Czech University of Life Sciences
Faculty of Environmental Sciences

Department of Water Resources and Environmental Modelling

Study Program: Landscape Engineering

Development of a research support tool for literature review facilitating data mining techniques

MASTER THESIS

Author: Aleksandr Kazantsev
Supervisor: doc. Ing. Mgr. Ioannis Markonis
Year: 2022

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Environmental Sciences

DIPLOMA THESIS ASSIGNMENT

Aleksandr Kazantsev

Landscape Engineering
Environmental Modelling

Thesis title

Development of a research support tool for literature review facilitating data mining techniques

---

Objectives of thesis

The thesis has a single objective. It aims to use machine learning algorithm(s) to classify the scientific results of published scientific research. It will achieve this by applying the text mining algorithm(s) in the abstracts of the presentations in the European Geophysical Union annual assembly (2011-2020).

Methodology

The following methodological steps have been chosen:

a. Meticulous literature review on the text mining algorithms and their applications in scientific publications.

b. Downloading of EGU abstracts and post-processing to transform them in readable format.

c. Application of the algorithm(s) for general classification of the abstracts.

d. Application of the algorithm(s) for investigation of specific reseach disciplines (one broad and one specialized).

e. Discussion of the recent trends of scientific research in geosciences over the last decade.

f. Investigation of the application of the algorithm(s) to other sets of abstracts. Case study: abstracts for a specilized topic from Web of Knowledge.

The proposed extent of the thesis

100 pages

Keywords

text mining, literature review, research trends, machine learning

---

Recommended information sources

https://arxiv.org/abs/1304.5457
https://jmlr.csail.mit.edu/papers/volume11/vinh10a/vinh10a.pd
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8378599/
https://www.tandfonline.com/doi/full/10.1080/13614533.2021.1918190?scroll=top&needAccess=true

---

Expected date of thesis defence

2021/22 SS – FES

The Diploma Thesis Supervisor

doc. Mgr. Ing. Ioannis Markonis, Ph.D.

Supervising department

Department of Water Resources and Environmental Modeling

Advisor of thesis

Ing. Nikola Šeborová

Electronic approval: 24. 11. 2020

prof. Ing. Martin Hanel, Ph.D.

Head of department

Electronic approval: 30. 11. 2020

prof. RNDr. Vladimír Bejček, CSc.

Dean

Prague on 30. 03. 2022

Author's Statement

I hereby declare that I have independently elaborated the diploma/final thesis with the topic of: Development of a research support tool for literature review facilitating data mining techniques and that I have cited all the information sources that I used in the thesis and that are also listed at the end of the thesis in the list of used information sources. I am aware that my diploma/final thesis is subject to Act No. 121/2000 Coll., on copyright, on rights related to copyright and on amendment of some acts, as amended by later regulations, particularly the provisions of Section 35(3) of the act on the use of the thesis. I am aware that by submitting the diploma/final thesis I agree with its publication under Act No. 111/1998 Coll., on universities and on the change and amendments of some acts, as amended, regardless of the result of its defence. With my own signature, I also declare that the electronic version is identical to the printed version and the data stated in the thesis has been processed in relation to the GDPR.

In Prague, 31^{th} March 2022 Aleksandr Kazantsev

Acknowledgment

I would like to express my sincere thanks to doc. Mgr. Ing. Ioannis Markonis, Ph.D, for leading this work and inspiring suggestions. I am deeply grateful to my consultant, Ing. Nikola Šeborová, for her support and help.

Aleksandr Kazantsev

Thesis Title:

Development of a research support tool for literature review facilitating data mining techniques

Author:	Aleksandr Kazantsev
Study Field:	Landscape Engineering
Study Program:	Environmental Modelling
Type of thesis:	Master Thesis
Supervisor:	doc. Ing. Mgr. Ioannis Markonis Department of Water Resources and Environmental Modelling
Consultant:	Ing. Nikola Šeborová Department of Water Resources and Environmental Modelling

Abstract: Numerous research papers and scholarly data have been released online over the past few years. The simplicity of accessing the data and tremendous growth of knowledge brings significant advantages to the research community, but it also creates the information overload problem, especially in academia. Therefore, there is a growing demand for the ability to segment and automatically analyze research papers in the research field. The unsupervised machine learning method of topic modelling is an automated method to extract information from scholarly data that has become increasingly popular. The application of topic modelling for generating topics to segment, explore and describe the Geophysical Research Abstracts (GRA) has been explored in this thesis. The topic modelling takes advantage of enormous amounts of text data to discover topics that run through a collection of documents by using statistical relationships between the terms in these documents. The Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF) algorithms and Okapi BM25 information retrieval model are implemented and evaluated. The effectiveness of the topic modelling is assessed by the ability to serve as a uniform categorization framework for research papers and by showing that algorithms can generate meaningful topics and keywords. Extensive data preparation and preprocessing of Geophysical Research Abstracts were required to apply topic models successfully. The results demonstrate that LDA and NMF algorithms could create topic models with meaningful topics and that topic modelling can be used for content analysis and potentially as an unsupervised categorization framework. However, it was found that NMF and LDA algorithms have different characteristics and should be applied in different usage cases. One of the significant weaknesses is that topics created by topic models, compared to manual methods, are more unreliable and could produce misleading results, which is an effect of its uncontrollable nature. On the other hand, its strength is the ability to analyze large amounts of text in a short time and at a low cost, deriving insights from many research papers. Topic modelling can complement other methods for content analysis or categorization, and it is a powerful method for aggregating and presenting the results to generate insights for efficiently analyzing and segmenting research papers.
Key words: topic modelling, NMF, LDA, topics, document keywords

Název práce:

Vývoj nástroje pro podporu výzumu pro přehled literatury usnadňující techniky pro těžbu dat

Autor:	Aleksandr Kazantsev
Studijní obor:	Krajinné inženýrství
Studijní program:	Environmentální modelování
Typ práce:	Diplomová práce
Vedoucí:	doc. Ing. Mgr. Ioannis Markonis Katedra vodních zdrojů a environmentálního modelování
Konzultant:	Ing. Nikola Šeborová Katedra vodních zdrojů a environmentálního modelování

Abstrakt: Během několika posledních let bylo online zveřejněno mnoho výzkumných prací a vědeckých údajů. Snadný přístup k datům a obrovský nárůst znalostí přináší významné výhody výzkumné komunitě, ale také vytváří problém přetížení informacemi, zejména v akademické sféře. Roste poptávka po schopnosti segmentovat a automaticky analyzovat výzkumné práce v oblasti vědy. Metoda strojového učení bez učitele a modelování témat je automatizovaná metoda pro získávání informací z odborných dat, která se stává populárnější. V této práci byla zkoumána aplikace modelování témat a jejich generování pro segmentaci, zkoumání a popis geofyzikálních výzkumných abstraktů (GRA). Modelování témat využívá obrovské množství textových dat k objevování témat, která procházejí sbírkou dokumentů pomocí statistických vztahů mezi pojmy v těchto dokumentech. Jsou implementovány a vyhodnoceny algoritmy Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF) a model získávání informací Okapi BM25. Efektivita tematického modelování se hodnotí podle schopnosti sloužit jako jednotný kategorizační rámec pro výzkumné práce a podle toho, algoritmy mohou generovat smysluplná témata a klíčová slova. K úspěšné aplikaci tematických modelů byla nutná rozsáhlá příprava dat a předzpracování geofyzikálních výzkumných abstraktů. Výsledky ukazují, že algoritmy LDA a NMF by mohly vytvářet modely témat se smysluplnými tématy a že modelování témat lze použít pro analýzu obsahu a potenciálně jako rámec kategorizace bez dozoru. Bylo však zjištěno, že algoritmy NMF a LDA mají různé charakteristiky a měly by být aplikovány v různých případech. Jednou z výrazných slabin je, že témata vytvořená tematickými modely jsou ve srovnání s manuálními metodami nespolehlivá a mohla by produkovat zavádějící výsledky, což je důsledkem jejich nekontrolovatelné povahy. Na druhou stranu, jeho silnou stránkou je schopnost analyzovat velké množství textu v krátkém čase a za nízkou cenu, na základě poznatků z mnoha výzkumných prací. Tématické modelování může doplňovat další metody pro analýzu obsahu nebo kategorizaci, a je to výkonná metoda pro agregaci a prezentaci výsledků, vytváření náhledů, efektivní analýzu a segmentaci výzkumných prací.

Klíčová slova: tématické modelování, NMF, LDA, téma, klíčová slova dokumentu

Contents

1 Introduction

With advances in information and computer technologies, we see the rapid emergence of research papers and scholarly data. Numerous research papers have been released online, and many archival materials have been digitized over the past few years. The tremendous growth of knowledge and simplicity of accessing the data brings significant advantages to the research community, but it also creates the information overload problem, especially in academia.[1] The increasing complexity of finding and categorizing the proper research papers has become even more demanding for researchers.[2]

Researchers spend many hours finding documents on specific topics; therefore, the demand for highly interpretative and convenient automated classification systems increases. However, the relations between papers are indistinct, and it is hard to accurately classify similar research papers based on keywords input from the user. There are numerous advanced techniques applied on the database level backed up with large-scale high computational machines to find the best matching documents in the shortest time based on the user input. For example, many research platforms, such as Google Scholar, ResearchGate, ScienceDirect, The SAO/NASA Astrophysics Data System (ADS), have successfully implemented an article/research recommendation system. The recommendation is based on the popularity among research communities and the content of an article. It is an essential tool in information retrieval and filtering, which helps identify related research articles from many publications.[3]

Recommendation systems are information filtering systems that analyze the behaviour of users to predict interests in information, products or services by employing data mining. With the ever-growing public information online, recommendation systems have proven to be an effective strategy to deal with information overload. For example, collaborative filtering works by utilizing the rating activities of items or users and content based works by comparing descriptions of items or profiles of users' preferences. Applications of recommendation systems are currently expanding beyond the commercial to include scholarly activities. Recommendation systems are more personalized and effective than the traditional keyword-based search technique

for massive amounts of scholarly data. Recommendation systems usually consider co-author relationships, researchers' interests and citation relationships to design the recommendation algorithms to provide the best matching results. However, it is sometimes hard to describe and summarize the search requirements if the researcher has no clear idea of what they are looking for and understanding of the topic, resulting in inappropriate keywords. In addition, for junior researchers with limited publishing experience, recommendation systems may advise unrelated articles that do not align with the area of research interests. On the contrary, the recommendation systems mainly recommend papers that align solely to their research interests for senior researchers with more substantial publication records.[4]

The unsupervised machine learning method of topic modelling is an automated method to extract information from scholarly data that has become increasingly popular. The topic modelling takes advantage of enormous amounts of text data and explores with the aim discover topics that run through a collection of documents by using statistical relationships between the terms in these documents. The topic consists of terms that are statistically related in the document collection. Topic modelling algorithms do not require labelling of the documents or any prior annotations, therefore reducing the time and costs of such projects.[5, 6, 7]

Topic modelling has broad applications in various contexts; however, scientific abstracts datasets are not widely researched. Topic modelling can yield valuable insights because topics are generated independently from human preconceptions and can potentially lead to unexpected but valuable results, such as relationships between research abstracts and hidden patterns in the data. The topic modelling has the potential to reduce vast data sources into meaningful topics, with interpretable and valuable results to a researcher.[8, 9] One of the goals in the scope of this thesis is to create an automated tool that creates a uniform categorization framework, independently from human preconceptions, for the research abstracts by utilizing the data from European Geosciences Union (EGU) and evaluate how topic models best can be used to create value in content analysis and categorization tasks in the scientific field.

2 Objectives

The main focus of this diploma thesis is on developing a research support tool used for the investigation of distinct research disciplines utilising data mining techniques on European Geophysical Union (EGU) annual assembly abstracts. Topic modelling was chosen as the primary technique for topic extraction, identifying patterns, retrieving information and organising the EGU abstracts. The topic is a recurring pattern of co-occurring words, and topic modelling is a method for tracing clusters of words in large bodies of texts [10]. Different topic modelling algorithms are introduced and evaluated in this thesis. Before analysing EGU abstracts, the pre-processing steps must be applied to the dataset. Therefore, various NLP techniques such as lemmatisation and various text representations are examined to improve the data quality and thus the accuracy and efficiency of the text mining process. One of the critical challenges of clustering text data is to evaluate the obtained results. Therefore, the following methodological steps have been chosen:

- Assessment of topic modelling algorithms and their applications in scientific publications.
- Acquisition, transformation and pre-processing of EGU abstracts.
- Application and evaluation of topic modelling algorithm for general clusterisation of abstracts.
- Application and evaluation of topic modelling algorithm for the investigation of specific research disciplines
- Discussion of the recent scientific research trends in geoscience over the last decade.

3 Literature review

This chapter provides the theoretical background, essential concepts from the field of machine learning, methods from natural language processing and the theory behind topic modelling will be presented which will serve as a basis for the concepts used throughout this thesis.

3.1 Machine Learning

Machine learning is a subset of artificial intelligence focused on algorithms that teach computers to learn from data. Machine learning can be classified into two main categories by the level of human intervention in the process: unsupervised learning and supervised learning.[2]

In supervised learning, the system can learn from data that has been labelled by humans so that it can make more accurate predictions. The most common application includes predictive analysis based on classification and regression problems. Supervised learning is excellent for learning complex patterns in data, but it relies on having a lot of manually labelled data. While unsupervised learning does not require labelled data, the range of potential applications is limited. A semi-supervised learning system can learn from labelled and unlabelled data, often seen as a more efficient method. It is based on the idea that a small amount of labelled data can improve the performance of an unsupervised learning model.[2]

Unsupervised learning is a type of machine learning where the computer is given data without any labels, meaning that the system is not given any feedback on its predictions. Therefore, the system must learn from data independently, without any human guidance, which is a more difficult task. Still, it can also lead to more accurate results since any human assumptions do not bias the system. There have been developed various types of algorithms such as Term Frequency-Inverse Document Frequency (TF-IDF), K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), Naïve Bayes classifier.[11] These methods classify documents based on the similarity of documents without predefined criteria.[11, 12] Un-

supervised learning is often used for data mining, information retrieval and pattern recognition.

3.2 Vector Space Model

After a pre-processing stage, the unstructured text is mathematically computable and manageable by text mining algorithms. Vector Space Model (VSM) is one of the most popular models. The VSM is an algebraic model based on similarity. Each text document of a collection C is represented as a vector of weighted features in an N-dimensional vector space, where N is the total number of unique terms occurring in the corpus, also called vocabulary. Each document d_j in a collection can be represented as a vector

$$d_j = (w_1, w_2, \cdots, w_N) \tag{3.1}$$

$$q = (w_1, w_2, \cdots, w_N) \tag{3.2}$$

where w_i is the weight of the term i in document j and $j \in 1 \cdots n, 1 \leq i \leq N$. As a result, we get term-document-matrix after joining these vectors.[13] Similarly, q is a query, and w_i is the count for a word in q.[14]

The vocabulary size can grow immensely, so it is essential only to store semantic meaning terms. Most elements inside a query and document will be equal to zero because vectors are highly sparse. For example, the text of a query is ”Standardized Precipitation Index”, where the vocabulary contains 5000 distinct terms. Therefore, the vector representation of this query will contain three ones, each located at the corresponding index for ”Standardized,” ”Precipitation” and ”Index,” and 4,997 zeros in every other index location.[14]

3.2.1 Bag-Of-Words

A Bag-Of-Words (BOW) is a data structure that stores a collection of words along with the number of times each word appears in a given text. This allows for rapid counting of word occurrences, useful for natural languages processing tasks such as sentiment analysis or machine translation. The BOW data structure is efficient, robust, and produces relatively good accuracy; however, the semantic representation is lost because the word order is not preserved. Semantically different sentences could have the exact representation model if the same words were used.[15]

Meaning of the word plays an essential part in distinguishing between various topics and grouping similar documents together in finding similar documents or topic modelling.

3.2.2 N-grams

A limitation of the bag-of-words model is its inability to represent idiomatic phrases of sequences of terms. N-grams are one way to eliminate this limitation. It is a sequence of items in any given sentence. For example, a unigram contains one item, a bigram has two items, a trigram consists of three items. The items can be words, bytes, characters or syllables. N-grams are commonly used in predictive analysis and identifying context because of their sequential nature. N-grams play an essential role in Statistical Natural Language Processing. For example, it helps in spelling correction, document clustering, language detection, authorship attribution, understanding context, automatic grading. In topic modelling, n-grams increase the model's accuracy to represent terms as real entities but can increase the dimensionality of the model. For example, the trigram "Geophysical Research Abstracts" is formed from the term sequence "Geophysical"," Research", and "Abstracts", but the terms can still exist individually, thus increasing the dimensionality of the corpus by one.[16]

3.2.3 Term Frequency-Inverse Document Frequency

Term Frequency-Inverse Document Frequency (TF-IDF) is a weighting scheme used to calculate the importance of a word or phrase within a document. It is often used to evaluate the relationship for each word in the collection of documents in text mining and information retrieval. The document d is represented as a vector of word frequencies t in term frequency $tf(t, d)$.

