Topic Modelling of Web Pages with Latent Dirichlet Allocation Methods

Akula Phanidhar
Malla Reddy University
Hyderabad, Telangana
2011CS020014@mallareddyuniversity.ac.in

Amara Kowshick
Malla Reddy University
Hyderabad, Telangana
2011CS020011@mallareddyuniversity.ac.in

Akuri Mahendrareddy
Malla Reddy University
Hyderabad, Telangana
2011CS020012@mallareddyuniversity.ac.in

Athnuri Nikitha
Malla Reddy University
Hyderabad, Telangana
2011CS020013@mallareddyuniversity.ac.in

Dr. G. Hariharan
Malla Reddy University
Hyderabad, Telangana
hariharan@mallareddyuniversity.ac.in

***Abstract*— Topic modelling with Latent Dirichlet Allocation (LDA) is a popular technique used in natural language processing to uncover hidden thematic structures within a collection of documents. When applied to web pages, LDA can help in identifying prevalent topics or themes across these pages.. This study delves into the utilization of Latent Dirichlet Allocation (LDA) methods to extract underlying topics within web pages, a fundamental pursuit in understanding the multifaceted landscape of online information. Web content analysis presents unique challenges owing to its diverse nature—comprising text, images, videos, and structured HTML elements—mandating rigorous preprocessing strategies to homogenize the data. By adapting the LDA model to accommodate these challenges, this research tackles the task of uncovering latent thematic structures prevalent across web content. Methodologically, the study explores parameter tuning and model adaptation to optimize LDA for web page analysis, navigating complexities such as varied content formats, noise, and inherent biases in web data. Addressing these intricacies involves parsing HTML, extracting meaningful textual information, and refining tokenization processes. Evaluating the fidelity and interpretability of discovered topics becomes pivotal, prompting the utilization of coherence scores, perplexity metrics, and human assessment to gauge the quality of generated topics.**

Additionally, this research confronts the dynamic nature of web content, proposing strategies like continuous model retraining and dynamic topic modeling to accommodate evolving trends and updates. Practical applications of the extracted topics span a spectrum of domains, encompassing content recommendation systems, user behavior analysis, sentiment analysis, targeted advertising, and the enhancement of search algorithms for improved relevance and user engagement. Supported by illustrative case studies, this study elucidates how LDA serves as a potent mechanism to distill coherent and meaningful topics from web pages, offering invaluable insights into the hidden structures within the vast expanse of online information. St

This comprehensive abstract encapsulates the depth and breadth of employing LDA for the analysis of web content, encompassing challenges, methodologies, evaluations, applications, and real-world implications.

I. INTRODUCTION

The exponential growth of digital content on the internet has rendered the task of understanding and organizing web-based information a formidable challenge. Web pages, diverse in content and structure, present a rich tapestry of data comprising textual information, multimedia elements, and HTML structures. Amidst this wealth of information, uncovering the underlying themes and topics embedded within web content is pivotal for various applications, ranging from enhancing search algorithms to personalized content recommendations and targeted advertising.

In response to this need, techniques from the realm of natural language processing and machine learning have emerged as potent tools for deciphering the latent structures inherent in web pages. One such technique, Latent Dirichlet Allocation (LDA), has garnered significant attention for its ability to extract latent topics from large corpora of text. LDA, a generative probabilistic model, offers a promising avenue for unraveling the thematic dimensions within web content.

However, applying LDA to web pages requires tailored methodologies that address the idiosyncrasies of web data, including its heterogeneous nature, noise, and structural complexities. Challenges also arise in adapting LDA to handle multimedia content, HTML tags, and dynamic updates within web pages. Successful implementation of LDA for web page analysis demands meticulous preprocessing, parameter tuning, and model adaptation to navigate these intricacies.

This study embarks on the exploration of employing LDA methods specifically tailored for web page topic modeling. By investigating the adaptation of LDA to the unique characteristics of web content and addressing the challenges inherent in this domain, this research aims to uncover coherent and meaningful topics embedded within web pages. The subsequent sections delineate the methodologies employed, challenges encountered, strategies devised for adaptation, evaluation metrics utilized, practical applications envisaged, and the broader implications of employing LDA for the analysis of web-based information.

This introduction sets the stage by highlighting the significance of uncovering latent topics in web content, introducing LDA as a potential solution, and outlining the challenges and goals of employing LDA for web page analysis.

II. LITERATURE SURVEY

1. Scalability and Efficiency
Studies focusing on scalability concerns and efficient algorithms for applying LDA to large-scale web corpora, considering computational resources and processing time.

2. Temporal Analysis
Research investigating the temporal dynamics of topics within web pages, addressing the evolution of themes over time and strategies to incorporate time-sensitive information into LDA-based models.

3. Semantic Understanding
Exploration of incorporating semantic information or external knowledge sources (such as ontologies or knowledge graphs) into LDA models for improved topic coherence and semantic understanding of web content.

