teacher university education year study man schools	work science research scientific year study laboratory investigation field sciences	temperature water air weather wind day great observation condition sea	energy atom theory ray heat gas light temperature current effect	year cent number increase average total large rate man table
Topic 16 Medicine	Topic 17 Science History	Topic 18 Scientific News	Topic 19 Philology	Topic 20 Culture
medical medicine health hospital war science public american scientific disease	american secretary section hold member university science december washington vice	year sir royal british director lecture hold congress die museum	science word type case term species rule statement writer fact	museum collection american history find man art people culture stone
Topic 21 Geology	Topic 22 Chemical Engineering	Topic 23 Scientific Writing	Topic 24 Cell Biology	Topic 25 Agriculture
rock foot valley river deposit water form lake bed find	chemical work industry chemist engineer chemistry great industrial coal engineering	work subject text general present reference student include discussion reader	cell egg form body organ tissue animal nerve normal process	soil plant food station agricultural agriculture experiment crop state forest
Topic 26 Education	Topic 27 Social Science	Topic 28 Astrophysics	Topic 29 Earth Science	Topic 30 Scientific Theory
university school year state assistant appoint harvard chemistry uni versity	state public institution government national work service tion board man	star observation observatory sun line earth result solar time great	survey state south north coast united island region california work	fact tion problem form theory method view science result condition

World History, Scientific News, Education, Scientific Writing, Social Science and Science History were discussed the most between 1900 and 1920. The probabilities of the topics are given in Table 9. It has been observed that the compilation is not very distinctive in this year's range, as the highest probability topics are those determined in the general category.

Table 9

Hot Topics Between 1900-1920 For Science Dataset

Topic	Number of Document	Probability
Scientific News	1372	0.0838
Education	1345	0.0822
Scientific Writing	933	0.0570
Social Science	900	0.0550
Science History	839	0.0513

Clusters extracted with k-means are shown in Table 10, and topics extracted with NMF are shown in Table 11 in 1900-1920 for Science Dataset.

Table 10

Clusters Between 1900-1920 For Science Dataset

Cluster	Top Terms
Topic 1	science, history, course, school, study, work, student, engineer, great, subject
Topic 2	form, foot, great, new, present, water, science, time, tion, rock
Topic 3	work, science, year, new, state, present, tion, time, great, use
Topic 4	work, science, great, time, year, tion, study, scientific, new, student
Topic 5	milk, culture, cent, bacteria, organism, water, medium, use, result, number
Topic 6	plant, form, specie, study, work, science, tion, result, time, case
Topic 7	water, solution, tion, use, result, form, science, method, time, work
Topic 8	water, solution, method, use, cent, result, tion, form, salt, chemical
Topic 9	science, scientific, great, work, life, time, tion, knowledge, fact, know

Table 10 (continued)

Topic 10	new, science, secretary, section, scientific, school, vice, year, educational, state
Topic 11	cell, form, egg, tion, case, body, tissue, result, surface, change
Topic 12	year, school, student, increase, gain, course, number, large, law, total
Topic 13	medical, school, work, year, student, course, state, time, medicine, education
Topic 14	specie, form, science, work, plant, water, tion, use, present, time
Topic 15	work, year, science, great, time, tion, new, scientific, state, institution
Topic 16	specie, type, form, new, genus, case, science, tion, plant, fish
Topic 17	earth, time, great, temperature, tion, science, result, light, year, use
Topic 18	year, state, work, new, science, school, present, tion, hold, scientific
Topic 19	science, work, new, year, state, time, use, tion, present, form
Topic 20	science, work, new, year, state, time, use, tion, present, form
Topic 21	school, student, work, course, year, science, teach, study, time, great
Topic 22	science, form, tion, time, great, theory, work, point, new, fact
Topic 23	line, observation, plate, time, great, result, magnitude, bright, use, light
Topic 24	atom, element, theory, weight, number, matter, science, particle, form, charge
Topic 25	medical, medicine, school, science, work, time, year, great, health, study
Topic 26	work, science, scientific, great, year, time, tion, new, institution, state
Topic 27	work, science, scientific, great, year, time, tion, new, institution, state
Topic 28	yellow, cocoon, mate, produce, lot, male, double, generation, race, female
Topic 29	water, ore, great, copper, rock, use, deposit, time, surface, large
Topic 30	solution, salt, water, cell, egg, experiment, case, live, osmotic, pressure

Table 11

Topics Which Are Extracted Using NMF Between 1900-1920 For Science Dataset

Topic	Top Terms
Topic 1	science, life, great, thing, scientific, think, time, mind
Topic 2	appoint, assistant, instructor, chemistry, professorship, chair, physics, agricultural, school

Table 11 (continued)

Topic 3	rock, foot, bed, deposit, mile, region, glacial, ice, limestone
Topic 4	plant, seed, botanical, botanist, grow, root, tree, fruit, study
Topic 5	section, session, hold, program, science, evening, membership, annual, joint
Topic 6	work, station, scientific, state, government, country, agricultural, industry, investigation, research
Topic 7	cell, tissue, nerve, form, chromosome, process, structure, substance, embryo, body
Topic 8	school, student, education, course, teach, high, educational, year, teacher, training
Topic 9	chemistry, chemical, chemist, industry, chem, laboratory, industrial, manufacture, analysis, product
Topic 10	cent, year, increase, total, population, production, average, number, rate, state
Topic 11	medical, medicine, hospital, clinical, physician, surgery, school, profession, anatomy, practise
Topic 12	specie, genus, type, form, generic, family, specimen, rule, new, nomenclature
Topic 13	museum, history, natural, specimen, collection, exhibit, art, museum, new, exhibition
Topic 14	work, text, student, subject, use, method, reference, treat, theory, new
Topic 15	soil, crop, nitrogen, wheat, fertilizer, fertility, sample, bacteria, phosphate, root
Topic 16	egg, membrane, sperm, fertilization, fertilize, embryo, permeability, experiment, development, cleavage
Topic 17	color, red, yellow, black, light, blue, eye, pink, dark, gray
Topic 18	institution, laboratory, trustee, building, room, carnegie, salary, student, faculty, year
Topic 19	male, female, sex, chromosome, sexual, pair, hybrid, produce, wing, character
Topic 20	atom, force, theory, body, unit, weight, motion, law, velocity, particle
Topic 21	engineering, engineer, electrical, civil, mechanical, mathematic, mining, industrial, engine
Topic 22	milk, culture, bacteria, organism, blood, food, medium, spore, infection, bacillus
Topic 23	weather, temperature, wind, air, meteorological, meteorology, pressure, rainfall, cyclone, balloon
Topic 24	year, royal, award, medal, lecture, expedition, hold, member, geological, age
Topic 25	bird, wing, life, sparrow, ornithologist, nest, wild, habit, game, flight
Topic 26	secretary, vice, section, new, member, tary, versity, secre, science, treasurer
Topic 27	fish, water, fishery, food, seal, fresh, coast, fauna, stream, supply
Topic 28	observatory, observation, solar, line, magnetic, spectrum, plate, instrument, inch, light
Topic 29	forest, tree, forestry, timber, state, land, forester, acre, wood, fire
Topic 30	health, public, service, state, food, fever, officer, tuberculosis, water, city

Education, which was one of the most talked about topics between 1900-1920, was also found with K-means and NMF. It has been observed that distinguishing words such as student, school appear in all three algorithms related to this topic.

Table 12

Top Terms Of Education Topic Between 1900-1920

LDA	K-Means	NMF
student school work teacher university education year study man schools	school student education course teach high educational year teacher training	school student work course year science teach study time great

Topics that are extracted by LDA and their labels between 1920-1940 are in Table 13.

Table 13

Topics And Labels Which Are Extracted Using LDA Between 1920-1940 For Science Dataset

Topic 1 Scientific Writing	Topic 2 Education	Topic 3 Academic Research	Topic 4 Astrophysics	Topic 5 Scientific Research
work science word subject publication text number writer term scientific	research work science student man medical year study school university	university state academy study science california research chemistry geology ohio	star sun earth light year observatory time observation air great	year work great time science life scientific early die death

Table 13 (continued)

Topic 6 Chemical Engineering	Topic 7 Medicine	Topic 8 Developmental Biology	Topic 9 Human Science	Topic 10 Other
industry chemical research engineering engineer year work industrial company laboratory	hospital year institute medical research director school american medicine chemistry	chromosome cell gene female sex egg species male case number	man science scientific day great world life thing time human	monthly scientific fig and progress and work the ing and
Topic 11 Mathematics	Topic 12 Plant Biology	Topic 13 Organismal Biology	Topic 14 Nuclear Physics	Topic 15 Ecology
theory mathematical law time number physics space motion system mathematics	plant soil tree seed root grow growth fruit leave find	cell culture medium tissue solution bacteria cent stain organism method	atom electron energy ray wave light element atomic particle charge	fish species insect bird animal water find large food sea
Topic 16 Scientific Theory	Topic 17 Research Methodology	Topic 18 Disease	Topic 19 Analytical Chemistry	Topic 20 Social Science
science scientific theory figure university price company lancaster laboratory advancement	problem study tion life fact science form change present process	disease animal virus health infection fever case blood medical day	tube glass water method apparatus place air small end fig	research scientific national science work year american state council congress
Topic 21 Chemistry	Topic 22 Anthropology	Topic 23 Geology	Topic 24 Neurophysics	Topic 25 Economic Science
acid solution water chemical method salt cent compound substance action	museum expedition work island collection study south american field north	water rock foot area surface sea earth river great region	nerve blood muscle brain effect response system cell nervous fiber	cent number rate result method increase test measure average table

Topic 26 Paleontology	Topic 27 Nutrition	Topic 28 Agriculture	Topic 29 Scientific News	Topic 30 Demography
fossil find man bone type species form skull animal tooth	vitamin animal rat diet growth cent day effect blood normal	forest state land area year agriculture agricultural soil increase country	section american session science hold secretary member program present general	year child man population age people race white country death

The most mentioned topics between 1920 and 1940 were medicine, political science, analytical chemistry, scientific writing and world history.