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{3.3}$$

Where $f_{t,d}$ is the raw count of a term in a document, divided by the total number of terms in document d.

Term frequencies (TF) consider all terms equally significant, making it impossible to assess the relevancy of a query. To measure how much information a word adds and the uniqueness of a term to the piece of content, the Inverse document frequency (IDF) is used with each frequency logarithmically scaled by the inverse ratio of documents containing the term (IDF).[17] A geometric distance function

is used to compare the similarity of vectors, and the vector space model slightly modifies the vector representation of documents and queries. In addition, the vector space model recognizes the limitations of only accounting for the frequency of terms inside a document. For example, the term ”model” may frequently appear inside a hydrological-themed corpus. So, the word ”model” holds less meaning than other words found in the corpus. Therefore, the vector space model considers the rarity of terms with respect to all other terms inside the corpus.[14]

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{3.4}$$

where, N is the total number of documents in the corpus $N = |D|$ divided by the $|\{d \in D : t \in d\}|$ number of documents in the corpus that contain the term t.

The term weight inside a query or document vector is the product between the inverse document frequency and term frequency:

$$w_{ij} = idf_i * tf_{ij} \tag{3.5}$$

3.2.4 Best Matching BM25

The Okapi BM25 information retrieval model assumes a bag of words interpretation for documents and queries. The base assumption is that occurrence of a query term in a document is an independent event and happens in a specified interval, the start and end of a document. Okapi BM25 was built as an approximation of the Poisson distribution because the Poisson distribution requires that the rate of occurrences for terms in a document is known ahead of time. The model takes two inputs, query q and document d_j, and loops through each term t_i that appears in the query and the document. The product of the inverse document frequency, the term frequency, and the query term frequency provides the score for a term. The overall score for a pair of vectors is the sum of all the values. Higher scores indicate that two vectors are similar to each other, and lower scores indicate that two vectors are dissimilar.[14, 18]

$$okapi(d_j, q) = \sum_{t_i \in q, d_j} idf_i \times tf_{ij} \times qtf_i \tag{3.6}$$

The inverse document frequency is a logarithmic function that gives a higher reward to terms that infrequently occur in the document collection.

$$idf_i = ln\frac{|D| - df_i + 0.5}{df_i + 0.5} \tag{3.7}$$

where idf_i is the inverse document frequency for a term i, $|D|$ is the number of documents in the corpus, and df_i is the number of documents in the corpus that contain the term i. The term frequency is a linear function that gives a higher score to terms that frequently occur in small documents and a lower score to the document d_j if the length of a document dl_j is longer than the average document length $avdl$ in the corpus. This is because longer documents have more opportunities to contain query terms, so it is essential to consider the difference between longer and shorter documents.[14]

$$tf_{ij} = \frac{(k_1 + 1)f_{ij}}{k_1(1 - b + b\frac{dl_j}{avdl} + f_{ij})} \tag{3.8}$$

where k_1 adjusts the weight of the term frequency with respect to the entire model and b adjusts the penalty score for document length.

The query term frequency for a term t_i is a linear function that gives a higher score for terms that appear multiple times in a query.

$$qtf_i = ln\frac{(k_2 + 1)f_i}{k_2 + f_i} \tag{3.9}$$

where f_i is the frequency of a term t_i in a query q and k_2 is a adjusts the influence of the query term frequency with respect to the entire model.

3.3 Data mining

Data mining and text mining are often complementary analytic processes; however, they handle different data types. Data mining deals with well-formatted and structured data, usually seen in databases. Knowledge Discovery in Databases (KDD) is the automated analysis and modelling of large data sets is called Knowledge Discovery in Databases (KDD). KDD is a process of identifying potentially useful, valid and understandable patterns in data.[19] The basic workflow of the KDD process consists of the following five phases:

- Data selection according to the objectives of the research.
- Pre-Processing phase includes handling errors or missing values and data cleaning, which is a fundamental step for data analysis. The data cleaning approach depends on the area of knowledge extraction.

- Data converted into the appropriate format required by the analysis method is done during the data transformation step. It includes dimensionality reduction and feature extraction to reduce the storage space and computation time.
- A data mining algorithm of choice is applied to extract data patterns.
- Interpret and evaluate the obtained results.

These five steps can be further extended based on the application and the overall goals. For example, clustering belongs to discovery methods that automatically identify data patterns.[19]

3.4 Text Mining

Text mining deals with unstructured textual data containing hidden information and underlying patterns, useful for research purposes. For example, valuable structured information can be uncovered from massive data using text mining algorithms. In addition, text mining techniques can classify or summarise unstructured data. It is possible to identify the various domains underlying the data through classification and clustering.[20]

3.5 Topic Modelling

Topic modelling is the unsupervised ML algorithm used for learning and extracting topics from documents. It is one of the most frequently used text mining techniques.

A corpus of documents can be explored based on their topics. Topic modelling is a statistical model that identifies the topics present in any given set of documents by identifying the most associated words. It can also connect words with similar meanings and distinguish between the various meanings depending on their context. In topic models, documents are categorised into themes that become the corpus's topics, viewed as a mixture of various topics. The topic is a multinomial distribution over words.[21]

However, it is hard to manually read large volumes of text and categorise it based on topics. An automated algorithm like Latent Dirichlet allocation (LDA) and Non-Negative Matrix Factorisation (NMF) requires minimum human intervention. The researcher has to input the number of topics to the algorithm, giving the topic probabilities of the words and the topic distribution of the corpus. A topic is a

collection of words that have semantic relatedness. This provides an idea of the distinct topics present in the collected data as an input model takes a document-word matrix where $DWM[i][j]$ equals to the number of occurrences of word i in a document j and a number of topics n_{topics} defined by the researcher. A topic modelling algorithm tries to find the co-occurrence of such patterns irrespective of the sentence's complexity. A model considers the corpus documents as a bag-of-words (BOW) from which the recurring co-occurrence patterns and topic distributions are found. The output will be a word-topic matrix and a topic-document matrix. The model outputs the words in each topic, making it difficult for a researcher to name the topics when they have minimum knowledge of the corpus domain.[20]

On the other hand, topic modelling is helpful in the automatic coding of a large corpus with minimum effort. It also paves the way for understanding the corpus from a different perspective. Topic modelling can also look at the data when applied to a small corpus. Finally, it helps analyse the text quicker, more efficiently, and objectively. A great way to explore the topics is to use visualisation. While topic modelling has its disadvantages, in that as a probabilistic model, it is not repeatable in ways required by more explanatory research, and its "accuracy" is challenging to evaluate, it offers a valuable mechanism for quickly summarizing and clustering large bodies of text in ways that can be used to guide further research and analysis.[22]

The origin of topic model algorithms is latent semantic analysis (LSA), also referred to as latent semantic indexing (LSI)[23]. The application of this algorithm on a text corpus requires that the corpus is first transformed into a document-term matrix, here denoted by X. LSA builds on singular value decomposition (SVD) to factorize this matrix X of the corpus into a set of component matrices. These matrices can be reduced to a lower rank and thus be an approximation of X when multiplied[24, 25]. One of these component matrices describes basis vectors, or eigen features, for describing X in possibly lower dimensions, while another component matrix represents a mapping of those bases to describe the data samples in X in the original dimensions[23]. The conceptual idea of LSA, and topic modelling in general, is to factorize the document-term matrix of the corpus into one matrix containing topic-term information (i.e. basis vectors) and another matrix containing document-topic information (i.e. the mapping between basis vectors and X), denoted by W and H in this thesis. The topic-term matrix H describes each topic as a weighted vector of length V, where each weight corresponds to the importance of a term in that topic. The document-topic matrix W describes each document as a weighted vector of length K, where each weight corresponds to the importance of a topic in that document. LSA is the basis for building other, more successful topic model algorithms. Most topic models share the exact composition of a topic-term matrix and a document-topic matrix, and the topic model algorithms aim to derive these

after some optimality objective. It is noteworthy that for dimensionality reduction applications, it is common to work with the transpose of X^T (a term-document matrix), thus requiring transposes W^T and H^T, which leads to other matrix operation orderings for factorization. The development of topic model algorithms originating from LSA has taken two different routes: one probabilistic approach and one that builds on linear algebra.

3.5.1 Latent Dirichlet Allocation - LDA

Latent Dirichlet allocation (LDA) is a generative probabilistic topic model algorithm and is one of the most popular methods for topic modelling. It is used to allocate documents to a specific topic group based on contained words within the topic.

LDA tries to capture the statistical structure using mixture distribution within a single document. The model should consider both the documents and words exchangeable, where every collection of an exchangeable random variable can be represented as a mixture distribution. A Bayesian model predicts the probability of an event based on prior knowledge. The documents are modelled to give a finite mixture of topics, producing an infinite mixture of topic probabilities.

The corpus is a collection of M documents $d = \{w_1, w_2, \cdots w_M\}$, where document d is a collection of N words $d = \{w_1, w_2, \cdots w_N\}$ and word is part of the vocabulary $V = \{1, \cdots V\}$. LDA also has two Dirichlet priors, α and β, are corpus-level variables which are sampled once during corpus generation and represents the per-document topic and word distribution, respectively. θ is a document-level variable which is sampled once for every document and represents the topic distribution per document. The dimensionality K 'topics' is supposed to be known and fixed.

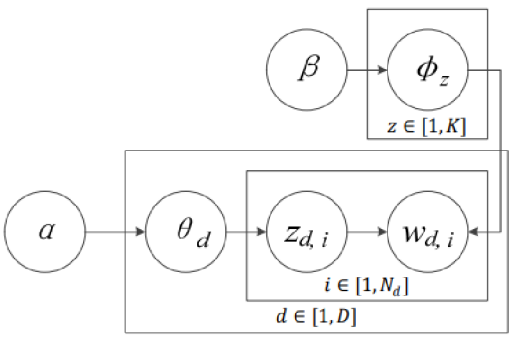


Figure 3.1: Graphical representation of LDA

LDA is a three-level hierarchical Bayesian model. The base assumption is that all documents are related, and documents with similar topics will use a matching set

Symbol	Description of Symbol
D	Collection of Documents
K	Collection of Topics
N_d	Length of Document d
w_d	The i th Word of Document d
z_d	The i th Topic of Document d
α	Dirichlet Prior Distribution of Topics on Documents in LDA
β	Dirichlet Prior Distribution of Words on Topics in LDA
θ_d	Polynomial Distribution of Topics on Documents d
ϕ_z	Polynomial Distribution of Words on Topic z

Table 3.1: LDA Symbols

of words. Each topic defines a multinomial distribution over the vocabulary and is assumed to have been drawn from a Dirichlet, $\beta_k \sim Dirichlet(\eta)$. There is a varying probability of document belonging to a particular topic group where each document contains a distribution of topics and probability of terms defining the topic group. Given the topics, LDA assumes the following generative process for each document d. First, draw a distribution over topics $\theta_d \sim Dirichlet(\alpha)$. For every N words w_n, choose a topic $(z_n) \sim Multinominal(\theta)$ and choose a word w_n from $p(w_n|z_n, \beta)$.[26, 20]

In this thesis, an online variational Bayes as an optimization-based algorithm is used. It is a deterministic alternative to sampling-based algorithms. This algorithm places several distributions over the latent variables instead of approximating the posterior with samples and then finds the distribution closest to the posterior with an optimization approach. Online variational Bayes is based on variational inference, called variational Bayes (VB). The idea in VB is to optimize the distribution to be close in Kullback-Leibler divergence to the posterior. However, VB requires a complete pass through the entire corpus each iteration and can therefore be very slow to apply if the corpus consists of many documents. Online variational Bayes was proposed to make this process more effective and is based on online stochastic optimization, which has been shown to produce suitable parameter estimates dramatically faster than traditional VB on large datasets.[26]

3.5.2 Non-Negative Matrix Factorization - NMF

By setting constraints on the matrix factorization process, the NMF can be described as an extension of LSA, thus differing from SVD. A property of SVD is that the basis vectors will be orthogonal to each other; to achieve this, some elements in the bases are forced to be negative. NMF usually takes a TF-IDF document-

term matrix as input. Some interpretative issues when considering the basis vectors describe features in X. For example, negative elements in the bases and mappings cause subtractions between columns, leading to a spread distribution of bases in describing a sample in X.

NMF can be described as an extension of LSA by imposing constraints on the matrix factorization process and thus differing from SVD, as there are significant issues with the SVD representation. NMF usually takes a TF-IDF document-term matrix as input. A property of SVD is that the basis vectors will be orthogonal to each other. Some elements in the bases are forced to be negative in achieving this. The data matrix X is factorized into matrices W and H that approximate X with the constraint that W and H only contain non-negative elements. This leads to improved interpretability due to non-negative representations of bases and encodings in W and H and an increased sparseness in these matrices as many elements are forced to zero. The topic encodings in H will be described by fewer and more distinguishable features, and the bases that describe assignments of topics to documents in W will also be fewer and more distinguished. As a result, the approximation of X by the product $W\ H$ will be of equal or lower rank K, with $(N+V)K \leq NV$.**??**

The algorithm for deducing W and H from X can be posed as an optimization problem

$$minD(X; W, H) \tag{3.10}$$

where the difference D between $W\ H$ and X is minimized and $W \geq 0, H \geq 0$. Frobenius norm is one of the most frequently adopted difference measures.

$$\|X - WH\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} ||X_{ij} - (WH)_{ij}|^2} \tag{3.11}$$

$$min\|X - WH\|_{\mathrm{F}} \tag{3.12}$$

Compared to standard topic modelling methods such as latent Dirichlet allocation (LDA), NMF essentially gives the same output types: A keyword-wise topic representation (the columns of W) and a topics document representation (the columns of H). However, the only difference is that the columns of W and H do not have a unit L1-norm, unlike the LDA outputs. Nonetheless, such a difference is negligible and can be manipulated via diagonal scaling matrices. Moreover, the column normalization on H does not affect the interpretation of each document in terms of its relative relationships to topics. In this sense, NMF can be used as an alternative to

topic modelling methods.[27, 24]

3.6 Natural Language Processing

Natural Language Processing (NLP) studies how computers can understand and process human language.[28] NLP, machine learning, and deep learning are subfields of artificial intelligence. NLP attempts to capture and process natural language using computer-based rules and algorithms. Processing human language is complex because "Language is highly ambiguous, ever-changing and evolving". NLP applications must be able to handle ambiguity, context and syntactic variations. Therefore various methods and results from linguistics are combined with artificial intelligence.[13] It includes language modelling, part-of-speech (POS) tagging, named entity recognition (NER), sentiment analysis, paraphrase detection, lemmatization and stemming.[29]

The natural language analysis consists of phonological, morphological, lexical, syntactic, semantic, pragmatic and discourse analysis. The grammatical process of word-formation used to express, for example, the number, case, gender or mood of a word is called inflexion. There are two ways to reduce inflexion forms stemming and lemmatization. They share the same idea but use different ways to achieve the result. Stemming is a more straightforward option; words are reduced to their stem using heuristics that cut off the end of words to achieve the correct base. The problem is that a stemming algorithm may cut off too much because it does not consider the word's context. However, the lemmatization algorithm relates different forms of the same word to their dictionary form - lemma. Therefore, it determines the part-of-speech (POS) essential to identifying the grammatical context. For example, to transform a sentence, *"Topic modelling is one of the most frequently used text mining techniques."* into a syntactic structure; a parser evaluates each sentence compared to formal grammar rules to provide the sentence structure. A semantic analyzer then uses the syntactic structure to establish a correct logic between words and sentences. This is done by determining the basic dependencies related to other words. The resulting word references structure is displayed in the figure 3.2.

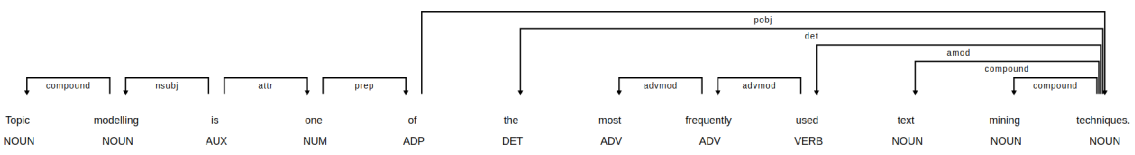


Figure 3.2: Dependency Structure

3.7 Text Pre-Processing

Collected data usually contains some noise. Therefore, the data should be pre-processed before transforming it into a form that computers can work. For example, images, text or videos require different pre-processing methods. Text pre-processing usually involves tokenization, filtering, normalization and lemmatization or stemming.[29]

Tokenisation is a complex process where text is broken down into smaller units called tokens. Tokens can be either word, characters, or n-gram characters. For example, one approach is to split up a sentence by spaces. However, all punctuation marks and brackets are not recognised as independent tokens.[29]

Numbers, whitespaces, symbols, punctuation and stopwords are filtered out from the text. Stopwords are frequently used words that do not contain much information, such as a, an or the. The spacy python library has a default list of 326 stopwords.

4 Methods

This chapter provides a detailed description of the dataset used in this thesis. It describes the process of crawling missing data and techniques used for labelling, pre-processing, feature representations, feature extraction, topic modelling methods, and the evaluation measures used to discuss their properties.