4. User Behavior and Interaction
Studies examining how LDA-derived topics from web pages intersect with user behavior, user-generated content, or interactions, leading to insights on user engagement and preferences.

5. Ethical and Bias Considerations
Discussions on ethical considerations in web content analysis, including biases introduced by LDA in topic extraction and strategies to mitigate potential biases in results.

6. Multimodal Topic Modeling
Research exploring approaches to fuse textual and non-textual (e.g., image, video) information in web pages for comprehensive multimodal topic modeling using LDA.

7. Spatiotemporal Analysis
Investigations into incorporating spatial information alongside temporal aspects, providing insights into geographic variations and temporal trends in topics across web pages.

8. Interactive and Dynamic Topic Modeling
Studies proposing interactive LDA-based models, allowing users to dynamically explore and refine topics based on feedback or evolving interests.

9. Cross-Domain Applications
Examination of how LDA-based topic modeling on web pages can be generalized or adapted for cross-domain applications beyond traditional web content, such as scientific articles, legal documents, or social media posts.

10. Open Challenges and Future Directions
Summarization of open challenges in the field, proposing potential avenues for future research, advancements, and interdisciplinary collaborations to further enhance LDA-based topic modeling on web pages.

III. METHODOLOGY

Existing System:

1. Data Collection and Preprocessing:
Web Scraping and Collection: Gather web pages from various sources (e.g., websites, forums, blogs) using web scraping techniques.
Text Extraction: Extract relevant textual content from web pages, handling HTML structures, and removing irrelevant information like navigation bars, ads, or boilerplate text.
Text Cleaning: Perform text preprocessing steps including tokenization, stop-word removal, stemming or lemmatization, and handling special characters or noise peculiar to web content.
Multimedia Handling: Address the inclusion of images, videos, or other non-textual elements (if relevant) using complementary techniques like image caption extraction or metadata incorporation.

2. LDA Model Application:
Vocabulary Creation: Generate a vocabulary of terms from the preprocessed text data.
Document-Term Matrix: Create a document-term matrix representing the frequency of terms in documents, preparing the data for LDA modeling.
Parameter Tuning: Experiment with different numbers of topics, alpha and beta hyperparameters, and iteration counts to optimize the LDA model for web page data.
LDA Model Training: Apply LDA algorithm to the prepared document-term matrix to extract latent topics from the web pages.

3. Evaluation and Refinement:
Topic Coherence: Evaluate the coherence of extracted topics using measures like coherence scores to ensure the interpretability and meaningfulness of topics.
Perplexity Evaluation: Calculate perplexity scores as an additional metric to assess the model's performance.
Model Refinement: Iterate on the LDA model by adjusting parameters based on evaluation metrics and domain-specific insights, aiming to improve topic quality.

4. Interpretation and Visualization:
Topic Interpretation: Analyze the top words or terms in each topic to interpret and label them based on the prevalent terms.
Visualization: Create visual representations like word clouds, topic-document distributions, or hierarchical topic structures to aid in understanding and presenting the extracted topics.

5. Application and Use Case Analysis:
Practical Application: Apply the extracted topics in practical scenarios such as content recommendation systems, information retrieval, or user behavior analysis.
Use Case Analysis: Explore how the identified topics align with specific use cases or industry needs, assessing the utility and relevance of the extracted topics.

6. Documentation and Reporting:
Documentation: Document the entire process including data collection, preprocessing steps, parameter settings, model training, evaluation metrics, and results.
Reporting: Present findings, insights, and limitations in a clear and concise manner, often through reports, visualizations, or academic papers.
es, perplexity metrics, and human assessment to

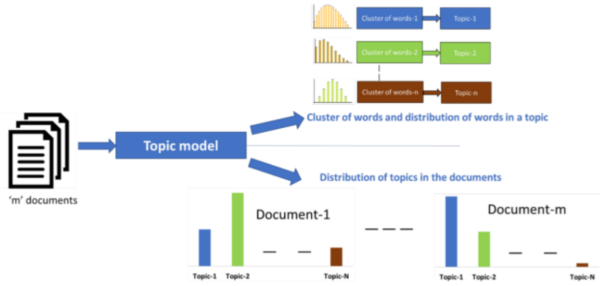
Proposed System:

Web Crawling: We are using techniques like web scraping to collect a diverse set of web pages from different domains using multiple web URL's.

