International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)
https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Topic Modeling for Web Page using LDA Algorithm and Web Content Mining: Testing and Evaluation

Noor Muneam Abbas[a], Raad Mahmood Mohammed[b], Hasan Aqeel Abbood[c], Mohammed Ali Mohammed[d*]

[a]*Computer Science Department, University of Technology, Baghdad, Iraq.*
[b,c,d]*College of Business Informatics, University of Information Technology and Communications (UOITC), Baghdad, Iraq.*
[a]*Email: 110128@uotechnology.edu.iq*
[d]*Email: mohammed.ali@uoitc.edu.iq*

Abstract

In recent years, the content of websites has become useful and is increasing rapidly, this information plays an important role in discovering various knowledge on the web. This paper aims to test and evaluate our previous work with the new dataset. The previous system applied the LDA Algorithm for Topic Modelling in Web content mining, which was tested and discussed on: different science content, a large dataset, and similarity value. According to the results on our new dataset (No. of rows: 298, No. of columns: 6, Computer, Mathematical, Physics, Chemistry Sciences), the system approves that the LDA algorithm is the best on the web content mining dataset.

Keywords: Topic Modeling; Web Content Mining; MRH dataset; LDA algorithm.

1. Introduction

Web Content Mining (WCM), one of the types, is defined as a significant process of extracting the knowledge and important information from web pages. WCM is a very important area due to the majority of the web content is text-based. WCM is a semi-structured web with two types: firstly, directly mine the content of pages. Secondly, improve the content search of other tools such as search engines. WCM is used to mine the data collected from web pages. There are many technologies used to mine such as Natural Language Processing (NLP) and Information Retrieval (IR) [1].

Received: 5/29/2025
Accepted: 7/12/2025
Published: 7/22/2025

---

* *Corresponding author.*

There are two methodologies to mine in WCM: the database and agent-based approach. The first approach helps in retrieving the semi-structured data from the web pages [2,3]. The three kinds of agents are customized web agents, information filtering categorizing agent and intelligent search agents [2]. Customized web agents try to find a web page in the user's profile. Information filtering categorizing agents reduce the user's time and effort in locating the relevant document through the specialized domain knowledge they possess. The filtering agent filters out irrelevant incoming documents and presents to the user only those documents that match the user's interest. Automatically, Intelligent search agents discover information according to a particular query utilizing user profiles [2,3,4].

Two main issues (managing problems, data querying) can be solved by database techniques for web services. There are three categories of tasks related to handling those problems: modeling and querying the web, information extraction and integration, and website construction and restructuring [5].

This paper aims to test and evaluate our previous research paper [6] using our new database [7]. In [6] we generate the topic model for a website using the LDA (Latent Dirichlet Allocation) algorithm and based on web content mining. Then, a comparison (according to similarity value) between the LDA and NMF (Non-negative Matrix Factorization) algorithms, shows that the LDA algorithm is the best algorithm for web content mining approach. So, the new test and evaluation will be applying an LDA algorithm to our new dataset (a larger number of diverse data with different situations).

In [7], our dataset is semi-structured type with a size of 4.05 MB. The dataset will be stored in CSV format (.csv), which can be opened in MS Excel. The dataset structure (shown in figure 1) contains many sheets (as a table in a database). The structure is designed as a relationship between the sheets in order to apply normalization, reduce the size, thus, for faster processing.

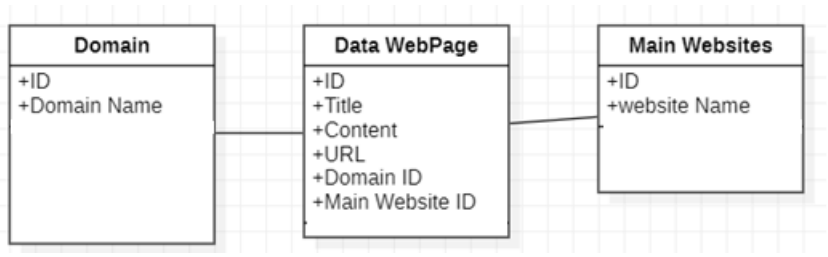

Figure 1: Structure of Dataset [7]

The total number of webpages (no. of rows) in the dataset is 298, broken into 144 in the Computer Science domain, 30 in the Chemistry Science domain, 38 in the Physics Science domain, and 86 in the Mathematics Science domain. we are going to use the following URLs: “geeksforgeeks.org” [8], “mathworld.wolfram.com” [9], “www.chemguide.co.uk” [10], “www.physicsclassroom.com” [11], and “socratic.org” [12] to create this dataset and save it as a CSV file.