Topic modelling methods applied to the dataset typically involve several pre-processing steps, as outlined in Figure 4.1. Data is first extracted from a source. Then, documents are extracted from the raw data set, consisting of text data. The textual elements are converted to lower case and then processed to remove standard punctuation, stop words. Next, text data are separated into tokens, and then the lemmatization technique is applied to each token. Next, feature representations of each document are created, followed by topic modelling methods.

In order to find which topic model provides the best results for the dataset, four metrics for unsupervised contexts have been used, and two feature representations. The variations at each step of the process are outlined in Table 4.1. The approach to each step is described in Figure 4.1.

4.1 European Geosciences Union

The European Geosciences Union (EGU) is a nonprofit interdisciplinary association of scientists founded in 2002. It is the leading organisation for Earth, planetary and space science research in Europe. The EGU publishes several diverse scientific journals that use an innovative open-access format. Also, EGU organises many meetings and activities. Activities include supporting early-career researchers, Geosciences Information For Teachers (GIFT) workshops, the EGU blogs, media services, the EGU blogs, awards and medals programme for outstanding scientists. The EGU General Assembly is the most widely known and largest European geosciences event. The first General Assembly of the EGU was held in 2004. Scientists from more than 100 countries regularly participate in EGU annual meetings. In 2019 more than 16,000 thousand scientists from all over the world participated in the event in person.

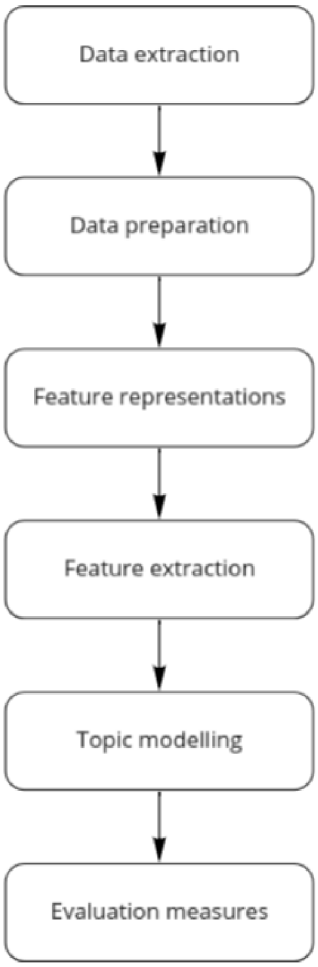


Figure 4.1: Process Pipeline

In 2020 more than 22,000 thousand scientists from 134 countries participated in the online event. A wide range of topics is covered during the meeting session, including planetary exploration, climate, volcanology, the Earth's internal structure and atmosphere, energy and resources. According to EGU General Assembly regulations, abstracts should be short (100–500 words), clear, concise, and written in English. It should not include any tables or figures. In addition, any mathematical symbols and equations must be typed in or embedded as images. Abstracts can be presented either as an oral, poster or during a PICO (Presenting Interactive Content) session by the author or co-authors.[30]

The EGU scientific activities are organised through scientific divisions encompassing all studies of the Earth and its environment and the solar system in general and Union-wide and Inter- and Transdisciplinary sessions (ITS). However, the vast majority of sessions at the EGU General Assembly are disciplinary sessions that allow participants to present and discuss their research with their peers. They cover the full spectrum of geosciences and space and planetary science. ITS was launched for the first time in 2016. It tackles a common theme through an inter-and transdisci-

Data set
EGU2009-2014
EGU2015-2021 Labeled
Feature representations
tf-idf BoW
BoW
Information Retrieval Method
BM25
Topic models
LDA
NMF
Evaluation Measures
NMI
AMI
ARI
c_v coherence measure

Table 4.1: Outline of the data set, feature representations and information retrieval methods, and extrinsic evaluation measures used in this thesis.

plinary combination of approaches, fostering cross-division links and collaborations. In addition, union-wide sessions are organised for all conference participants at the General Assembly.[30]

Figure 4.2: Sessions Division

The EGU organisation consist of scientific divisions, committees, and councils. There are eight committees with administrative functions and 22 scientific divisions responsible for scientific activities related to the Earth, planetary and space sciences. Figure 4.2 shows that Hydrological Sciences (HS), Atmospheric Sciences (AS), Soil System Sciences (SSS) and Climate: Present, Past, Future (CL) are four dominant scientific divisions in the years from 2015 to 2021. Another ten divisions include Union-wide and Inter- and Transdisciplinary sessions.

The HS Divisions includes all aspects of the terrestrial hydrological cycle, including surface water, precipitation, soil water, groundwater and its relationships between hydrology and soils and interactions with the atmospheric part of the hydrological cycle and between geomorphology and hydrology. The division also covers the hydrosphere and the biosphere. Furthermore, how hydrological processes are observed, quantitatively computed, and the division addresses forecasted. Finally, management and operation of water resources by societies in various parts of the world are also within the division's realm. [31]

AS include studies of the atmosphere composition, aerosol and cloud physics, gas-particles interactions and chemical reaction kinetics studied in the labs. The research division covers the large-scale dynamical, meteorological processes and systems in the atmosphere such as global atmosphere circulation and cyclones to the small scale turbulent mixing. Moreover, they cover the time frame from centuries in connection with climate research to seconds in the context of fast chemistry. [31]

Soil is the basis of life on Earth and the interface between the crust and atmosphere. The SSS aims to coordinate the EGU scientific programme on Soil Science and related activities. Furthermore, the SSS contributes actively with EGU by promoting scientific interchange and disseminating activity carried out by members.[31]

CL includes the study of any climate archive from rocks to ocean cores, speleothems, ice cores, chronicles, to instrumental records. CL division is very interdisciplinary and covers climate variations on all time scales. It pools from many disciplines and has many co-organised and co-listed sessions with other divisions at the general assembly. Besides observations, the division covers climate modelling on all time scales from the deep past to the future. CL main focus on the climate on Earth but may also expand other planets or the sun.[31]

To ensure the quality of the sessions and make them comprehensive, each division consists of multiple distinct fields within the broad area. Each year the members of the Subdivision Committees meet during the EGU General Assembly and prepare the draft programme for next year's meeting. From 2015 to 2021, 33 divisions and 3690 unique subdivisions have been identified.

The accepted abstracts from the General Assemblies 2005–2019 of the European Geosciences Union (EGU) are published in Geophysical Research Abstracts (GRA) conference series. In addition, the abstracts underwent an access review by the session conveners. As a result, it links the annual conference programmes listing programme groups, included sessions, and their contributions. The abstracts and site content are licensed under the Creative Commons Attribution 4.0 License, which gives rights to copy and redistribute the material in any medium or format and remix, transform, and build upon the material for any purpose as long as the author and source are properly cited. Thus, CC BY facilitates scientific knowledge dissemination, transfer, and growth.[30]

4.2 Data extraction

The abstracts are available on the Geophysical Research Abstracts (GRA) website from 2005 to 2019, and from 2020, abstracts and related presentation materials

become part of the EGUsphere. In GRA and EGUsphere, abstracts are available as Portable Document Format (PDF) files. Also, most of the EGU General Assembly abstracts are indexed in The SAO/NASA Astrophysics Data System (ADS), and many abstracts from open-access peer-reviewed journals are available in EBSCO.

The ADS is a digital library portal for researchers in astronomy and physics, operated by the Smithsonian Astrophysical Observatory (SAO) under a NASA grant. The ADS maintains three bibliographic collections containing more than 15 million records covering publications in general science, astronomy and astrophysics, physics, including all arXiv e-prints. Abstracts and full-text of major astronomy and physics publications are indexed and searchable through the ADS search form. In addition, the ADS tracks citations and usage of its records to provide advanced discovery, evaluation capabilities and access pointers to many external resources, including electronic articles available from publishers' websites, data catalogues and data sets hosted by external archives.[32]

All available data fields for the EGU publications have been queried in ADS.

RangeIndex: 178711 entries, 0 to 178710
Data columns (total 31 columns):

	Column	Non-Null Count	Dtype
0	bibcode	178711 non-null	object
1	abstract	178498 non-null	object
2	aff	178708 non-null	object
3	alternate bibcode	79 non-null	object
4	arxiv class	59 non-null	object
5	author	178708 non-null	object
6	bibstem	178711 non-null	object
7	database	178711 non-null	object
8	doctype	178711 non-null	object
9	first author	178708 non-null	object
10	id	178711 non-null	int64
11	identifier	178711 non-null	object
12	keyword	80 non-null	object
13	orcid pub	178708 non-null	object
14	page	178709 non-null	object
15	pub	178711 non-null	object
16	pubdate	178711 non-null	object
17	title	178708 non-null	object
18	year	178711 non-null	int64

RangeIndex: 178711 entries, 0 to 178710			
19	read count	178711 non-null	int64
20	property	178711 non-null	object
21	citation count	178711 non-null	int64
22	indexstamp	178711 non-null	object
23	volume	10 non-null	float64
24	orcid other	2405 non-null	object
25	orcid user	1253 non-null	object
26	bibgroup	230 non-null	object
27	doi	4 non-null	object
28	copyright	13 non-null	object
29	grant	12 non-null	object
30	data	1 non-null	object

Table 4.2: Indexed GRA in ADS

From table 4.2, we can see that there are available 178711 entries. However, many fields contain null values. After dropping columns with null values and rows with empty abstracts and titles, we got 178492 remaining non-null entries.

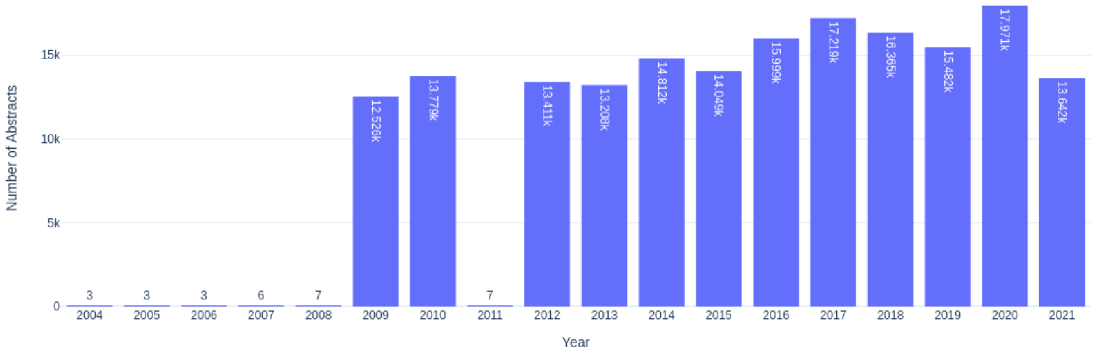


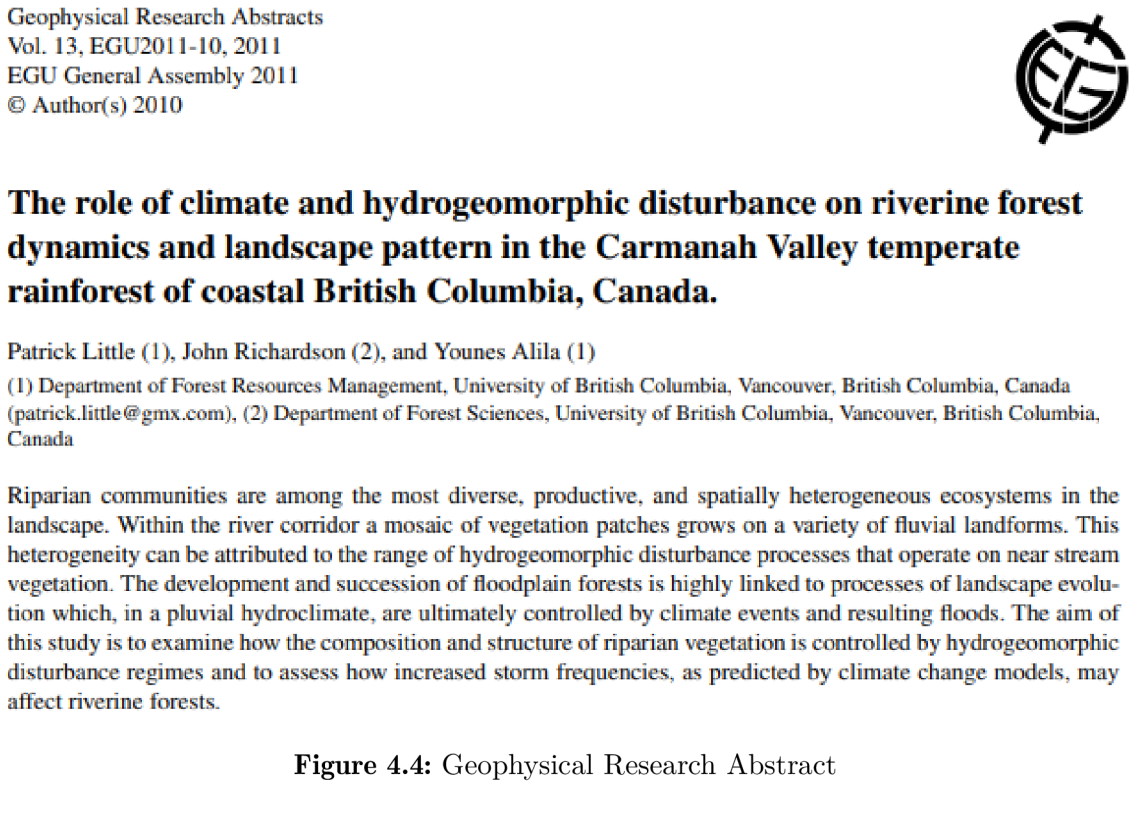
Figure 4.3: Abstracts count by year

Figure 4.3 shows the number of retrieved abstracts from ADS for each year. The data is available for the period from 2004 to 2021. However, data from 2004 to 2008 and 2011 are not indexed. In the scope of this research, we will work with the data from 2009 to 2021. The missing data for the year 2011 have been crawled from GRA website.

Pdfminer.six information extracting tool was used to parse PDF documents. Pdfminer.six is a community maintained fork of the original PDFMiner. Unlike other PDF-related tools, it focuses entirely on getting and analyzing text data. It is possible to obtain the exact location of text in a page and other information such as fonts or lines directly from the source code of the PDF. A PDF document

consists of a collection of objects and associated structural information that describes the appearance of one or more pages as a single self-contained sequence of bytes. PDFMiner.six attempts to reconstruct some of those structures by using heuristics on the positioning of characters. GRA has a logical structure where all PDF files consist of a header holding the information about GRA itself, license and id of the document, title, authors and their affiliation and the abstract. A total of 13787 entries have been obtained after data from GRA.

Author and affiliation in PDF files have the same font size and are identified as a single information block.



Geophysical Research Abstracts
Vol. 13, EGU2011-10, 2011
EGU General Assembly 2011
© Author(s) 2010

The role of climate and hydrogeomorphic disturbance on riverine forest dynamics and landscape pattern in the Carmanah Valley temperate rainforest of coastal British Columbia, Canada.

Patrick Little (1), John Richardson (2), and Younes Alila (1)

(1) Department of Forest Resources Management, University of British Columbia, Vancouver, British Columbia, Canada (patrick.little@gmx.com), (2) Department of Forest Sciences, University of British Columbia, Vancouver, British Columbia, Canada

Riparian communities are among the most diverse, productive, and spatially heterogeneous ecosystems in the landscape. Within the river corridor a mosaic of vegetation patches grows on a variety of fluvial landforms. This heterogeneity can be attributed to the range of hydrogeomorphic disturbance processes that operate on near stream vegetation. The development and succession of floodplain forests is highly linked to processes of landscape evolution which, in a pluvial hydroclimate, are ultimately controlled by climate events and resulting floods. The aim of this study is to examine how the composition and structure of riparian vegetation is controlled by hydrogeomorphic disturbance regimes and to assess how increased storm frequencies, as predicted by climate change models, may affect riverine forests.

Figure 4.4: Geophysical Research Abstract

Named-entity recognition (NER) from spaCy python library was used for classification of authors and affiliation.

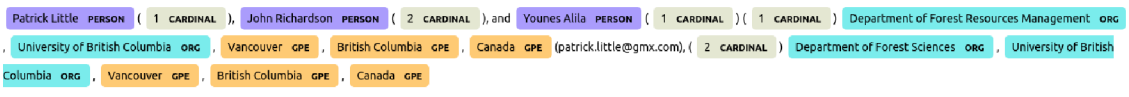


Figure 4.5: Named-entity recognition

No ground truth topic labels exist for this data set, so topic divisions and subdivisions were scraped for 2015 to 2021 with the BeautifulSoup4 python package. Beautiful Soup is a Python library for pulling HTML and XML files data. It provides an idiomatic way of navigating, searching, and modifying the parse tree. HTML (HyperText Markup Language) is the most fundamental building block of the Web, which defines the meaning and structure of web content. [33] The GRA website has

well defined HTML structure, so it was possible to extract abstracts ids, authors, title, type of presentation, division and subdivision by parsing HTML classes. A full list of all the division can be found in Appendix B.

