Data and Metadata. 2024; 3:282
doi: 10.56294/dm2024282

ORIGINAL

Data lake management using topic modeling techniques

Gestión de lagos de datos mediante técnicas de modelado temático

Mohamed Cherradi[1] ✉, Anass El Haddadi[1]

[1]Data Science and Competetive Intelligence Team (DSCI), ENSAH, Abdelmalek Essaâdi, University (UAE) Tetouan. Morocco.

Cite as: Cherradi M, El Haddadi A. Data lake management using topic modeling techniques. Data and Metadata. 2024; 3:282. https://doi.org/10.56294/dm2024282

Submitted: 07-10-2023 **Revised:** 09-01-2024 **Accepted:** 14-04-2024 **Published:** 15-04-2024

Editor: Prof. Dr. Javier González Argote

ABSTRACT

With the rapid rise of information technology, the amount of unstructured data from the data lake is rapidly growing and has become a great challenge in analyzing, organizing and automatically classifying in order to derive the meaningful information for a data-driven business. The scientific document has unlabeled text, so it's difficult to properly link it to a topic model. However, crafting a topic perception for a heterogeneous dataset within the domain of big data lakes presents a complex issue. The manual classification of text documents requires significant financial and human resources. Yet, employing topic modeling techniques could streamline this process, enhancing our understanding of word meanings and potentially reducing the resource burden. This paper presents a comparative study on metadata-based classification of scientific documents dataset, applying the two well-known machine learning-based topic modelling approaches, Latent Dirichlet Analysis (LDA) and Latent Semantic Allocation (LSA). To assess the effectiveness of our proposals, we conducted a thorough examination primarily centred on crucial assessment metrics, including coherence scores, perplexity, and log-likelihood. This evaluation was carried out on a scientific publications corpus, according to information from the title, abstract, keywords, authors, affiliation, and other metadata aspects. Results of these experiments highlight the superior performance of LDA over LSA, evidenced by a remarkable coherence value of (0,884) in contrast to LSA's (0,768).

Keywords: Data Lake; Big Data; Machine Learning; Topic Modeling; LDA; LSA.

RESUMEN

Con el rápido aumento de la tecnología de la información, la cantidad de datos no estructurados del lago de datos está creciendo rápidamente y se ha convertido en un gran desafío para analizar, organizar y clasificar automáticamente con el fin de obtener la información significativa para un negocio impulsado por datos. El documento científico tiene texto sin etiquetar, por lo que es difícil vincularlo adecuadamente a un modelo temático. Sin embargo, la elaboración de una percepción de temas para un conjunto de datos heterogéneo dentro del dominio de los grandes lagos de datos presenta un problema complejo. La clasificación manual de documentos de texto requiere importantes recursos financieros y humanos. Sin embargo, el empleo de técnicas de modelado temático podría agilizar este proceso, mejorando nuestra comprensión de los significados de las palabras y reduciendo potencialmente la carga de recursos. Este artículo presenta un estudio comparativo sobre la clasificación basada en metadatos de un conjunto de datos de documentos científicos, aplicando los dos enfoques de modelado temático basados en aprendizaje automático más conocidos, el Análisis de Dirichlet Latente (LDA) y la Asignación Semántica Latente (LSA). Para evaluar la eficacia de nuestras propuestas, realizamos un examen exhaustivo centrado principalmente en métricas de evaluación cruciales, como las puntuaciones de coherencia, perplejidad y log-verosimilitud. Esta evaluación se llevó a cabo sobre un corpus de publicaciones científicas, según la información del título, el resumen, las palabras clave, los autores, la afiliación y otros aspectos de los metadatos. Los resultados de estos experimentos destacan el rendimiento superior de LDA sobre LSA, evidenciado por un notable valor de coherencia de (0,884) en contraste con el de LSA (0,768).

Palabras clave: Data Lake; Big Data; Machine Learning; Topic Modeling; LDA; LSA.

INTRODUCTION

Recently, topic modeling has gained significant traction in computer science, text mining, ad-hoc information retrieval, historical document analysis, computational social science, and comprehension of scientific publications.[1] Generative probabilistic models enable the automatic discovery of latent topics and extraction of valuable insights from extensive sets of unstructured data. Since its introduction, it has captured the interest of researchers from various fields, generating considerable enthusiasm and engagement. This comprehensive technique provides a deeper understanding of their field of study and expertise, shedding light on their abilities.[2,3] Thus, when researchers propose scientific papers within their areas of knowledge, it becomes essential to model their topic interests in academic social networks as a critical step. In this context, a researcher profile developed using non-observable variables based on the articles that pique their preference using the topic modelling technique, which enables the system to capture knowledge about its area of expertise and competencies to predict such needs in terms of pertinent research articles. Therefore, topic modeling can be used to efficiently recognize the several scientific fields. Thereby, this research looks towards methods for effectively organizing, analyzing, retrieving, and finding insights in text corpora. Further, topic modelling is a brand-new, incredibly powerful technique for automatically categorizing documents and summarizing enormous amounts of textual information in a big data lake.[4] By providing a simple way to evaluate vast amounts of unlabeled text data and reveal the hidden connections between items and themes, topic modelling plays a crucial role and is helpful in online digital libraries for the generation of supplemental metadata.[5] It simplifies the process of finding the relevant data by cataloguing documents on pertinent topics and making them searchable with different keywords.

Among the widely used topic modeling approaches, we mention the Latent Dirichlet Allocation (LDA) and the latent semantic analysis (LSA). Moreover, through semantic analysis and topic extraction from a researcher's publications, topic modeling facilitates the gathering and analysis of data pertaining to their activities. Nonetheless, modeling heterogeneous data sources is never an easy task since it's perpetually challenging for a human to identify topics in large data. Therefore, finding useful information among the massive textual case data is an extremely challenging and time-consuming task. Since scientific documents are so specialized, they are characterized by the use of a particular vocabulary, some are innovative and have a direct impact on policy decisions.[6] Further, it is quite challenging to select a technique that would be effective for an application in terms of extracting semantic topics, more coherent topics, to give a deeper insight of dataset or corpus. Hence, the suitable approach to the assessment of topic-based models is difficult to find. Although, some research on the comparative analysis of several topic modeling strategies has been done in the literature.[7,8,9] However, the existing studies have been conducted for specific application domains. This constitutes a significant limitation, as we were unable to locate any systematic attempt to compare and empirically evaluate topic modeling approaches with the goal of identifying a generalized topic model approach that is appropriate for data lake systems that handle heterogeneous data sources.