The remaining sections of the paper are structured as follows: Section 2 presents related works. Section 3 showed the previously proposed system. Section 4 shows the result and discussion with our dataset. section 5

presents the system workbench. Finally, Section 6 concludes the paper.

2. Related Works

This section will briefly describe the previous studies on topic modelling algorithms and then summarize the key aspects of each study as shown in Table 1. Bhat and his colleagues (2019) [13] proposed two variants of Latent Dirichlet allocation (LDA) based on DNN, 2NN Deep LDA, and 3NN Deep LDA. Yang and his colleagues (2020) [14] proposed a distribution topic model, referred to as Named Entity Topic Model (NETM), to extract web content popularity growth factors. Then compared the NETM and LDA model. Youngseok and his colleagues (2021) [15] propose a web page ranking method using topic modelling for effective information collection and classification. The proposed method is applied to the document ranking technique to avoid duplicate crawling when crawling at high speed. Hamza H.M and his colleagues (2023) [16] present the HTML Topic Model (HTM) as an innovative topic model. The input values a HTML tags, and the output values understand the structure of web pages. the benefit of this model is to learn coherent topics in web content data. Simra Shahid and his colleagues (2023) [17] present HyHTM- a Hyperbolic geometry-based Hierarchical Topic Model that addresses some limitations by incorporating hierarchical information from hyperbolic geometry to explicitly model hierarchies in topic models.

Table 1: Summarizing the key aspects of each study

Study	Proposed Method	Focus/Improvement	Key Findings
Bhat and his colleagues (2019) [13]	2NN Deep LDA, 3NN Deep LDA	LDA with Deep Neural Networks for improved topic learning	2NN Deep LDA is faster than LDA and 3NN, providing better topic learning accuracy.
Yang and his colleagues (2020) [14]	Named Entity Topic Model (NETM)	Extracting web content popularity growth factors	NETM outperforms LDA in accuracy but does not account for the HTML structure of web pages.
Youngseok and his colleagues (2021) [15]	Web document ranking using topic modelling	Efficient document classification and redundant crawling elimination	Proposed method enhances document ranking and ensures efficient classification while avoiding duplication.
Hamza H.M and his colleagues (2023) [16]	HTML Topic Model (HTM)	Incorporating HTML tags into topic modeling	HTM outperforms LDA and Correlated Topic Model in topic coherence when applied to web content.
Simra Shahid and his colleagues (2023) [17]	HyHTM (Hyperbolic Geometry-based Hierarchical Topic Model)	Modeling hierarchical relationships among topics	HyHTM better attends to parent-child topic relationships, outperforming four baseline models.

3. The Proposed System

In [6], the design and implementation of the proposed topic model with a web content mining system, which is

used to create a topic name from the content of a web page. the system consists of the following steps: preprocessing step, the content will be cleaned, and preparing for to next step. Bag of words, Compute the count of the total occurrences of the most frequently used words. Apply the LDA algorithm to get the weighted words for the document. Topic Label, the system uses the Gemini [18] chatbot to generate a label for the LDA words. Figure 2 shows the flowchart, and Algorithm 1 shows the algorithm steps of the system [6].

