









ORIGINAL

Data Lake Management System based on Topic Modeling

Sistema de gestión de lagos de datos basado en el modelado de temas

Amine El Haddadi¹ , Oumaima El Haddadi¹ , Mohamed Cherradi¹ , Fadwa Bouhafer¹ , Anass El Haddadi¹ , Ahmed El Allaoui¹ 

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

Cite as: El Haddadi A, El Haddadi O, Cherradi M, Bouhafer F, El Haddadi A, El Allaoui A. Data Lake Management System based on Topic Modeling. Data and Metadata. 2023;2:183. <https://doi.org/10.56294/dm2023183>


Submitted: 09-08-2023

Revised: 17-10-2023

Accepted: 27-11-2023

Published: 28-12-2023

Editor: Prof. Dr. Javier González Argote 

Guest Editor: Yousef Farhaoui 

Note: Paper presented at the International Conference on Artificial Intelligence and Smart Environments (ICAISE'2023).

ABSTRACT

In an environment full of competitiveness, data is a valuable asset for any company looking to grow. It represents a real competitive economic and strategic lever. The most reputable companies are not only concerned with collecting data from heterogeneous data sources, but also with analyzing and transforming these datasets into better decision-making. In this context, the data lake continues to be a powerful solution for storing large amounts of data and providing data analytics for decision support. In this paper, we examine the intelligent data lake management system that addresses the drawbacks of traditional business intelligence, which is no longer capable of handling data-driven demands. Data lakes are highly suitable for analyzing data from a variety of sources, particularly when data cleaning is time-consuming. However, ingesting heterogeneous data sources without any schema represents a major issue, and a data lake can easily turn into a data swamp. In this study, we implement the LDA topic model for managing the storage, processing, analysis, and visualization of big data. To assess the usefulness of our proposal, we evaluated its performance based on the topic coherence metric. The results of these experiments showed our approach to be more accurate on the tested datasets.

Keywords: Data Lake; Big Data; Business Intelligence; LDA; Topic Modeling.

RESUMEN

En un entorno lleno de competitividad, los datos son un activo valioso para cualquier empresa que quiera crecer. Representa una verdadera palanca económica y estratégica competitiva. Las empresas más reputadas no sólo se preocupan de recopilar datos procedentes de fuentes de datos heterogéneas, sino también de analizar y transformar estos conjuntos de datos para mejorar la toma de decisiones. En este contexto, el lago de datos sigue siendo una potente solución para almacenar grandes cantidades de datos y proporcionar análisis de datos para apoyar la toma de decisiones. En este artículo, examinamos el sistema inteligente de gestión de lagos de datos que aborda los inconvenientes de la inteligencia empresarial tradicional, que ya no es capaz de hacer frente a las demandas impulsadas por los datos. Los lagos de datos son muy adecuados para analizar datos de diversas fuentes, sobre todo cuando la limpieza de datos requiere mucho tiempo. Sin embargo, la ingesta de fuentes de datos heterogéneas sin ningún esquema representa un problema importante, y un lago de datos puede convertirse fácilmente en un pantano de datos. En este estudio, implementamos el modelo temático LDA para gestionar el almacenamiento, procesamiento, análisis y visualización de big data. Para valorar la utilidad de nuestra propuesta, evaluamos su rendimiento basándonos en la métrica de coherencia temática. Los resultados de estos experimentos mostraron que nuestro enfoque es más preciso

en los conjuntos de datos probados.

Palabras clave: Data Lake; Big Data; Inteligencia de Negocio; LDA; Modelado Temático.

INTRODUCTION

The idea of a data lake has grown in popularity as an efficient solution to manage massive amounts of disparate data sources. Indeed, the use of big data technology helps companies drastically boost their business. Thus, business intelligence is a sophisticated strategy for processing raw data into actionable information. ⁽¹⁾ By finding new possibilities, highlighting possible pitfalls, uncovering new business insights, and improving decision-making processes, business intelligence helps organizations to increase their productivity. ⁽²⁾ As a result, most industries consider BI to be a high priority. However, conventional business intelligence concentrates on structured data only, ignoring very important information hidden in unstructured data. Therefore, this could lead to an incomplete view of the environment, which can lead to ineffective decision-making. Moreover, conventional business intelligence systems can perform various levels and types of structured data analysis, but they aren't intended to manage unstructured data. Big Data poses significant challenges for these systems due to the heterogeneity of data, which can be structured or unstructured. As a result, traditional business intelligence is severely limited in its ability to benefit from big data advantages. Therefore, we need to rethink how we harness the power of business intelligence in terms of how is ingested, stored, and analyzed.