Recognizing the limitations within the existing literature, a comprehensive and systematic benchmark, along with empirical evaluations of topic modeling approaches tailored for data lake systems handling heterogeneous data sources, has been notably lacking. Addressing this critical gap, our study introduces the application of topic modeling within the context of data lake systems, underscoring its relevance in the expansive field of machine learning. A topic is the central idea or theme of a discussion, concept, conversation, or collection of documents. It can also be explained as a collection of words that have the same or related meanings. A topic can be broken down into many different levels of detail, such as a single sentence, paragraph, entire digital library collection, or more.[10,11] Moreover, topic modelling, an unsupervised method that examines a set of documents, reveals word and phrase patterns, and then automatically groups the words and phrases that most accurately represent the set of documents. It generate a collections of phrases and words that they believe to be related in a manner that helps in understanding the relevance of these linkages. While topic models have found their place in the realm of data analytics systems, their application within a data lake system remains underexplored. Our primary objective is to assess the applicability of well-established approaches within the context of a data lake. This paper conducts a thorough comparative analysis of topic modeling methods, identifying latent Dirichlet allocation as a prominent choice for the task. Through empirical study, we specifically compare two distinct approaches: LDA and LSA. The primary focus of this investigation is to evaluate the interpretability of these methods and ascertain their reliability as effective tools for modeling the scientific specializations of researchers.

The remainder of this article is as follow: Section 2 offers essential background information, laying the foundation for a comprehensive understanding of the subsequent content. In Section 3, we detail the proposed

methodology, followed by the presentation of results and discussions in Section 4. Lastly, Section 5 encapsulates the conclusion and outlines potential future perspectives.

Related Works

Since the term "data lake" was first used in the industry, it was initially introduced by James Dixon (CTO of Pentaho) as a solution that manages raw data from various sources and meets a range of customer requirements. [12] They can accommodate a robust ecosystem for generating innovative, data-driven business decisions and recognize a variety of data sources. There are significant suggestions for data lake architecture,[13] as well as comparisons with data warehouses and publications that go through its concept, components, and problems. [14,15] From 2019, data lakes have increased their contributions to both the business and academic communities. However, the majority of data lake approaches are abstract, based on a particular use case or a specific layer of data lake architecture, and do not provide a comprehensive perspective from data extraction to information retrieval. In the current academic landscape, topic modeling techniques are being widely employed to classify documents. These techniques are classified into two categories.[3] It concerns probabilistic and no-probabilistic topic modeling. Figure 1 shows a classification of the different techniques of topic modeling.

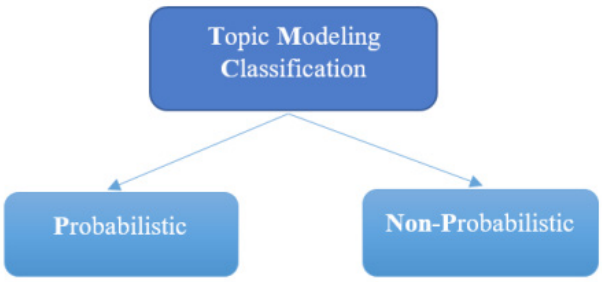

Figure 1. Categorization of topic modeling techniques

Our proposal is related to the most recent work in the domain of topic modeling interests since it applies techniques like LDA and LSA in the context of data lake systems. There are several studies and papers on various subject modeling techniques, including LSA and LDA, among others. Table 1, shows a comparative study between these two algorithms based on their advantages, disadvantages, and approaches kind. Since, several academics have been interested in employing topic modeling and text mining to discover the hidden topics in a corpus of documents. The authors[16] summarizes that topic modeling is a superior tool for examining latent topics in the archive. As a result, several academics have recommended topic modeling research to identify key topics that run through the corpus of scientific papers. These studies employ a variety of approaches and algorithms, including LDA, LSA, NMF, etc. The conventional techniques (such as BERTopic, Top2Vec, pLSA, Correlated Topic Model (CTM)) for analyzing data are designed for small volumes, indicating their inadequacies in the face of enormous information overload.[17] Thereby, LDA and LSA topic models assist the analysis of massive volumes of unlabeled text.

Table 1. Comparison of topic modelling models

Parameter	LDA	LSA
Approach	Dirichlet probability	Probabilistic technique, Expectation maximization
Merit	Unsupervised generative model	Works with synonyms
Demerit	Suitable for large data only	High compute time

In Amami et al.[18] the researcher's profile is derived using LDA from the articles they have assessed. The resulting profile is a vector where each component is associated with a topic. In Younus et al.[19] the authors utilized the tweeter-LDA algorithm, a variation of LDA proposed by Yu et al.[20] to extract subjects of interest from the tweets of young researchers. Dai et al.[21] proposed latent topic extraction through LDA with matrix factorization as a combined approach for suggesting citation and bibliographic references to scholars. Uys et al.[22] employed the LDA model to evaluate 62 documents on health-related issues, aiming to understand various fitness-related data categories. Similarly, the LDA topic modeling approach was employed for Vanamala et al.[23] to extract properties from log data, focusing on SQL injection threats. Exploring the field of related contexts, Costa Silva et al.[24] examines the analysis of scientific papers from PNAS (Proceedings of the National Academy of Sciences of the United States of America) using the LDA approach. The results were satisfactory, as the identified scientific disciplines correlated with the uncovered themes. Another study Lamba et al.[25] review the categorization of scientific unstructured text materials using majority

standard topic modeling methods based on the complete text, organizing terms into a sequence of themes.

Shifting towards the programming industry, Barua et al.[26] employed LDA on a popular Q&A website to identify challenges and popular themes in the programming industry. The findings revealed a wide range of topics discussed within the developer community. Mathkunti et al.[27] presented a technique combining the LDA algorithm and SVM, using LDA as a feature to create a representation with fewer topic dimensions. Subsequently, the data was categorized using SVM while retaining important semantic information and significantly reducing features. However, this study primarily utilized large-scale external datasets to uncover hidden topics and employed a semi-supervised learning approach to address short and sparse topics. Further, Chen et al.[28] examined linked topic and dynamic topic models using a sample of JSTOR archives articles.

