Demystifying Playwright Originality using Topic Modeling as a tool for textual comparison

Helena Aleluya Jose
Computer Science Department at Earlham College
Hcjose22@earlham.edu
Earlham College
Richmond, Indiana, USA

ABSTRACT

This capstone project aims to develop a novel approach for analyzing and comparing playwrights' unique voices and styles by applying topic modeling techniques, specifically Latent Dirichlet Allocation (LDA), on a corpus of theater texts. By leveraging advanced natural language processing (NLP) methods, the project will preprocess and prepare a diverse collection of plays for topic modeling, implement custom algorithms tailored for dramatic literature, and analyze the discovered topics to identify recurring themes, character archetypes, and narrative structures prevalent across different playwrights' works. Interactive visualization tools will be developed to facilitate the exploration and interpretation of these insights, enabling literary scholars and critics to understand the creative processes better and the influences that shape dramatists' voices.

Keywords

Topic modeling techniques, textual analysis, LDA model, Digital Humanities

1. INTRODUCTION

Theater plays are rich sources of literary expression, reflecting their authors' unique voices and styles. Understanding the thematic patterns, character archetypes, and narrative structures prevalent in a playwright's works can provide valuable insights into their creative process and literary influences. However, manually analyzing and comparing these elements across a large corpus of plays can be exhausting, often leading to a lack of interpretability and transparency in the analysis process.

This project aims to leverage advanced natural language processing (NLP) techniques, specifically topic modeling with Latent Dirichlet Allocation (LDA), to automate and enhance the analysis of theater texts, uncovering similarities and differences in the voices of various dramatists. As Blei et al. (2003) discussed, LDA is a generative probabilistic model that allows for the discovery of abstract "topics" within a collection of documents. This project seeks to uncover the latent thematic structures, character dynamics, and narrative arcs across different plays and playwrights by applying LDA to theater texts.

2. Related Work

Topic modeling has proven to be a valuable technique for uncovering latent themes and patterns in large text corpora across various domains. The paper "Topic Modeling with Latent Dirichlet Allocation" by Dylan Leeman provides a comprehensive overview of the LDA algorithm and its applications in text analysis. Leeman discusses the challenges of applying topic modeling to specialized domains like literary texts, such as the need for custom preprocessing techniques to handle unique formatting conventions (e.g., stage directions, character lists) and the incorporation of advanced NLP methods like named entity recognition (NER) to identify and annotate named entities like characters, locations, and organizations within the texts (Schmidt, 2019).[10]

Several studies have explored the application of topic modeling techniques to literary analysis. For instance, the Victoria and Albert Museum (V&A) in London has undertaken a project titled "Mapping the London Stage: Digital Approaches to Theatre History," which used topic modeling to analyze their collection of historical play texts and trace the evolution of theatrical themes and literary borrowings over time (Brock & Tanner, 2017) [2]. This project demonstrated the potential of topic modeling for gaining new insights into the history of theater and literary traditions.

While traditional LDA algorithms are effective for general text analysis, researchers have proposed domain-adapted topic models specifically tailored for literary and narrative texts. Buntine and Mishra (2014) [3] developed the Poisson Decomposition Model, which accounts for the burstiness, and repetition often observed in fictional narratives. Alternatively, Jacobs and Dermau (2019) used neural topic modeling with the TopicRNN architecture to capture sequential dependencies and contextual information inherent in narrative arcs.[5]

This project will draw inspiration from existing work on interactive visualization tools for topic modeling to facilitate the interpretation and exploration of the discovered topics and patterns. Systems like Parallel Topical Pies (Chuang et al., 2012)[4] and TopicReveal (Liu et al., 2018) [8] allow users to explore and compare topics across multiple text collections simultaneously, enabling a deeper understanding of the underlying themes and their relationships. Additionally, performance optimization techniques like distributed computing frameworks (Zaharia et al., 2010) [11] and specialized topic modeling libraries (Řehůřek & Sojka, 2010) [9] may be employed to handle large corpora efficiently.

3. Design and Implementation

The algorithm for analyzing playwrights' voices and styles using Latent Dirichlet Allocation (LDA) begins with an extensive data collection phase, where a diverse corpus of theater texts, encompassing various playwrights and genres, is gathered. Following data collection, the preprocessing stage involves several steps to prepare the text for analysis. This includes tokenization, where the text is segmented into individual words or phrases, removal of stopwords such as common articles and prepositions, lemmatization or stemming to reduce words to their base form, and handling special formatting such as stage directions and character lists. Additionally, named entity recognition (NER) is employed to identify characters, locations, and organizations within the text, ensuring they are treated as distinct entities during analysis.

The custom adaptation of LDA for dramatic literature involves several key components. First, topics are defined to capture thematic patterns, character dynamics, and narrative structures specific to plays. Topics are initialized based on common themes found in dramatic literature or prior knowledge of playwrights' styles. The model is then trained using specialized preprocessing techniques and inference algorithms like Gibbs Sampling [7] or Variational Inference [9] to estimate topic distributions for each document and word-topic assignment.