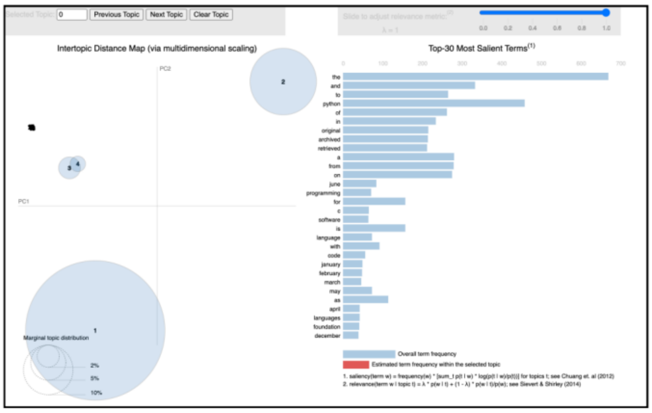
HTML Parsing: The HTML content of web pages is parsed to extract text content, removing markup tags, scripts, and other non-text elements.

Text Cleaning: The extracted text is cleaned by removing stop words, special characters, and performing stemming or lemmatization to reduce noise in the data.

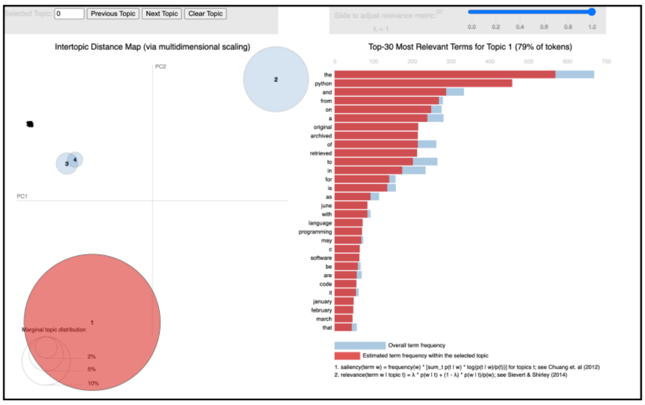
Bag-of-Words(Bow) Representation: The cleaned text is transformed into a numerical vector space models, where each document is represented as a vector of word frequencies or coherence scores.



IV. ANALYSIS AND RESULTS



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

V. CONCLUSION

In conclusion, Latent Dirichlet Allocation (LDA) methods for topic modelling of web pages offer a robust framework for organizing and extracting insights from the vast web content landscape. LDA facilitates efficient content categorization, aiding in improved information retrieval, content recommendation, and trend analysis. While challenges like noisy data and model interpretability persist, LDA's value in enhancing web-based user experiences and content management is evident. Continuous learning and ethical data practices remain essential as web content evolves.

In the future the exponential growth of online information, effective techniques for organizing and extracting meaningful insights from web content have become increasingly crucial. LDA, a probabilistic generative model, emerged as a powerful framework for uncovering latent topics within large document collections.

VI. REFERENCES

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

[2] "Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(suppl 1), 5228–5235.

[3] Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. The Annals of Applied Statistics, 1(1), 17–35.

[4] Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking LDA: Why Priors Matter. Advances in Neural Information Processing Systems, 22, 1973–1981.

[5] Steyvers, M., & Griffiths, T. (2007). Probabilistic Topic Models. In Latent Semantic Analysis: A Road to Meaning. Lawrence Erlbaum Associates..

[6] Blei, D. M., & McAuliffe, J. D. (2007). Supervised topic models. Advances in Neural Information Processing Systems, 20, 121–128.

[7] Zhang, J., Ghahramani, Z., & Yang, Y. (2009). Flexible latent variable models for multi-task learning. Proceedings of the 26th Annual International Conference on Machine Learning, 1193–1200.

[8] Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 50–57.

[9] Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476), 1566–1581.

[10] Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. Proceedings of the 17th ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining, 448–456..

[11] Das, R., Zaheer, M., & Dyer, C. (2015). Gaussian LDA for topic models with word embeddings. Advances in Neural Information Processing Systems, 28, 2597–2605.

[12] Seroussi, Y., Zhang, Z., & Ouyang, R. (2018). Weakly supervised aspect detection for online reviews using generative models. Proceedings of the 27th International Conference on World Wide Web, 1219–1227.

[13] Andrzejewski, D., Zhu, X., & Craven, M. (2009). Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. Proceedings of the 26th Annual International Conference on Machine Learning, 25–32.

[14] Peng, J., & Li, J. (2014). Incorporating topic correlation and hierarchy for online community discovery. Information Sciences, 284, 135–149..

[15] Heinrich, G. (2005). Parameter estimation for text analysis. Technical report, University of Leipzig.

[16] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. Proceedings of the ACM International Conference on Multimedia, 675–678.

[17] Ruokolainen, T., Salakoski, T., & Ginter, F. (2008). Comparing self-organizing maps and LDA for topic modeling of clinical narratives. Proceedings of the 2nd International Workshop on Health Document Text Mining and Information Analysis, 25–32.

[18] Suominen, H., Zhou, L., Hanlen, L., & Ferraro, G. (2010). Combining generative and discriminative representation learning for medical information retrieval. Journal of the American Medical Informatics Association, 17(4), 423–428.

[19] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 262–272.

[20] Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. Advances in Neural Information Processing Systems, 22, 288–296.