METHODOLOGY

In this section, we present a comprehensive methodology for the topic modeling process. We begin by introducing a design proposal that outlines the key steps, including data preprocessing, hyper-parameter selection tuning, and data encoding. Subsequently, we delve into the topic modeling techniques using the Latent Dirichlet Allocation (LDA) algorithm and the Latent Semantic Analysis (LSA) algorithm, discussing their underlying principles, methodologies. To evaluate the quality of the topic models, we employ a set of evaluation metrics that assess coherence, distinctiveness, and overall performance.

Design Proposal for the Topic Modeling Process

When we encounter more textual material every day and more information becomes available, it is getting harder to find what we're seeking for. Therefore, in order to organize, analyze, search for, and uncover the hidden insights in any sizable body of textual data, we need the right tools and methodologies. Topic modeling is the term for these techniques highly effective method for automatically classifying documents, doing unsupervised analyses of huge document groups, comprehending enormous amounts of information in any large group from unstructured data, and summarizing massive amounts of textual data.[28,29] By offering a quick technique to assess enormous amounts of unlabeled text and point out the hidden connections between objects as well as topics expressed in documents. Then, topic modeling plays a vital role in scientific databases by aiding in the development of additional metadata,[30] ensuring accurate and effective text processing and classification. The categorization of data lake resources demands special attention from both users and digital libraries. By grouping documents with related topics together and making them searchable using a variety of keywords, it makes it easier for users to find the necessary information. Further, information retrieval operations are used to search for and access textual material like books, documents, and studies. Therefore, topic modeling techniques use an algorithm that tries to search for specific topics using a set of words in order to identify semantic themes in a collection of documents. These most frequent words listed by topic enable for the interpretation of texts based on their usage in various topics and their relative weights within various topics.

As part of our contribution, we compare LSA and LDA using a corpus of academic papers. Additionally, we investigate the impact of bi-gram collocation and NLP stages, including the ideal number of topics, on the performance of LSA and LDA. This comparison provides valuable insights into the effectiveness and applicability of LDA and LSA in extracting meaningful topics from academic content. Thus, the methodology is set up so that topic modeling strategies may be compared by using the same datasets for both implementations. During the topic modeling process, the text can be interpreted as a number of different topics, and each topic consists of a collection of frequently occurring words. Moreover, topics can be found by linking words with related meanings. Figure 2, depicts how the two information retrieval techniques (LDA and LSA) employ several strategies to automatically produce themes in the text corpus.

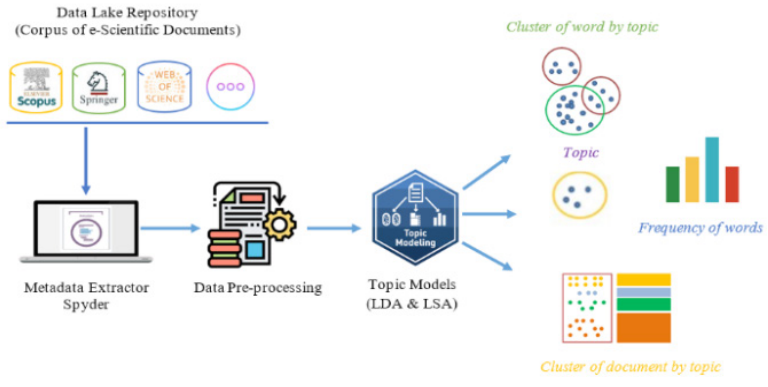

Figure 2. Proposed topic modeling-based architecture for data lake repository management

The experimental model diagram illustrates the stages of data analysis, starting with the input (raw data), followed by data preprocessing, data vectorization conversion, and comparison of two topic modeling methods in identifying the document similarity within-corpus and with the unseen document to categorize the word associations and coherence score as a measure for topic comparison and goodness of the topic model.

Data preprocessing

Serves as the initial stage of text mining, converting raw data into actionable insights through various techniques. It plays a critical role in converting human language text into a machine-understandable format. This process efficiently organizes unstructured text and preserves essential keywords for subject categorization. Recognizing its significance, we acknowledge its importance in enabling effective analysis and interpretation of textual data, facilitating deeper insights and comprehensive exploration of the content. In natural language texts, common usage of prepositions, pronouns, and ambiguous words adds complexity to the data preprocessing phase, making it a time-intensive step. This critical stage involves preparing the data to extract valuable insights from unstructured text. Data preprocessing encompasses cleaning and organizing raw data to make it suitable for creating and training topic models. Real-world data often lacks specific attribute values, exhibits inconsistent trends, and contains errors, outliers, and incompleteness. Data preparation comes into play to clean, format, and organize the raw data, enabling immediate use by topic models. Then, The CRoss Industry Standard Process for Data Mining (CRISP-DM) depicted in figure 3, serves as the cornerstone process model for data mining, establishing a robust foundation for the entire data mining process.

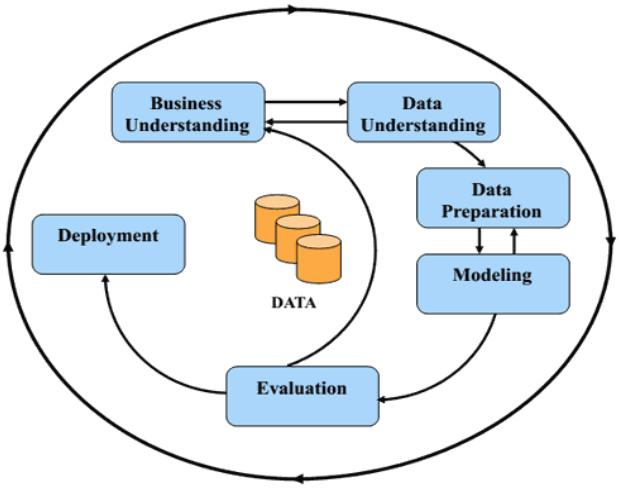

Figure 3. CRISP-DM methodology for Data Lake

Hyper-parameter selection tuning

To identify the optimal combination for the topic model, we have used the hyper-parameter grid search technique. This approach involves evaluating perplexity and log-likelihood measures to assess the performance of each model. Grid search systematically constructs multiple topic models by exploring various combinations of the given values. Therefore, it is crucial to tune the hyper-parameters as they significantly influence the overall performance of a topic model. One commonly employed technique for hyper-parameter tuning is utilizing the efficient GridSearchCV method from the Scikit-Learn package. This method systematically explores all feasible combinations of hyper-parameters to get the ideal configuration of values within the parameter search space. However, it is important to consider that this approach can be computationally intensive and time-consuming since it evaluates every possible combination in the grid. Therefore, alternative approaches, such as randomized search or Bayesian optimization, may be considered to mitigate resource consumption and reduce the tuning time.