The emergence of big data requires a new data management paradigm. Companies have recently requested the ability to analyze any type of data, but traditional solutions are unable to keep up. As a result of big data, new technologies such as data lakes have emerged, allowing businesses to store and manage massive volumes of heterogeneous datasets. However, in the absence of data schema, a data lake must be supported by metadata to prevent the lake from transforming into a data swamp, which means useless data or not actionable data, or also known as undocumented data. ⁽³⁾ In this paper, we implement the LDA algorithm to prove the usefulness of a business data lake as a data repository capable of storing and processing enormous amounts of data from various sources. Indeed, the business data lake technique simplifies the enterprise data warehouse abbreviated EDW processing and decreases its complexity. When compared to a standard EDW, this technique enhances the ability to respond to business needs while also prolonging the life of EDW systems.

Data lakes are one of the challenging concepts that have evolved throughout the big data age. Despite the fact that Data Lake is a brand-new concept with revolutionary ideas, it faces numerous challenges. ^(4,5) However, data lake research is significant because of the potential to store heterogeneous amounts of data as-is without any predefined schema. This paper presents the application of LDA topic modeling as a data mining technique for discovering the main topics of a set of documents stored in the lake. Thereby, discovering the high-level topic models that exist within a collection of documents enables companies to uncover the challenging issues, concerns, environmental feedback, face competitors, and therefore gives the company a better advantage to deliver the right product to the right person at the right time.

Given the challenge of new raw data formats flowing inside the data lake and the considerable variability of such data, the solution to these problems is not simple. The main contribution examined in this paper is to explore the content documents' abstracts by discovering the latent topics that run through a corpus by analyzing the words of the collection resources stored in the lake. Therefore, we explore the topic modelling techniques to maintain metadata about the data lake's informational content. We specifically focus on addressing the Latent Dirichlet Allocation (LDA) algorithm to extract the hidden topics in the heterogeneous data lake's resources. Data consumers will benefit from the topics uncovered since they will be able to find the data they need in the enormous amounts of data stored in the data lakes for analysis purposes. Information discovery to identify, locate, integrate, and re-engineer data requires a significant effort and time, accounting for over 70 % of all time spent on data analytics projects, ⁽⁶⁾ which obviously needs to be reduced.

The remainder of this paper is organized as follows: Section 1 examines the background information required to understand the remainder of the paper. Section 2 presents the proposed methodology. Section 3 is devoted to the analysis of the results. Finally, in Section 7, we draw a conclusion to our study and provide research perspectives.

Related works

In this section, we describe the necessary background of the study to enable readers to fully understand the rest of the paper. We started by introducing the data lake concept and these different challenges. Afterwards, we examine the research on topic modeling, these areas of application, and these various technologies

Data lake

Since the term "data lake" was first used in the industry, it has been developing. It was originally introduced by Pentaho CTO James Dixon in 2010 as a solution that manages raw data from one source and serves a variety of user requirements.⁽⁷⁾ It is in contrast with data warehouses or data mart solutions, where strict data extraction, transformation, and cleaning procedures are required and the structure and intended use of the data must be predefined and fixed. But with data lakes, it is possible to bypass the pricey common operations of data warehouses, including data transformation, by keeping raw data in its native format, without any processing, and storing it "as-is".

More discussions about the structures, purposes, and uses of data lakes occurred in 2015. According to a study,⁽⁸⁾ IBM has improved the issue of data lake governance and put forth the challenge of data swamp, which attempts to make the raw data consumable. Data swamp is the process of treating raw data in the data lake in a series of operations that are often performed when the raw data is first created. These operations are as follows: (a) choose the datasets that may be pertinent to the business profit; (b) explore the legal requirements in the selected datasets; (c) load the input data into the storage systems for the data lake without any time-consuming data transformations; (d) maintain the heterogeneous dataset by describing its descriptive and semantic metadata; (e) the data must be prepared for analytical applications; and (f) allow different users to explore and use the data from the lake by eventually providing them with data visualization. Thus, data lakes also face other challenges, like data maintenance and security. Indeed, Marty (Marty, 2015) has suggested a specific method of data access with an emphasis on data security and event management. For example, not all of the sensitive user data from a human resource management system is physically stored in the data lake; some of it is still kept in the original data repositories. Given sparsely formatted personal data, the architecture is presented to store, process, analyze, and query the data. Additionally, a proposal highlighting the significance of data security.^(9,11)

Since 2016, the popularity of data lakes has exploded in both the business and academic communities. There are high-level proposals about data lake architecture,^(2,3,4,11) as well as comparisons with data warehouses,⁽¹²⁾ and publications that discuss its concept, components, and issues. However, a number of companies have created data lakes as commercial solutions, including IBM and Cloudera, Google, Microsoft, Azure, SAP, Amazon AWS, Snowflake, and Oracle.⁽¹³⁾ Additionally, Teradata provides the fundamental capabilities for data input, metadata management, and data governance through its open-source data lake building platform.^(14,15,16) It can be used by developers to build up specialized features for their own data lakes. Thus, Delta Lake from Databricks is another open-source data lake that provides a storage layer consistent with the Apache Spark APIs.