The extracted data for 2011 was concatenated with data obtained from ADS, resulting in 192250 non-null values. After merging the divisions and subdivisions to the original data frame, only 176030 remained. Therefore, there are 94465 non-null entries for 2015 to 2021 and 81565 without predefined divisions for 2009 to 2014.

4.3 Pre-Processing

In order to execute the KDD process, a Python3 package was developed and all steps are documented and can be reproduced in the Jupyter Notebooks. The basic software and libraries used in this thesis are:

- Dependency manager and basic software: poetry, Jupyter Notebook
- Data Pre-Processing: pandas, spaCy, NumPy
- Data Transformation and Data Mining: gensim, scikit-learn
- Visualization: plotly

After loading the data sets, it is stored in a pandas data frame. First, the rows with null values and empty abstracts are dropped. Then, it is cleaned up and pre-processed before the EGU data set is converted into a machine-readable format. The conversion into tokens and the pre-processing is mainly done using the open-source libraries spaCy and regular expressions (re). spaCy is written in Python and Cython. SpaCy was chosen because of its claimed accuracy for the syntactic analysis and its high performance. First, a language model containing language-specific rules must be loaded for the tagging, parsing and entity recognition process. The library spaCy provides different pre-trained language models. The small-sized English model trained on written web text that includes syntax, vocabulary, entities and word vectors was used in this thesis. After loading the English model, an NLP object containing the processing pipeline (tagger, parser, ner) is received. During processing, spaCy first generates tokens and then separates words by whitespace characters and applies exception rules and prefix or suffix rules. Then, the adjectives, adpositions, pronouns, conjunctions, symbols, numerals, determiners, particles and spaces are removed. The removed POS classes are described in Appendix A. Typical examples of kept POS classes are nouns or adjectives, depending on the application. In the context of topic modelling, Martin and Johnson [34] showed that a nouns

only dataset produced the most meaningful topics. They suggest that reducing the articles to nouns may be advantageous since this improves the topics' semantic coherence and yields more interpretable topics. Next, each token is reduced to its lower case base form using the lemmatization tool provided in spaCy. Since stop words do not contain significant meaning, they are removed in the next step. Next, since some noise is usually present in the corpus after performing, the words with lengths less than three are removed. Finally, the processed abstracts are saved in a doc object stored in pandas DataFrame.

The bigrams, trigrams and quadgrams are identified and concatenated back to the dataframe with an underscore and considered a single word. Bigrams are phrases containing two words, like 'civil engineering', where 'civil' and 'engineering' are more likely to co-occur rather than appear separately. Trigrams are phrases containing three more likely co-occur, for example, 'vegetation index ndvi'. Likewise, quadgrams are occurrences like 'palmer drought severity index'. A pointwise Mutual Information (PMI) score was used to identify the top 1000 significant bigrams, trigrams and quadgrams that have a noun like structures and occur at least 50 times in the corpus. The randomly sampled ngrmas example is presented in table 4.3.

index	**bigrams**	**trigrams**	**quadgrams**
849	null hypothesis	heterogeneous porous medium	micro rain radar mrr
1181	auroral oval	ascend descend orbit	oceanographic data centres nodc
1016	campi flegrei	contrib mineral petrol	springer verlag berlin heidelberg
588	scanning radiometer	national centers environmental	yu explanation endogenous activity
1917	standard deviation	atmospheric sounding mipas	quantum cascade laser absorption
670	neutral atom	stratospheric polar vortex	gpp ecosystem respiration reco
674	degree celsius	indonesian tsunami early	irish ice sheet biis
1231	instituto dom	positive matrix factorization	directory thesaurus search tool
327	imaging spectroradiometer	absorption spectroscopy doas	akaike information criterion aic
292	bohemian massif	electrical resistivity tomography	extended kalman filter ekf

Table 4.3: EGU sampled ngrams

After pre-processing the text data, the features are generated. Before converting the corpus, which contains all documents into vectors, a mapping dictionary between each word in a document and a unique id must be generated. The open-source natural language processing library gensim, implemented in Python and Cython, is used to build the document vectors. Dictionary class gensim.corpora.dictionary is used to generate the BOW model. The function doc2bow() is used for representation using term frequency encoding, which counts the number of occurrences of each word, converts it to its integer id stored in the dictionary, and returns it as a sparse vector. In order to generate TF-IDF encoded tokens a corresponding model is build using models.TfidfModel(), unlike the regular corpus, TF-IDF downweights tokens that frequently appears across documents. TF-IDF is computed by multiplying a local component like term frequency (TF) with a global component, inverse document

frequency (IDF), and optionally normalizing the result to unit length. As a result, the frequently occurring words across documents will get downweighted.

4.4 Feature extraction

The Best Matching (BM25) function is a ranking function that ranks a group of documents depending on the keywords that appear in each document. In this thesis, the implemented Gensim library version 3.8.3 was used. The BM25 function obtains the score for each (word, document) pair to rank documents. This function is a family of scoring functions. The BM25 function is an information retrieval formula function, which belongs to the BM family of retrieval models, and determines the weight of a term t in document d. All documents are scored against the query, and only documents with a positive score remain in the corpus to reduce the number of features and improve topic modelling accuracy.

The time complexity of BM25 is $O(m \times avgdl)$, where m is the number of documents and $avgdl$ is the average document length. It is swift and produces good results.

4.5 Topic modelling

Unsupervised learning does not require upfront categorization work and provides a way to view broad patterns across large bodies of texts, patterns which could be used in subsequent research to create labelled data. Topic models generate a high-level overview of a body of literature, using word frequency and co-occurrence patterns to identify different topics or subjects of discourse. Topic modelling algorithms are optimized for information retrieval and summary problems. LDA and NMF algorithms were used to find patterns within a corpus, an algorithmic technique that groups words into an arbitrary number of topics based on the probability of their co-occurrence within documents. The documents are passed to the algorithm with no contextual or category information. Topic models provide extensive information describing their respective corpora, which can be used to identify patterns across the documents and identify content on particular themes.

The number of topics and some additional parameters controlling how the algorithm processes the data are required input information for LDA and NMF algorithms. The number of topics was set to 5, 10, 15, 20, 25 and 30, and the number of unique labels in the evaluation data is equal to the number of topics in each pass. The LDA topic model was trained with ten passes and a chunk size of 15,800, and

the NMF model was also trained with ten passes and a maximum of ten iterations per batch. For other hyper-parameters in LDA and NMF, the default values were used in the gensim package.

4.6 Topics over time

Graphs of topic prevalence over time are used to identify spikes and depict the relationship between the various topics in a corpus. However, topic prevalence over time is not a measure returned with the standard modelling tools such as LDA, NMF. Instead, it was computed by combining the model data with external metadata and aggregating the model results. The average of topic weights per year was calculated to compute topic significance over time. It is equally important to understand how these topics have changed over time. Given the potential utility of breaking topics down through time, it is possible to measure topic presence through time with the following steps:

- Extract individual document topic proportions as determined by the LDA or NMF models. Gensim LDA and NMF models can classify the specific relative proportions for all topics within each document.
- Calculate the yearly average given the entire sample of text documents for each topic.
- Visualize topic weights in a time-series plot.

From the model, the topic distribution for each given document, the normalised minimum probability is extracted. The average topic weight is computed by adding all of the weights for a given topic in a time period and divided by the total number of documents in that time period.[35] The document topic proportion weights from the model are extracted and merged with the original data frame. Then, the yearly time-series weights for each topic are created with the individual document topic proportion data. Then using the group by function in pandas, the yearly average for each topic is taken. The final data frame includes topic weight information for every unique document in corpus. For time-series visualization of topics from 2009 to 2021, the top 3 keywords that define the topic are concatenated.

4.7 Evaluation measures

Topic modelling presents unique challenges in that the algorithm is probabilistic, not deterministic. Therefore, the weights assigned to topics within documents

and words that constitute a topic will vary with each run on the same corpus. As a result, there is no "ground truth" in topic modelling techniques against which to evaluate the results. However, topic modelling techniques are good at identifying general themes and patterns in the corpus.[36, 37] A growing range of strategies is used to evaluate and improve the quality of topic models to make the results more stable and coherent. In addition, these strategies provide valuable mechanisms for increasing confidence that the model provides a useful abstraction of the underlying literature.

Intrinsic and extrinsic measures are usually used for evaluation of document clustering methods. Intrinsic measures, such as cluster separation and cohesion, do not require a ground truth label. Instead, such measures describe the variation within clusters and between clusters. However, they are dependent on the feature representations used, so they do not give comparable results for methods that use different feature sets. Extrinsic measures require a ground truth label but can be compared across methods. Standard extrinsic measures include precision, recall and F1, but these are dependent on the ordering of cluster labels to ground-truth labels.[38] Measures such as the mutual information and Rand index are more appropriate in this case as they are independent of the absolute values of the labels.

Perplexity and coherence are intrinsic evaluation metrics widely used for language model evaluation. Perplexity captures how surprised a model is of new data it has not seen before and is measured as the normalized log-likelihood of a held-out test set. In addition, the perplexity metric measures how probable some new unseen data is given the model learned earlier and well does the model represent or reproduce the statistics of the held-out data. The smaller the perplexity, the more precise is the model. However, recent studies have shown that predictive likelihood and human judgment are often not correlated.[39]

The concept of topic coherence combines several measures into a framework to evaluate the coherence between topics inferred by a model. Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between semantically interpretable topics and artefacts of statistical inference. C_v measure is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity. C_v topic coherence and human evaluation are highly correlated. Therefore, coherence measure can be used to compare difference topic models based on their human-interpretability. C_v topic coherence is essentially an index measure of the co-occurrence of the words extracted by the topic model. If those words from the same topic co-occur often, the model is well performed.

As a result, our research would apply C_v coherence as the evaluation measurement of our topic model.[39]

Mutual information measures the mutual dependence between two discrete random variables. It quantifies the reduction in uncertainty about one discrete random variable is given knowledge of another. High mutual information indicates a significant reduction in uncertainty. For two discrete random variables X and Y with joint probability distribution $p(x, y)$, the mutual information, $MI(X, Y)$, is given by

$$\mathrm{MI}(X, Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \tag{4.1}$$

A commonly used measure is the normalised mutual information (NMI), which normalises the MI to take values between 0 and 1, with 0 representing no mutual information and 1 being agreement. This is useful to compare results across methods and studies. NMI is given by

$$\mathrm{NMI(X, Y)} = \left(\frac{\mathrm{MI(X, Y)}}{\sqrt{\mathrm{H(X)H(Y)}}} \right) \tag{4.2}$$

where $H(X)$ and $H(Y)$ denote the marginal entropies, given

$$H(U) = -\sum_{i=1}^{n} P(x_i) \log(P(x_i)) \tag{4.3}$$

This value of the mutual information and also the normalized variant is not adjusted for chance and will tend to increase as the number of different labels (clusters) increases, regardless of the actual amount of "mutual information" between the label assignments.

The Rand index is a pair counting measure for the similarity between the labels and clusters. It also takes values between 0 and 1, 0 representing random labelling and 1 representing identical labels. Given a set of elements $S = o_1 \ldots, o_n$ and two partitions of S to compare, $X = X_1 \ldots, X_r$ and $Y = Y_1 \ldots, Y_s$, the Rand index represents the frequency of times the partitions X and Y are in agreement over the total number of observation pairs. Using the expected value, the adjusted mutual information can then be calculated using a similar form to that of the adjusted Rand index (RI). If X is a ground truth class assignment and Y the clustering a, the number of pairs of elements that are in the same set in X and in the same set in Y. b, the number of pairs of elements that are in different sets in X and in different sets in Y. The unadjusted Rand index is then given by:

$$\text{RI} = \frac{a+b}{S_2^{n_{samples}}} \tag{4.4}$$

For extrinsic clustering evaluation measures to be useful for comparison across methods and studies, such measures need a fixed bound and a constant baseline value. Both the NMI and the RI are scaled to have values between 0 and 1, so satisfy the first condition. However, it has been shown that both measures increase monotonically with the number of labels, even with an arbitrary cluster assignment.[38] This is because the mutual information and Rand index do not have a constant baseline, implying that these measures are not comparable across clustering methods with different clusters. Adjusted versions of the MI and RI have been proposed to account for this. The adjusted rand index, ARI, adjusts the RI by its expected value:

$$\text{ARI} = \frac{\text{RI(X,Y)} - E[\text{RI(X,Y)}]}{\max(\text{RI(X,Y)}) - E[\text{RI(X,Y)}]} \tag{4.5}$$

where where E[RI(X,Y)] denotes the expected value of $RI(X, Y)$. The ARI takes values between 0 and 1, with 1 representing identical partitions, and is adjusted for the number of partitions in X and Y. Using the expected value, the adjusted mutual information can then be calculated using a similar form to that of the adjusted Rand index:

$$\text{AMI} = \frac{\text{MI(X,Y)} - E[\text{MI(X,Y)}]}{\text{mean}(H(X), H(Y)) - E[\text{MI(X,Y)}]} \tag{4.6}$$

where E[MI(X,Y)] represents the expected value of the MI.[40] The AMI takes values between 0 and 1, with 1 representing identical partitions adjusted for the number of partitions used. The best measures to ensure comparative evaluations are the AMI and the ARI.

AMI is the preferable measure when the labels are unbalanced, and there are small clusters, while the ARI should be used when the labels have large and similarly sized volumes.[41] The AMI, ARI, and NMI measures are used in the thesis. Many previous studies have reported the NMI measure, so it was included it in evaluation for comparison purposes. Given the data and methods of this study, it is likely that the AMI is more appropriate than the ARI, as Table **??** and Figure 4.2 show that the distribution of documents across labels is unbalanced.

5 Results

This chapter presents the findings from the experiments of building and evaluating meaningful topic models. First, the results from the general clusterisation of abstracts with NMF and LDA algorithms will be given. Then, the following section presents the results of feature extraction and topic modelling results on the subset of data for one domain with a specified keyword with a Best Matching 25 algorithm. Next, the recent scientific research trends in geosciences over the last decade will be presented. Finally, the experiment results will be presented for other sets of abstracts.

5.1 General topic modelling of the abstracts

The EGU dataset is unbalanced in document size, measured by computing the number of tokens in each document.

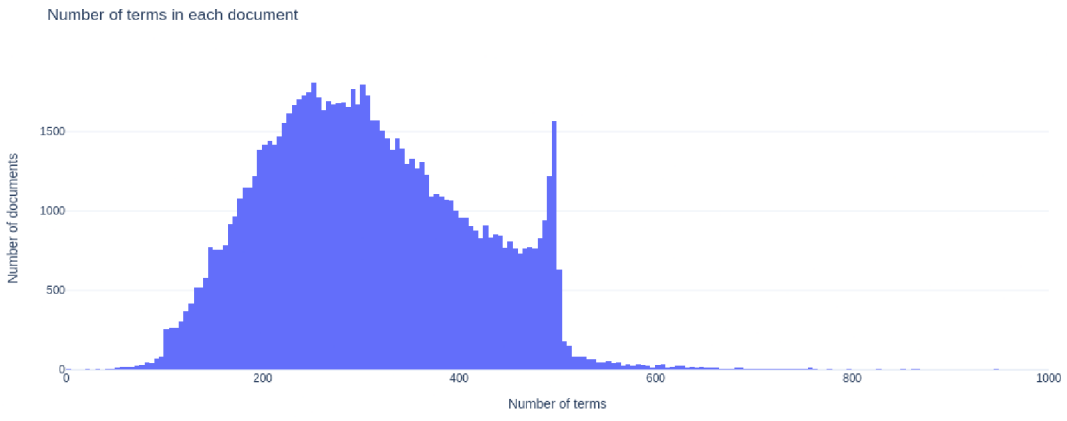


Figure 5.1: Number of terms in each document before preprocessing

As illustrated in Figure 5.2, the number of tokens contained in each document after preprocessing varies from 1 to 1537 per document. The average number of tokens in a document is 175. Compared to the number of tokens before preprocessing in Figure 5.1 the number of tokens varies from 2 to 2622, with an average of 309

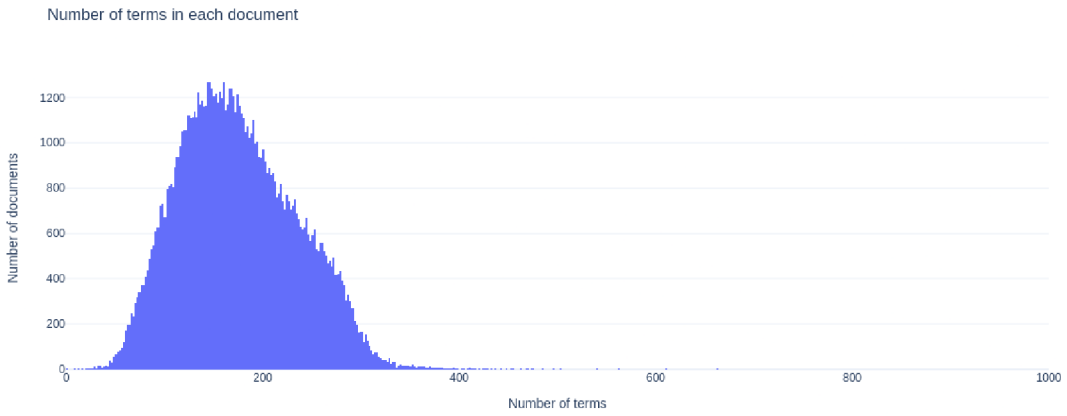


Figure 5.2: Number of terms in each document after preprocessing

tokens. Cleaning and preprocessing the tokens reduces the data set almost by half. A significant number of features are removed from the corpus mainly due to the presence of many stop words.