Data encoding

The transformation of textual data into a format amenable to automated processing requires data encoding. A widely employed approach is the use of a bag-of-words matrix, where each document in a corpus is represented by a token vector, and the matrix entries denote the word frequencies. Yet, several techniques, such as OneHotEncoding, CountVectorizer and TF-IDF, can be employed for data encoding. In our experiment, we have adopted TfidfVectorizer to encode the preprocessed data. This technique harnesses Term document-Inverse Frequency efficiently converts text into a meaningful numerical representation, enabling the application of topic models.

https://doi.org/10.56294/dm2024282

Topic Modeling with LDA algorithm

LDA, an unsupervised machine learning technique, finds extensive application in document modeling, classification, and information retrieval. The LDA perspective considers each document as a multinomial distribution across k topics, and each topic is represented as a multinomial distribution across words. Through the process of matrix factorization, LDA decomposes the co-occurrence matrix (documents, terms) into two lower-dimensional matrices: (documents, topics) and (topics, words), as shown in the figure 4.

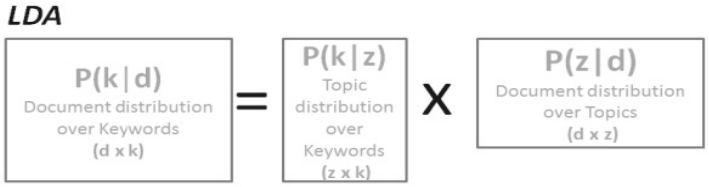

Figure 4. Matrix Decomposition of the LDA Algorithm

Moreover, LDA serves as a generative model, based on the idea that documents are constructed from a mixture of different topics, where each topic is represented by a distribution over a given vocabulary.

Topic Modeling with LSA algorithm

LSA (Latent Semantic Analysis), an unsupervised machine learning technique. It serves as an automated approach for indexing and retrieval by establishing a semantic network that connects terms and documents. LSA leverages sophisticated mathematical analysis to accurately deduce complex relationships. At its core, LSA applies relationship constraints to documents and utilizes the algebraic technique known as Single Value Decomposition (SVD), as depicted in the figure 5. The main objective of LSA is to partition a matrix of words and documents into separate matrices representing topics and their corresponding documents. This representation significantly reduces the dimensions of the original matrix.

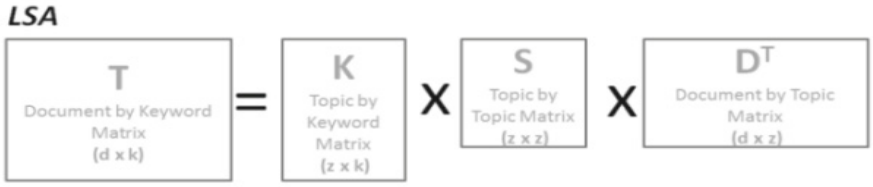

Figure 5. Matrix Decomposition of the LSA Algorithm

Indeed, the corpus is initially represented using LSA as a term-document matrix. The entries of this matrix can be populated with either term frequency values. However, the term-document matrix representation poses several challenges, mainly due to its high dimensionality and the presence of term noise. To address these challenges, one potential solution is to employ the singular value decomposition (SVD) method. The primary advantage of using SVD in LSA is its capability to express the term-document matrix as linearly independent components.

Evaluation metrics

Topic accuracy within topic modeling refers to the precision and alignment of the extracted topics with the latent themes present in the analyzed corpus. Assessing topic accuracy is of utmost importance, guarantees the reliability and relevance of the generated topics, enabling deeper understanding and analysis of complex textual data. Indeed, the selection of various parameters, including the number of inferred topics and the values of the hyper-parameters, profoundly affects the effectiveness of a topic model. The Gibbs sampling technique offers valuable insights into choosing appropriate hyper-parameter values.

Number of topics

To determine the appropriate number of topics, we considered the top four topics for the two approaches. The selection was based on their prevalence scores, which are arranged in descending order. These chosen topics effectively capture the majority of the data points and are regarded as high-quality representations. It is essential to avoid overfitting the model by selecting an excessive number of topics. Conversely, opting for too few topics may lead to underfitting and improper cluster formation. Striking the right balance is crucial to achieve optimal results and meaningful insights.

https://doi.org/10.56294/dm2024282

Log-likelihood score and Perplexity

Topic coherence serves as an evaluation metric to assess the relevance of topics by examining the coherence of word groupings within the model-generated topics. Multiple metrics exist to quantify coherence, each offering unique methodologies. The analyzed techniques for topic coherence calculate a sum of a confirmation measure over pairs of words, utilizing the top words of each topic as input. This facilitates the construction of a topic coherence matrix, providing insight into the semantic similarity among words within a given topic. Coherence measures, including U_{Mass} and *UCI*, can be employed to evaluate topic models, with *UCI* typically utilized for intrinsic measures and U_{Mass} for extrinsic measures.

Interpretability

After determining the optimal number of topics through perplexity and coherence measures and ensuring the topic modeling coherence. It is vital to effectively visualize the topics to ensure a meaningful interpretation. To achieve this, we will employ two-gram term frequencies and word cloud visualization. This technique offers a comprehensive representation of the words that constitute each topic and provides valuable insights into the proximity and relationships between topics. Moreover, it serves as a valuable tool to assess the performance and accuracy of the topic model, facilitating the identification and implementation of potential improvements, if necessary.