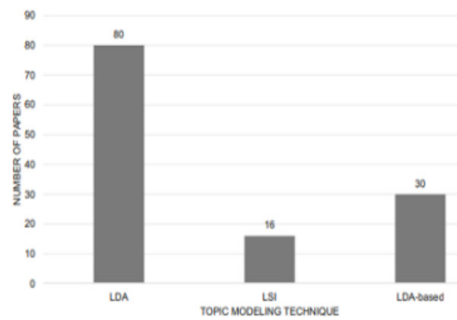
Topic modeling

In computer science, natural language processing (NLP) is a demanding study area that enables computers to understand human language processing in text documents. Topic modeling approaches are strong, intelligent algorithms that are frequently used in NLP to identify topics and mine data from unordered documents.⁽¹⁷⁾ In a broad sense, topic modeling techniques based on LDA have been used in information retrieval, social media analysis, text mining, and natural language processing. For instance, topic modeling focused on social media analytics makes it easier to comprehend how people react and converse in online communities and to extract insightful patterns from their interactions in addition to the content they post on social media websites. Another research group concentrated on topic modeling in software engineering, which allowed them to extract topics from source code and visualize software similarity.⁽¹⁸⁾ Figure 1 illustrates the different areas of application. Meanwhile, LDA is used as a straightforward method to calculate the degree of similarity between source files and determine how each document is distributed across subjects. The authors showed that this method can be useful for software refactoring and project organization. Indeed, there are several topic modeling techniques. Latent Dirichlet Allocation (LDA) is one of the most well-known and widely used methods for topic modeling.^(1,6,19) This model may be applied successfully to a number of document types, including collections of news articles, policy documents, social media posts, software engineering, political science, medicine, linguistics, and tweets.

Due to the fact that topic modeling techniques can be highly beneficial and efficient in natural language processing to semantic mining and latent finding in documents and datasets. Therefore, our motivation is to research topic modeling methodologies in numerous fields with the coverage of multiple features, including models, tools, datasets, and applications. Thus, the significance of topic modeling will grow over time across a variety of fields. In accordance with prior research, we present a categorization of existing approaches to topic models based on the LDA model and in a variety of subject areas, including social media, software engineering, geographic, politics, healthcare, and dialectal science. Table 1, examine some important studies from computational linguistics studies with topic modeling methodology.



(a) The various applications of topic modeling.



(b) Number of papers per topic modelling approach (Silva et al., 2021).

Figure 1. Topic modeling approaches and their field of use

| Table 1. Exemplary LDA-based topic modeling achievements | | |
|--|--|-------------------------------|
| Research proposal | Research Objective | Experiment dataset |
| (Heintz et al., 2013) | A technique for language discovery and a resource for conceptual metaphors. | Wikidata |
| (Lui et al., 2014) | An approach that can locate documents with various languages. | ALTW2010 |
| (Levy and Franklin, 2014) | Investigate Political Conflict for Trucking Industry. | Portal for online regulations |
| (Zirn and Stuckenschmidt, 2014) | Proposed a strategy for analyzing political documents on multiple dimensions | National elections |
| (Zhang et al., 2017) | Discovering user preference distribution | Yelp dataset |
| (Zhang et al., 2015) | Geo-locational cluster discovery | Reuters-21587 |
| (Yang et al., 2017) | Malicious Android app detection | Malicious dataset |
| (Yonggan Li et al., 2016) | Examining the emotions analysis | Twitter dataset |

METHODS

In this section, we will describe the different procedures, as well as the LDA model that we used to overcome the data swamp issue.

Problem statement

Ranking the documents in the search space with a field of expertise as an input query is the challenge of locating significant resources in the data lake. Indeed, finding documents with expertise in a particular field is a routine operation that has a wide range of uses. We suggest a topic modeling strategy for this goal. One of the most well-liked methods for topic modeling that has been effectively used in numerous text mining problems is latent Dirichlet allocation (LDA). It gained popularity by being designated as a sub-task in the information retrieval.

A key objective of an expert discovering system is to assess the likelihood that candidate C is a specialist based on the input query Q. The system can rank the candidates according to this likelihood by computing it for each candidate in the search space. As a result, the primary challenge in expert discovery is precisely estimating $P(C|Q)$.

The goal of topic modeling is to make a coarse-grained representation of the documents and words in a collection of documents. Each document is presented as a probabilistic mixture of topics once a set of topics are extracted from a corpus. Furthermore, topics are probability distributions over words in a document.