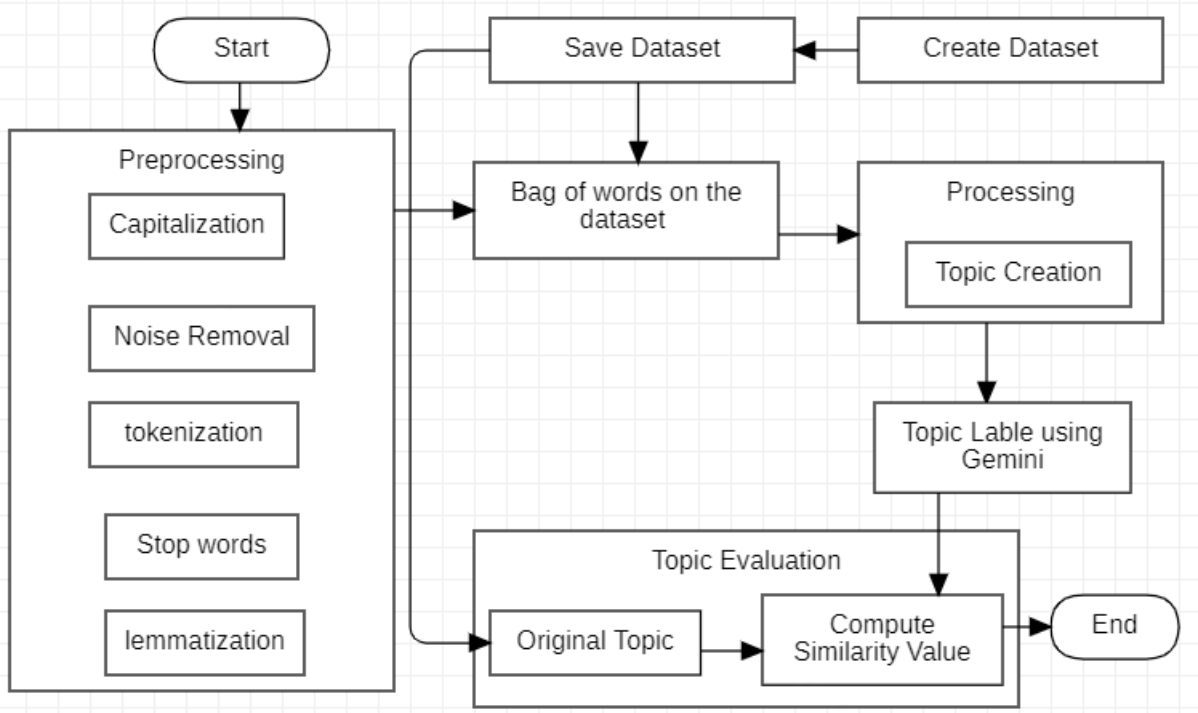

Figure 2: proposed system [6]

Table

Algorithm 1: Proposed Methodology [6]	
Input:	Document
Output:	Topic_Name
Steps:	
1.	Preprocessing Step (Tokenization, Stop words, Lemmatization, Data Cleaning, etc.). Compute the Bag of words.
2.	Processing Topic Creation using the LDA algorithm.
3.	Topic Labeling using Gemini with the top 5 words for a document
4.	Compute the Similarity value between (generate topic, original topic).
5.	

4. Result and Discussion

The system [6] will be tested on the dataset [7] and also compared with the NMF algorithm. The sample result of topics name for computer science is shown in Table 2 for the NMF algorithm and Table 3 for the LDA algorithm. And the sample result of topics name for a physics science is shown in Table 4 for the NMF algorithm and Table 5 for the LDA algorithm. And the sample result of topics name for a chemistry science is shown in Table 6 for the NMF algorithm and Table 7 for the LDA algorithm. Finally, the sample result of topics

name for a mathematics science is shown in Table 8 for the NMF algorithm and Table 9 for the LDA algorithm.

Table 2: Topic Name for computer science using the NMF algorithm

No	Original Topic Page	NMF Words	Gemini Topic Label
1	Backend Development	development, backend, web, application, developer	Backend Web Application Development
2	Machine Learning Tutorial	learning, machine, data, tutorial, algorithm	Machine Learning Algorithms Tutorial
3	Software Engineering Tutorial	software, model, engineering, development, tutorial	Software Model Engineering Development Tutorial

Table 3: Topic Name for computer science using the LDA algorithm

No	Original Topic Page	LDA Words	Gemini Topic Label
1	Backend Development	development, backend, web, application, developer	Backend Web Application Development
2	Machine Learning Tutorial	learning, machine, data, tutorial, algorithm	Machine Learning Algorithms Tutorial
3	Software Engineering Tutorial	software, model, engineering, development, tutorial	Software Model Engineering Development Tutorial

Table 4: Topic Name for physics science using the NMF algorithm

No	Original Topic Page	NMF Words	Gemini Topic Label
1	Solids, liquids and gases	gas, liquid, solid, change, simple	Simple State Changes: gases, liquids, Solids
2	transition metals	metals, chemistry, transition, menu, ion	transition metals Ion Chemistry
3	Electrolysis	electrolysis, calculation, menu, avogadro, constant	Electrolysis Calculation