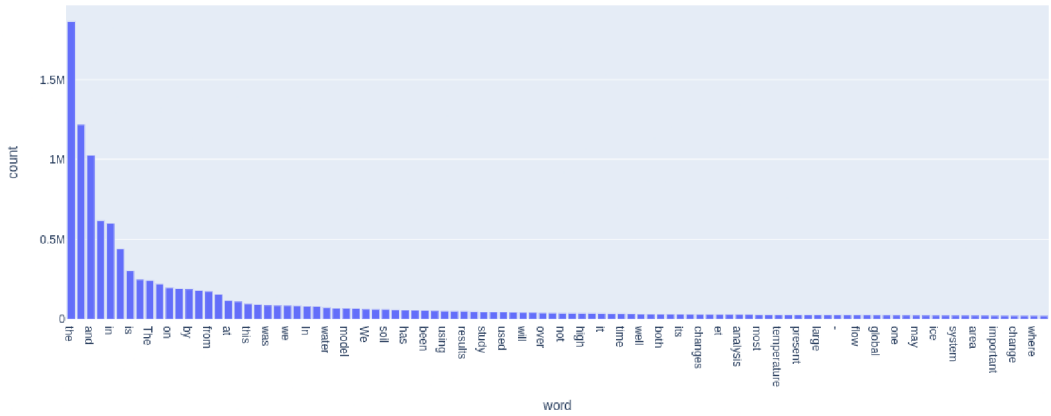


Figure 5.3: Top 100 tokens in corpus before preprocessing

The dataset before preprocessing contains many stop words, for example, *the, and, in* and many others.

The topic modelling techniques were applied to the dataset with preprocessed abstracts and abstracts containing bigrams, trigrams, and quadgrams with a different number of predefined topics. Some disadvantages of methods used in the topic modelling steps have already been presented in the chapters about theoretical basics. For example, the topic modelling result of not processed data is usually more inaccurate than the preprocessed data. Furthermore, TF-IDF should provide the best results for topic modelling. Therefore, the topic modelling results for all five data settings and the different representations and evaluation measures were generated based on the EGU data set with the different number of divisions to confirm

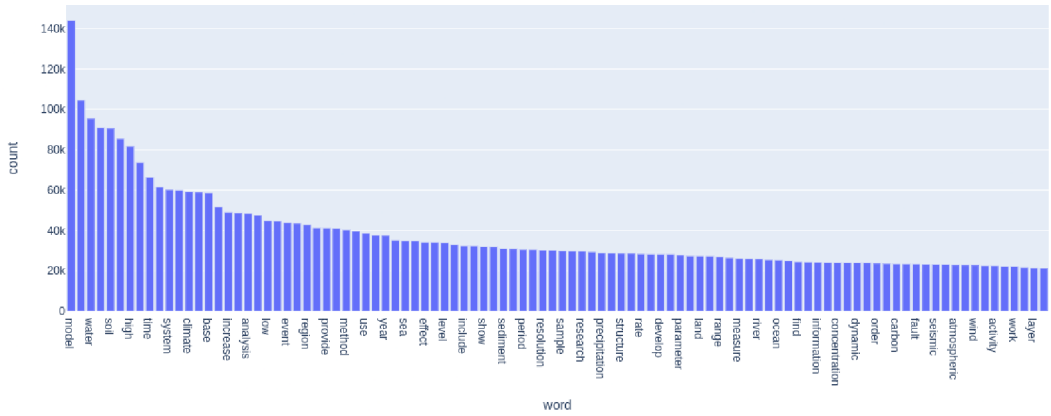


Figure 5.4: Top 100 tokens in corpus after preprocessing

or disprove these claims. In addition, the results were evaluated for topic modelling accuracy. Dirichlet hyperparameter for the document-topic density (alpha) and Dirichlet hyperparameter for the term-topic density (eta) produced the models with the most meaningful topics when they were set to auto, meaning the algorithm learns asymmetric priors from the corpus. The optimal number of passes was set to ten, with ten iterations for the LDA model. For the NMF model, the initial hyperparameter for passes was set to ten, initialisations of the W (w_max_iter) and H (h_max_iter) matrices were set to ten set to default values in these experiments.

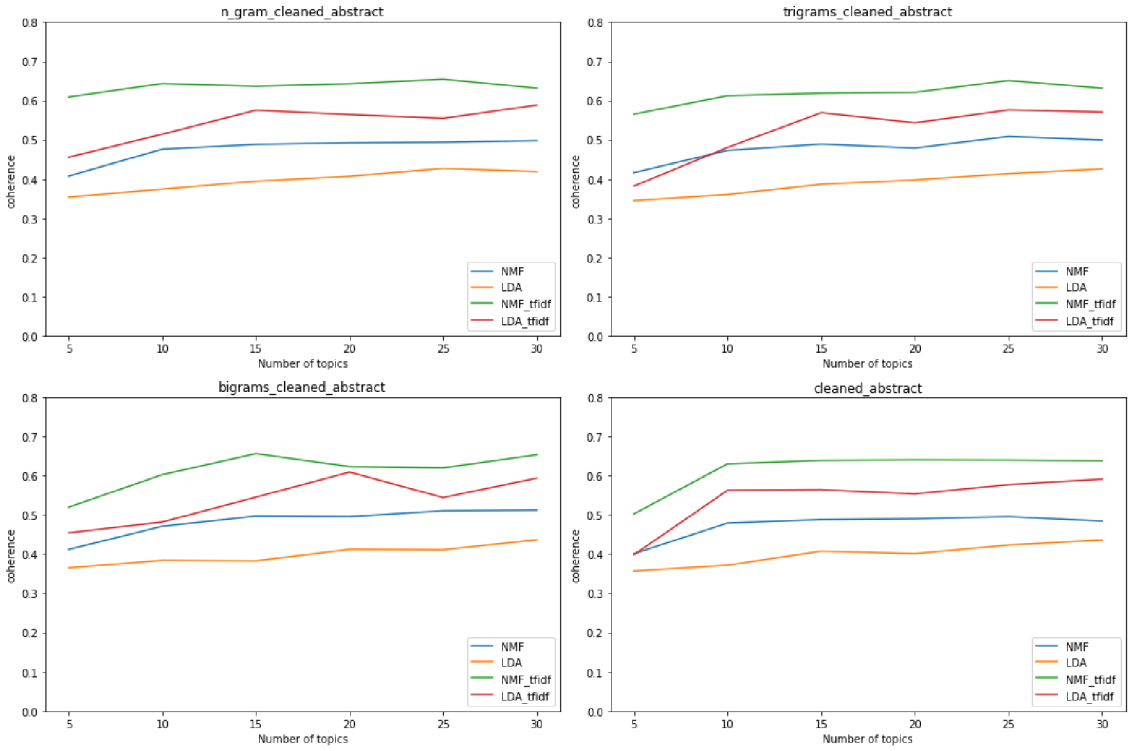


Figure 5.5: Coherence measure

It is evident from Figure 5.5 that TF-IDF BoW has a higher coherence score

than BoW structure for both LDA and NMF models. Furthermore, the combination of TF-IDF BoW and NMF has a higher score for each number of topics and a different number of ngrams. For example, the highest value is 15 topics with concatenated bigrams with a coherence of 0.656, closely followed by 25 and 30 topics with concatenated ngrams and concatenated bigrams with coherence values of 0.654 and 0.653, respectively.

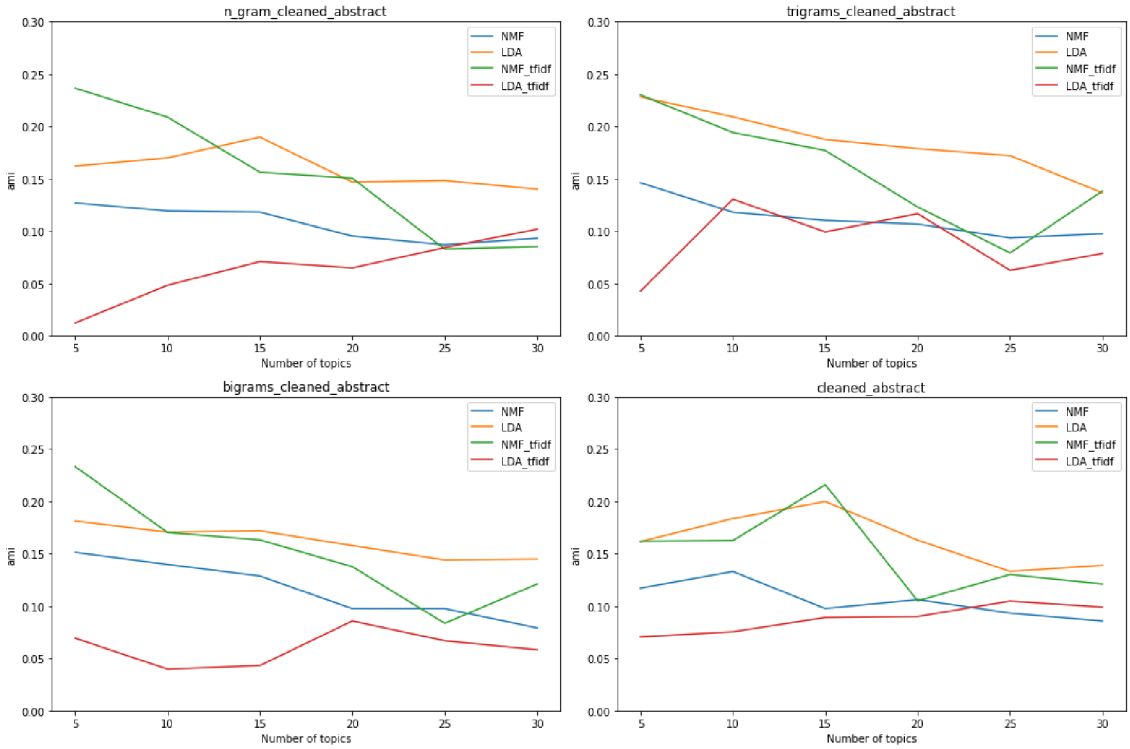


Figure 5.6: AMI measure

From Figure 5.6, we can see that the AMI score is slightly decreasing with an increasing number of topics. The best value of AMI is for five topics for TF-IDF BoW NMF model and quadgrams, which is 0.256.

The NMI measure with LDA Bow generally has a better score, closely followed by NMF TF-IDF BoW with the increasing number of topics. The best value is for 15, 25 and 30 topics for cleaned abstracts where NMI equals 0.34, 0.33 and 0.328

The ARI measure is the preferred measure where the labels have large volumes and are balanced.[41]. This dataset was relatively balanced (given in Table **??**), so the ARI is the more appropriate performance measurement than the NMI and AMI. The preprocessed abstracts and 15 topic numbers produced the best results with ARI values of 0.348 and 0.331, for TF-IDF BoW NMF and LDA models respectively. Interestingly, some methods had a relatively significant drop in score between the NMI and AMI measures, indicating that the chance adjustment of the AMI is essential.

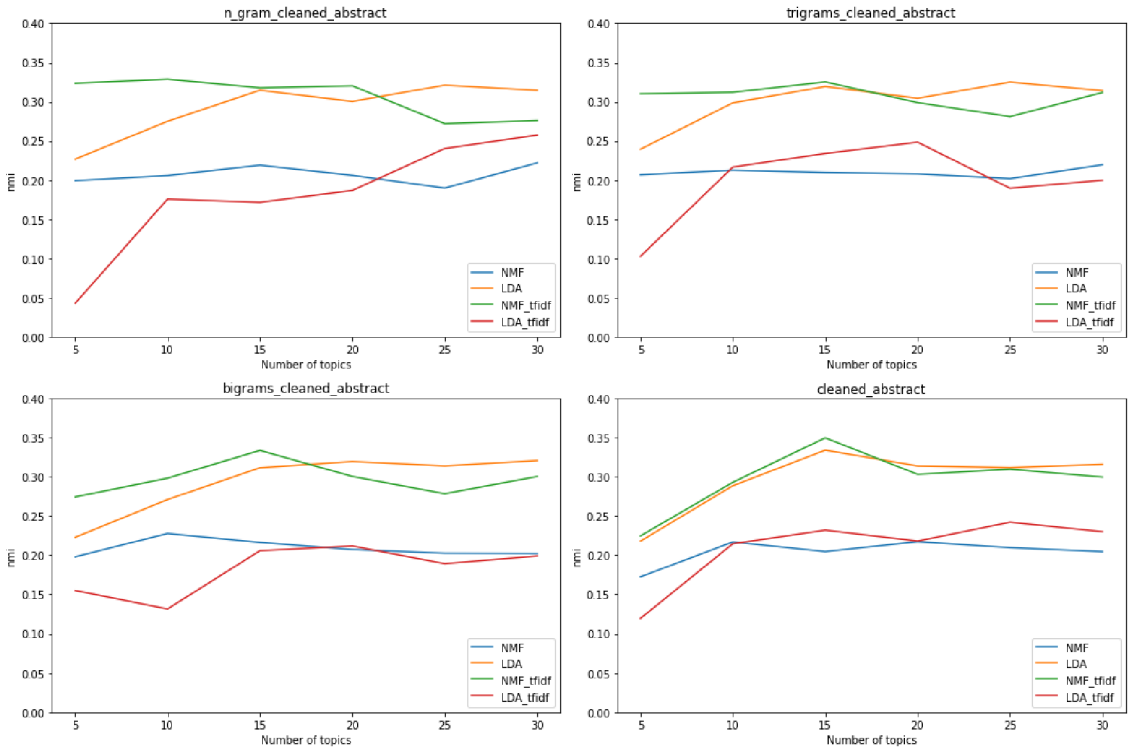


Figure 5.7: NMI measure

Topic models create topics of terms that frequently co-occur, which makes this topic reasonable, and it can still achieve high coherence scores as these metrics are also based on term co-occurrences. The NMF model was superior in producing coherent topics than the LDA model, especially on fewer topics, and TF-IDF BOW feature representation was better than BOW. The coherence scores for all models increases with an increasing number of topics. Topics from the LDA and NMF models learned from the dataset are presented. This selection of topics is presented to compare NMF and LDA models and show how the topics from the respective algorithms change when the feature representation changes. In addition, a few topics with the top five most heavily weighted terms were chosen as illustrative examples. A complete list of all topics for all models with feature representation with the best coherence score can be found in Appendix C.

Out of 30 topics from TF-IDF NMF model 8 random topics are sampled and presented in figure 5.9. Model produced semantically coherent topics. Topic 13, 9, 26 and 7 highly likely belong to Hydrological science division, topic 2 belongs to Ocean science division, topic 27 belongs to Soil System Sciences division. Topics offer clear semantic interpretation.

8 random topics are sampled and presented in figure 5.9 for TF-IDF LDA model. An interesting observation in the LDA topics is that topics are less interpretable despite high coherence score. This is because the TF-IDF LDA model tends to give shorter words more weight, leading to many keyword abbreviations.