LDA Algorithm for Data Lakes

Input: training data D , the number of topics K , Dirichlet parameters α and β
Output: topic assignment matrix Z , topic-document matrix M , word-topic matrix N

1. For all topics k in $[1, K]$ do
2. sample mixture components $k \sim \text{Dir}(\beta)$
3. End for
4. For all documents m in $[1, M]$ do
5. sample mixture proportion $m \sim \text{Dir}(\alpha)$
6. sample document length $N_m \sim \text{Poiss}(\varphi)$
7. For all words n in $[1, N_m]$ do
8. sample topic index $Z_{m,n} \sim \text{Mult}(\theta_m)$
9. sample term for word $W_{m,n} \sim \text{Mult}(\varphi_k)$
10. End for
11. End for

Figure 2. LDA algorithm pseudo-code for data lakes

One of the most well-liked topic modeling strategies nowadays is LDA, which was first introduced by (Blei et al., 2003). LDA is a probabilistic generative model. The fundamental concept is that the documents are depicted as arbitrary mixtures of latent topics, where a topic is defined by a distribution across words. When using word probabilities from LDA, the topics' highest probability words typically provide a good indication of what the topic is. LDA makes the assumption that each topic and document has a consistent Dirichlet prior distribution. Figure 2 depicts the pseudo code for determining the LDA parameters

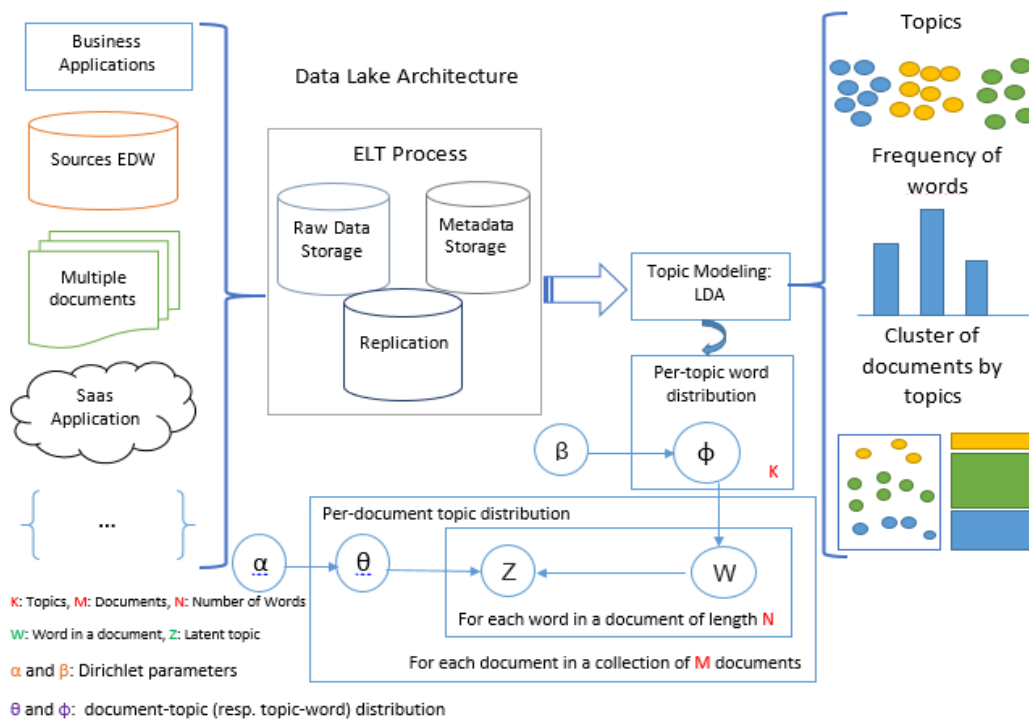
Proposed framework research

To pinpoint the main documents in the data lake, topic analysis of heterogeneous resources, a data mining technique, was used in this study. A massive amount of data kept in electronic and digital records offers great opportunities and has a significant impact on data lake knowledge discovery, information extraction, and analytical reasoning. Yet, a popular method for studying massive data sets is topic modeling. Meanwhile, one of the most well-liked topic modeling techniques, latent dirichlet allocation (LDA), collects vocabulary from a document corpus to create latent "topics." However, given the complexity of the data involved and the difficulty in distributing the computation across several computing nodes, building meaningful topic models with enormous document collections that comprise millions of pages and billions of tokens is tough. Therefore, a number of data processing frameworks have been created in recent years, including Spark, Mallet, and Gensim, to solve the challenges of analyzing massive amounts of unlabeled text from many areas in a scalable, effective way. In this paper, we propose a prototypical implementation based on the LDA algorithm implemented using the Gensim framework for the extraction of topics that represent the accumulation of heterogeneous data stored in lakes. The proposed framework for this investigation is presented holistically in figure 3.