RESULTS AND DISCUSSIONS

In this section, we present the results of our experiments and engage in a comprehensive discussion. We begin by describing the experiment setup, which outlines the parameters and configurations employed in our study. Subsequently, we delve into the details of the experiment dataset, highlighting its characteristics and composition. Moving on to the core of our findings, we present the experiment results, which are further divided into several sub-sections. Firstly, we examine the optimal number of topics, aiming to identify the most suitable number for our topic model. Next, we explore topic terms frequency, investigating the distribution and occurrence of terms within each topic. We then analyze the document topics frequency, providing insights into the prevalence and distribution of topics across the document collection. Furthermore, we evaluate the performance of our topic model. Finally, we entail a comprehensive discussion, synthesizing our results, addressing implications, and presenting key observations and insights.

Experiment setup

In our experiment setup, we have implemented our solution using the Python programming language, utilizing the NLTK (Natural Language Toolkit) module for natural language processing tasks such as tokenization, stemming, and stop-words removal. For topic modeling, we relied on Gensim, a robust and scalable open-source library known for its various capabilities, including creating corpora, generating document representations, identifying topics, and examining semantic structures. To perform data analytics, preprocessing, document comparison, grid search, and visualization, we leveraged important Python libraries such as Pandas, NumPy, SciPy, Scikit-learn, and Matplotlib. Our models were trained using Google Colab, which provided us with a powerful computing environment featuring a 2,2 GHz CPU, 16 GB of RAM, and 128 GB of SSD. This allowed us to extract valuable information from our data lake repository by combining topic modeling algorithms with their parameter configurations.

Experiment datasets

Our research focused on a carefully curated corpus of scientific documents spanning the period from 2018 to 2023. The corpus consisted of publicly available texts sourced from reputable research publications within the data lake domain. To construct the data lake document corpus, we relied on a dataset comprising earlier research papers obtained from well-established scientific databases, covering a broad range of proposals related to data lake concepts. The Selenium package was employed for automated extraction of critical metadata attributes from scientific articles, including title, abstract, affiliation, authors, keywords, and publisher details. Using Spacy in conjunction with regular expressions, we efficiently parsed the collection of documents, facilitating seamless navigation, efficient tag searching, and content editing across heterogeneous data sources. The parsed data was then transformed into a CSV file for further processing, enabling the identification of text topics and preservation of keywords for effective organization of unstructured content. Rigorous text preprocessing and necessary dataset refinements were performed to ensure the relevance of our topic modeling approach. The data was subsequently partitioned into an 80 % training set and a 20 % testing set for comprehensive experimental analysis.

Experiment results

The results of our topic modeling analysis using the LDA and LSA algorithms revealed valuable insights

into the underlying themes present in our dataset. Figure 6.a illustrates the top ten keywords extracted using the 2-grams technique, providing a comprehensive view of the most significant terms within the corpus of documents. These keywords include terms such as 'data lake management,' 'big data analytics,' 'data integration,' and 'data governance,' which are crucial in understanding the primary focus areas within the data lake domain. Furthermore, figure 6.b displays the word cloud visualization of the four identified topics, offering a visual representation of the frequency and prominence of terms within each topic. This visualization aids in identifying key concepts and their interrelationships, highlighting the core themes that emerge from the topic modeling analysis. Overall, the figures demonstrate the effectiveness of both the LDA and LSA algorithms in uncovering meaningful topics and providing valuable insights into the dataset.

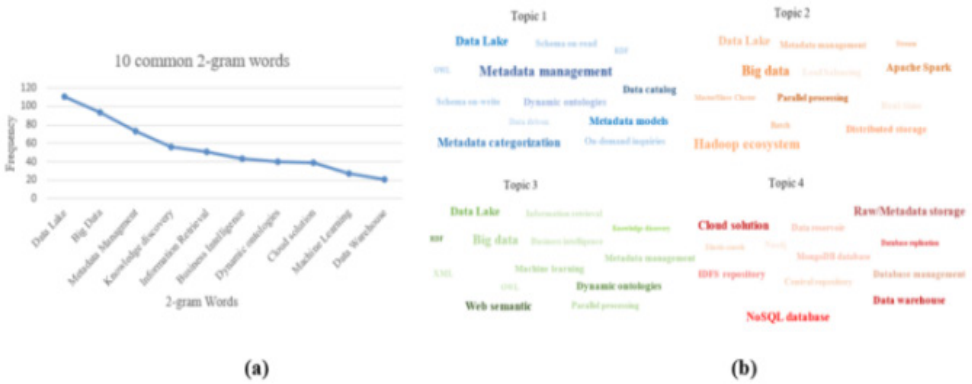

Figure 6. Top keywords extracted from data lake repository

Optimal number of topics

In our exploration to determine the optimal number of topics, we experimented with topic models by adjusting the number of topic parameters. As illustrated in figure 7, the coherence scores exhibit an upward trend as the number of topics surpasses 4. Within the range of 4 to 12 topics, there is a gradual increase in coherence score. However, it is important to consider the trade-off between higher coherence scores and the risk of either underfitting or overfitting the data. Based on Gensim's LDA model, a reasonable choice of 4 topics strikes a balance between coherence and model performance.

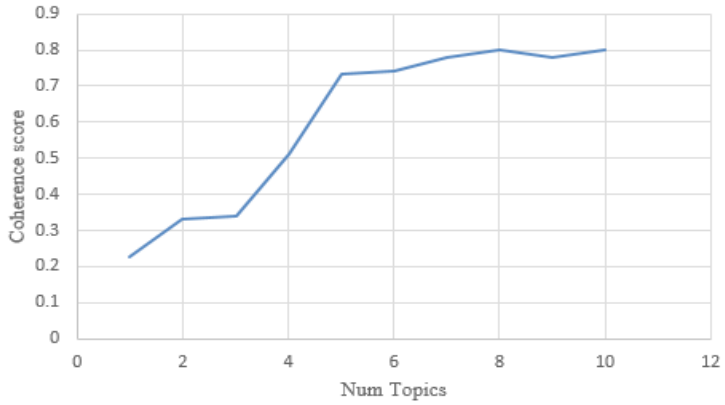

Figure 7. Determining the Optimal Number of Topics

The consistent results derived from both LDA and LSA techniques highlight the superiority of selecting 4 topics as the input parameter compared to other tested numbers.