Table 5: Topic Name for physics science using the LDA algorithm

No	Original Topic Page	LDA Words	Gemini Topic Label
1	Solids, liquids and gases	Diffusion in gases, liquids, and Solids Mixtures	gas, liquid, solid, mixture, diffusion
2	transition metals	metals, chemistry, ion, menu, transition	transition metals Ion Chemistry
3	Electrolysis	electrolysis, menu, calculation, basic, introduction	Basic Electrolysis Calculation

Table 6: Topic Name for chemistry science using the NMF algorithm

No	Original Topic Page	NMF Words	Gemini Topic Label
1	Interference	wave, interference, phase, amplitude, destructive	Destructive Wave Interference
2	Quantization of Energy	energy, photon, physic, quantization, quantum	Quantum Photon Energy
3	Electric Force	force, electric, field, point, path	Electric Field Force Along a Path

Table 7: Topic Name for chemistry science using the LDA algorithm

No	Original Topic Page	LDA Words	Gemini Topic Label
1	Interference	interference, amplitude, phase, destructive	Destructive Interference
2	Quantization of Energy	energy, photon, quantization, physic, quantum	Quantum Photon Energy Quantization
3	Electric Force	force, electric, field, point, energy	Electric Field Force and Point Energy

Table 8: Topic Name for mathematics science using the NMF algorithm

No	Original Topic Page	NMF Words	Gemini Topic Label
1	Divided Difference	difference, first, sometimes, point, function	First Difference of a Function at a Point
2	BBP-Type Formula	formula, 8k, bailey, type, 129	Bailey 8K Type 129 Formula
3	Central Difference	delta, 3f, central, difference, integer	Central Difference Integer Delta Formula (3f)

Table 9: Topic Name for mathematics science using the LDA algorithm

No	Original Topic Page	LDA Words	Gemini Topic Label
1	Divided Difference	difference, function, divided, sometimes, point	Divided Difference Function at a Point
2	BBP-Type Formula	bailey, formula, type, pp, bbp	Bailey Formula PP and BBP Type
3	Central Difference	delta, difference, central, interval, involve	Central Difference Delta Intervals

And the evolution step, Table 10 and Figure 3 show that the LDA algorithm is the best among the NMF algorithm according to the cosine similarity value, the similarity will be computed between the original topic and the generated topic.

Table 10: Similarity values between topics

No	Original Title	NMF Title	Similarity Value	LDA Title	Similarity Value
1	Backend Development	Backend Web Application Development	0.7071	Backend Web Application Development	0.7071
2	Machine Learning Tutorial	Machine Learning Algorithms Tutorial	0.8660	Machine Learning Algorithms Tutorial	0.8660

3	Software Engineering Tutorial	Software Model Engineering Development Tutorial	0.7745	Software Model Engineering Development Tutorial	0.7745
4	Solids, liquids and gases	Simple State Changes: gases, liquids, Solids	0.7071	Diffusion in gases, liquids, and Solids Mixtures	0.8164
5	transition metals	transition metals Ion Chemistry	0.7071	transition metals Ion Chemistry	0.7071
6	Electrolysis	Electrolysis Calculation	0.7071	Basic Electrolysis Calculation	0.5773
7	Interference	Destructive Wave Interference	0.5773	Destructive Interference	0.7071
8	Quantization of Energy	Quantum Photon Energy	0.4082	Quantum Photon Energy Quantization	0.7071
9	Electric Force	Electric Field Force Along a Path	0.6324	Electric Field Force and Point Energy	0.6324
10	Divided Difference	First Difference of a Function at a Point	0.3535	Divided Difference Function at a Point	0.7071
11	BBP-Type Formula	Bailey 8K Type 129 Formula	0.5163	Bailey Formula PP and BBP Type	0.7745
12	Backend Development	Central Difference Integer Delta Formula (3f)	0.7071	Central Difference Delta Intervals	0.7071