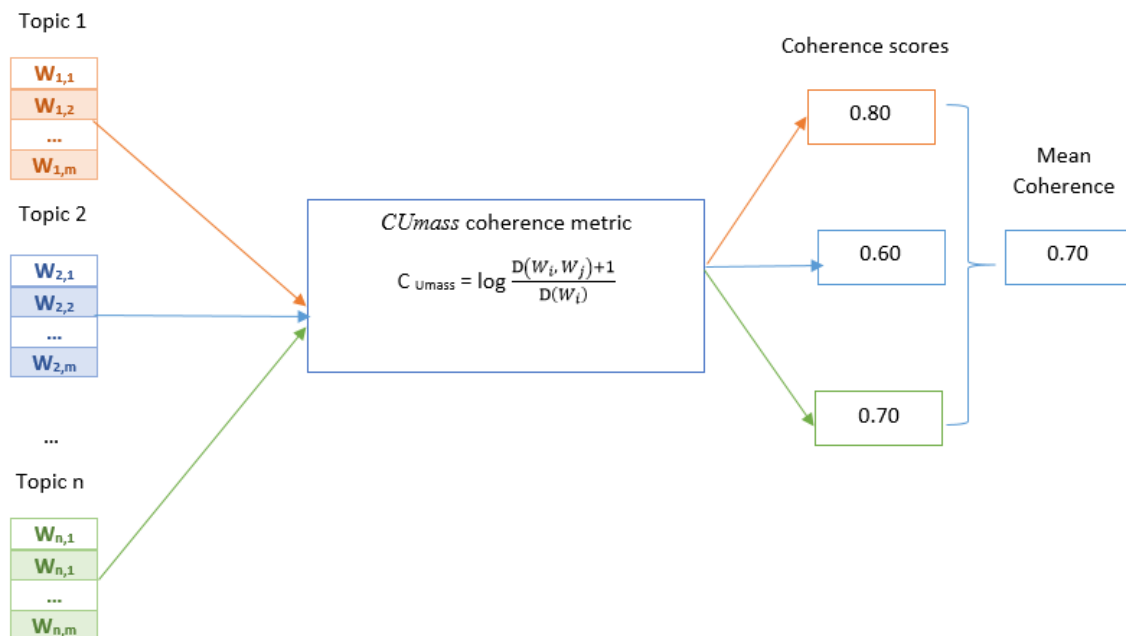
Evaluation metrics

Probabilistic topic models, like LDA, are well-liked text analysis methods because they give the corpus both a predictive and latent topic representation. Since these models' unsupervised training processes make it difficult to evaluate such assumptions, there is a long-standing belief that the latent space they uncover is generally relevant and valuable. But it's also critical to be able to evaluate various models and methodologies and determine whether a trained model is undeniably good or incredible. An impartial evaluation of the quality would be necessary. A single metric that can be optimized and compared would be the ideal way to gather this data. There are several approaches frequently used for the evaluation of such a model, such as LDA. Among these approaches are eyeballing models, including top N words and topics/documents; intrinsic evaluation metrics including capturing model semantics and topics interoperability; human judgement explains the compatibility of the extracted topics with the content of the document; extrinsic evaluation metrics analyze the model's performance in completing predetermined tasks. In this article, we'll explore more about topic coherence, an intrinsic assessment metric, and how to apply it to statistically support the model choice.

Before gaining an understanding of topic coherence, we briefly examine the perplexity metric. Perplexity is a common statistic for language model evaluation and one of the intrinsic evaluation metrics as well. It is expressed as the normalized loglikelihood of a held-out test set and measures how shocked a model is by new data that it has never seen before. By concentrating on the log-likelihood metric, we can consider the perplexity measure as a method for evaluating how likely some new unexpected data is given the previously taught model. Nevertheless, recent research has revealed that the correlation between prediction likelihood (alternatively, perplexity) and human judgment is frequently absent, and even occasionally somewhat anticorrelated.



Therefore, issues produced by optimizing for perplexity cannot be understood by humans. This constraint on perplexity measurement spurred more research aimed at simulating human judgment and, consequently, Topic Coherence. The idea of topic coherence integrates several measurements into a framework to assess the coherence between topics identified by a model. The CUMass coherence metric is among the most frequently used coherence metrics. It builds word content vectors based on their co-occurrences and then computes the score using cosine similarity and normalized pointwise mutual information (NPMI). Figure 4 illustrates the topic coherence measurement process.



Where $D(w_i, w_j)$ represents the number of times the terms w_i and w_j appear together in documents, and $D(w_i)$ represents the number of times the word w_i appears on its own. A better coherence score is indicated by a higher number.

RESULTS

In this section, we will present the main findings of the research proposal. We start by describing the setup for the experiment. Then, we investigate the results analysis.

Experiments Setup

The goal of this study is to discover the most recent data lake resources and eliminate worthless data by identifying the significant issues, such as data swamps. For that, information gathered over a four-year period (2018-2022) from scientific databases such as IISI, Scopus, Web of Science, etc. Using the search engine, the information was chosen based on the following standards: it had to be indexed, published by a credible publisher, and subject to worldwide peer review, as shown in table 2.

Table 2. The list of abstract papers considered and the satisfaction of the selection criteria

| No | Paper type | # of research paper | Reputed publisher | International peer-reviewed |
|----|------------|---------------------|-------------------|-----------------------------|
| 1 | Conference | 9 | X | X |
| 2 | Journal | 6 | X | X |

Further, the experiments are conducted on a device whose hardware and software specs are listed in table 3.