Topic terms frequency

The pursuit of topic term frequency stands as a key objective within the framework of a topic modeling algorithm. Then, table 2 showcase the word distributions within the top four topic clusters. Each topic establishes connections to one or more documents, with the assigned mixing percentages based on word frequencies in the corpus. Notably, words within a single topic frequently exhibit resemblances in meaning. Moreover, the identification of appropriate labels for each topic can be achieved through meticulous analysis of the assigned words. For instance, upon evaluating the words, topic 1 can be accurately labeled as "data catalog".

Examining table 2 reveals the grouping of contributions related to the management of data lakes, focusing on the concept of a metadata model, within topic 1. Thereby, proposals pertaining to the manipulation of data

lakes using the big data ecosystem, including Hadoop, Spark, and other relevant technologies, are clustered in topic 2. Furthermore, documents that explore data lakes from the perspectives of machine learning and semantic web are consolidated in topic 3. Lastly, topic 4 encompasses proposals related to the manipulation of data lakes using data warehouses and NoSQL solutions.

Table 2. Top ten two-gram terms using topic modelling techniques

Topics	Terms
T1	0,031*(Data catalog) + 0,029*(Data driven) + 0,023*(Data lake) + 0,021*(Metadata categorization) + 0,018*(Metadata management) + 0,016*(Metadata model) + 0,015*(Metadata schema) + 0,015*(On-demand inquiries) + 0,014*(Schema on-read) + 0,014*(Schema on-write)
T2	0,059*(Big data ecosystem) + 0,053*(Data Lake) + 0,048*(Hadoop) + 0,043*(Distributed storage) + 0,043*(Parallel processing) + 0,037*(Apache Spark) + 0,031*(Metadata management) + 0,028*(Stream/Batch/Real-time processing) + 0,024*(Master/Slave cluster) + 0,023*(Load balancing)
T3	0,068*(Business intelligence) + 0,061*(Data lake) + 0,058*(Dynamic ontologies) + 0,054*(Information retrieval) + 0,045*(Knowledge discovery) + 0,039*(Machine learning) + 0,039*(Metadata management) + 0,037*(Big data) + 0,036*(RDF & OWL) + 0,031* (Semantic web)
T4	0,081*(NoSQL database) + 0,076*(Cloud solution) + 0,069*(Data lake) + 0,064*(Data replication) + 0,061*(Data reservoir) + 0,057*(Data warehouse) + 0,055*(Data/Metadata storage) + 0,052*(Database management) + 0,049*(HDFS repository) + 0,46*(MongoDB database)

Document topics frequency

Ensuring optimal document topic frequency is the main aim in the design and implementation of topic modeling algorithms. It consists to create interpretable representations of documents that facilitate the exploration of topics within an unlabeled document collection. Table 3 serves as an illustrative example of such interpretable document representations, wherein each document is represented by a distribution of topics. For instance, document 1 is predominantly composed of topic 2 with a weight of 0,976, topic 4 with a weight of 0,016, and negligible weights assigned to other topics.

Table 3. Topics distribution over each document

Document Index	Topic1	Topic2	Topic3	Topic4
1	0	0,976	0	0,016
2	0	0	0,794	0,201
3	0,018	0,290	0,683	0
4	0	0,894	0	0,061
5	0,020	0,092	0	0,0871
6	0,017	0,028	0,111	0,796
7	0,125	0,810	0	0,025
8	0	0	0,063	0,893
9	0,793	0	0,215	0
10	0	0,998	0	0

Additionally, table 3 reveals that document vectors often contain numerous zero values. This suggests that only a limited selection of topics are present in a given document, aligning with the idea that most documents cover a relatively small number of topics. These findings greatly aid in enhancing the comprehensibility of document vectors for human interpretation. For example, document 10 demonstrates a topic score of 0,998 for topic 2 and zero scores for all other topics. Based on this information, it can be inferred that topic 2 is primarily assigned to document 10. Accordingly, within the framework of a topic model, each document represents a collection of multiple topics. However, typically, only one topic takes precedence and occupies the foreground.

Topic models evaluation

To assess the performance of the topic models, topic coherence serves as a valuable metric. It measures the contextual relationship and relative distance between terms within topics, providing insights into the perplexity and semantic coherence of topic clusters. Utilizing these metrics helps distinguish semantically meaningful topics and accurately determine the optimal number of topics by assigning coherence scores that reflect the interpretability of each topic. Figure 8 showcases the topic coherence measurements between the LDA and LSA models.

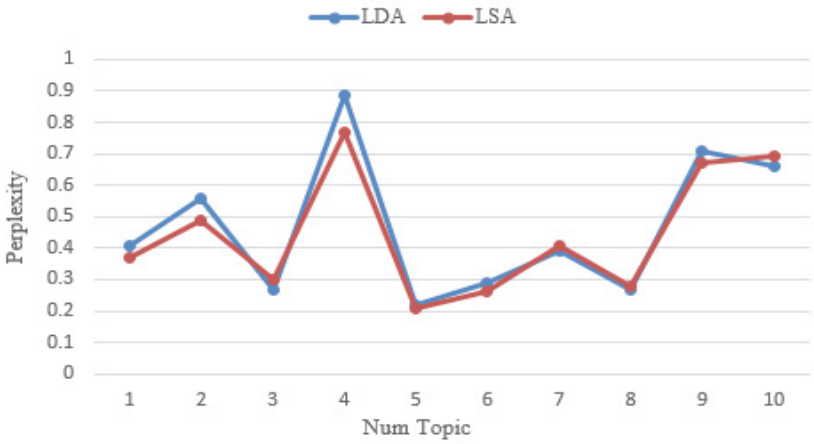

Figure 8. A Comparative Analysis of LDA and LSA Algorithms in Terms of Topic Coherence Perplexity

We employed topic coherence, a metric that measures word similarity within each topic. Higher coherence perplexity scores indicate more coherent and well-defined topics. Comparing the two models using figure 8, we observed that the LDA model outperformed the LSA model, reporting a coherence score of 0,884 compared to 0,768. Furthermore, the LDA model exhibited more significant and informative words for each topic, resulting in a more concise and coherent representation of the topics within the data lake corpus.