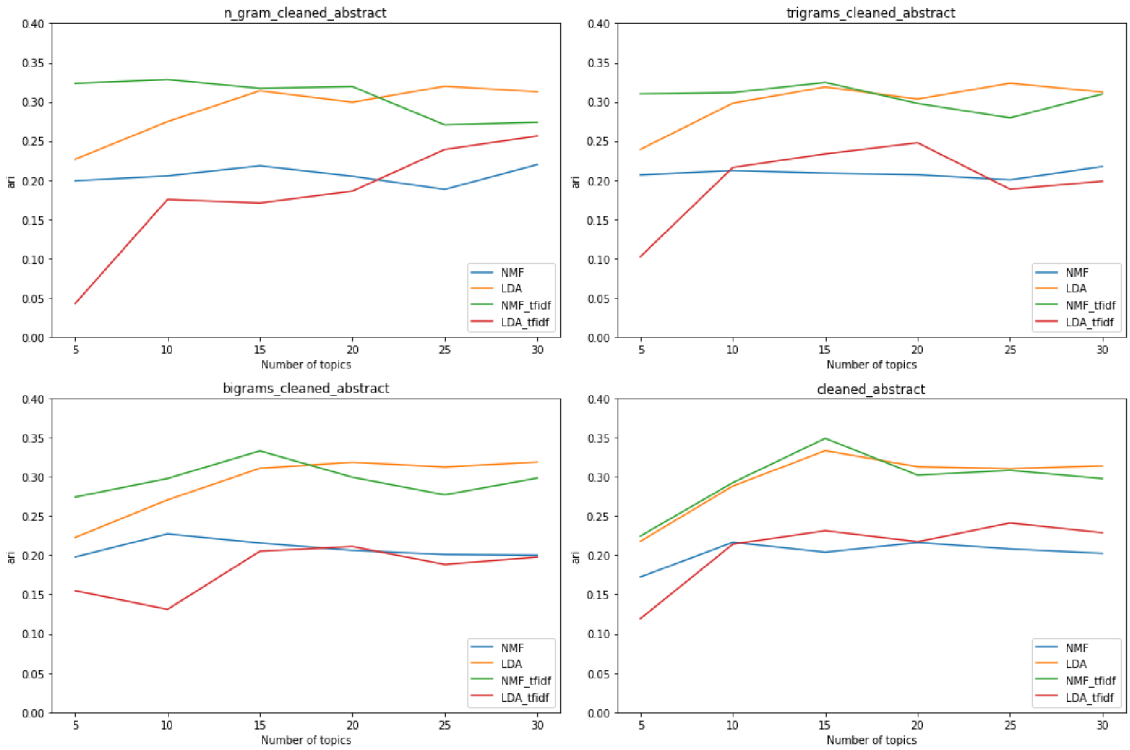


Figure 5.8: ARI measure

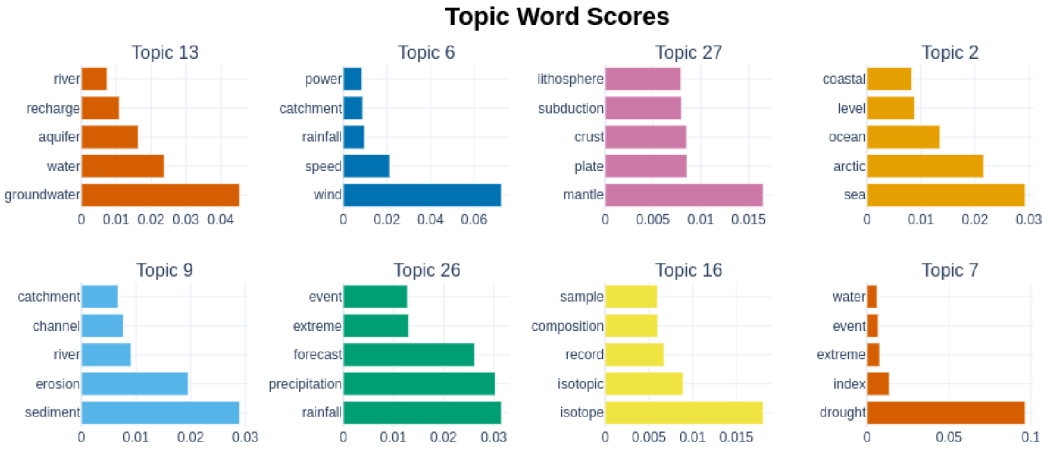


Figure 5.9: TF-IDF NMF topic example

TF-IDF NMF and LDA produced almost identical topics. However, there was some indication that NMF tended to produce more diverse topics than LDA through the experiments. NMF topics produced broader topics with attention to the concepts related to specific segments in the data, while LDA has many overlapping keywords that broadly fit the whole dataset, with less regard to specific patterns in smaller segments in the data. In contrast, TF-IDF LDA produces more topics that are hard to interpret because they tend to be too specific.

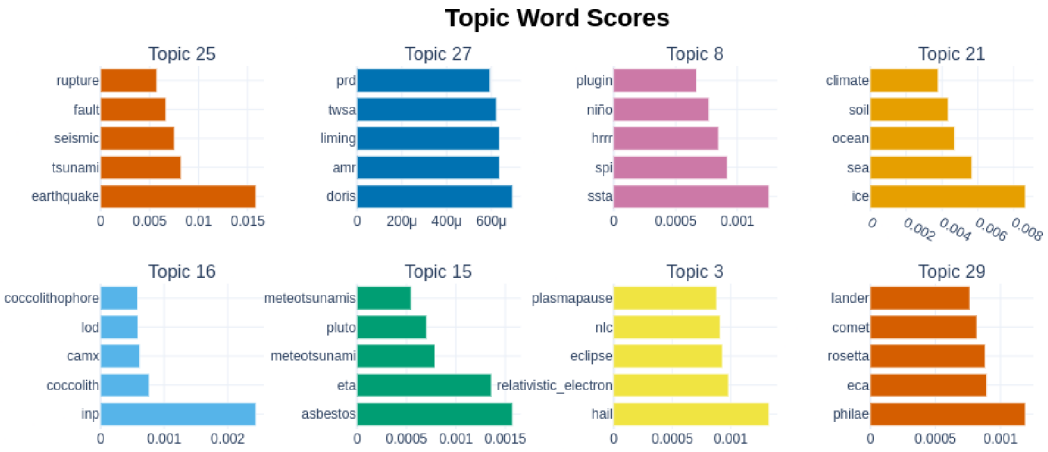


Figure 5.10: TF-IDF LDA topic example

5.2 Investigation of specific research discipline

The Hydrological Sciences division contain most of the abstracts, and it is evident from Figure fig:NMF topic example that potentially many of the topics fall under that division. Therefore, the test keyword ”drought” was chosen to investigate the specific research discipline in this broad HS division. After extracting features with the BM25 algorithm, the number of divisions has been reduced from 30 to 21 and subdivisions from 2286 to 634 from table 5.1 the comparison for the top 10 divisions is presented.

index	division	count	count after BM25 feature extraction
0	HS – Hydrological Sciences	6399	669
1	AS – Atmospheric Sciences	5033	229
2	SSS – Soil System Sciences	4008	222
3	CL – Climate: Past, Present, Future	3643	174
4	NH – Natural Hazards	3507	149
5	BG – Biogeosciences	2531	73
6	TS – Tectonics & Structural Geology	2327	50
7	OS – Ocean Sciences	1576	16
8	GM – Geomorphology	1541	15
9	GMPV – Geochemistry, Mineralogy, Petrology	1525	11
10	ERE – Energy, Resources and the Environment	1438	10

Table 5.1: Division of EGU2014-2021 subset before and after feature extraction

The optimal number of topics have been chosen by looking at the highest coherence score figure 5.11, which is equal to 0.544 for six topics.

The results for TF-IDF NMF model 6 topics are presented in figure 5.12.

One of the practical applications of topic modelling is to determine to which topic a given document belongs. First, the topic number with the highest percentage

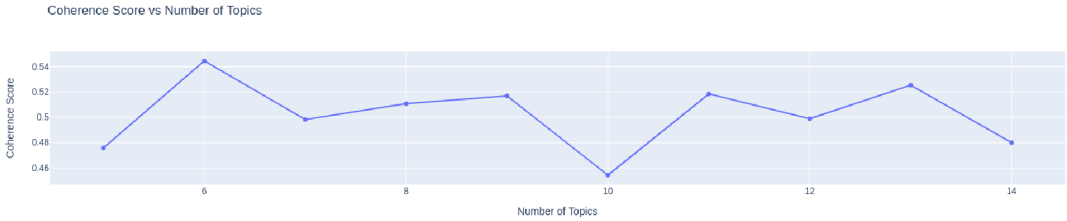


Figure 5.11: Coherence score vs number of topics

Figure 5.12: BM25 and TF-IDF NMF on subset

contribution in that document was discovered. Then, the documents with the highest contribution to identifying specified keywords in the topics of the model were classified. However, keywords may not be enough to make sense of a topic. Therefore, the next step is to find the documents a given topic has contributed to the most and infer the topic by reading that document. For example, in the research "Assessment of crop physical drought vulnerability in Sub-Saharan Africa" by Yang, Hong, Kamali, Bahareh and Abbaspour, Karim was identified to have the highest document topic probability to such keywords as "index", "crop", "yield", "precipitation" and "moisture".

	author	title	abstract	doc_topic_probabily	topic_keywords
348	[Yang, Hong, Kamali, Bahareh, Abbaspour, Karim]	Assessment of crop physical drought vulnerability in Sub-Saharan Africa	Crop yields exhibit known responses to droughts. However, quantifying crop drought vulnerability is often not straightforward, because it is interwoven with different components that are not all known on a practical spatial scale. This study aims to develop a physical Crop Drought Vulnerability Index (CDVI) through linking the Drought Exposure Index (DEI) with the Crop Failure Index (CFI) in Sub-Saharan Africa. Two different DEIs, namely DEIP and DEIR, were derived from cumulative distribution functions fitted to precipitation and residual of precipitation and potential evapotranspiration, respectively. The DEIP-X and DEIR-X were calculated for different time scales (i.e., X = 1, 3, 6, 9 and 12 months). Similarly, CFI was calculated by fitting a cumulative distribution function to maize yield simulated using the Environmental Policy Integrated Climate (EPIC) model. Using a power function, curves were fitted to CFI and DEI relations resulting in five different shapes, each explaining a specific class of vulnerability. The results indicated that in Central Africa the highest correlation was found between CFI and DEIR-1, while this was not the case for other parts of Africa, where CFI was strongly correlated to DEIP-3 and DEIP-6. Our findings show that some Southern African countries, the West-Sahelian strip, and parts of Eastern Africa are highly vulnerable to drought, whereas CDVI is low in Central Africa because of relatively high rainfall and rare occurrence of crop water stress. The proposed methodology provides complementary information on quantifying different degrees of vulnerabilities and can be applied to different regions and scales.	1.0	index, crop, yield, precipitation, moisture, agricultural, soil, spi, ndvi, spatial

Figure 5.13: Document topic probability

5.3 Topics over Time

The EGU dataset was preprocessed from 2009 to 2021, and topics were identified with the TF-IDF NMF model. The created dataset contains 176023 non-null entries. The overall view of the data is presented in table 5.2 by computing basic statistics.

Max:	1.0
Min:	0.103
Average:	0.453
Median:	0.423
Most frequent value:	0.384

Table 5.2: Summary Statistics

The topics range from 100% of the tokens in a document to 10%, with an average of 45% and a median value of 42%. However, the most frequent value is near 38%, indicating that the data predominantly describes topics with a significant presence in the documents. The average topic weight is computed by adding all of the weights for a given topic in a time period and dividing by the total number of documents in that time period, resulting in the average weight of the topic over all documents in the corpus presented in table 5.3.

index	year	topic_id	doc_topic_probability	total_docs	average_weight	topic_label
0	2009	0	154.057280	12525	0.012300	0_flow_debris_catchment
1	2010	0	144.260902	13780	0.010469	0_flow_debris_catchment
2	2011	0	24.655163	13789	0.001788	0_flow_debris_catchment
3	2012	0	140.567896	13414	0.010479	0_flow_debris_catchment
4	2013	0	137.582450	13220	0.010407	0_flow_debris_catchment
...	...	...	...	...	...	...
385	2017	29	197.330777	16073	0.012277	29_earthquake_seismic_tsunami
386	2018	29	193.575732	15422	0.012552	29_earthquake_seismic_tsunami
387	2019	29	157.828447	14617	0.010798	29_earthquake_seismic_tsunami
388	2020	29	61.478314	7520	0.008175	29_earthquake_seismic_tsunami
389	2021	29	109.662857	12012	0.009129	29_earthquake_seismic_tsunami

Table 5.3: The average weight of topics over time

The top five and bottom five topics determined by topic proportion within the aggregate corpus produced by TF-IDF NMF are visualised in Figure 5.14 and Figure 5.15 respectively. For example, topic number 15, which dominates in proportions for most of the time series and has been rising from 2017 to 2020, consists of the following keywords: 'model, datum, method, parameter, approach, information, system, user, uncertainty, data'. The keywords show that the topic holds general information and does not provide meaningful insights into data. By looking into the division data, topic 15 has 1190 documents from Hydrological Sciences and 1089 from Earth and Space Science Informatics, the rest of the divisions, are evenly distributed. Topic 11 with the top keywords "soil, moisture, content, erosion, organic, land, property, water, soc, agricultural' clearly belongs to the Soil System Science division. It has

been developing steadily over the years, with peak values in 2020. Topic 9 provides insightful information describing sea surface temperature (SST) in the Atlantic and Pacific oceans, and other keywords are 'variability, circulation, precipitation'. Topic 9 is highly likely to hold documents from Atmospheric Sciences, Climate: Past, Present, Future and Ocean Sciences. Topic 8 and 2 declined in recent years. Both topics hold documents from Tectonics, and Structural Geology division, where topic 2 focuses on Geochemistry, Mineralogy, Petrology and Volcanology with keywords 'rock, deformation, shear, stress, fracture, strain, grain, fluid, pressure, pore' and tropic 8 focuses on Geodynamics with the keywords 'mantle, crust, plate, subduction, lithosphere, crustal, continental, slab, margin, lithospheric'.

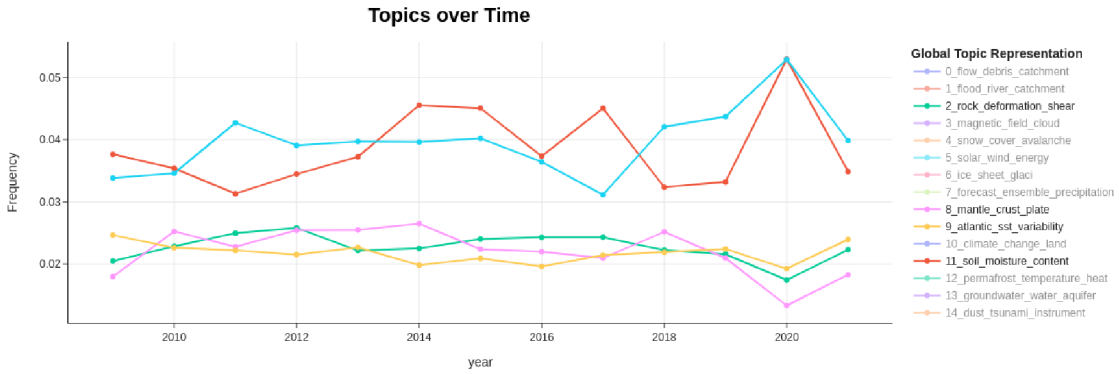


Figure 5.14: TF-IDF NMF topic trends top 5

Examining the bottom-5 topics yields interesting, albeit unsurprising, topic patterns over time. Although all five topics are from a division that contains a smaller number of research papers, it is worthy of attention that the TF-IDF NMF model was able to determine specific keywords. For example, topic 18, with the keywords 'lake, core, water, holocene, record, sediment, glacial, glaci, lacustrine, dam' highly likely describes Stratigraphy, Sedimentology & Palaeontology division. Figure 5.15 shows that division was increasing until the year 2016, then experienced a sharp decline in 2017, and from then, it continues to rise steadily. From the top keyword 'permafrost' in topic 12, it is evident that it belongs to the Cryospheric Sciences division. Keywords in Topic 4 do not define the division clearly, but it is apparent that the topic focuses on snow. Both topics 12 and 4 have been steady over the past years. Topic 16 belongs to the Natural Hazards division with the keywords 'fire, burn, forest, wildfire, vegetation, fuel, area, severity, post, emission'. Topic 14 has 'dust, tsunami, instrument, radar, mars, particle, satellite, emission, mission, mineral' keywords and is highly likely to belong to the Planetary & Solar System Sciences division. According to the results, topic 16 is the least discussed topic over the years, with the lowest value in 2011.

The top five and bottom five topics determined by topic proportion within the

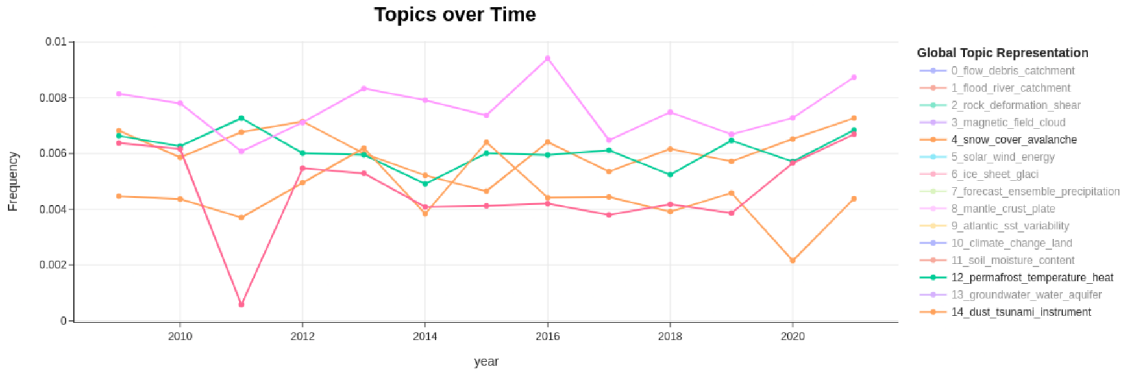


Figure 5.15: TF-IDF NMF topic trends bottom 5

aggregate corpus produced by LDA are visualised in Figure 5.16 and Figure 5.17 respectively. Again, same as in the TF-IDF NMF model, topic 29 dominates in proportions for most of the time series, except in 2020. The LDA model provided more descriptive keywords for the most dominant topic: 'model, precipitation, climate, datum, forecast, simulation, base, rainfall, resolution, scale' and topic highly likely to share the proportion of documents from Hydrological Sciences, Climate: Past, Present, Future and Atmospheric Sciences. Figure reffig:LDA topic trends top 5 shows that all topics rise steadily over the years except for topic 26, which had a sharp decline in the year 2020. Topic 21 contains keywords 'soil, organic, plant, sample, high, content, microbial, different, increase, matter' and falls under Soil System Sciences and Biogeosciences division. The top ten keywords in topic 14 were identified as 'flood, landslide, risk, area, event, hazard, water, impact, system, management'. The keywords in topic 14 describe Natural Hazards and Hydrological Sciences division. The LDA and TF-IDF NMF models identified the Hydrological Sciences and the Soil System Sciences divisions as the most dominant topics from 2009 to 2021. This is most likely because HS and SSS are two divisions with the highest number of abstracts.