Table 3. Experiments device specification

| Hardware Specifications | Software Specifications |
|--|--------------------------------------|
| CPU: Intel(R) Core(TM) i7-10510U CPU @ 1,80GHz | Operating System (OS) : Windows 10 |
| RAM: 16GB | Framework implementation: Gensim |
| DISK: 1TO (HDD), 512 (SSD) | Programming language: Python |
| GPU: NVIDIA GeForce MX250 | Storage space: Data lakes repository |

Results Analysis

The primary goal of this investigation is to identify the hidden topics stored in the lake. For this, the top 10 keywords from a pool of 498 that appear in 15 research paper abstracts are chosen using semantic analysis. However, determining the ideal number of topics is extremely challenging, especially if there is no prior information about the data. In this paper, we propose to use topic coherence to determine the optimal number of topics across two fixed LDA hyper-parameters validation $\alpha= 0,1$ and $\beta= 0,89$. This evaluation measurement aids in reaching the highest accuracy and decreasing processing time. Figure 5 shows the optimal number of topics.^(20,21,22)

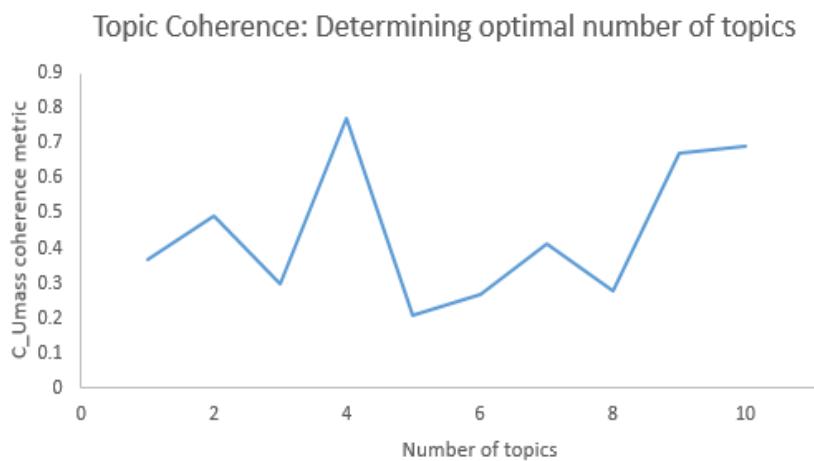


Figure 5. Selecting the optimum number of LDA topics

On the other hand, the key to success in LDA is to set the correct value of LDA hyper-parameters to find the optimal number of topics that will take a reasonable amount of processing time with high consistency, as we can see in table 4. Although there are many inventive ways to approach the selection procedure, in this research paper, we chose the values that produced the highest CUMass score for $K = 4$.⁽²³⁾

| Table 4. The ideal LDA hyper-parameters provide the optimum number of topics | | | |
|--|---------|--------|-------------|
| α | β | CUMass | # of topics |
| 0,1 | 0,89 | 0,672 | 4 |
| 0,13 | 0,90 | 0,657 | 4 |
| 0,72 | 0,92 | 0,628 | 4 |
| 0,81 | 0,93 | 0,597 | 4 |

Meanwhile, the frequency of words is visualized by using the bar chart illustrated in figure 6 to understand the data analysis results intuitively.⁽²⁴⁾