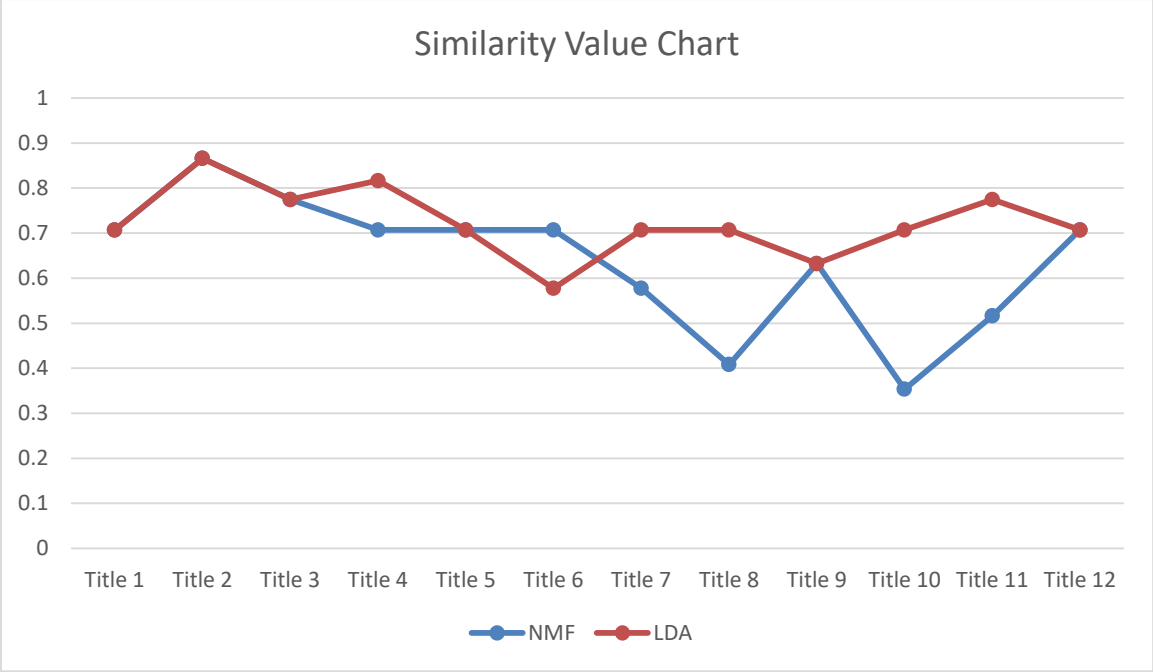

Figure 3: Similarity values between topics

Cosine Similarity [19]: This metric measures the cosine of the angle between two vectors (the original title and the generated title). A cosine similarity score of 1 indicates perfect similarity, while a score of 0 means no similarity. **Analysis for each Title**:

The comparison of title generation using LDA and NMF shows that, in general, LDA performs better than NMF in generating titles that are more semantically aligned with the original titles, as evidenced by the cosine similarity values. Both models perform well in capturing the general topics, but LDA tends to generate more precise and contextually relevant titles in most cases.

For example, for the title "Solids, liquids and gases", LDA generates a title "Diffusion in gases, liquids, and Solids Mixtures" with a similarity of 0.8164, which is higher than NMF's title "Simple State Changes: gases, liquids, Solids" with a similarity of 0.7071. This indicates that LDA is better at understanding the specific relationship between the concepts, while NMF produces a more general and less accurate title.

In cases like "Machine Learning Tutorial", "Software Engineering Tutorial", and "Backend Development", both LDA and NMF generate similar titles with high cosine similarity, indicating that both models perform well with straightforward or general topics. However, LDA consistently outperforms NMF in more complex or technical cases.

The title "Electrolysis" provides an interesting case where NMF performs better than LDA. The similarity values are 0.7071 for NMF's generated title ("Electrolysis Calculation") and 0.5773 for LDA's generated title ("Basic Electrolysis Calculation"). This indicates that NMF's output, "Electrolysis Calculation", is closer in meaning to the original title, "Electrolysis", compared to LDA's "Basic Electrolysis Calculation". One possible

reason for this could be that NMF produces a title that is more concise and directly related to the core concept of "Electrolysis".

Overall, the results suggest that LDA is the more effective model for generating topic titles compared to NMF, particularly when the topic is technical or when more specific terms are involved. While NMF performs well in simpler cases, LDA's ability to capture detailed relationships and generate more relevant titles makes it the superior choice for this task.

5. System Workbench

In this section, we will be describing the system requirements such as Central Processing Unit (CPU) and Memory for our approach using LDA and NMF algorithms. According to the reviewing of the CPU (shown in figure 4) and memory usage (shown in figure 5) for LDA algorithm, the Average CPU required is 24.025% which approximate one core (64-bit core Intel(R) Core(TM) i5-6440HQ processor) And the Average Memory required is 1.729 GB.