Table 14

Hot Topics Between 1920-1940 For Science Dataset

Topic	Number of Document	Probability
Medicine	3320	0.1506
Social Science	1236	0.0561
Analytical Chemistry	1124	0.0510
Scientific Writing	894	0.0406
Demography	887	0.0402

Clusters extracted with k-means are shown in Table 15, and topics extracted with NMF are shown in Table 16 in 1920-1940.

Table 15

Clusters Between 1920-1940 For Science Dataset

Cluster	Top Terms
Topic 1	science, scientific, work, new, great, time, year, tion, social, life
Topic 2	blood, cell, time, tion, study, result, science, body, protein, form

Table 15 (continued)

Topic 3	forest, tree, year, plant, land, great, time, state, specie, large
Topic 4	plant, soil, seed, science, use, water, time, study, great, form
Topic 5	session, section, science, present, report, hold, new, study, year, program
Topic 6	section, session, science, new, present, hold, report, year, vice
Topic 7	science, year, work, new, use, state, time, study, scientific, present
Topic 8	use, water, time, science, plant, result, cent, study, method, form
Topic 9	age, child, year, test, life, cent, ability, time, great, education
Topic 10	science, year, work, new, state, scientific, study, time, present, research
Topic 11	cell, tissue, form, growth, organism, gene, cancer, change, process, science
Topic 12	science, scientific, work, great, new, time, human, fact, knowledge, method
Topic 13	year, great, state, increase, time, population, work, use, land, country
Topic 14	use, solution, water, method, temperature, study, tion, result, form, time
Topic 15	theory, time, science, great, number, tion, light, law, point, fact
Topic 16	science, work, scientific, new, year, time, great, study, us, tion
Topic 17	medical, medicine, science, year, school, work, student, health, great, time
Topic 18	atom, element, particle, number, charge, great, hydrogen, time, form, weight
Topic 19	year, time, use, new, great, science, state, know, work, light
Topic 20	section, science, session, secretary, report, hold, present, year, vice, new
Topic 21	time, great, year, work, science, new, day, know, study, scientific
Topic 22	time, theory, wave, space, light, motion, law, point, science, atom
Topic 23	year, new, science, state, work, school, medical, hold, chemistry, study
Topic 24	evolution, mutation, gene, specie, change, new, tion, cell, process, form
Topic 25	specie, form, type, time, study, tion, great, race, science, know
Topic 26	soil, land, erosion, plant, crop, use, water, great, cent, year
Topic 27	section, session, hold, science, present, program, afternoon, morning, joint, secretary
Topic 28	water, surface, great, salt, time, form, use, temperature, change, solution
Topic 29	earth, rock, great, surface, crust, time, mountain, depth, form, tion
Topic 30	plant, rust, resistant, wheat, crop, race, year, stem, certain, new

Table 16

Topics Which Are Extracted Using NMF Between 1920-1940 For Science Dataset

Cluster	Top Terms
Topic 1	science, life, human, great, knowledge, thing, social, think, scientific, fact
Topic 2	secretary, member, vice, treasurer, annual, executive, science, officer, section, state
Topic 3	solution, protein, reaction, salt, compound, ion, cent, sodium, concentration, hydrogen
Topic 4	session, section, program, hold, afternoon, evening, symposium, morning, joint, dinner
Topic 5	atom, element, particle, theory, hydrogen, helium, molecule, charge isotope, structure
Topic 6	chemistry, chemical, chemist, industry, industrial, laboratory, company, chemistry, organic
Topic 7	tube, glass, use, fig, apparatus, rubber, inch, air, end, plate
Topic 8	health, public, service, state, fever, child, officer, food, population, education
Topic 9	medical, medicine, hospital, school, physician, new, clinical, surgery, building, patient
Topic 10	work, scientific, government, state, country, institution, service, industry, research, international
Topic 11	specie, bird, fish, genus, form, insect, specimen, evolution, find, type
Topic 12	museum, expedition, natural, history, new, exhibit, collection, bird, specimen, field
Topic 13	plant, seed, tree, root, grow, growth, garden, fruit, botanical, crop
Topic 14	medal, award, royal, work, gold, recognition, honor, presentation, annual, member
Topic 15	science, scientific, new, price, illustrate, advertisement, figure, copy, edit, text
Topic 16	forest, tree, forestry, land, timber, fire, acre, state, forester, conservation
Topic 17	cell, tissue, organism, growth, live, medium, protoplasm, culture, structure, form
Topic 18	engineering, engineer, electrical, company, new, industry, mechanical, civil, industrial, technology
Topic 19	virus, mosaic, strain, tobacco, inoculation, inoculate, protein, culture, infection, brain
Topic 20	year, age, death, die, old, emeritus, life, recent, work, memorial
Topic 21	soil, crop, land, erosion, agricultural, root, moisture, fertilizer, nitrogen, culture
Topic 22	appoint, assistant, school, associate, resign, agricultural, state, station, physics, chair
Topic 23	light, wave, earth, theory, motion, observatory, velocity, line, distance, solar
Topic 24	male, female, chromosome, sex, egg, gene, genetic, population, number, rate
Topic 25	water, fish, temperature, surface, air, stream, ice, fishery, heavy, fresh

Table 16 (continued)

Topic 26	vitamin, rat, oil, liver, milk, food, growth, yeast, cod, fat
Topic 27	student, course, work, teach, school, study, teacher, subject, text, field
Topic 28	fellowship, grant, fellow, study, year, award, work, research, foundation, application
Topic 29	stain, cent, slide, color, solution, use, blue, method, dye, section
Topic 30	blood, extract, adrenal, rat, normal, tissue, hormone, gland, injection, tumor

Medicine, one of the most talked about topics between 1920-1940, was also caught with K-means and NMF. It has been observed that distinguishing words such as medical and health appeared in all three algorithms related to this topic. The Analytical Chemistry topic also was captured only by NMF and LDA, and 6 of the 10 words with the highest probability were found to be common.

Table 17

Top Terms Of Medicine Topic Between 1920-1940

LDA	K-Means	NMF
hospital	medical	school
year	medicine	student
institute	science	work
medical	year	course
research	school	year
director	work	science
school	student	teach
american	health	study
medicine	great	time
chemistry	time	great

Table 18

Top Terms Of Analytical Chemistry Between 1920-1940

LDA	NMF
tube	tube
glass	glass
water	use
method	fig
apparatus	apparatus
place	rubber
air	inch
small	air

end fig	end plate
------------	--------------

Topics that are extracted by LDA and their labels between 1940-1960 are in Table 19.

Table 19

Topics And Labels Which Are Extracted Using LDA Between 1940-1960 For Science Dataset

Topic 1 Scientific News	Topic 2 Research Methodology	Topic 3 Scientific Theory	Topic 4 Chemistry	Topic 5 Science History
research national work scientific state service public government council fund	problem study method tion change form individual factor development result	theory century scientific great work early year monthly time write	chemical chemistry industry product research chemist oil year work material	section american session hold science secretary member program present general
Topic 6 Scientific Writing	Topic 7 Electromagnetic	Topic 8 Paleontology	Topic 9 Disease	Topic 10 Organismal Biology
science scientific american price company figure laboratory advancement text hill	sound radio current power electric room laboratory high instrument time	find day man time long good body case leave hand	virus culture animal disease bacteria organism day serum infection medium	blood cell animal tissue nerve effect normal muscle gland body
Topic 11 Scientific Research	Topic 12 Plant Biology	Topic 13 Education	Topic 14 Academic Research	Topic 15 Analytical Chemistry
work year student science research great man time scientific study	plant soil seed tree grow crop root growth insect fruit	university year school medical appoint director die medicine research sir	university state california study academy science ohio illinois iowa michigan	tube method glass water fig place apparatus small plate surface

Table 19 (continued)

Topic 16 Biochemistry	Topic 17 Mathematics	Topic 18 Nutrition	Topic 19 Developmental Biology	Topic 20 Medicine
acid solution water cent reaction compound protein substance concentration form	theory number law time mathematical point system space tion physical	vitamin animal cent rat growth diet food day milk experiment	species cell egg form chromosome female find sex number gene	medical disease health medicine public hospital study patient year school
Topic 21 Social Science	Topic 22 Earth Science	Topic 23 Engineering	Topic 24 Cell Biology	Topic 25 Ecology
science man scientific world human life thing fact nature knowledge	water sea fish weather temperature air coast ocean great foot	institute american award engineering medal research engineer director company university	science subject publication work word number writer term list present	forest bird animal man area find species tree year wild
Topic 26 Nuclear physics	Topic 27 Geology	Topic 28 Economic Science	Topic 29 Anthropology	Topic 30 Astrophysics
atom energy ray electron element wave particle atomic charge mass	rock earth area geological valley foot, surface deposit river time	year, increase population cent age man rate great country people	museum expedition american collection work congress south island international study	light star sun earth year observatory time observation solar line

The most mentioned topics between 1940 and 1960 were geology, nutrition, astrophysics, chemistry and social science.

Table 20

Hot Topics Between 1940-1960 for Science Dataset

Topic Label	Number of Document	Probability
Geology	2211	0.1003
Nutrition	1573	0.0714
Astrophysics	1295	0.0588
Chemistry	1088	0.0494
Social Science	1074	0.0487

Clusters extracted with k-means are shown in Table 21 and topics extracted with NMF are shown in Table 22 in 1940-1960.