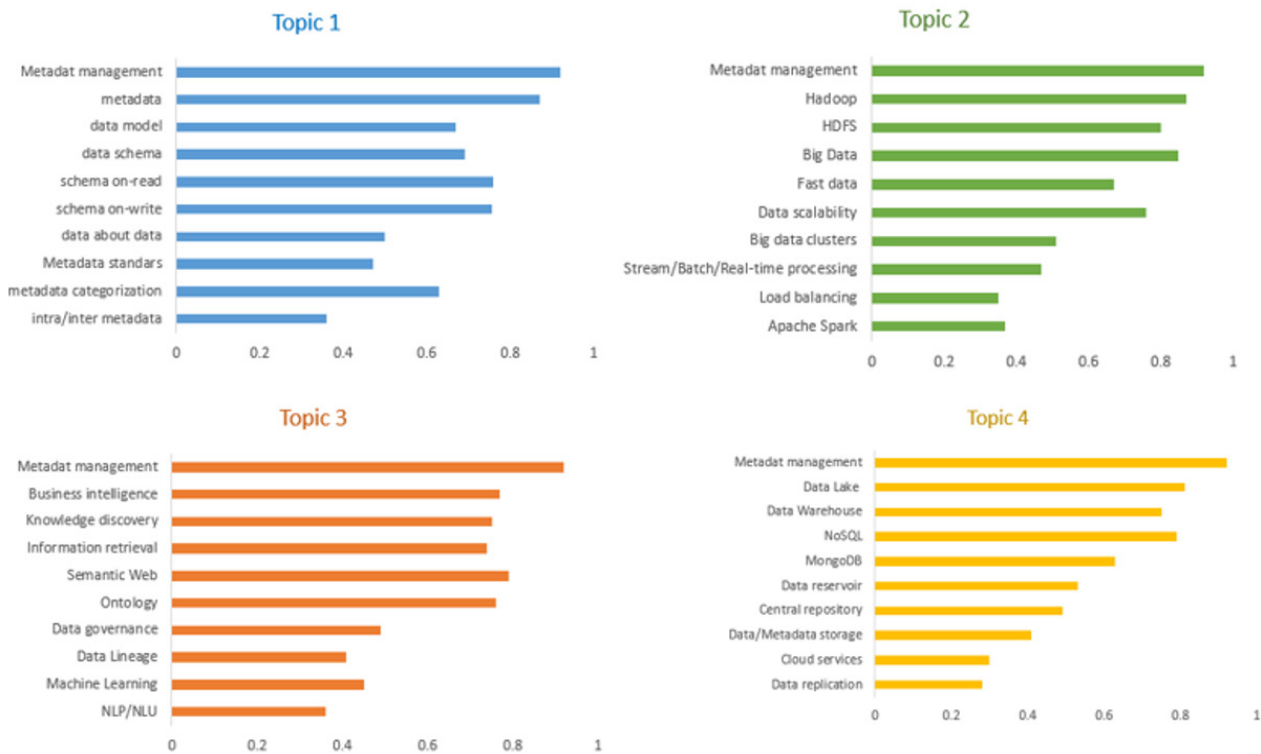


Figure 6. Words per topic in data lake documents

According to the results, the main “top 10” keywords in the data lake system cover the topics of ‘metadata management’ rather than ‘Data Catalog’, ‘NoSQL databases’, ‘Cloud Services’, ‘Hadoop Framework’, ‘Ontology’, ‘Data Governance’, ‘Semantic Web’, ‘Machine Learning algorithms’, ‘Big Data ecosystem’, and ‘Business Intelligence’. Therefore, the LDA algorithm fit the need for data analysis and representation of latent topics which gives us a clear view of the content of the lakes and subsequently avoids the data lake issues called data swamp.^(12,15,20,21,22)

CONCLUSION

In this paper, we investigate the importance of LDA for modeling data lakes by providing their classification into four categories, avoiding the data swamp issue. Indeed, the generated topics can be grouped into the reference category and used as a tool to make it easier for authors to understand the documents gathered in the lake. Thus, we addressed topic extraction techniques, performance measurements, and estimating inference parameters. Furthermore, the LDA approach has been successfully implemented as a topic modeling algorithm and offers better performance in terms of topic coherence. An important future perspective is derived through the use of labeled documents. It consists of establishing the label of each of the determined topics implicit information in the context of the documents to create better predictions.

REFERENCES

1. Fang, H., 2015. Managing data lakes in big data era: What’s a data lake and why has it become popular in

data management ecosystem. 2015 IEEE International Conference on Cyber Technology in Automation.

2. Kim, M., Kim, D., 2022. A Suggestion on the LDA-Based Topic Modeling Technique Based on ElasticSearch for Indexing Academic Research Results. *Applied Sciences* 12, 3118. <https://doi.org/10.3390/app12063118>

3. Suriarachchi, I., Plale, B., 2016. Crossing analytics systems: A case for integrated provenance in data lakes, in: 2016 IEEE 12th International Conference on E-Science (e-Science). Presented at the 2016 IEEE 12th International Conference on e-Science (e-Science), pp. 349-354. <https://doi.org/10.1109/eScience.2016.7870919>

4. Silva, C.C., Galster, M., Gilson, F., 2021. Topic modeling in software engineering research. *Empir Software Eng* 26, 120. <https://doi.org/10.1007/s10664-021-10026-0>

5. Yeh, wei-chih, McIntosh, S., Sobolevsky, S., Hung, P., 2017. Big Data Analytics and Business Intelligence in Industry. *Information Systems Frontiers* 19. <https://doi.org/10.1007/s10796-017-9804-9>

6. Inmon, B., 2016. *Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump*, First edition. ed. Technics Publications, Basking Ridge, NJ.

7. Yang, X., Lo, D., Li, L., Xia, X., Bissyandé, T.F., Klein, J., 2017. Characterizing malicious Android apps by mining topic-specific data flow signatures. *Information and software technology* 27-39.

8. Cherradi, M., El Haddadi, A., Routaib, H., 2022. Data Lake Management Based on DLDS Approach. pp. 679-690. https://doi.org/10.1007/978-981-16-3637-0_48

9. Cherradi, M., El Haddadi, A., 2022a. Grover's Algorithm for Data Lake Optimization Queries. *International Journal of Advanced Computer Science and Applications* 13, 568-576. <https://doi.org/10.14569/IJACSA.2022.0130866>

10. Terrizzano, I.G., Schwarz, P., Roth, M., Colino, J.E., 2015. *Data Wrangling: The Challenging Journey from the Wild to the Lake*. Presented at the Conference on Innovative Data Systems Research.