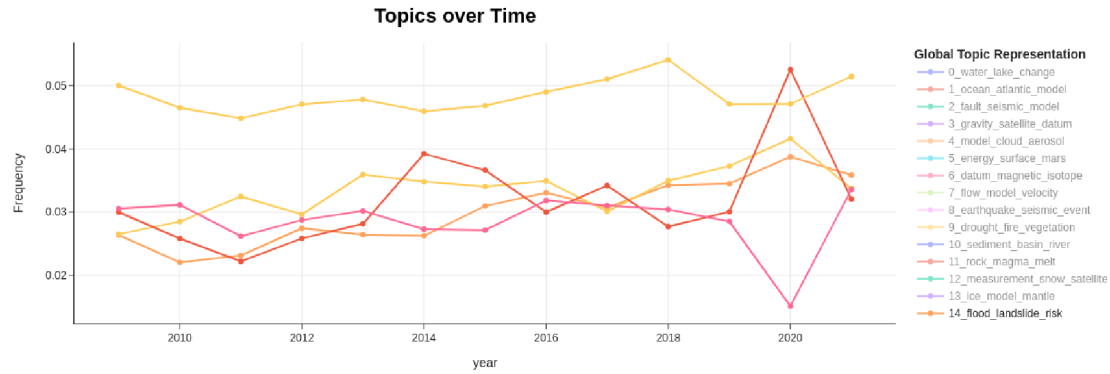


Figure 5.16: LDA topic trends top 5

Examining the bottom-5 topics for the LDA model, it is noticeable that all topics are steadily declining over the past years. For example, topic 13 belongs to the Geodynamics and the Cryospheric Sciences divisions with the keywords 'ice, model, mantle, sheet, subduction, plate, slab, lithosphere, continental, shelf'. Interestingly, the TF-IDF NMF model captured the same GD and CR divisions trends. The steepest decline has topic 5 with the keywords 'energy, surface, mars, planet, crater, solar, system, earth, planetary, mission', which describes the Planetary & Solar System Sciences (PS) division. Again, it is noticeable that TF-IDF NMF and LDA models captured the same trend with the decline of the PS division.

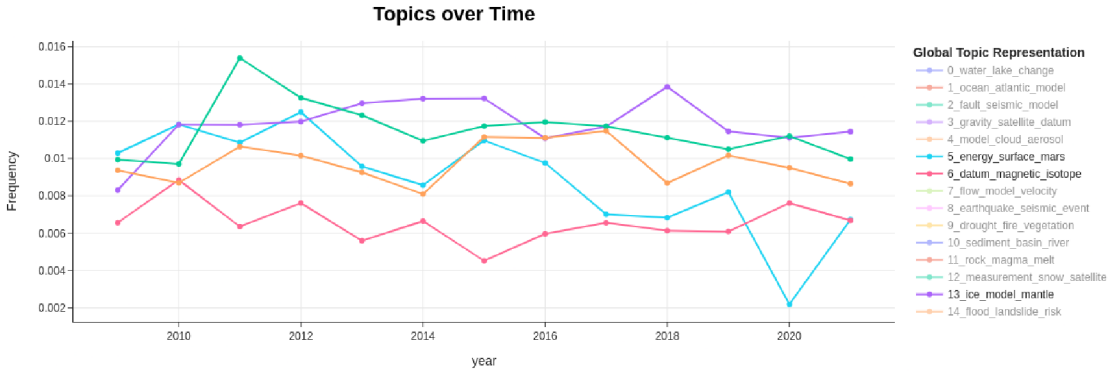


Figure 5.17: LDA topic trends bottom 5

6 Discussion

This chapter discusses the results of the experiments in this thesis. The application of topic modelling for generating topics to segment, exploration and description the Geophysical Research Abstracts (GRA) has been studied in this thesis. The effectiveness of the topic modelling is assessed by the ability to serve as a uniform categorization framework for research papers and by showing that algorithms can generate meaningful topics and keywords. Supported by this context, this thesis aims to understand the validity, viability of use, and limitations of such a topic model.

The experiment results on the EGU dataset verified that data preparation is essential for the successful application of topic modelling, and it is crucial for producing coherent topics and accurately assigning research papers to them. The first step in the data preprocessing step is dimensionality reduction, which is done by filtering out infrequent terms and removing stopwords. The results show that the corpus was significantly reduced, which helps produce less noisy topics affecting topic quality and increased computations performance. However, these findings are not explicitly stated in the results and were found after conducting experiments on the topic modelling algorithms. Lemmatisation is a powerful technique that reduces dimensionality and provides human-readable and interpretable terms. The nouns were selected using the spaCy part of speech (POS) library because nouns are the most relevant POS class in topics, both from topic modelling algorithms and human domain experts, as indicated in previous research. [34, 7]. Ngrams is another powerful feature selection technique, as indicated in research by Mikolov et al.[42] It was found that forming bigrams produces the best coherence score for all models and a different number of topics. Results show that ngrams can be a valuable data preparation method and can produce a valuable result and increase the coherence of topics. However, it is not recommended to form quadgrams because it is computationally expansive and does not increase the topics' interpretability because ngrams were rarely present in the top ten keywords. For both LDA and NMF models, the data was represented as a Bag of Words (BOW), and TF-IDF weighed BOW. The TF-IDF representation weighs uncommon terms higher, adding more prior information than the more straightforward term frequency counts for both LDA and NMF.

Results show that the TF-IDF LDA model produced many keywords focused on the short terms, mostly accounting for the abbreviations. LDA is a generative method that samples common terms in a corpus, conditioned on terms that frequently occur together to create representative topics that mirror the corpus in a probabilistic sense. NMF is a dimensionality reduction technique aimed at finding a lower subspace that accurately describes the most significant and diverse patterns in the data, limited by the number of dimensions or topics it can factorise the corpus into[16]. This emphasis in the thesis is to show the difference in the resulting topics between LDA and NMF. LDA generally delivers more stable and coherent topics than NMF, as stated in the research by Stevens et al.[43] and Mifrah S. and Benlahmar E. [44]. This thesis confirms that the LDA model is more stable in topic coherence than NMF. On average, however, NMF outperforms LDA in the sense of coherence, especially for models of a small number of topics. NMF is better at classifying topics, which is indicated by the findings through ARI, AMI and NMI metrics. A significant finding through analysing topics over time is that NMF tends to find topics that represent specific, distinct patterns in data segments, while LDA has generated more overlapping topics but with more meaningful keywords. Therefore, it is less likely that LDA will generate topics relevant only to specific patterns and data segments.

Depending on the use cases, a topic model can extract a different number of topics. All evaluation metrics are intended to increase the topic's interpretability by humans. Therefore, a higher K value in topics will give more granular results, while fewer topics will result in broader topics. It is possible to find an optimal number of topics by optimising coherence metrics. The main finding in the thesis indicates that with a low number of topics, NMF and TF-IDF weighted NMF produced more coherent results, while LDA and TF-IDF weighted LDA produced more interpretable results at around 15 topics with the peak values at 30 topics. Running models on research papers with many divisions generally produced topics with a broader span of concepts, resulting in less granular topics on specific concepts. The research should decide the level of granularity before application. The Best Matching (BM25) should be applied first to explore a specific topic and retrieve relevant documents related to the query keywords, and then the number of topics K can be chosen following the same logic. The query example keyword was "drought" in the results, which yielded documents mainly related to the Hydrological Science (HS) division, thus resulting in highly granular topics in HS subject, reducing the model's conceptual span.

Topics models are data-dependent, meaning that the characteristics of the data determine the results of the topic model. The concepts related to topics are determined by how these concepts co-occur in research papers. LDA and NMF algorithms behave differently in this regard. However, both models generalise well when learned with datasets made up of multiple divisions. Exploring trends in topics over time,

it was found that topics were meaningfully interpretable and did not degrade quantitatively. The LDA and TF-IDF NMF models identified the same trends where Hydrological Sciences and the Soil System Sciences divisions are the most dominant topics and Planetary & Solar System Sciences (PS) division is less dominant from 2009 to 2021, which means that the diversity of the topics and concepts is determined by how diverse the research abstracts in the corpus are.

Given the results, it is clear that topic modelling is a powerful method for deriving insights from a tremendous amount of research papers. Furthermore, being an unsupervised machine learning method, topic modelling does not require predefined labels. Topic models use unlabeled data as input which is one of its key strengths, but this can also be a significant weakness as it gives uncontrollable models. As a result, the topic modelling sometimes yields incorrect topics with little semantic meaningfulness, meaning that automated methods could not replace manual analysis. However, it can complement other methods for content analysis or categorisation, and it is a powerful method for aggregating and presenting the results to generate insights for efficiently analysing and segmenting research papers. The thesis results confirm that many unexpected topics were formed, but they still provided valuable results for understanding the content of the research papers. Results also confirm that topic modelling is a suitable tool for investigating trends, and it is possible to derive insights from the research paper abstracts that contributed most to the formation of the specific keywords describing the topic.

7 Conclusion

It is challenging to evaluate topic models objectively because evaluation methods are tightly correlated with human judgment. Furthermore, the data on which topics models are learned is the primary determinant of the topic's usefulness, so data acquisition and preprocessing are paramount in topic modelling applications. Therefore, a flexible data preparation framework is required to apply models successfully.

The LDA and NMF models produced different topic distributions on the same datasets. LDA tends to mirror the entire dataset better with more topics, while NMF finds and represents specific patterns in the dataset to capture more variations and performs well with a different number of topics. There is no optimal choice between the two algorithms, and depending on the use case, the researcher should consider the different strengths and weaknesses of the models. The same concept applies to the number of searched topics, as the granularity level in topics is often more important than higher metric scores. The conclusion topics created by topic models, compared to manual methods, are more unreliable and could produce misleading results, which is an effect of its uncontrollable nature. On the other hand, its strength is the ability to analyze large amounts of text in a short time and at a low cost, deriving insights from many research papers. Topic modelling can complement other content analysis or categorization methods, and it is a powerful method for aggregating and presenting the results of research paper abstracts.

The thesis has some unanswered questions which hold potential for future research. For example, the degree to which the TF-IDF representation determines the differences between LDA and NMF is unclear. In addition, the BOW representation used by LDA and NMF has some limitations, and it would be interesting to evaluate different data representations to enhance topic modelling results.

Bibliography

1. XIA, Feng; LIU, Haifeng; LEE, Ivan; CAO, Longbing. Scientific Article Recommendation: Exploiting Common Author Relations and Historical Preferences. *IEEE Transactions on Big Data*. 2016, roč. 2, č. 2, pp. 101–112. Available from DOI: 10.1109/TBDATA.2016.2555318.

2. BAFNA, Prafulla; PRAMOD, Dhanya; VAIDYA, Anagha. Document clustering: TF-IDF approach. In: 2016, pp. 61–66. Available from DOI: 10.1109/ICEEOT.2016.7754750.

3. PATRA, Braja Gopal; MAROUFY, Vahed; SOLTANALIZADEH, Babak; DENG, Nan; ZHENG, W. Jim; ROBERTS, Kirk; WU, Hulin. A content-based literature recommendation system for datasets to improve data reusability – A case study on Gene Expression Omnibus (GEO) datasets. *Journal of Biomedical Informatics*. 2020, roč. 104, p. 103399. ISSN 1532-0464. Available from DOI: https://doi.org/10.1016/j.jbi.2020.103399.

4. ZHU, Jie; PATRA, Braja G.; YASEEN, Ashraf. Recommender system of scholarly papers using public datasets. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*. 2021, roč. 2021, pp. 672–679. ISSN 2153-4063. Available also from: https://pubmed.ncbi.nlm.nih.gov/34457183. 3477705[PII].

5. BLEI, David M. Probabilistic Topic Models. *Commun. ACM*. 2012, roč. 55, č. 4, pp. 77–84. ISSN 0001-0782. Available from DOI: 10.1145/2133806.2133826.

6. GRIMMER, Justin; STEWART, Brandon M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*. 2013, roč. 21, č. 3, pp. 267–297. Available from DOI: 10.1093/pan/mps028.

7. JACOBI, Carina; ATTEVELDT, Wouter; WELBERS, Kasper. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*. 2015, roč. 4, pp. 1–18. Available from DOI: 10.1080/21670811.2015.1093271.

8. SURJANDARI, Isti; ROSYIDAH, Asma; ZULKARNAIN, Zulkarnain; LAOH, Enrico. Mining Web Log Data for News Topic Modeling Using Latent Dirichlet Allocation. In: 2018, pp. 331–335. Available from DOI: 10.1109/ICISCE.2018.00076.

9. LIU, Qian; CHEN, Qiuyi; BA; SHEN, Jiayi; WU, Mbbs; SUN, Mbbs; MING, Wai-Kit. data analysis and visualization of newspaper articles on thirdhand smoke. 2019.

10. BLEI, David M. Probabilistic topic models. *IEEE Signal Process. Mag.* 2012, roč. 27, pp. 55–65.

11. KIM, Sang-Woon; GIL, Joon-Min. Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences*. 2019, roč. 9, č. 1, p. 30. ISSN 2192-1962. Available from DOI: 10.1186/s13673-019-0192-7.

12. AGGARWAL, Charu C; ZHAI, ChengXiang. A survey of text classification algorithms. In: *Mining text data*. Springer, 2012, pp. 163–222.

13. JURAFSKY, D.; MARTIN, J.H.; NORVIG, P.; RUSSELL, S. *Speech and Language Processing*. Pearson Education, 2014. ISBN 9780133252934. Available also from: https://books.google.cz/books?id=Cq2gBwAAQBAJ.

14. LABOUVE, Eric. RELEVANCE ANALYSIS FOR DOCUMENT RETRIEVAL. 2019.

15. KNOTH, Stefanie. *Topic Explorer Dashboard : A Visual Analytics Tool for an Innovation Management System enhanced by Machine Learning Techniques*. 2020. MA thesis. Linnaeus University, Department of computer science a media technology (CM).

16. SVENSSON, Karin; BLAD, Johan. *Exploring NMF and LDA Topic Models of Swedish News Articles*. 2020. UPTEC STS, č. 20037. ISSN 1650-8319.

17. MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. *Introduction to Information Retrieval*. Cambridge University Press, 2008. Available from DOI: 10.1017/CBO9780511809071.

18. ROBERTSON, Stephen E.; ZARAGOZA, Hugo. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 2009, roč. 3, pp. 333–389.

19. FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. *AI Mag.* 1996, roč. 17, pp. 37–54.

20. JAYARAMAN, Dhivya. N-gram based keyword topic modelling for Canadian Longitudinal Study on Aging survey data. 2018.

21. JIANG, Yichen; JIA, Aixia; FENG, Yansong; ZHAO, Dongyan. Recommending academic papers via users' reading purposes. *RecSys'12 - Proceedings of the 6th ACM Conference on Recommender Systems*. 2012. Available from DOI: 10.1145/2365952.2366004.

22. WALKER, Daniel; LUND, William; RINGGER, Eric. Evaluating Models of Latent Document Semantics in the Presence of OCR Errors. In: 2010, pp. 240–250.

23. DEERWESTER, Scott C.; DUMAIS, Susan T.; LANDAUER, Thomas K.; FURNAS, George W.; HARSHMAN, Richard A. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* 1990, roč. 41, pp. 391–407.

24. CASALINO, Gabriella; MENCAR, Corrado; NICOLETTA, Del. Non Negative Matrix Factorizations for Intelligent Data Analysis. In: 2016, pp. 49–74. ISBN 978-3-662-48330-5. Available from DOI: 10.1007/978-3-662-48331-2_2.

25. SHAHNAZ, Farial; BERRY, Michael W.; PAUCA, V.Paul; PLEMMONS, Robert J. Document clustering using nonnegative matrix factorization. *Information Processing & Management*. 2006, roč. 42, č. 2, pp. 373–386. ISSN 0306-4573. Available from DOI: https://doi.org/10.1016/j.ipm.2004.11.005.

26. HOFFMAN, Matthew D.; BLEI, David M.; BACH, Francis R. Online Learning for Latent Dirichlet Allocation. In: *NIPS*. 2010.

27. KUANG, Da; CHOO, Jaegul; PARK, Haesun. Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering. In: 2014.

28. WEHLE, Hans-Dieter. Machine Learning, Deep Learning, and AI: What's the Difference? In: 2017.

29. BEUMER, Lisa. *Evaluation of Text Document Clustering Using K-Means*. 2020. Available also from: http://search.proquest.com.infozdroje.czu.cz/dissertations-theses/evaluation-text-document-clustering-using-em-k/docview/2407620660/se-2?accountid=26997. PhD thesis. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2021-05-20.