Table 21

Clusters Between 1940-1960 For Science Dataset

Cluster	Top Terms
Topic 1	protein, virus, enzyme, reaction, study, use, tion, form, result, cell
Topic 2	time, use, field, fig, tion, high, light, result, large, value
Topic 3	use, result, study, time, effect, day, water, test, tion, solution
Topic 4	science, scientific, scientist, new, work, time, year, great, fact, problem
Topic 5	social, science, scientist, scientific, human, natural, study, tion, problem, physical
Topic 6	year, time, use, plant, new, great, state, work, science, study
Topic 7	new, science, study, state, medical, school, laboratory, year, work, field
Topic 8	new, science, year, state, work, study, school, program, scientific, research
Topic 9	section, science, company, sound, session, aaas, wheeler, symposium, statistic, new
Topic 10	science, use, new, work, study, present, time, tion, year, research
Topic 11	science, session, aaas, section, program, new, symposium, annual, scientific, hold
Topic 12	section, session, science, program, hold, present, new, symposium, secretary, aaas
Topic 13	new, use, work, science, field, study, present, tion, problem, time
Topic 14	new, annual, science, univ, assoc, state, dept, international, ave, scientific
Topic 15	new, year, science, school, state, medical, laboratory, study, award, work

Table 21 (continued)

Topic 16	gene, genetic, mutation, different, tion, effect, study, type, mutant, new
Topic 17	science, room, section, new, company, session, aaas, exhibit, symposium, auditorium
Topic 18	scientific, science, work, government, year, problem, scientist, state, new, program
Topic 19	science, time, new, theory, scientific, social, tion, human, work, use
Topic 20	sample, date, age, submit, collect, wood, site, level, peat, depth
Topic 21	science, scientific, year, scientist, work, new, state, time, tion, field
Topic 22	age, year, fig, contribution , line, time, datum, present, rate, curve
Topic 23	effect, study, tion, use, cell, new, plant, science, growth, method
Topic 24	soil, plant, calcium, water, use, surface, fertilizer, erosion, process, great
Topic 25	cell, culture, tissue, study, use, tion, normal, time, growth, result
Topic 26	school, high, science, student, education, year, program, test, study, work
Topic 27	water, use, time, great, land, surface, form, large, ice, foot
Topic 28	session, room, new, joint, aaas, science, section, symposium, seience, plant
Topic 29	snow, ice, foot, melt, water, density, wind, meter, depth, high
Topic 30	coast, level, land, coastal, time, bering, fault, year, line, form

Table 22

Topics Which Are Extracted Using NMF Between 1940-1960 For Science Dataset

Topic	Top Terms
Topic 1	medical, medicine, health, school, hospital, public, physician, service, clinical, tropical
Topic 2	blood, rat, day, patient, injection, plasma, dog, liver, normal, effect
Topic 3	science, social, scientist, scientific, human, think, fact, life, knowledge, people
Topic 4	tube, use, glass, fig, apparatus, solution, pressure, air, method, temperature
Topic 5	program, government, scientific, scientist, agency, service, federal, public, state, country
Topic 6	session, aaas, symposium, hold, program, afternoon, arrange, evening, annual, science
Topic 7	plant, seed, root, growth, fruit, tree, spray, leave, leaf, ppm
Topic 8	cell, tissue, red, fig, stain, cellular, normal, culture, membrane, chromosome

Table 22 (continued)

Topic 9	year, age, museum, work, death, die, old, time, history, life
Topic 10	virus, strain, infection, mosaic, tobacco, particle, inoculation, antibody, passage, test
Topic 11	section, secretary, stain, tissue, officer, science, vice, slide, membership, psychology
Topic 12	school, student, education, teacher, high, teach, course, science, program, mathematic
Topic 13	univ, annual, dept, assoc, ave, inst, new, international, nual, symposium
Topic 14	enzyme, reaction, solution, compound, activity, chem, phosphate, sodium, extract, concentration
Topic 15	appoint, award, engineering, chemistry, company, medal, laboratory, engineer, vice, chief
Topic 16	science, editorial, postage, aaas, wolfe, advertising, lancaster, manuscript, executive, issue,
Topic 17	award, fellowship, grant, program, year, application, foundation, announce, fellow, study
Topic 18	soil, crop, forest, land, fertilizer, erosion, conservation, agricultural, calcium, farmer
Topic 19	food, oil, production, industry, cent, population, product, increase, use, industrial
Topic 20	sample, percent, table, curve, age, value, fig, datum, rate, average
Topic 21	vitamin, liver, rat, nutrition, yeast, deficiency, folic, vita, ascorbic, assay
Topic 22	illus, plate, new, review, univ, chapman, philosophical, textbook, chemistry, scientific
Topic 23	tumor, cancer, tissue, growth, malignant, liver, incidence, normal, transplant, carcinoma
Topic 24	Earth, Light, Solar, Particle, Wave, Satellite, Theory, Star, Space, temperature
Topic 25	nuclear, radiation, power, radioactive, laboratory, physics, reactor, test, effect, isotope
Topic 26	culture, medium, growth, strain, organism, test, bacteria, grow, day, embryo
Topic 27	new, illustrate, museum, history, univ, chapman, publication, philosophical, review, science
Topic 28	protein, amino, nitrogen, fraction, plasma, globulin, peptide, methionine, acid, synthesis
Topic 29	water, foot, rock, fish, ice, surface, deposit, depth, temperature, geological
Topic 30	gene, genetic, specie, chromosome, population, male, sex, mutation, female, evolution

Astrophysics, one of the most talked about topics between 1940-1960, was also caught with the NMF. Biochemistry topic was captured with all three algorithms, but it was observed that the words found with K-means were not as distinctive as in NMF and LDA algorithms.

Table 23

Top Terms Of Astrophysics Topic Between 1940-1960

LDA	NMF
light	earth
star	light
sun	solar
earth	particle
year	wave
observatory	satellite
time	theory
observation	star
solar	spaced
line	temperature

Table 24

Top Terms Of Biochemistry Topic Between 1940-1960

LDA	K-Means	NMF
acid	protein	enzyme
solution	virus	reaction
water	enzyme	solution
cent	reaction	compound
reaction	study	activity
compound	use	chem
protein	tion	phosphate
substance	form	sodium
concentration	result	extract
form	cell	concentration

Topics that are extracted by LDA and their labels between 1960-1980 are in Table 25.

Table 25

Topics and Labels Which Are Extracted Using LDA Between 1960-1980 For Science Dataset

Topic 1 Education	Topic 2 Hardware Technology	Topic 3 Geology	Topic 4 Scientific Research	Topic 5 Chemistry
----------------------	-----------------------------------	--------------------	-----------------------------------	----------------------

Topic 25 (continued)

student university year school education woman social medical child faculty	system service circle reader computer card instrument write information time	sea year fig ocean area sediment depth core deep geol	university american univ soc research washington state science conf annual	acid reaction compound chem chemical acids form group carbon product
Topic 6 Disease	Topic 7 Geochemistry	Topic 8 Metallurgy	Topic 9 Neurophysics	Topic 10 Organic Chemistry
patient drug disease blood study human dose animal normal effect	temperature water pressure rate surface solution heat high flow oxygen	sample rock lunar material percent apollo iron crater grain surface	cell potential nerve response fig muscle fiber membrane physiol neurons	sample water concentration percent air ddt carbon lead table result
Topic 11 Other	Topic 12 Technology	Topic 13 Neuroscience	Topic 14 Astrophysics	Topic 15 Developmental Biology
esect ill irl alld diserent olf ithe arld state tlhe	research science program problem national technology scientific development scientist system	rat activity effect brain cyclic increase concentration amp hour control	earth solar time field fig space source surface region atmosphere	species female male animal bird insect males adult behavior day
Topic 16 Organismal biology	Topic 17 Academic Research	Topic 18 Animal Testing	Topic 19 Scientific Theory	Topic 20 Social Science
gene dna chromosome genetic cell mutant type strain population hybrid	illus university wiley science series academic chemistry symposium study research	response animal subject test day stimulus group eye experiment visual	work science tion theory problem study present man field subject	science research year program national social report scientist state member
Topic 21 Genetic	Topic 22 Nuclear Energy	Topic 23 Electrophysics	Topic 24 Data Science	Topic 25 Healthcare

Topic 25 (continued)

protein acid enzyme rna amino chain activity fig synthesis dna	energy power cost year oil water plant nuclear fuel percent	energy electron field particle state high ion radiation laser atom	data, number time rate tion model result change difference analysis	cell virus mouse serum antigen tumor antibody human day cancer
Topic 26 Plant biology	Topic 27 Cell Culture	Topic 28 Biochemistry	Topic 29 Earth science	Topic 30 Paleontology
plant species food seed growth soil population area forest crop	cell culture medium hour fig percent tissue membrane growth day	structure fig crystal form electron surface ray pattern plane layer	science aaas editorial dec arrange editor news session office washington	site early area date find south year fig stone bone

The most mentioned topics between 1960 and 1980 were Scientific Theory, Political Science, Technology, Animal Testing and Neuroscience.

Table 26

Hot Topics Between 1960-1980 For Science Dataset

Topic Label	Number of Document	Probability
Scientific Theory	3307	0.1485
Social Science	2302	0.1034
Technology	1199	0.0539
Animal Testing	1011	0.0454
Neuroscience	865	0.0389

Clusters extracted with K-means are shown in Table 27, and topics extracted with NMF are shown in Table 28 for 1960-1980.