Discussions

To establish the underscore and novelty of our work, we compare our findings and methodologies with relevant existing studies. For instance, Bellaouar et al.[31] and Mohammed et al.[32] proposed a topic modeling approach for analyzing healthcare data. Their study focused on extracting meaningful topics from a large corpus on e-books. While their work delved into specific data, our research expands beyond this domain and applies topic modeling to a curated corpus of scientific documents within the data lake context. By doing so, we contribute to the understanding of topic modeling techniques in a different domain, showcasing the versatility and applicability of these methods. Nonetheless, within the context of benchmark studies, Kherwa et al.[33] and Kalepalli et al.[34] explored topic modeling techniques comprehensively. While their study focused on random configuration models, our research expands the application of topic modeling by specifically targeting the optimal parameters. By adapting topic modeling techniques to the data lake domain, we highlight the potential of these methods for uncovering insights from unstructured data in a different context. Further, Maarif[35] and Daud et al.[36] investigated the application of topic modeling in the analysis of customer reviews and sentiment analysis. Their study employed LDA to extract topics from online customer reviews and analyze the sentiment associated with each topic. In contrast, our research explores the utilization of topic modeling techniques for analyzing a curated corpus of scientific documents within the data lake domain. This demonstrates the adaptability of topic modeling techniques across different discipline and highlights the unique challenges and opportunities associated with data lake analysis.

CONCLUSIONS

In summary, this research significantly contributes to the application of topic modeling techniques in data lake analysis, showcasing their effectiveness in extracting valuable insights from unstructured scientific documents. By comparing our findings with existing studies, we have established the originality and uniqueness of our research, positioning it as a significant advancement in the field of data lake analysis using topic modeling techniques. The utilization of the LDA and LSA algorithms provided a comprehensive evaluation of their performance in revealing latent topics within the curated corpus of scientific publications. Our results highlighted the superiority of LDA in generating more coherent and meaningful topics, thereby enhancing our understanding of the data lake domain. Nevertheless, it is crucial to acknowledge that topic modeling remains a dynamic and evolving area of research. To further enhance the effectiveness and applicability of topic modeling in data lake analysis, several factors warrant consideration. One promising avenue for future research is the evaluation of topic models' performance on diverse datasets from various domains. Assessing the scalability and generalizability of the proposed topic modeling techniques can provide additional insights into their robustness and effectiveness in different contexts. Furthermore, exploring alternative algorithms and techniques beyond LDA and LSA could offer valuable perspectives. Advanced models, such as Hierarchical Dirichlet Processes (HDP) and Non-negative Matrix Factorization (NMF), have shown promising results in certain domains and merit investigation for their potential application in data lake analysis. By recognizing the ongoing nature of topic

modeling research and exploring improvement avenues, we anticipate contributing to its evolution in data lake analysis, revolutionizing data-driven decision-making across domains.

REFERENCES

1. Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of Topic Models. Foundations and Trends® in Information Retrieval, 11, 143-296. https://doi.org/10.1561/1500000030

2. Boussaadi, S., Aliane, D. H., & Abdeldjalil, P. O. (2020). The Researchers Profile with Topic Modeling. IEEE International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), 1-6. https://doi.org/10.1109/ICECOCS50124.2020.9314588

3. Kherwa, P., & Bansal, P. (2018). Topic Modeling : A Comprehensive Review. ICST Transactions on Scalable Information Systems, 7, 159623. https://doi.org/10.4108/eai.13-7-2018.159623

4. Anupriya, P., & Karpagavalli, S. (2015). LDA based topic modeling of journal abstracts. 2015 International Conference on Advanced Computing and Communication Systems, 1-5. https://doi.org/10.1109/ICACCS.2015.7324058

5. Newman, D., Noh, Y., Talley, E., Karimi, S., & Baldwin, T. (2010). Evaluating topic models for digital libraries. Proceedings of the 10th Annual Joint Conference on Digital Libraries, 215-224. https://doi.org/10.1145/1816123.1816156

6. Syed, S., & Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation (p. 174). https://doi.org/10.1109/DSAA.2017.61

7. Wörner, J., Konadl, D., Schmid, I., & Leist, S. (2021, juin 14). Comparison of topic modelling techniques in marketing—Results from an analysis of distinctive use cases. European Conference on Information Systems (ECIS).

8. Hua, T., Lu, C.-T., Choo, J., & Reddy, C. (2020). Probabilistic Topic Modeling for Comparative Analysis of Document Collections. ACM Transactions on Knowledge Discovery from Data, 14, 1-27. https://doi.org/10.1145/3369873

9. Vayansky, I., & Kumar, S. (2020). A review of topic modeling methods. Information Systems, 94, 101582. https://doi.org/10.1016/j.is.2020.101582

10. Cherradi, M., & El Haddadi, A. (2023). EMEMODL : Extensible Metadata Model for Big Data Lakes. International Journal of Intelligent Engineering and Systems, 16. https://doi.org/10.22266/ijies2023.0630.18

11. Zagan, E., & Danubianu, M. (2020). Data Lake Approaches : A Survey. International Conference on Development and Application Systems (DAS), 189-193. https://doi.org/10.1109/DAS49615.2020.9108912

12. Cherradi, M., & El Haddadi, A. (2022). Data Lakes : A Survey Paper. In Innovations in Smart Cities Applications, Vol. 5 (p. 823-835). Springer. https://doi.org/10.1007/978-3-030-94191-8_66

13. Cherradi, M., EL Haddadi, A., & Routaib, H. (2022). Data Lake Management Based on DLDS Approach. Proceedings of Networking, Intelligent Systems and Security, 679-690. https://doi.org/10.1007/978-981-16-3637-0_48

14. Cherradi, M., & El Haddadi, A. (2022). A Scalable framework for data lakes ingestion. Procedia Computer Science, 215, 2022, 809-814. https://doi.org/10.1016/j.procs.2022.12.083