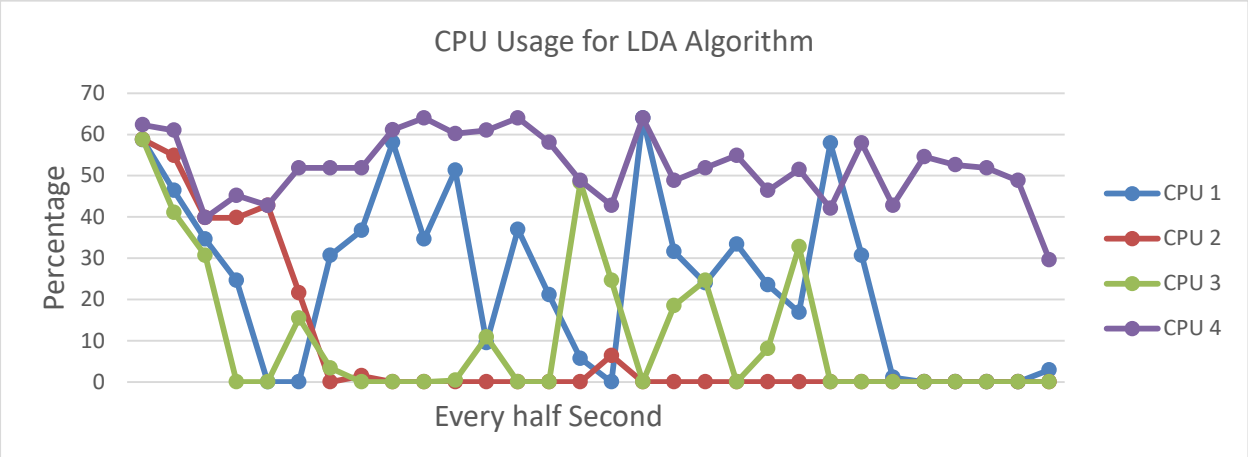

Figure 4: CPU usage for the LDA algorithm

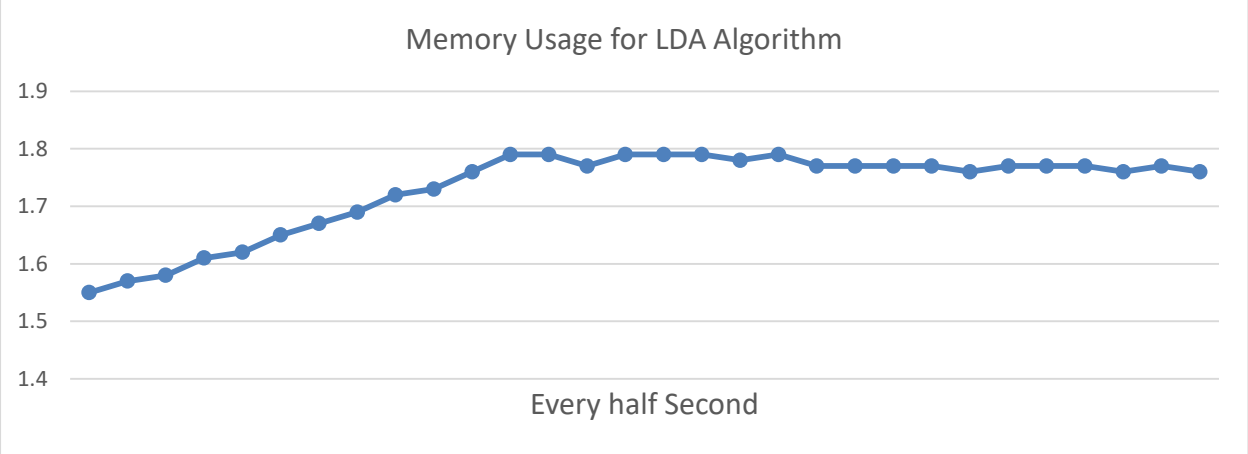

Figure 5: Memory Usage for the LDA algorithm

Also, according to the reviewing of the CPU (shown in figure 6) and memory usage (shown in figure 7) for NMF algorithm, the Average CPU required is 30.129% which approximate one core (64-bit core Intel(R) Core(TM) i5-6440HQ processor) And the Average Memory required is 1.451 GB.