Table 27

Clusters Between 1960-1980 For Science Dataset

Cluster	Top Terms
Topic 1	activity, fig, percent, day, tion, use, study, effect, result, rat
Topic 2	illus, new, science, study, chemistry, symposium, theory, physics, translate, tion
Topic 3	use, study, tion, effect, time, test, new, drug, result, change
Topic 4	specie, population, plant, male, new, tion, number, large, time, fig
Topic 5	water, surface, temperature, fig, use, tion, sample, high, pressure, low
Topic 6	new, illus, service, science, use, instrument, scientific, time, available, high
Topic 7	science, arrange, new, study, problem, program, social, use, symposium, present
Topic 8	cell, fig, percent, tion, use, culture, activity, study, result, hour
Topic 9	fig, sample, use, time, tion, high, surface, datum, result, temperature
Topic 10	use, new, year, state, tion, cost, time, percent, problem, increase
Topic 11	science, new, year, state, program, tion, use, work, study, time
Topic 12	leader, discussion, study, chemistry, vice, tion, structure, cell, effect, reaction
Topic 13	fig, time, temperature, surface, tion, high, use, large, earth, line
Topic 14	cell, antigen, protein, membrane, antibody, fig, tion, form, virus, specific
Topic 15	new, service, science, use, illus, information, time, state, scientific, instrument
Topic 16	field, magnetic, solar, plasma, fig, time, earth, line, region, fusion
Topic 17	gene, cell, protein, genetic, mutant, tion, result, synthesis, enzyme, fig
Topic 18	science, associate, assistant, section, state, year, fellow, aaas, term, health
Topic 19	reaction, fig, state, tion, atom, enzyme, form, structure, ion, use
Topic 20	population, rate, birth, family, woman, percent, year, age, tion, fertility
Topic 21	illus, new, science, study, symposium, chemistry, tion, theory, translate, academic
Topic 22	preside, science, state, program, section, session, new, school, tion, versity
Topic 23	new, univ, assoc, dept, annual, inst, ave, science, Congr, state
Topic 24	protein, chain, amino, cell, fig, bind, sequence, tion, pro, study
Topic 25	science, new, use, tion, study, state, year, work, time, research

Topic 27 (continued)

Topic 26	student, school, year, faculty, science, education, new, medical, time, study
Topic 27	cell, culture, virus, fig, percent, tion, use, hour, study, result
Topic 28	science, scientific, new, scientist, tion, social, program, technology, year, state
Topic 29	chain, light, sequence, heavy, region, antibody, protein, residue, gene, human
Topic 30	fig, year, time, site, age, tion, early, use, large, date

Table 28

Topics Which Are Extracted Using NMF Between 1960-1980 For Science Dataset

Topic	Top Terms
Topic 1	scientist, science, scientific, social, political, policy, country, public, people, technology
Topic 2	cell, culture, medium, growth, hour, tissue, fig, lymphocyte, tumor, percent
Topic 3	illus, new, symposium, translate, chemistry, interscience, plate, prentice, academic, science
Topic 4	univ, assoc, dept, inst, ave, annual, new, Congr, engineer, international
Topic 5	rat, brain, day, injection, hour, lesion, serotonin, effect, weight, food
Topic 6	theory, work, field, review, physics, problem, topic, discussion, text, subject
Topic 7	stimulus, visual, eye, trial, light, stimulation, test, subject, msec, evoke
Topic 8	lunar, rock, sample, crater, soil, grain, surface, glass, iron, plagioclase
Topic 9	chromosome, gene, genetic, mutant, population, hybrid, strain, mutation, human, frequency
Topic 10	editorial, advertising, manager, phone, science, editor, sale, ave, assistant, view
Topic 11	virus, tumor, viral, infection, cancer, leukemia, particle, strain, culture, human
Topic 12	student, faculty, education, teach, campus, school, institution, year, course, educational
Topic 13	protein, chain, amino, peptide, hemoglobin, residue, sequence, synthesis, ribosome, bind
Topic 14	enzyme, activity, liver, phosphate, fraction, percent, extract, minute, concentration, hour
Topic 15	sediment, age, year, fig, earthquake, ridge, date, fault, site, depth
Topic 16	plant, specie, soil, food, seed, population, crop, growth, forest, leaf
Topic 17	health, medical, care, patient, service, hospital, physician, cancer, hew, program

Topic 28 (continued)

Topic 18	male, female, sex, behavior, specie, pheromone, sexual, adult, egg, mate
Topic 19	school, medicine, state, medical, emeritus, versity, engineering, education, vice, city
Topic 20	program, federal, budget, year, agency, government, policy, fiscal, science, support
Topic 21	membrane, potential, nerve, lipid, ion, axon, current, fig, electrode, solution
Topic 22	antigen, antibody, anti, lymphocyte, tumor, rabbit, patient, immune, graft, donor
Topic 23	water, temperature, ice, surface, pressure, sample, air, fish, vapor, concentration
Topic 24	atom, ion, reaction, hydrogen, temperature, gas, bond, oxygen, structure, molecule
Topic 25	fuel, power, oil, cost, nuclear, industry, use, system, price, gas
Topic 26	drug, lsd, effect, use, patient, methadone, heroin, ddt, abuse, agent
Topic 27	woman, child, family, sex, birth, percent, population, age, fertility, career
Topic 28	sleep, brain, stage, deprivation, hour, period, wake, wave, night, wakefulness
Topic 29	cyclic, cyclase, hormone, nucleotide, kinase, adenylyl, phosphodiesterase, theophylline, receptor, adenylylate
Topic 30	aaas, science, session, arrange, program, section, symposium, annual, member, public

Technology, which was one of the most talked about topics between 1960-1980, was also captured with k-means and NMF, and similar words were obtained in all three algorithms.

Table 29

Top Terms Of Technology Topic Between 1960-1980

LDA	K-Means	NMF
research	science	scientist
science	scientific	science
program	new	scientific
problem	scientist	social
national	tion	political
technology	social	scientific
scientific	program	country
development	technology	public
scientist	year	people
system	state	technology

Topics that are extracted by LDA and their labels between 1980-2000 are in Table 30.

Table 30

Topics And Labels Which Are Extracted Using LDA Between 1980-2000 For Science Dataset

Topic 1 Geophysics	Topic 2 Education	Topic 3 Plant Biology	Topic 4 Healthcare	Topic 5 Energy Systems
fig rock area year ridge earthquake geol earth north age	illus science wiley university series academic chemistry symposium study chairman	plant growth seed soil grow leave light leaf nitrogen bacteria	cell virus mouse serum antigen tumor antibody culture day human	energy power nuclear oil cost year plant air fuel water
Topic 6 Social Science	Topic 7 Electromagnetic	Topic 8 Chemistry	Topic 9 Paleontology	Topic 10 Geochemistry
science man social work scientific time scientist human life fact	energy electron field particle high state ray laser radiation fig	reaction structure compound chem acid bond form group atom ion	species animal primate group mammal evolution large fossil body form	temperature water pressure rate fig surface concentration high solution increase
Topic 11 Molecular Biology	Topic 12 Scientific Research	Topic 13 Hardware Technology	Topic 14 Astrophysics	Topic 15 Developmental Biology
protein activity enzyme acid percent fraction concentration minute fig hour	university american univ soc science research washington state annual conf	service circle system reader science card instrument research write computer	sample rock lunar material percent iron grain surface ratio apollo	species female insect male population males bird ddt sex adult

Table 30 (continued)

Topic 16 Disease	Topic 17 Earth science	Topic 18 Genetic	Topic 19 Scientific News	Topic 20 Neurophysics
drug patient disease blood study dose effect human animal normal	water sea ocean year lake marine sample fish sediment area	dna rna gene protein chain chromosome cell acid genetic synthesis	research science university year program national scientist state member report	cell potential nerve response fiber fig muscle physiol membrane neurons
Topic 21 Technology	Topic 22 Animal Testing	Topic 23 Medicine	Topic 24 Cell Culture	Topic 25 Education
research science problem technology country system program development united world	animal day group rat test response behavior hour control sleep	rat cyclic effect activity amp increase brain hormone concentration control	cell fig membrane culture tissue electron hour medium section surface	student woman year population child percent study school age number
Topic 26 Geography	Topic 27 Scientific Theory	Topic 28 Geology	Topic 29 Optic	Topic 30 Scientific Writing
earth solar time data space source fig atmosphere surface field	tion study work theory system problem present field research information	area site date early year forest south north find stone	fig light eye response visual subject time tion stimulus pattern	science editorial editor view esect news ill include comment washington

The most mentioned topics between 1980-2000 were Scientific News, Scientific Theory, Social Science, Molecular Biology, Disease.

Table 31

Hot Topics Between 1980-2000 For Science Dataset

Topic Label	Number of Document	Probability
Scientific News	2866	0.1287
Scientific Theory	2234	0.1003
Social Science	1376	0.0618
Molecular Biology	1035	0.0465
Disease	879	0.0395

Clusters extracted with K-means are shown in Table 32, and topics extracted with NMF are shown in Table 33 in 1980-2000.

Table 32

Clusters Between 1980-2000 For Science Dataset

Cluster	Top Terms
Topic 1	gene, cell, sequence, fig, use, site, protein, region, contain, human
Topic 2	use, science, new, year, high, study, time, datum, state, increase
Topic 3	fig, use, result, time, high, tion, datum, surface, science, sample
Topic 4	position, experience, molecular, program, opportunity, application, candidate, send, cell, include
Topic 5	leader, discussion, cell, structure, protein, molecular, tion, gene, study, new
Topic 6	cell, gene, protein, tumor, expression, ibid, fig, region, specific, type
Topic 7	position, molecular, experience, program, opportunity, application, cell, science, send, candidate
Topic 8	residue, structure, bind, protein, helix, fig, site, chain, peptide, domain
Topic 9	cell, fig, protein, gene, line, result, expression, use, bind, human
Topic 10	fault, earthquake, fig, slip, event, stress, plate, large, year, time
Topic 11	position, molecular, experience, application, opportunity, program, send, cell, candidate, include
Topic 12	reaction, chem, rate, state, fig, use, enzyme, product, bind, bond
Topic 13	protein, cell, bind, membrane, fig, sequence, biol, gene, site, domain
Topic 14	cell, fig, protein, use, result, current, effect, activity, receptor, tion

Table 32 (continued)

Topic 15	male, female, mutant, courtship, gene, fly, sex, court, genetic, specific
Topic 16	sequence, protein, cell, fig, gene, clone, region, amino, use, human
Topic 17	receptor, cell, bind, protein, fig, sequence, domain, type, gene, contain
Topic 18	science, year, new, use, work, study, time, research, tion, state
Topic 19	hiv, cell, virus, infection, viral, culture, fig, science, human, study
Topic 20	protein, cell, fig, bind, sequence, activity, contain, gene, amino, peptide
Topic 21	year, fig, time, datum, water, change, age, rate, temperature, ice
Topic 22	position, experience, molecular, application, program, opportunity, candidate, send, equal, include
Topic 23	cell, protein, antigen, receptor, fig, gene, type, result, ibid, induce
Topic 24	surface, fig, phy, state, structure, use, atom, phase, high, temperature
Topic 25	gene, cell, expression, sequence, fig, protein, plant, use, region, site
Topic 26	illus, new, science, symposium, physics, study, diego, theory, plate, scientific
Topic 27	fig, time, image, datum, field, high, large, solar, use, region
Topic 28	chromosome, gene, genetic, sequence, marker, human, clone, use, cell, region
Topic 29	bind, protein, site, fig, cell, sequence, contain, gene, transcription, use
Topic 30	splice, site, fig, mrna, protein, pre, reaction, sequence, bind, cell