11. Ashish, T., Ben, S., 2016. *Architecting Data Lakes* [Book] [WWW Document]. URL <https://www.oreilly.com/library/view/architecting-data-lakes/9781492042518/> (accessed 2.12.23).

12. Heintz, I., Gabbard, R., Srivastava, M., Barner, D., Black, D., Friedman, M., Weischedel, R., 2013. Automatic Extraction of Linguistic Metaphors with LDA Topic Modeling.

13. Alnoukari, M., 2022. From Business Intelligence to Big Data: The Power of Analytics. pp. 823-841. <https://doi.org/10.4018/978-1-6684-3662-2.ch038>

14. Zhang, L., Sun, X., Zhuge, H., 2015. Topic discovery of clusters from documents with geographical location. *Concurrency and Computation: Practice and Experience* 27. <https://doi.org/10.1002/cpe.3474>

15. Romero-Carazas R. Prompt lawyer: a challenge in the face of the integration of artificial intelligence and law. *Gamification and Augmented Reality* 2023;1:7-7. <https://doi.org/10.56294/gr20237>.

16. Gonzalez-Argote J. A Bibliometric Analysis of the Studies in Modeling and Simulation: Insights from Scopus. *Gamification and Augmented Reality* 2023;1:5-5. <https://doi.org/10.56294/gr20235>.

17. Gonzalez-Argote D, Gonzalez-Argote J, Machuca-Contreras F. Blockchain in the health sector: a systematic literature review of success cases. *Gamification and Augmented Reality* 2023;1:6-6. <https://doi.org/10.56294/gr20236>.

18. Madera, C., Laurent, A., 2016. The next information architecture evolution: the data lake wave, in: *Proceedings of the 8th International Conference on Management of Digital EcoSystems, MEDES*. Association for Computing Machinery, New York, NY, USA, pp. 174-180. <https://doi.org/10.1145/3012071.3012077>

19. Cherradi, M., El Haddadi, A., 2023. DLDB-Service: An Extensible Data Lake System, in: Ben Ahmed, M., Abdelhakim, B.A., Ane, B.K., Rosiyadi, D. (Eds.), *Emerging Trends in Intelligent Systems & Network Security*,

Lecture Notes on Data Engineering and Communications Technologies. Springer International Publishing, Cham, pp. 211-220. https://doi.org/10.1007/978-3-031-15191-0_20

20. Zhang, Y., Chen, M., Huang, D., Wu, D., Li, Y., 2017. iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. *Future Generation Computer Systems* 66, 30-35. <https://doi.org/10.1016/j.future.2015.12.001>

21. Cherradi, M., El Haddadi, A., 2022b. Data Lakes: A Survey Paper. pp. 823-835. https://doi.org/10.1007/978-3-030-94191-8_66

22. Dixon, J., 2010. Pentaho, Hadoop, and Data Lakes | James Dixon's Blog [WWW Document]. URL <https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/> (accessed 2.12.23).

23. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L., 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimed Tools Appl* 78, 15169-15211. <https://doi.org/10.1007/s11042-018-6894-4>

24. Levy, K., Franklin, M., 2014. Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry. *Social Science Computer Review* 32, 182-194. <https://doi.org/10.1177/0894439313506847>

25. Yonggan Li, Xueguang Zhou, Yan Sun, Huanguo Zhang, 2016. Design and implementation of Weibo sentiment analysis based on LDA and dependency parsing. *China Commun.* 13, 91-105. <https://doi.org/10.1109/CC.2016.7781721>

26. Ruzgas, T., Bagdonavičienė, J., 2017. Business Intelligence for Big Data Analytics. *International Journal of Computer Applications Technology and Research* 6, 001-008. <https://doi.org/10.7753/IJCATR0601.1001>

27. Cherradi, M., El Haddadi, A., Routaib, H., 2021. Moroccan Data Lake Healthcare Analytics for Covid-19. <https://doi.org/10.5220>

FINANCING

No financing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Amine El Haddadi, Oumaima El Haddadi, Mohamed Cherradi, Fadwa Bouhafer, Anass El Haddadi, Ahmed El Allaoui.

Research: Amine El Haddadi, Oumaima El Haddadi, Mohamed Cherradi, Fadwa Bouhafer, Anass El Haddadi, Ahmed El Allaoui.

Drafting - original draft: Amine El Haddadi, Oumaima El Haddadi, Mohamed Cherradi, Fadwa Bouhafer, Anass El Haddadi, Ahmed El Allaoui.

Writing - proofreading and editing: Amine El Haddadi, Oumaima El Haddadi, Mohamed Cherradi, Fadwa Bouhafer, Anass El Haddadi, Ahmed El Allaoui.