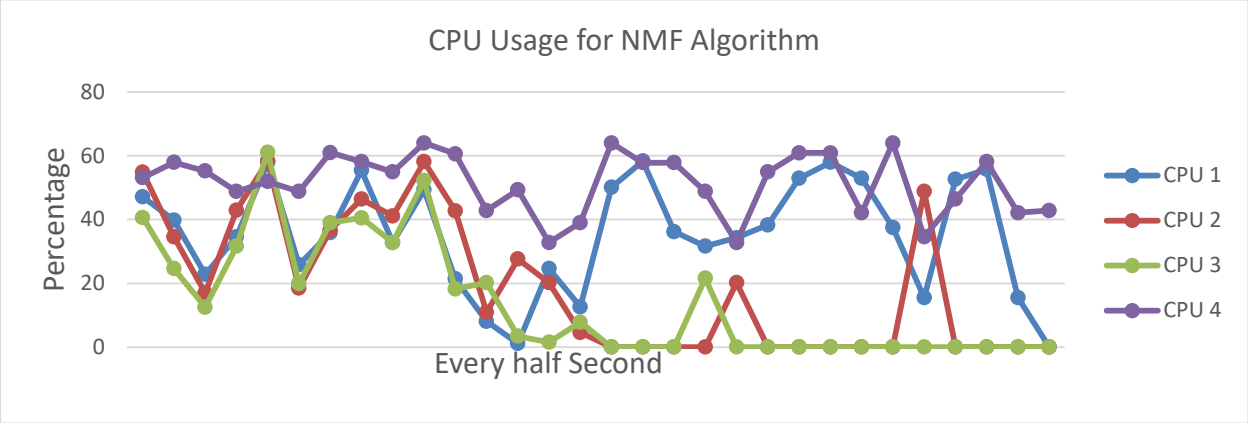

Figure 6: CPU usage for NMF algorithm

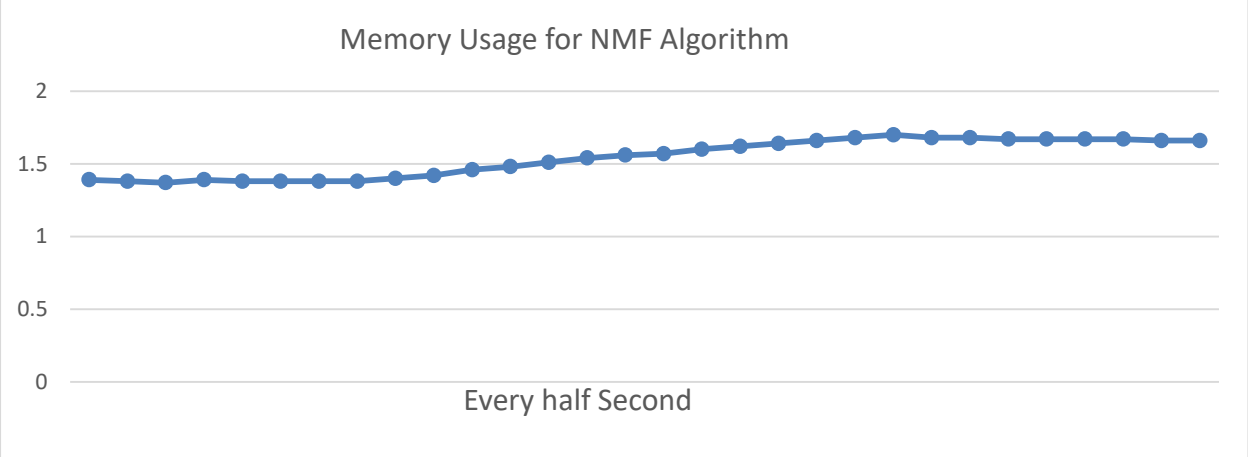

Figure 7: Memory Usage for the NMF algorithm

Finally, according to this workbench and the values are approximately equal, which is based on the average value, so we conclude that the LDA and NMF algorithms require approximately the same CPU and memory usage.

6. Conclusion

This study evaluates the performance of the LDA algorithm for topic modeling on a new dataset with diverse science content, including computer, mathematical, physics, and chemistry topics. The evaluation, based on cosine similarity between original titles and those generated by LDA and NMF, shows that LDA consistently outperforms NMF, particularly for more complex or technical topics. While both models perform similarly for simpler topics, LDA demonstrates superior performance for domain-specific content, generating more relevant titles. The results from this dataset (298 rows, 6 columns) confirm that LDA is the most effective topic modeling

algorithm for web content mining, making it an ideal tool for knowledge discovery across diverse scientific fields.