Table 33

Topics Which Are Extracted Using NMF Between 1980-2000 For Science Dataset

Topic	Top Terms
Topic 1	cell, antigen, fig, anti, antibody, expression, culture, hour, line, growth
Topic 2	work, scientific, scientist, science, social, theory, history, problem, people, researcher
Topic 3	temperature, phase, fig, pressure, wave, magnetic, field, phy, particle, high
Topic 4	protein, kinase, membrane, amino, domain, yeast, biol, fusion, sequence, subunit
Topic 5	vitae, experience, employer, position, curriculum, opportunity, candidate, molecular, application, send
Topic 6	gene, chromosome, genetic, expression, mutation, cell, mutant, embryo, strain, genome
Topic 7	hiv, infection, drug, viral, infect, trial, patient, blood, cell, immune

Table 33 (continued)

Topic 8	plant, specie, forest, population, crop, soil, seed, leaf, tree, food
Topic 9	aaas, science, section, member, annual, scientific, executive, electorate, technology, engineering
Topic 10	ice, climate, warm, ozone, temperature, water, global, year, change, greenhouse
Topic 11	galaxy, universe, cluster, light, cosmic, hole, matter, gas, radio, object
Topic 12	brain, neuron, rat, visual, nerve, stimulus, neuronal, cortical, synaptic, sensory
Topic 13	manager, advertising, editorial, sale, recruitment, reprint, staff, science, traffic, correspondent
Topic 14	cancer, patient, drug, tumor, risk, trial, breast, health, clinical, therapy
Topic 15	peptide, residue, helix, bind, chain, structure, amino, domain, loop, sequence
Topic 16	rock, age, water, year, earth, sample, impact, sediment, fossil, crust
Topic 17	earthquake, fault, quake, seismic, slip, stress, event, parkfield, plate, earth
Topic 18	receptor, bind, kinase, tyrosine, egf, domain, insulin, nmda, hormone, affinity
Topic 19	budget, year, program, science, agency, funding, federal, fund, cut, plan
Topic 20	bind, site, fig, transcription, promoter, mutant, activity, splice, lane, wild
Topic 21	company, industry, patent, technology, government, japanese, market, drug, country, policy
Topic 22	doe, fusion, nuclear, ssc, weapon, lab, physics, accelerator, physicist, particle
Topic 23	woman, male, female, sex, child, gender, career, social, sexual, health
Topic 24	atom, bond, chem, reaction, structure, metal, ion, surface, cluster, molecule
Topic 25	student, education, school, science, teach, faculty, program, teacher, minority, engineering
Topic 26	current, voltage, membrane, fig, potential, conductance, ion, solution, subunit, calcium
Topic 27	virus, viral, infection, antibody, lav, patient, infect, immune, antigen, strain
Topic 28	sequence, clone, probe, human, chromosome, pcr, genome, region, fig, primer
Topic 29	image, surface, stm, fig, optical, light, afm, feature, use, sample
Topic 30	space, shuttle, station, launch, mission, satellite, solar, flight, rocket, earth

Disease, which was one of the most talked about topics between 1980-2000, was also caught with K-means and NMF.

Table 34

Top Terms of Disease Topic between 1980-2000

LDA	K-Means	NMF
drug patient disease blood study dose effect human animal normal	hiv cell virus infection viral culture fig science human study	hiv infection drug viral infect trial patient blood cell immune

Topics that are extracted by LDA and their labels between 2000-2020 are in Table 35.

Table 35

Topics and Labels Which Are Extracted Using LDA Between 2000-2020 For Science Dataset

Topic 1 Metallurgy	Topic 2 Ecology	Topic 3 Education	Topic 4 Plant Biology	Topic 5 Medical Genetics
surface material structure science layer crystal film fig image temperature	bird animal body fish wing insect sound fly eye light	science social human student www university, people usa scientific study	species plant forest ecosystem area soil fish science population conservation	protein bind structure domain site fig complex dna rna helix
Topic 6 Astrophysics	Topic 7 Scientific Writing	Topic 8 Biochemistry	Topic 9 Organismal Biology	Topic 10 Geology
star solar , mass space ray galaxy astrophys planet energy field	univ science www fax aas editor research news manager cell	cell protein fig signal bind receptor membrane biol activity kinase	gene fig cell plant expression protein rna mutant type transcription	ice change climate ocean year sea temperature water science record

Table 35 (continued)

Topic 11 Electrophysics	Topic 12 Social Science	Topic 13 Disease	Topic 14 Molecular Biology	Topic 15 Organismal Biology
phys state energy field electron quantum rev spin atom lett	science research national year scientist university program government social week	mouse cancer disease patient cell tumor drug study gene age	university chair research molecular biology card circle vice development experience	cell fig stem development expression actin biol type embryo embryos
Topic 16 Nuclear physics	Topic 17 Data Science	Topic 18 Neurophysics	Topic 19 Academic Research	Topic 20 Technology
energy water fuel emission cost gas nuclear carbon world power	fig data time rate model support result line increase number	neurons brain response activity neuron cell cortex neurosci control synaptic	research university position biology application candidate program experience opportunity science	protein www drug technology research company cell data gene dna
Topic 21 Research Methodology	Topic 22 Scientific Research	Topic 23 World History	Topic 24 Cell Biology	Topic 25 Geophysics
system model science tion process time study nature change provide	year researcher university work find team time scientist study lab	human year site date africa early university ago ancient african	cell mouse fig response immune antigen infection receptor science type	earth mantle science rock surface earthquake geophys mars fault crater
Topic 26 Healthcare	Topic 27 Developmental Biology	Topic 28 Paleontology	Topic 29 Genetic	Topic 30 Chemistry
virus hiv disease vaccine infection health drug aids human infect	population female species male genetic selection evolution males sex evolutionary	fossil tree early evolution species science group taxa bone million	gene genome sequence dna human protein chromosome region sequences genetic	chem reaction bond water chemical chemistry oxygen ion product metal

The most mentioned topics between 2000-2020 were Social Science, Scientific Research, Education, Metallurgy and Electrophysics.

Table 36

Hot Topics Between 2000-2020 For Science Dataset

Topic Label	Number of Document	Probability
Science News	3283	0.1419
Scientific Research	2598	0.1123
Education	1708	0.0738
Metallurgy	990	0.0428
Electrophysics	967	0.0418

Clusters extracted with K-means are shown in Table 37, and topics extracted with NMF are shown in Table 38 in 2000-2020.

Table 37

Clusters Between 2000-2020 For Science Dataset

Cluster	Top Terms
Topic 1	specie, science, population, fig, use, year, datum, study, time, rate
Topic 2	position, application, candidate, program, science, faculty, include, molecular, opportunity, teach
Topic 3	chair, cell, new, vice, science, position, application, molecular, program, candidate
Topic 4	fig, neuron, activity, brain, science, time, use, effect, stimulus, cell
Topic 5	position, application, program, molecular, candidate, experience, include, cell, teach, science
Topic 6	position, application, program, candidate, science, include, faculty, experience, molecular, opportunity
Topic 7	position, application, candidate, program, faculty, teach, include, molecular, opportunity, tenure
Topic 8	gene, genome, sequence, protein, fig, cell, expression, science, use, human
Topic 9	position, application, program, candidate, science, experience, molecular, include, cell, opportunity
Topic 10	position, experience, molecular, candidate, application, program, cell, include, opportunity, science
Topic 11	change, ice, year, fig, climate, water, science, time, increase, datum
Topic 12	state, fig, phy, field, time, use, science, spin, atom, phase

Table 37 (continued)

Topic 13	univ, science, aaas, cell, use, com, new, protein, manager, technology
Topic 14	position, application, candidate, program, science, experience, include, opportunity, molecular, center
Topic 15	fig, science, use, surface, high, time, low, large, structure, result
Topic 16	position, application, science, candidate, program, include, experience, center, cell, opportunity
Topic 17	fig, cell, protein, type, gene, science, expression, mutant, wild, bind
Topic 18	position, application, candidate, program, molecular, science, experience, teach, include, opportunity
Topic 19	cell, protein, science, gene, use, new, study, bind, human, tion
Topic 20	cell, fig, protein, science, gene, expression, use, type, result, bind
Topic 21	gene, sequence, genome, fig, science, cell, protein, human, use, genetic
Topic 22	gene, genome, sequence, human, protein, number, assembly, family, region, chromosome
Topic 23	position, molecular, experience, program, science, candidate, cell, application, include, mail
Topic 24	bind, fig, domain, structure, site, protein, residue, helix, form, science
Topic 25	cell, fig, protein, science, stem, type, gene, expression, use, tissue
Topic 26	science, year, new, use, study, time, work, research, tion, researcher
Topic 27	position, experience, molecular, candidate, application, cell, program, include, science, opportunity
Topic 28	science, univ, com, use, cell, new, protein, gene, service, aaas
Topic 29	experience, position, molecular, cell, candidate, scientist, program, include, development, science
Topic 30	protein, cell, fig, bind, science, gene, domain, use, biol, interaction

Table 38

Topics Which Are Extracted Using NMF Between 2000-2020 For Science Dataset

Cluster	Top Terms
Topic 1	budget, science, plan, week, agency, year, government, scientist, say, panel
Topic 2	cell, stem, embryo, antigen, tissue, expression, clone, immune, receptor, gene
Topic 3	position, candidate, application, vitae, experience, curriculum, employer, faculty, opportunity, program
Topic 4	gene, genome, sequence, chromosome, genetic, human, expression, protein, genomic, transcription
Topic 5	univ, aaas, science, manager, sale, pcr, com, service, senior, inst