30. TRANI, Olivia. *First Timers' Guide to the EGU General Assembly* [https://cdn.egu.eu/static/latest/meetings/ga/EGU-General-Assembly-First-Timers-Guide.pdf]. [B.r.]. Accessed: 5-3-2022.

31. *How to vEGU – Networking (part 3): Early Career Scientist networking events!* 2021. Available also from: https://blogs.egu.eu/geolog/2021/04/13/how-to-vegu-early-career-scientist-newtorking-events/.

32. *About ADS* [https://ui.adsabs.harvard.edu/about/]. [B.r.]. Accessed: 10-3-2022.

33. RICHARDSON, Leonard. Beautiful soup documentation. *April*. 2007.

34. MARTIN, Fiona; JOHNSON, Mark. More Efficient Topic Modelling Through a Noun Only Approach. In: *ALTA*. 2015.

35. WIERINGA, Jeri. A Gospel of Health and Salvation: Modeling the Religious Culture of Seventh-day Adventism, 1843-1920. In: 2019.

36. BELFORD, Mark; MAC NAMEE, Brian; GREENE, Derek. Stability of Topic Modeling via Matrix Factorization. *Expert Syst. Appl.* 2018, roč. 91, č. C, pp. 159–169. ISSN 0957-4174. Available from DOI: 10.1016/j.eswa.2017.08.047.

37. MIMNO, David; WALLACH, Hanna M.; TALLEY, Edmund; LEENDERS, Miriam; MCCALLUM, Andrew. Optimizing Semantic Coherence in Topic Models. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 262–272. EMNLP '11. ISBN 9781937284114.

38. CURISKIS, Stephan A.; DRAKE, Barry; OSBORN, Thomas R.; KENNEDY, Paul J. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing & Management*. 2020, roč. 57, č. 2, p. 102034. ISSN 0306-4573. Available from DOI: https://doi.org/10.1016/j.ipm.2019.04.002.

39. RÖDER, Michael; BOTH, Andreas; HINNEBURG, Alexander. Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. 2015.

40. VINH, Nguyen Xuan; EPPS, Julien; BAILEY, James. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*. 2010, roč. 11, pp. 2837–2854.

41. ROMANO, Simone; VINH, Nguyen Xuan; BAILEY, James; VERSPOOR, Karin. Adjusting for chance clustering comparison measures. *The Journal of Machine Learning Research*. 2016, roč. 17, č. 1, pp. 4635–4666.

42. MIKOLOV, Tomas; SUTSKEVER, Ilya; CHEN, Kai; CORRADO, G.s; DEAN, Jeffrey. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*. 2013, roč. 26.

43. STEVENS, Keith; KEGELMEYER, Philip; ANDRZEJEWSKI, David; BUTTLER, David. Exploring Topic Coherence over many models and many topics. In: 2012.

44. MIFRAH, Sara; BENLAHMAR, EL Habib. Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus. *International Journal of Advanced Trends in Computer Science and Engineering*. 2020. Available from DOI: 10.30534/ijatcse/2020/231942020.

List of Figures

List of Tables

Appendices

A Part of Speech (POS) tags

POS	DESCRIPTION	EXAMPLES
ADJ	adjective	big, old, green, incomprehensible, first
ADP	adposition	in, to, during
ADV	adverb	very, tomorrow, down, where, there
AUX	auxiliary	is, has (done), will (do), should (do)
CONJ	conjunction	and, or, but
CCONJ	coordinating conjunction	and, or, but
DET	determiner	a, an, the
INTJ	interjection	psst, ouch, bravo, hello
NOUN	noun	girl, cat, tree, air, beauty
NUM	numeral	1, 2017, one, seventy-seven, IV, MMXIV
PART	particle	's, not,
PRON	pronoun	I, you, he, she, myself, themselves, somebody
PROPN	proper noun	Mary, John, London, NATO, HBO
PUNCT	punctuation	., (,), ?
SCONJ	subordinating conjunction	if, while, that
SYM	symbol	
VERB	verb	run, runs, running, eat, ate, eating
X	other	sfpksdpsxmsa
SPACE	space	

Table 1: part-of-speech tag describtion

B EGU General Assembly Divisions

Division	Number of abstracts
HS – Hydrological Sciences	12758
AS – Atmospheric Sciences	10183
SSS – Soil System Sciences	7958
CL – Climate: Past, Present, Future	7132
NH – Natural Hazards	7071
BG – Biogeosciences	5120
TS – Tectonics & Structural Geology	4704
OS – Ocean Sciences	3181
GM – Geomorphology	3054
GMPV – Geochemistry, Mineralogy, Petrology	3035
ERE – Energy, Resources and the Environment	2829
CR – Cryospheric Sciences	2793
ST – Solar-Terrestrial Sciences	2720
SSP – Stratigraphy, Sedimentology & Palaeontology	2532
SM – Seismology	2481
GD – Geodynamics	2464
PS – Planetary & Solar System Sciences	2355
NP – Nonlinear Processes in Geosciences	2327
G – Geodesy	2145
GI – Geosciences Instrumentation & Data Systems	1796
ITS – Inter- and Transdisciplinary Sessions	1650
EMRP – Earth Magnetism & Rock Physics	1510
ESSI – Earth & Space Science Informatics	1342
EOS – Education and Outreach Sessions	613
EOS – Educational and Outreach Symposia	456
US – Union Symposia	117
ML – Medal Lectures	72
MAL – Medal and Award Lectures	45
SEV – Side events	8
SCS – Science and Society	6
KL – Keynote Lectures	4
GL – Lectures for a general geoscience audience	2

Table 2: Count of abstracts per division in the EGU 2015-2021 dataset.

C Top keywords for different models

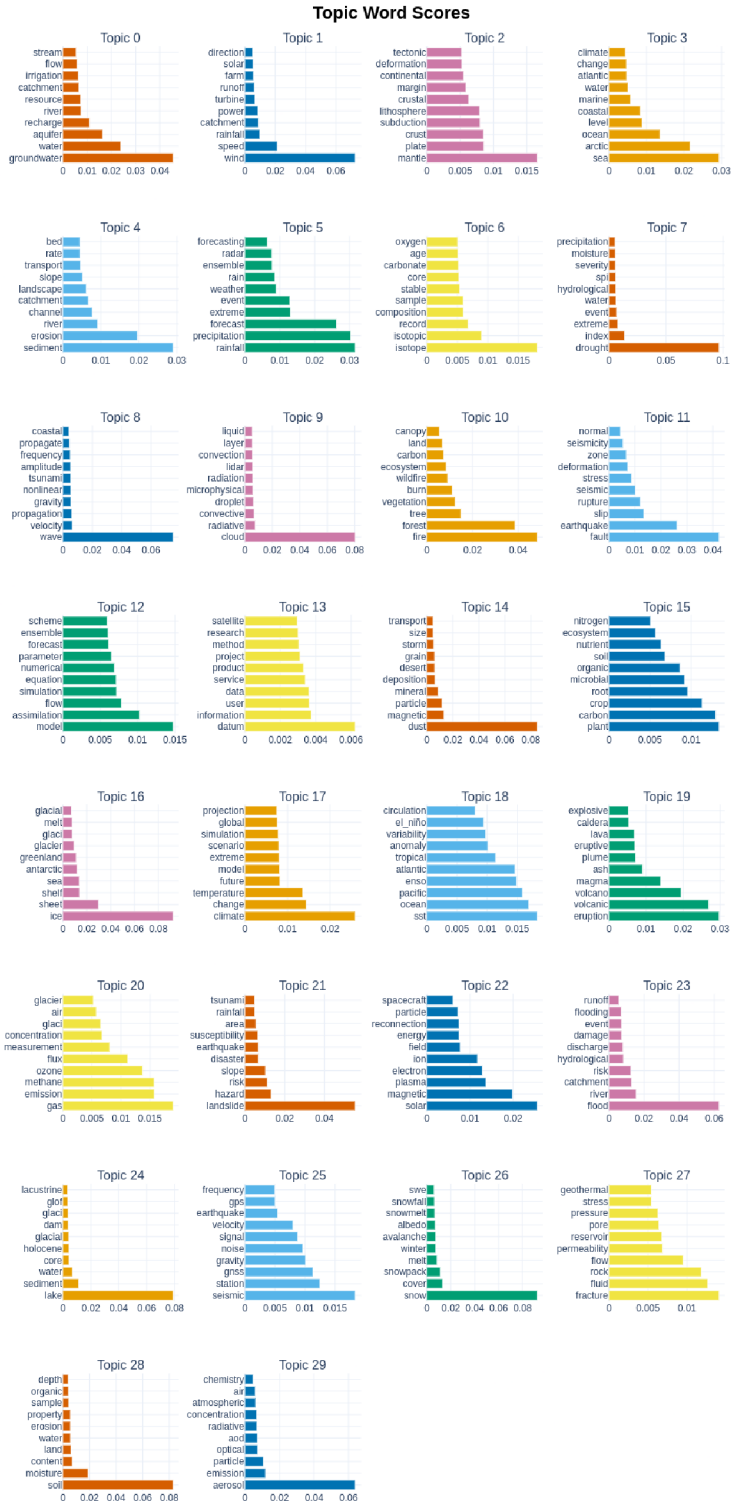


Figure 1: TF-IDF NMF model: 30 topics

Figure 2: NMF model: 30 topics

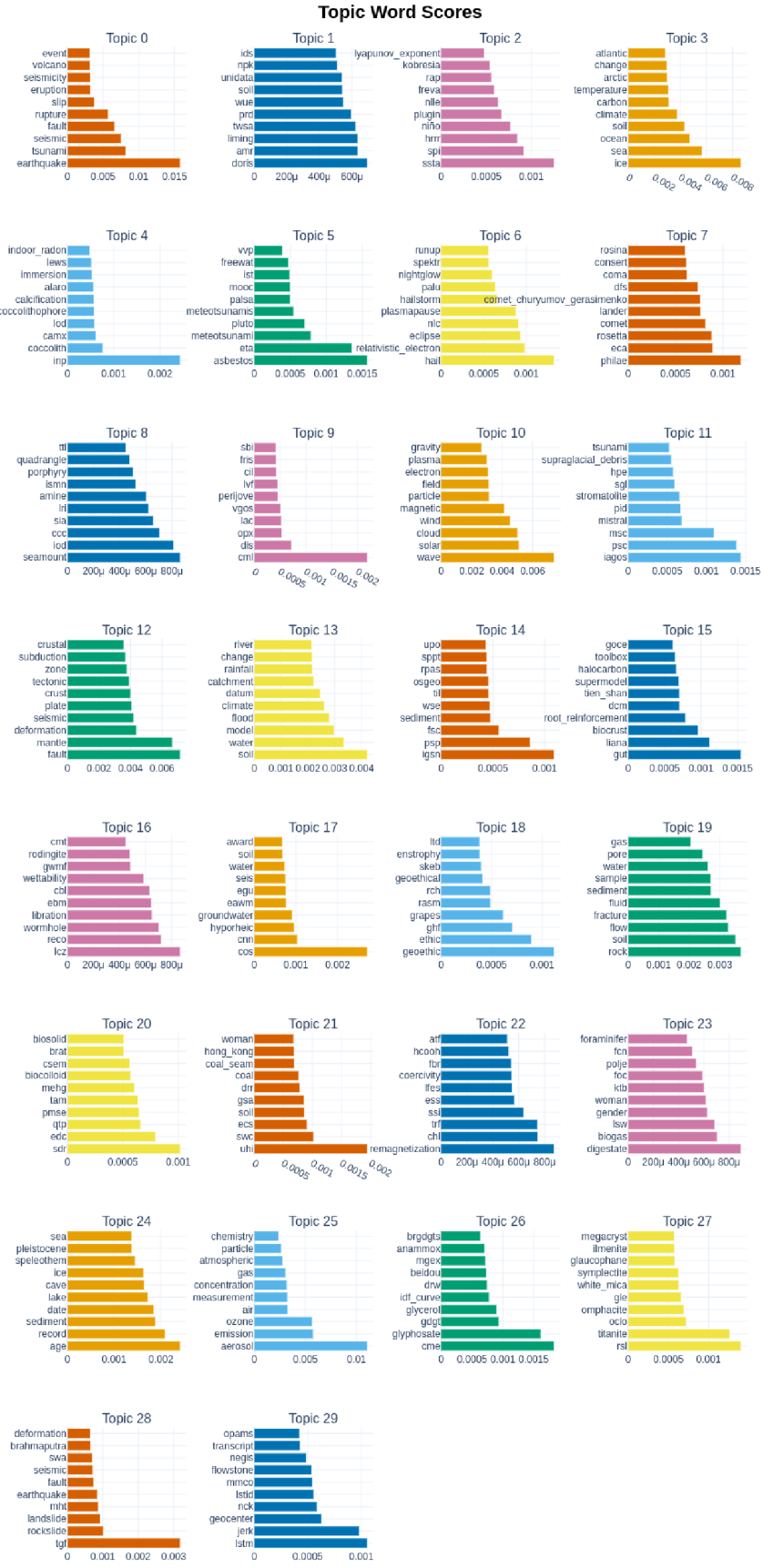


Figure 3: TF-IDF LDA model: 30 topics

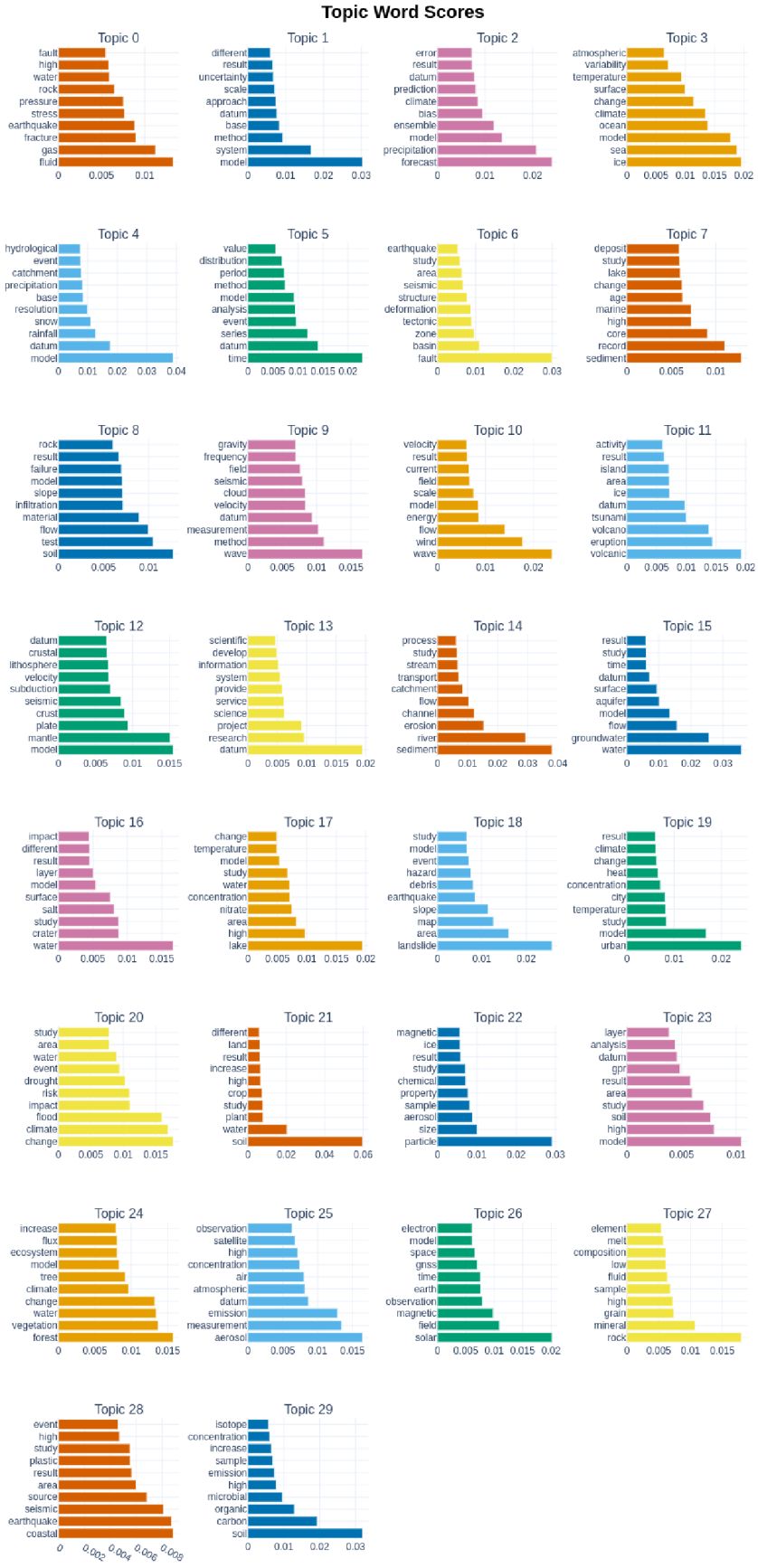


Figure 4: LDA model: 30 topics