15. Cherradi, M., Bouhafer, F., & El Haddadi, A. (2023). Data Lake Governance using IBM-Watson Knowledge Catalog. Scientific African, 21, e01854. https://doi.org/10.1016/j.sciaf.2023.e01854

16. Li, Z., Shang, W., & Yan, M. (2016). News text classification model based on topic model. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 1-5. https://doi.org/10.1109/ICIS.2016.7550929

17. Jipeng, Q., Zhenyu, Q., Yun, L., Yunhao, Y., & Xindong, W. (2019). Short Text Topic Modeling Techniques, Applications, and Performance : A Survey (arXiv:1904.07695). arXiv. https://doi.org/10.48550/arXiv.1904.07695

18. Amami, M., Faiz, R., Stella, F., & Pasi, G. (2017). A graph based approach to scientific paper recommendation. Proceedings of the International Conference on Web Intelligence, 777-782. https://doi.org/10.1145/3106426.3106479

19. Younus, A., Qureshi, M. A., Manchanda, P., O'Riordan, C., & Pasi, G. (2014). Utilizing Microblog Data in a Topic Modelling Framework for Scientific Articles' Recommendation. In L. M. Aiello & D. McFarland (Éds.), Social Informatics (Vol. 8851, p. 384-395). Springer International Publishing. https://doi.org/10.1007/978-3-319-13734-6_28

20. Yu, K., Zhang, B., Zhu, H., Cao, H., & Tian, J. (2012). Towards Personalized Context-Aware Recommendation by Mining Context Logs through Topic Models (Vol. 7301, p. 443). https://doi.org/10.1007/978-3-642-30217-6_36

21. Dai, T., Zhu, L., Cai, X., Pan, S., & Yuan, S. (2018). Explore semantic topics and author communities for citation recommendation in bipartite bibliographic network. Journal of Ambient Intelligence and Humanized Computing, 9(4), 957-975. https://doi.org/10.1007/s12652-017-0497-1

22. Uys, W., Du Preez, N., & Uys, E. W. (2008). Leveraging unstructured information using topic modelling (p. 961). https://doi.org/10.1109/PICMET.2008.4599703

23. Vanamala, M., Yuan, X., & Roy, K. (2020). Topic Modeling And Classification Of Common Vulnerabilities And Exposures Database. In International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD) (p. 5). https://doi.org/10.1109/icABCD49160.2020.9183814

24. Costa Silva, C., Galster, M., & Gilson, F. (2021). Topic modeling in software engineering research. Empirical Software Engineering, 26. https://doi.org/10.1007/s10664-021-10026-0

25. Lamba, M., & Margam, M. (2022). Topic Modelling and its Application in Libraries : A review of specialized literature. World Digital Libraries, 15, 105-120. https://doi.org/10.18329/09757597/2022/15207

26. Barua, A., Thomas, S. W., & Hassan, A. E. (2014). What are developers talking about? An analysis of topics and trends in Stack Overflow. Empirical Software Engineering, 19(3), 619-654. https://doi.org/10.1007/s10664-012-9231-y

27. Mathkunti, N., & Rangaswamy, S. (2020). Machine Learning Techniques to Identify Dementia. SN Computer Science, 1. https://doi.org/10.1007/s42979-020-0099-4

28. Chen, T.-H., Thomas, S. W., & Hassan, A. E. (2016). A survey on the use of topic models when mining software repositories. Empirical Software Engineering, 21(5), 1843-1919. https://doi.org/10.1007/s10664-015-9402-8

29. Kunsabo, J., & Dobša, J. (2022). A Systematic Literature Review on Topic Modelling and Sentiment Analysis. In International Scientific Conference Central European Conference on Information and Intelligent Systems.

30. Albalawi, R., Yeap, T., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data : A Comparative Analysis. Frontiers in Artificial Intelligence, 3. https://doi.org/10.3389/frai.2020.00042

31. BELLAOUAR, S., BELLAOUAR, M. M., & GHADA, I. E. (2021). Topic Modeling : Comparison of LSA and LDA on Scientific Publications. 2021 4th International Conference on Data Storage and Data Engineering, 59-64. https://doi.org/10.1145/3456146.3456156

32. Mohammed, S. H., & Al-augby, S. (2020). LSA & LDA topic modeling classification : Comparison study on e-books. Indonesian Journal of Electrical Engineering and Computer Science, 19(1), Article 1. https://doi.org/10.11591/ijeecs.v19.i1.pp353-362

33. Kherwa, P., & Bansal, P. (2021). A Comparative Empirical Evaluation of Topic Modeling Techniques. In D. Gupta, A. Khanna, S. Bhattacharyya, A. E. Hassanien, S. Anand, & A. Jaiswal (Éds.), International Conference on Innovative Computing and Communications (p. 289-297). Springer. https://doi.org/10.1007/978-981-15-5148-2_26

34. Kalepalli, Y., Tasneem, S., Phani Teja, P. D., & Manne, S. (2020). Effective Comparison of LDA with LSA for Topic Modelling. International Conference on Intelligent Computing and Control Systems (ICICCS), 1245-1250. https://doi.org/10.1109/ICICCS48265.2020.9120888

35. Maarif, M. R. (2022). Summarizing Online Customer Review using Topic Modeling and Sentiment Analysis. JISKA (Jurnal Informatika Sunan Kalijaga), 7(3), Article 3. https://doi.org/10.14421/jiska.2022.7.3.177-191

36. Daud, S., Ullah, M., Rehman, A., Saba, T., Damaševičius, R., & Sattar, A. (2023). Topic Classification of Online News Articles Using Optimized Machine Learning Models. Computers, 12(1), Article 1. https://doi.org/10.3390/computers12010016

FINANCING

None.

CONFLICT OF INTEREST

None.

AUTHORSHIP CONTRIBUTION

Conceptualization: Mohamed Cherradi, Anass El Haddadi.
Data curation: Mohamed Cherradi, Anass El Haddadi.
Research: Mohamed Cherradi, Anass El Haddadi.
Project management: Mohamed Cherradi, Anass El Haddadi.
Resources: Mohamed Cherradi, Anass El Haddadi.
Supervision: Mohamed Cherradi, Anass El Haddadi.
Writing - original draft: Mohamed Cherradi, Anass El Haddadi.