During the analysis phase, the discovered topics will be examined in detail. This includes identifying the most significant words associated with each topic to uncover recurring themes and patterns. Moreover, character names are analyzed to identify archetypes, and topics are explored in relation to narrative arcs and structural elements such as exposition, rising action, climax, and resolution. Interactive visualization tools will be developed to facilitate exploration and comparison of topics across playwrights and plays, allowing users to delve into individual texts for deeper analysis.

Performance optimization techniques are implemented to ensure efficient processing and scalability. This includes utilizing parallel processing and distributed computing frameworks to speed up computation and employing memory-efficient data structures and algorithms to handle large text corpora and model parameters. [7]

Validation of the algorithm involves comparing the identified topics with existing literary analysis and expert evaluations, with iterative improvements driven by user feedback and advancements in natural language processing and topic modeling techniques. The algorithm's documentation provides detailed information on implementation specifics and findings, contributing valuable insights into the creative processes and literary influences shaping playwrights' voices and styles in dramatic literature.

4. Timeline

Weeks 1-2: Exploration of Topic Modeling Tools

a) Explore existing topic modeling tools and resources such as Gensim, spaCy, and LDA implementations in Python.
b) Review relevant literature and research papers on topic modeling techniques and applications in textual analysis.

Week 3: Data Acquisition

a) Acquire relevant theater text datasets from sources such as Project Gutenberg, OpenText, or specialized theater archives.
b) Explore platforms like Kaggle, UCI Machine Learning Repository, or OpenML for additional datasets if needed.

Weeks 4-5: Data Preprocessing

a) Handle missing values and outliers in the acquired theater text datasets.
b) Perform initial data cleaning and normalization using Python libraries such as Pandas and NumPy.
c) Preprocess the text data for topic modeling, including tokenization, stopword removal, and lemmatization.

Weeks 6-7: Framework Development and Model Training

a) Develop the project framework using Python programming language.
b) Utilize libraries such as Gensim or spaCy for topic modeling implementation.
c) Train preliminary Latent Dirichlet Allocation (LDA) models on preprocessed theater text data.

Weeks 8-9: Topic Modeling Integration and Evaluation

a) Integrate LDA topic modeling into the project framework.
b) Explore topic visualization techniques to aid in interpretation.

Weeks 10-11: Interpretation and Analysis

a) Analyze the discovered topics to identify recurring themes, character archetypes, and narrative structures across different playwrights' works.
b) Evaluate the effectiveness of the topic modeling approach in capturing playwrights' voices and styles.
c) Conduct initial comparison and exploration of topics using Jupyter Notebook or similar tools.

Week 12: Refinement and Validation

a) Interpret the results of the topic modeling analysis and identify any areas for improvement.
b) Refine the program framework and topic modeling models based on insights gained from the analysis.
c) Validate the findings by comparing them with existing literary analysis and expert evaluations.

Week 13: Final Paper and Presentation Preparation

a) Compile the final paper documenting the project, including the methodology, results, discussion, and conclusion.
b) Prepare the final program and poster summarizing the project for presentation.
c) Review and finalize all materials for submission.

Week 14: Final Paper, Program, and Poster Due

5. ACKNOWLEDGMENTS

Special thanks to Charlie Peck for his guidance in crafting this proposal.

6. REFERENCES

[1] Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). https://www.aclweb.org/anthology/N19-4010/

[2] Brock, A., & Tanner, S. (2017). Modeling the History of Shakespearean Intertextuality. In J. Estill, D. Carnegie, and A. Murphy (Eds.), Metadata and Semantics Research (pp. 95-108). Springer, Cham.

[3] Buntine, W., & Mishra, S. (2014). Experiments with non-parametric topic models. Proceedings of the 20th ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 881-890). https://dl.acm.org/doi/10.1145/2623330.2623711

[4] Chuang, J., Ramage, D., Manning, C., & Heer, J. (2012). Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. https://idl.cs.washington.edu/papers/model-driven-vis

[5] Jacobs, A. M., & Dermau, L. F. (2019). Neural Text Modeling with Topic-Guided Generative Adversarial Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3811-3821). https://aclanthology.org/D19-1397/

[6] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. https://aclanthology.org/N16-1030/

[7] Leeman, Dylan. "Topic Modeling with Latent Dirichlet Allocation." Bachelor's Thesis, Department of Computer Science, Earlham College, 2016. [Online]. Available: https://portfolios.cs.earlham.edu/wp-content/uploads/2016/09/dylan-leeman-final_paper.pdf

[8] Liu, S., Wang, X., Chen, J., Zhu, J., & Guo, B. (2018). TopicReveal: An exploratory tool for topic modeling. Journal of Visual Languages & Computing, 46, 50-63. https://www.sciencedirect.com/science/article/pii/S1045926X17301638

[9] Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. https://radimrehurek.com/gensim/lrec2010_final.pdf

[10] Schmidt, B. M. (2019). Preprocessing for Analyzing Drama. In Nuanced Text Data Transfer And Curation Information Access For Imprecise Corpus Data Outside Traditional Horizons (pp. 1-25). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-22714-8_1

[11] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster Computing with Working Sets. Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing. https://www.usenix.org/legacy/event/hotcloud10/tech/full_papers/Zaharia.pdf