Table 38 (continued)

Topic 6	age, sample, isotope, sediment, earth, record, rock, iron, datum, year
Topic 7	protein, bind, domain, membrane, residue, receptor, structure, kinase, interaction, biol
Topic 8	researcher, say, year, team, new, people, work, find, scientist, lab
Topic 9	drug, trial, patient, clinical, health, company, risk, patent, child, therapy
Topic 10	specie, population, diversity, extinction, bird, conservation, biodiversity, fish, ecosystem, ecology
Topic 11	galaxy, universe, hole, dark, cluster, matter, black, light, dwarf, disk
Topic 12	brain, neuron, memory, stimulus, synaptic, activity, receptor, visual, cortical, neuronal
Topic 13	climate, warm, change, global, temperature, greenhouse, atmospheric, record, sst, year
Topic 14	plant, arabidopsis, crop, seed, root, soil, rice, leaf, mutant, pathogen
Topic 15	ice, sheet, greenland, melt, shelf, year, polar, glacial, record, warm
Topic 16	earthquake, fault, slip, seismic, quake, wave, geography, plate, stress, earth
Topic 17	student, education, school, teach, science, teacher, program, faculty, learn, woman
Topic 18	virus, flu, pandemic, infection, outbreak, strain, viral, human, avian, bird
Topic 19	science, letter, scientific, social, public, human, policy, issue, health, life
Topic 20	solar, planet, space, earth, comet, mission, asteroid, disk, planetary, dust
Topic 21	coral, fish, fishery, reef, fishing, ecosystem, protect, decline, calcification, spawn
Topic 22	hiv, infection, country, epidemic, prevalence, people, infect, health, immune, viral
Topic 23	male, female, woman, sex, chromosome, sperm, egg, sexual, mate, population
Topic 24	water, surface, deep, flow, soil, liquid, fish, sediment, depth, vapor
Topic 25	chem, reaction, bond, atom, metal, structure, molecule, catalyst, surface, chemistry
Topic 26	phy, spin, field, magnetic, state, phase, temperature, atom, wave, frequency
Topic 27	cancer, tumor, mutation, patient, breast, gene, risk, growth, lung, therapy
Topic 28	fuel, emission, nuclear, power, hydrogen, gas, cost, weapon, oil, country
Topic 29	forest, tree, land, soil, tropical, deforestation, fire, ecosystem, amazon, timber
Topic 30	particle, physicist, physics, lhc, accelerator, cern, detector, matter, collider, neutrino

Additionally, topics of Science Magazine within 10 years was also extracted with LDA. Education, Human Science, Geology, Electricity, Ecology topics were the most discussed topics between 1880-1890. History, Geology, Research Methodology, Science News, Chemistry topics were the most discussed topics between 1890-1900. Education, Science Writing, Geology, Science History, Scientific News topics were the most discussed topics between 1900-1910. Education, Scientific News, Scientific Research, Scientific News, Human Science, Astrophysics topics were the most discussed topics between 1910-1920. Education, Research Methodology, Science History, Scientific News, Human Science Astrophysics topics were the most discussed topics between 1920-1930. Medicine, Analytical Chemistry, engineering, Human Science, Geography Astrophysics topics were the most discussed topics between 1930-1940. Medicine, chemistry, Social Science, Genetic, Science History Astrophysics topics were the most discussed topics between 1940-1950. Scientific Theory, Education, Medicine, Energy Systems, Scientific Research topics were the most discussed topics between 1950-1960. Scientific Theory, Scientific News, Genetic, Disease, Astrophysics topics were the most discussed topics between 1960-1970. Scientific Theory, Scientific News, Technology, Social Science, Organismal Biology topics were the most discussed topics between 1970-1980. Scientific Research, Scientific Theory, Genetic, Energy Systems, Healthcare topics were the most discussed topics between 1980-1990. Scientific News, Academic Research, Electrophysics, Disease, Astrophysics topics were the most discussed topics between 1990-2000.

4.2.2 British journal of the history of science dataset results. Documents were collected from the BJHS from 1960 to 2020. The documents were divided into 10-year parts and the results were analyzed. LDA were used.

For BJHS, the number of topics was chosen as 20, since it was observed that the number of topics was tested from 10 to 100 in 20-year interval, and it was observed that the topics were best separated. The topics, which were extracted in 20-year intervals, were manually labeled in 4 main categories and divided into 27 sub-categories, and the changes in the topics were monitored over the years. Manually determined categories for BJHS are given in table 39.

Topics which are extracted by LDA and their labels between the years of 1960-1980 are in Table 40 and also hot topics for these years are showed in Table 41.

Table 39

BJHS Dataset Categories

Biological Science	Pyhsical Science	Social Science	Scientific Method
Medicine	Environmental	Philosophy	Scientific Writing
Healthcare	Technology	Education	Science history
Evolution	Geology	Social science	Scientific news
Biological	Physics	Culture	Scientific research
Paleontology	Chemistry	Human science	
Biochemistry	Engineering		
Disease	Nature		
	Mathematical		
	Electromagnetic		
	Industry		
	Astrophysics		

Table 40

Topics and Labels Which Are Extracted Using LDA between 1960-1980 for BJHS Dataset

Topic 1 Philosophy	Topic 2 Environmental	Topic 3 Scientific Research	Topic 4 Other	Topic 5 Other
view nature claim theory fact world idea philosophy natural understand	history water environment colonial environmental india forest land state south	research war science scientist nuclear national american scientific international work	ref hist sci isbn science history index bibl ill century	del los las una spanish spain por para con madrid
Topic 6 Medicine	Topic 7 Engineering	Topic 8 Scientific Writing	Topic 9 Healthcare	Topic 10 Science History

Table 40 (continued)

medical health disease medicine patient hospital woman child case doctor	technology engineer design engineering work system company industrial industry iron	biology university index species isbn human ill biological provide include	drug patent trial industry pharmaceutical company american health test pharmacy	newton early text work century body bacon translation boyle motion
Topic 11 Geology	Topic 12 Physics	Topic 13 Chemistry	Topic 14 Other	Topic 15 Education
geological earth geology rock sea survey expedition figure map weather	theory model case quantum time system physics physical law problem	research chemistry cell chemical study work laboratory molecular university structure	die den das eine zur als auf ein sich dem	year royal work time university letter british student member school
Topic 16 Technology	Topic 17 Other	Topic 18 Evolution	Topic 19 Biological	Topic 20 Social Science
science social research study technology scientific knowledge work public practice	paris dan une sur par est pour france qui french	darwin natural museum species history animal human collection specimen naturalist	biology selection genetics population genetic organism biological evolution theory gene	science history century scientific university study work modern social knowledge

Table 41

Hot Topics Between 1960-1980 For BJHS Dataset

Topic	Number of Document	Probability
Physics	427	0.0963
Technology	396	0.0893
Education	375	0.0846
Medicine	316	0.0713
Philosophy	310	0.0699

Topics which are extracted by LDA and their labels between the years of 1980-2000 are in Table 42 and also hot topics for these years are showed in Table 43.

Table 42

Topics and Labels Which Extracted Using LDA Between 1980-2000 For BJHS Dataset

Topic 1 Engineering	Topic 2 Electromagnetic	Topic 3 Medicine	Topic 4 Other	Topic 5 Technology
engineer engineering design patent work technology company iron industry production	theory physics experiment field force ray energy electron magnetic heat	history science review century american woman university medicine historical medical	hist sci index univ illus bibl science rev med phil	science social scientific study knowledge scientist research technology work practice
Topic 6 Nature	Topic 7 Paleontology	Topic 8 Mathematical	Topic 9 Evolution	Topic 10 Biochemistry
man work life time great write nature idea science view	work museum year letter time history south collection island natural	mathematical mathematics motion number work method observation set theory time	darwin species evolution natural selection theory biology animal evolutionary huxley	soc research cell work chem study molecular virus reaction chemical
Topic 11 Geology	Topic 12 Physics	Topic 13 Science History	Topic 14 Other	Topic 15 Industry
rock geology earth geological age figure geologist fig geol area	theory model case quantum time system physics physical law problem	newton century boyle lavoisier natural early text work experiment royal	die den das eine zur als auf ein sich dem	people industry government change country year development world system policy
Topic 16 Education	Topic 17 Scientific News	Topic 18 Disease	Topic 19 Other	Topic 20 Scientific Research

Table 42 (continued)

royal	history	medical	dan	research
year	technology	medicine	une	university
sir	illustration	disease	est	science
art	bibliography	health	par	laboratory
council	review	patient	sur	war
member	american	hospital	qui	work
university	industry	physicians	pour	scientific
lecture	tion	drug	paris	year
work	univ	clinical	pas	scientist
design	industrial	case	son	american

Table 43

Hot Topics Between 1980-2000 For BJHS Dataset

Topic	Number of Document	Probability
physical	483	0.1135
technology	396	0.0930
nature	362	0.0851
education	321	0.0754
industry	293	0.0688

Topics which are extracted by LDA and their labels between the years of 2000-2020 are in Table 44 and also hot topics for these years are showed in Table 45.