References

[1] D. Navadiya and R. Patel, "Web Content Mining Techniques – A Comprehensive Survey," *Int. J. Eng. Res. Technol. (IJERT)*, vol. 1, no. 10, pp. 1–6, 2012. [Online]. Available: https://doi.org/10.17577/IJERTV1IS10269

[2] S. A. Inamdar and G. N. Shinde, "An agent based intelligent search engine system for web mining," *Res., Reflections and Innovations in Integrating ICT in Educ.*, 2000. [Online]. Available: https://doi.org/10.20935/AcadNano7479

[3] K. R. Srinath, "An overview of web content mining techniques," *Int. Res. J. Eng. Technol. (IRJET)*, vol. 4, no. 11, pp. 1258–1261, 2017. [Online]. Available: https://www.irjet.net/archives/V4/i11/IRJET-V4I11335.pdf

[4] R. H. Salman, M. Zaki, and N. A. Shiltag, "A studying of web content mining tools," *Al-Qadisiyah J. Pure Sci.*, vol. 25, no. 2, 2020. [Online]. Available: https://doi.org/10.29350/2411-3514.1202

[5] D. Florescu, A. Levy, and A. Mendelzon, "Database techniques for the World-Wide Web: A survey," *ACM SIGMOD Rec.*, vol. 27, no. 3, pp. 59–74, 1998. [Online]. Available: https://doi.org/10.1145/290593.290605

[6] M. A. Mohammed, R. M. Mohammed, and H. A. Abbood, "Topic modeling for web page using LDA algorithm and web content mining," *J. Educ. Pure Sci.*, vol. 15, no. 3, in press, 2025.

[7] M. A. Mohammed, H. A. Abbood, and R. M. Mohammed, "MRH: A large-scale text dataset for web content mining," *J. Port Sci. Res.*, vol. 8, no. 4, pp. 321–326, 2025. [Online]. Available: https://doi.org/10.36371/port.2025.4.2

[8] GeeksforGeeks. [Online]. Available: https://www.geeksforgeeks.org/. Accessed: May 3, 2025.

[9] Wolfram MathWorld. [Online]. Available: https://mathworld.wolfram.com. Accessed: May 4, 2025.

[10] Chemguide. [Online]. Available: https://www.chemguide.co.uk. Accessed: May 6, 2025.

[11] The Physics Classroom. [Online]. Available: https://www.physicsclassroom.com. Accessed: May 7, 2025.

[12] Socratic. [Online]. Available: https://socratic.org. Accessed: Mar. 7, 2025.

[13] K. Sharma, G. Shrivastava, and V. Kumar, "Web mining: Today and tomorrow," in *Proc. 3rd Int. Conf.*

Electron. Comput. Technol., 2011, vol. 1, pp. 399–403. [Online]. Available: https://doi.org/10.1109/ICECTECH.2011.5941631

[14] Y. Yang, Y. Liu, X. Lu, J. Xu, and F. Wang, "A named entity topic model for news popularity prediction," *Knowl.-Based Syst.*, vol. 208, p. 106430, 2020. [Online]. Available: https://doi.org/10.1016/j.knosys.2020.106430

[15] Y. Lee and J. Cho, "Web document classification using topic modeling based document ranking," *Int. J. Electr. Comput. Eng.*, vol. 11, pp. 2386–2392, 2021. [Online]. Available: https://doi.org/10.11591/ijece.v11i3.pp2386-2392

[16] H. H. Altarturi, M. Saadoon, and N. B. Anuar, "Web content topic modeling using LDA and HTML tags," *PeerJ Comput. Sci.*, vol. 9, p. e1459, 2023. [Online]. Available: https://doi.org/10.7717/peerj-cs.1459

[17] S. Shahid *et al*., "HyHTM: Hyperbolic geometry based hierarchical topic models," *arXiv preprint*, arXiv:2305.09258, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.09258

[18] G. Team *et al*., "Gemini: A family of highly capable multimodal models," *arXiv preprint*, arXiv:2312.11805, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2312.11805

[19] J. Ye, "Cosine similarity measures for intuitionistic fuzzy sets and their applications," *Math. Comput. Model.*, vol. 53, no. 1–2, pp. 91–97, 2011. [Online]. Available: https://doi.org/10.1016/j.mcm.2010.07.022