Quantifying Themes in Theatre Using Topic Modeling (LDA)

Helena Aleluya Jose
Computer Science Department
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.

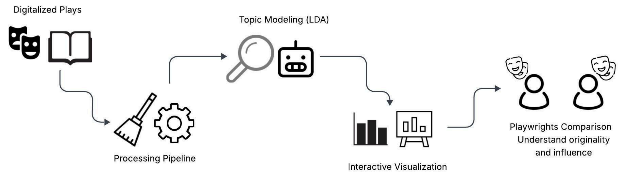

Image 1: Graphical Abstract

2. Literature Review

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. Project Purpose

The central purpose of this project is twofold. First, it seeks to adapt topic modeling methodologies to theatrical texts, addressing the challenges posed by their unique structural conventions. Second, it aims to provide scholars in literature and theatre studies with computational tools that enable scalable comparative analysis of playwrights' voices. The broader contribution lies in demonstrating how computational methods can augment interpretive practices in the humanities.

4. 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.

5. Usage and Application

The system is designed for use by literary scholars, critics, and students. Users provide digitized plays as input; the system processes the corpus and outputs topic distributions alongside visualization interfaces. The interface allows users to:

- Examine topic prevalence across plays or authors.
- Identify recurrent thematic clusters.
- Analyze character archetypes through preserved named entities. Applications include comparative authorship studies, the exploration of intertextual influences, and pedagogical support in literature curricula.

6. Methods

The primary dataset will consist of a curated corpus of theater texts, including plays from multiple playwrights and genres to ensure diversity in style and voice. Sources may include publicly available repositories such as Project Gutenberg, British Library, Folger Shakespeare library and specialized theater archives. Texts will be collected in digital format (e.g., TXT, PDF) and standardized for preprocessing. Metadata such as playwright, year, genre, and location will also be recorded to enable comparative analysis across authors and periods.

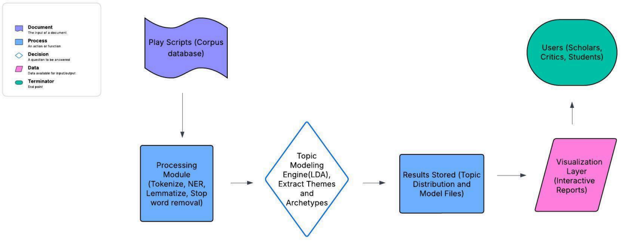
Document
The input of a document
Process
An action or function
Decision
A question to be answered
Data
Data available for input/output
Terminator
End point
Play Scripts (Corpus database)
Processing Module (Tokenize, NER, Lemmatize, Stop word removal)
Topic Modeling Engine(LDA), Extract Themes and Archetypes
Results Stored (Topic Distribution and Model Files)
Visualization Layer (Interactive Reports)
Users (Scholars, Critics, Students)

Image 2: Data Architecture Diagram

7. Software and Tools

The project will be developed primarily in Python, drawing on a suite of tools that support text processing, topic modeling, and visualization. Core NLP tasks, such as tokenization, lemmatization, and named entity recognition, will be handled using spaCy, with NLTK assisting in additional preprocessing steps like stopword refinement. For topic modeling, the project will rely on Gensim's implementation of Latent Dirichlet Allocation (LDA). Data preparation and normalization will be carried out using Pandas and NumPy. Visual exploration of the results will be supported through Matplotlib, Plotly, and PyLDAvis, enabling both static and interactive representations of topics. Development, experimentation, and documentation will take place within Jupyter Notebook.

8. Analysis Plan

Data Preprocessing: Texts will be cleaned to remove stage directions, punctuation, and irrelevant symbols. Stopwords will be removed, and words will be lemmatized. NER will identify characters, locations, and other entities, which will be tagged as separate tokens.

Topic Modeling: Custom adaptations of LDA will be applied to capture thematic patterns, character archetypes, and narrative structures unique to dramatic literature. The model will be trained using techniques like Gibbs Sampling or Variational Inference. The number of topics will be tuned for coherence and interpretability.

Topic Analysis: Discovered topics will be examined to identify recurring themes, character dynamics, and narrative structures. Topic distributions will be compared across playwrights to analyze stylistic differences and similarities.

Visualization: Interactive visualization tools will be developed to explore topics, compare playwrights, and examine topic evolution across texts. Features may include topic similarity measures and word clouds for each topic.

Validation: The model's outputs will be validated by comparing discovered topics to existing literary analyses and expert evaluations. Iterative refinement will be conducted based on insights gained and feedback received.

Documentation and Reporting: All preprocessing steps, model parameters, analysis results, and visualization outputs will be documented. The final deliverables will include a written report, interactive visualizations, and the fully processed dataset for future research.

9. Results

The topic modeling analysis identified five recurring thematic clusters across the four selected plays: *A Raisin in the Sun*, *Fences*, *Macbeth*, and *Doctor Faustus*. Using Latent Dirichlet Allocation (LDA), each play was represented as a mixture of overlapping topics based on word co-occurrence patterns.

Topic	Keywords
1	demand, skin, cross, tremble, needed, devotion, blind
2	Mama, ruth, don, ain, money,son, door
3	Doctor, text, master, sweet, spirits, horse, art
4	Lady, shall, fear, speak, son, castle, nature
5	Troy, rose, door, told, kitchen, august

Table 1: Extracted Topics and Representative Keywords

Each topic aligns with recognizable motifs in dramatic literature. For instance, Topic 2 relates to domestic and generational tension in *A Raisin in the Sun* and *Fences*, while Topics 3 and 4 reflect moral and existential struggles central to *Macbeth* and *Doctor Faustus*. Topic 1, with words like "devotion" and "cross," suggests recurring moral or spiritual conflict shared across works.