Table 44

Topics and Labels Which Are Extracted Using LDA Between 2000-2020 For BJHS Dataset

Topic 1 Philosophy	Topic 2 Physics	Topic 3 Medicine	Topic 4 Culture	Topic 5 Evolution
theory science problem term case philosophy fact time view system	theory force heat light particle physics electron maxwell thomson motion	medical medicine history american, pharmacy healt physicians hospital philadelphia university	technology history illustration bibliography review american museum culture century tion	darwin species natural animal man nature theory lyell evolution selection

Table 44 (continued)

Topic 6 Technology	Topic 7 Social Science	Topic 8 Science History	Topic 9 Scientific Research	Topic 10 Scientific News
science university scientific research social scientist study technology history work	soc work research lond year proc university study war time	newton motion body galileo line law hooke time force equal	history review university century van del sciences russian summary medicine	sir council art royal examination award fellow design member year
Topic 11 Astrophysics	Topic 12 Education	Topic 13 Human science	Topic 14 Chemistry	Topic 15 Geology
table earth star india sun planet year indian observation time	year country industry people time great work problem education good	work man letter year write royal time great early century	chemical chemistry chem gas chemist acid work element air substance	rock sea north area fig geology geol deposit island geological
Topic 16 Engineering	Topic 17 Other	Topic 18 Other	Topic 19 Disease	Topic 20 Other
water design work air engine power glass material machine engineer	die den das zur berlin eine german auf leipzig ist	paris sur dan une par est qui sciences pour histoire	disease blood medical physiology animal case research med effect study	hist science sci illus bibl rev index univ med history

Table 45

Hot Topics Between 2000-2020 For BJHS Dataset

Topic	Number of Document	Probability
Human science	314	0.1341
Education	267	0.1141
Philosophy	265	0.1132
Technology	196	0.0837
Scientific News	188	0.0803

Chapter 5

Results And Discussion

In this chapter, the change of topics over the years has been examined and interpreted. The agriculture topic was discussed only between 1900-1940. The topic of animal testing was only discussed between the years 1960-1980. Ecology was discussed frequently until the 1960s, and then continued to be spoken in 2000 and later. While Evolution was one of the most discussed topics between 1880-1900, it lost its popularity later than. Mathematics, one of the important basic sciences, was discussed frequently until the 1960s, but then lost its popularity. The topic of zoology was discussed only between 1880-1900. Technology topic has been one of the most discussed topics since 1960. The topic of healthcare and genetics has been one of the most talked about topics since 1960. The topic of molecular biology has been one of the most discussed topics since 1980. Hardware Technology was talked about frequently between 1960 and 2000.

For the Science dataset, one of the most discussed topics in the every year range was seen as social science. The change in the words of this topic over the years can be seen in Table 46. Disease topic is also one of the popular topics every year range, and the change of words is shown in 47.

Table 46

The Change Of Social Science Topic's Words Over The Years

1880 1900	1900 1920	1920 1940	1940 1960	1960 1980	1980 2000	2000 2020
state work year country report united government survey public states	state public institution government national work service tion board main	research scientific national science work year american state council congress	research national work scientific state service public government council fund	science research year program national university report scientist state member	research science university year program national scientist state member report	science research national year scientist university program government million week

Table 47

The Change Of Disease Topic's Words Over The Years

1880 1900	1900 1920	1920 1940	1940 1960	1960 1980	1980 2000	2000 2020
disease case health fever death water time animal medical result	disease bacteria animal case blood organism culture find fever day	disease animal virus health infection fever case blood medical day	virus culture animal disease bacteria organism day serum infection medium	patient drug disease blood study human dose animal normal effect	drug patient disease blood study dose effect human animal normal	mouse cancer disease patient cell tumor drug study gene age

Table 48

Change Of Chemistry Topic's Words Over The Years

1880 1900	1900 1920	1920 1940	1940 1960	1960 1980	1980 2000	2000 2020
gas acid chemical solution water form substance liquid metal chemistry	acid solution water chemical method salt cent compound substance action	acid solution water chemical method salt cent compound substance action	chemical chemistry industry product research chemist oil year work material	research chemistry cell chemical study work laboratory molecular university structure	reaction structure compound chem acid bond form group atom ion	chem reaction bond water chemical chemistry oxygen ion product metal

Table 49

Change Of Geology Topic's Words Over The Years

1880 1900	1900 1920	1920 1940	1940 1960	1960 1980	1980 2000	2000 2020
foot water inch place large small end stone surface iron	rock foot valley river deposit water form lake bed find	water rock foot area surface sea earth river great region	rock earth area geological valley foot, surface deposit river time	sea year fig ocean area sediment depth core deep geol	area site date early year forest south north find stone	ice change climate ocean year sea temperature water science record

Table 50

Change Of Plant Biology Topic's Words Over The Years

1880 1900	1900 1920	1920 1940	1940 1960	1960 1980	1980 2000	2000 2020
plant experiment year soil food cent station milk seed result	plant species tree botanical find form leave flower flora grow	plant soil tree seed root grow growth fruit leave find	plant soil seed tree grow crop root growth insect fruit	plant species food seed growth soil population area forest crop	plant growth seed soil grow leave light leaf nitrogen bacteria	species plant forest ecosystem area soil fish science population conservation

The changing of the probabilities of the topics over the years is shown in Figure 5. In Figure 5, the possibilities of plant biology, disease, paleontology topics which are evaluated under biological science are shown. While it was observed that the popularity of the topic of Disease increased in 1920-1940, it was observed that it decreased towards the 2000s. It was observed that the Paleontology topic showed a similar trend with the disease topic, and its probabilities increased between 1920-1940. Plant biology, on the other hand, showed the opposite trend, its probabilities decreased between 1920 and 1940, and increased between 2000 and 2020.

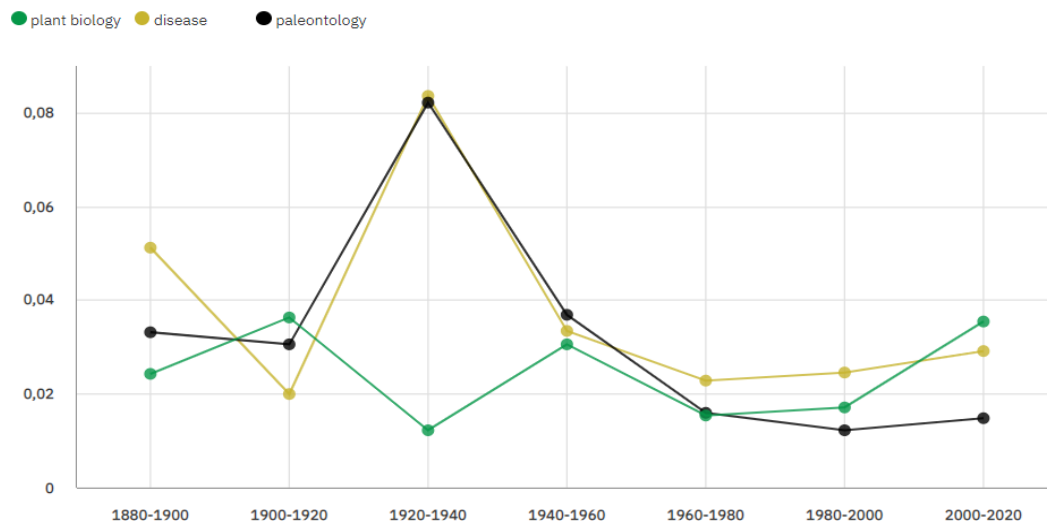


Figure 5. Probabilities of plant biology, disease, paleontology over the years

The probabilities of the astrophysics, chemistry and geology topics which are evaluated under Physical Science are shown in the figure below. It was seen that these 3 topics were mostly examined between the years 1940-1960, while their popularity decreased towards the 2000s, it was seen that it increased again in the years 2000-2020.

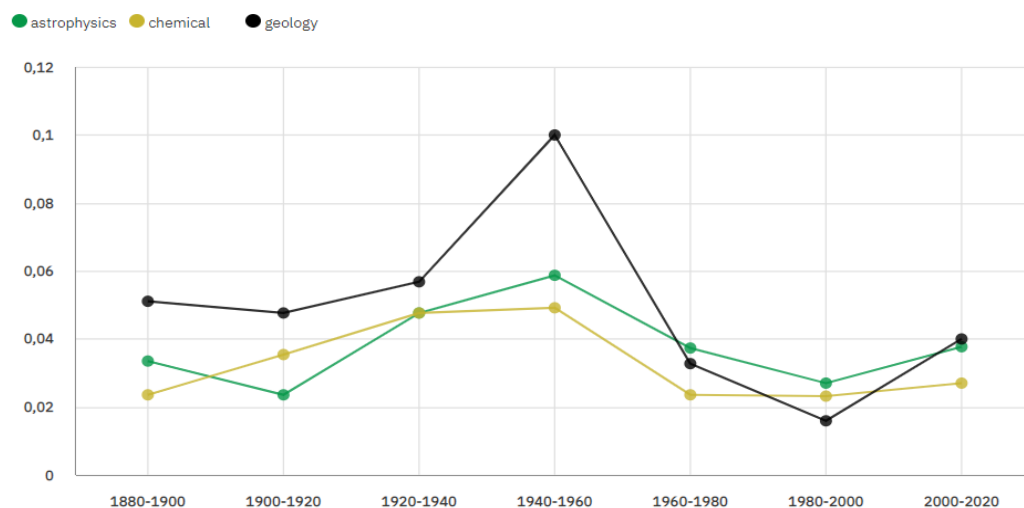


Figure 6. Probabilities of astrophysics, chemistry and geology over the years

Technology, physics, geology, medicine, evolution, science history, education topics were discussed in the BJHS dataset every year. The topics of Astrophysics and disease were discussed only between 2000-2020. While the industry topic was discussed between 1980-2000, the engineering topic was discussed between 2000-

2020. In the BJHS dataset, the changes in the words of the 3 most talked about topics over the years are shown in the Table 51, Table 52 and Table 53.

Table 51

Change Of Medicine Topic's Words Over The Years

1960 1980	1980 2000	2000 2020
medical health disease medicine patient hospital woman child case doctor	history science review century american woman university medicine historical medical	medical medicine history american, pharmacy healt physicians hospital philadelphia university

Table 52

Change Of Technology Topic's Words Over The Years

1960 1980	1980 2000	2000 2020
science social research study technology scientific knowledge work public practice	science social scientific study knowledge scientist research technology work practice	science university scientific research social scientist study technology history work

Table 53

Change Of Physics Topic's Words Over The Years

1960 1980	1980 2000	2000 2020
theory model case quantum time system physics physical law problem	theory model case quantum time system physics physical law problem	theory force heat light particle physics electron maxwell thomson motion

The popularity of the technology topic continued to increase until 2000. Physics, one of the important basic sciences, was mostly discussed between the years 1960-1980. The Evolution topic was most talked about between 1980-2000. Medicine and geology topics were mostly discussed between the years 1960-1980, and then their probabilities being mentioned decreased.