Play	Dominant Theme
A Raisin in the Sun	Domestic tension and aspiration
Fences	Legacy, labor, and deferred dreams
Macbeth	Ambition and moral consequence
Doctor Faustus	Knowledge, morality, and fear

Table 2: Topic Distribution per Play

Overall, the model demonstrates that while each play presents distinct cultural and historical contexts, the underlying themes, faith, ambition, family, and moral consequence, appear repeatedly. These findings support the idea that certain human concerns persist across different periods, genres, and authorships.

10. Discussion

The topic modeling results reveal that Latent Dirichlet Allocation was able to uncover meaningful thematic clusters across the four selected plays, demonstrating the model's ability to detect high-level structures even within highly stylized dramatic texts. Themes such as generational aspiration, moral conflict, ambition, and spiritual negotiation recurred across the corpus, suggesting that the playwrights, despite distinct historical and cultural contexts, are bound by shared human concerns. Topic 2, for instance, captured domestic and intergenerational tensions prevalent in *A Raisin in the Sun* and *Fences*, while Topics 3 and 4 reflected the metaphysical and moral anxieties central to *Doctor Faustus* and *Macbeth*. These outputs align closely with long-established literary interpretations, indicating that the LDA model is capable of approximating key thematic contours recognized by scholars.

However, the results also expose critical limitations of the method. LDA assumes a bag-of-words structure, flattening the inherently dynamic nature of dramatic literature [21]. Plays rely heavily on sequential tension, character-driven arcs, and dialogic relationships, elements that LDA cannot fully capture. As a result, some topics conflate linguistic patterns that are structurally similar but contextually distinct. For example, Topic 1 contains religious or moral vocabulary that manifests differently in *Fences* than in *Doctor Faustus*, yet the model treats them as equivalent because of shared lexical patterns. This confirms the critiques raised in digital humanities scholarship: standard topic models risk oversimplifying narrative form, collapsing nuanced emotional and dramaturgical structures into statistical clusters.

Another challenge arises from character names and stage directions, even with NER tagging, these features can dominate topics in ways not always semantically meaningful [20]. The prominence of names like "Troy" and "Rose" in Topic 5 reflects the narrative weight of these characters, but also highlights the difficulty of balancing their lexical frequency with thematic significance. As Schmidt (2019) and Leeman (2016) note, dramatic texts require special preprocessing to avoid distortion due to the interplay of dialogue, character labeling, and formatting structures. These distortions are reflected in the uneven topic distributions and the partial blending of unrelated motifs across plays.

Overall, while the model successfully surfaces core thematic motifs, its interpretability depends heavily on domain knowledge and critical mediation [22]. The results are best understood not as definitive representations of each playwright's voice, but as computational approximations that must be contextualized within broader dramaturgical analysis. The experiment demonstrates the potential of LDA for comparative theatre studies, while also revealing the contours of its limitations when dealing with narrative complexity, symbolic density, and the multimodal nature of dramatic texts.

11. Future Work

Future iterations of this project would benefit from methodological and architectural enhancements designed to address the structural challenges of dramatic literature. First, incorporating neural topic models, such as TopicRNN or contextualized embedding-driven approaches, may better capture sequential relationships and narrative arcs that LDA's bag-of-words assumption cannot accommodate [5]. These models can recognize how a theme evolves across scenes, enabling more nuanced analysis of rising action, conflict, and resolution. Second, expanding the corpus to include a broader selection of playwrights, genres, and historical periods would enhance the model's generalizability and reduce corpus bias. Including plays from non-Western traditions, marginalized voices, or contemporary works could illuminate how thematic structures shift across cultural contexts, an area particularly important for theatre studies. Third, future work should incorporate hierarchical models (e.g., hLDA) or Poisson-based models specifically designed for bursty narrative text, which could improve topic coherence in dialogue-heavy works [3]. Additionally, integrating richer metadata, such as scene boundaries, character networks, sentiment arcs, and dramaturgical function, could allow for multi-layered analyses that move beyond word frequency to capture structural and emotional dimensions.

Interactive visualization tools should also be expanded. Deploying dynamic dashboards with topic evolution timelines, character-topic networks, and cross-play comparisons would support more intuitive interpretability for scholars and students [22]. These tools could be paired with expert-informed validation pipelines, enabling theatre scholars to annotate and contextualize topics in real time. Finally, future work should address ethical and epistemological considerations. Topic models risk reinforcing canonical narratives or obscuring subtextual elements that may not manifest lexically. Collaborative evaluation with domain experts, sensitivity to marginalized voices within the corpus, and careful interpretive framing will be essential to ensure responsible computational literary analysis.

12. ACKNOWLEDGMENTS

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

13. 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

[12] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

[13] Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. https://spacy.io

[14] McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. *Proceedings of the 9th Python in Science Conference*, 51–56.

[15] Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). *Array programming with NumPy*. *Nature*, 585(7825), 357–362.

[16] Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment*. *Computing in Science & Engineering*, 9(3), 90–95.

[17] 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*, 45–50.

[18] Plotly Technologies Inc. (2015). *Collaborative data science*. Montréal, QC. https://plotly.com

[19] Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. (2016). *Jupyter Notebooks – a publishing format for reproducible computational workflows.* In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 87–90.

[20] Schmidt, Benjamin M. *"Preprocessing for Analyzing Drama." Nuanced Text Data*… Springer, 2019

[21] Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History.* University of Illinois Press, 2013.

[22] Chang, Jonathan, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David Blei. "Reading Tea Leaves: How Humans Interpret Topic Models." *Advances in Neural Information Processing Systems* (NIPS), 2009.