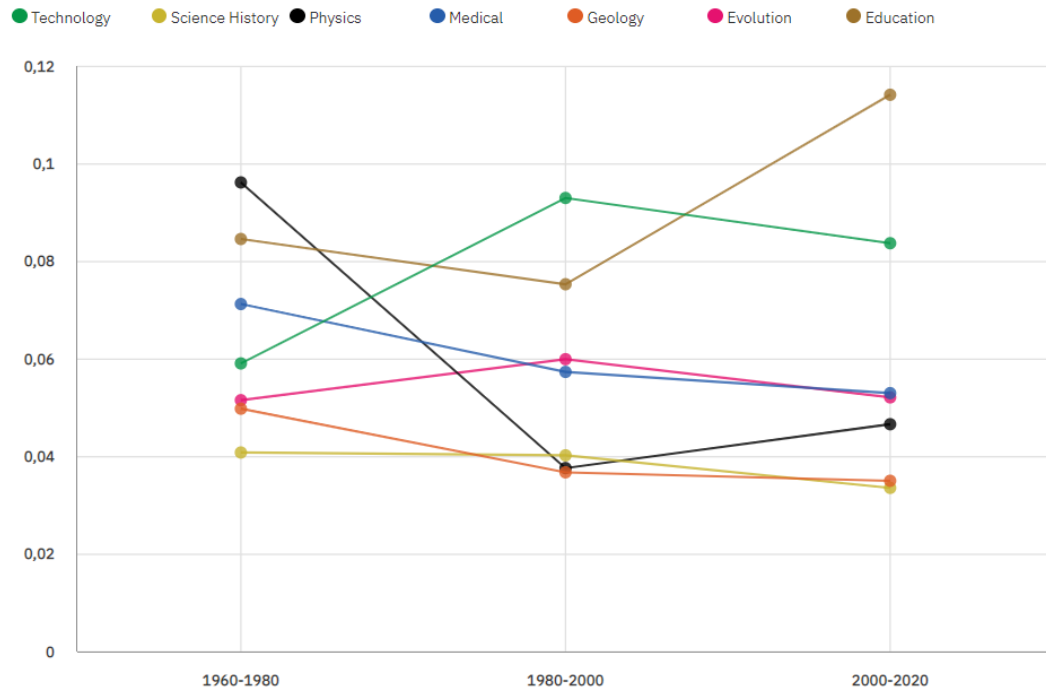


Figure 7. Probabilities of BJHS dataset's topics over the years

Chapter 6

Conclusion

One of the most important techniques to be used for analysis of large text datasets is topic modeling. Within the scope of this thesis, the 153,109 article which are collected from the Science Magazine was examined with LDA, NMF and K-means in 10-year and 20-year intervals. And then, the trends of topics was tried to be analyzed. Also for 12,570 article which are collected from the BJHS journal, topic models were extracted with LDA and the trends of topic was followed.

While applying the LDA method, the most important points were how to determine the number of topics and how to label the extracted topics. In order to determine the number of topic, numbers of topics were tested from 10 to 100 in clusters of 10 in each dataset and the number of topic that best explained the corpus were selected. While the optimal number of topic was determined as 30 for the Science dataset divided into 20-year intervals, the optimal number of topic was found to be 20 in the Science dataset divided into 10-year intervals. The optimal number of topics in the BJHS dataset was found to be 20.

While assigning the topic labels, 62 sub-categories under 4 main categories were determined manually for Science Magazine. For BJHS Journal, 27 sub-categories were determined under 4 main categories. These categories were determined both by considering the general topics of the journals and by examining the words that appeared in the topic models.

Topics were extracted with LDA and NMF. Then, clustering was done with K-means. The number of clusters are given the same number of topics in the topic model. And 10 words with the highest TF-IDF score in each cluster were selected and it was decided that the related cluster was involved to those words. The parameter α was chosen as 50/topic and β is set to 0.1 by default. Then, the topics and words extracted with LDA, NMF and K-Means were compared. It has been tried to be understood whether similar topics have been caught. It has been observed that LDA and NMF generally extract similar topics with similar words. It was seen that the K-means algorithm was able to extract less specific words, and it was seen that it could not obtain some topics which are seen with LDA and NMF at all.

Afterwards, it was tried to determine the most talked about ones among the topics extracted with LDA. Therefore, the topic which has the highest probability in each document was taken into consideration and evaluated as the dominant topic of that document. Then, based on the number of documents assigned to the dominant topics, the most mentioned topics were tried to be caught.

The results which are obtained and the changing of the most discussed scientific topics in Science and BJHS journals were tried to be examined in Results and Discussion Chapter. In the Science dataset, it has been observed that topics such as plant biology, astrophysics, paleontology, social science, chemistry, geology, education, disease, medicine, ecology are discussed every two decades intervals. In addition, over the years, it has been observed that some interesting topics specific to those years have been captured. In the BJHS dataset, it has been seen that the topics of technology, physics, geology, medicine, evolution, science history, education are discussed every year.

REFERENCES

- Bittermann, A., Fischer, A. (2018). How to Identify Hot Topics in Psychology Using Topic Modeling. *Zeitschrift für Psychologie*, 226, 3-13.
<https://doi.org/10.1027/2151-2604/a000318>
- Blei, D. (2012). Probabilistic Topic Models. *Communications of the ACM*, 55, 77-84.
- Blei, D. M., Lafferty, J. D. (2006). Dynamic Topic Models. *Proceedings of the 23rd International conference on Machine learning*, 113-116.
- Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Boyd-Graber, J., Hu, Y., Minmo, D. (2017). *Applications of Topic Models*. Colorado: New Foundations and Trends.
- Breiman L. (2003). What Is Text Mining?
<https://people.ischool.berkeley.edu/~hearst/text-mining.html> Accessed: 2022-08-01.
- Carron-Arthur, B., Reynolds, J., Bennett, K., Bennett, A., Griffiths, K. (2016). What's all the talk about? Topic modelling in a mental health Internet support group. *BMC Psychiatry*, 16, 367. <https://doi.org/10.1186/s12888-016-1073-5>
- Choo, J., Lee, C., Reddy, C. K., Park, H. (2013). Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE transactions on visualization and computer graphics* 19, 12(2013), 1992–2001.
- Choo, J., Lee, C., Reddy, C. K., Park, H. (2015). Weakly supervised nonnegative matrix factorization for user-driven clustering. *Data Mining and Knowledge Discovery* 29, 6 (2015), 1598–1621.

- Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., & Harshman, R.A. (1990). Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.*, 41, 391-407.
- DiMaggio, P., Nag, M., & Blei, D.M. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41, 570-606.
- Erlin, M. (2017). Topic Modeling, Epistemology, and the English and German Novel. *Journal of Cultural Analytics* 2 (2). <https://doi.org/10.22148/16.014>
- Griffiths, T. L., Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences (PNAS)*.
<https://doi.org/10.1073/pnas.0307752101>
- Hofmann, T. (2013). Probabilistic Latent Semantic Analysis. *Appears in Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*.
<https://doi.org/10.48550/arXiv.1301.6705>
- Kim, H., Choo, J., Kim, J., Reddy, C. K., Park, H. (2015). Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 567–576.
- Kuang, D., Choo, J., Park, H. (2014). Nonnegative matrix factorization for interactive topic modeling and document clustering. M. Celebi (Ed.), *Partitional Clustering Algorithms*. Springer.
- MacMillan, K., Wilson, J. D. (2017). Topic supervised non-negative matrix factorization. <https://arxiv.org/abs/1706.05084>. Accessed: 2022-08-01.
- Martin, F., Johnson, M. (2015). More Efficient Topic Modelling Through a Noun Only Approach. *In Proceedings of the Australasian Language Technology Association Workshop*, 111-115.

- Mimno, D. (2012). Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage*, 5(1), 1–19. <https://doi.org/10.1145/2160165.2160168>
- Ranaei, S., Suominen, A., Dedehayir, O. (2017). A topic model analysis of science and technology linkages: A case study in pharmaceutical industry. *2017 IEEE Technology & Engineering Management Conference (TEMSCON)*, 49-54.
- Suominen, A., Toivanen, H., (2016). Map of Science with Topic Modeling: Comparison of Unsupervised Learning and Human-Assigned Subject Classification, *Journal of the Association for Information Science and Technology*, vol. 67, no. 10, 2464–2476.
- Székely, N., vom Brocke, J., (2017). What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. *PLoS ONE* 12(4). <https://doi.org/10.1371/journal.pone.0174807>
- Talley, E.M., Newman, D., Mimno, D., Herr, BW, 2nd., Wallach, H, M., Burns, G, A., Leenders, A, G., McCallum, A. (2011). *Database of NIH grants using machine learned categories and graphical clustering*. *Nat Methods*. Jun;8(6):443-4.
- Thomas, S, W., Adams, B. , Hassan, A, E., Blostein, D. (2012). Studying software evolution using topic models. *Science of Computer Programming*, Vol 80, Part B, 457-479. <https://doi.org/10.1016/j.scico.2012.08.003>
- Wehrheim, L. (2019). Economic history goes digital: topic modeling the Journal of Economic History. *Cliometrica*, 13, 83-125.
- Wei, X., Croft, W, B., LDA-based document models for ad-hoc retrieval. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 178-185.

Yang, M, S., Sinaga, K, P. (2019). A Feature-Reduction Multi-View k-Means Clustering Algorithm. *IEEE Access*, vol. 7, 114472-114486.

Yau, CK., Porter, A., Newman, N., Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics* 100, 767–786.
<https://doi.org/10.1007/s11192-014-1321-8>

