Topic Modeling *One Hundred Years of Solitude* with LDA

1. Introduction

This research explores thematic structures in *One Hundred Years of Solitude* through Latent Dirichlet Allocation (LDA). Employing a probabilistic generative approach, the study identifies recurring topics across the text, offering insights into its narrative structure and thematic composition. Beyond literary analysis, this methodology demonstrates potential applications in various fields such as policy-making, private sector decision-making, and advertising. This work contributes to computational literary analysis and highlights the interdisciplinary value of unsupervised learning techniques for textual data exploration.

Gabriel García Márquez's *One Hundred Years of Solitude* is celebrated for its rich themes and complex narrative structure. This study applies topic modeling to analyze the latent thematic structures within the novel. By leveraging LDA, an unsupervised machine learning algorithm, the analysis aims to uncover patterns in word co-occurrences to explore underlying topics. The broader motivation for this study lies in demonstrating how such computational techniques can be applied not only to literature but also to other domains such as policymaking, customer insights, and semantic search.

For instance, policymakers could employ topic modeling to analyze vast amounts of textual data, such as public comments on legislation or surveys, to identify major areas of concern or public sentiment. Similarly, private companies can use this method for targeted advertising by analyzing customer reviews or feedback to discover the key topics that resonate with their audience. In advertising, topic modeling can help create customer personas by clustering user interests and behavior patterns, improving personalization in marketing campaigns.

2. Research question

This research seeks to address the following question: How can topic modeling, specifically Latent Dirichlet Allocation (LDA), uncover the latent thematic structures in the fictional books, and what insights can this approach offer about its narrative complexity and recurring motifs?

3. Data

The dataset comprises the full text of *One Hundred Years of Solitude*, preprocessed to remove non-textual elements such as formatting tags and extraneous symbols. The preprocessing steps included tokenization, where the text was split into individual words or tokens, and the removal of stop words—common terms such as "and," "the," and "of," which do not contribute significantly to thematic analysis. Lemmatization, which converts words to their root forms, was not applied in this analysis. This decision was based on the fact that lemmatization is not always necessary for LDA and BoW approaches, which rely on word co-occurrence patterns rather than root forms. The lack of lemmatization does not significantly impact results unless the vocabulary contains many inflected forms. Additionally, in the context of literary analysis, lemmatization might oversimplify the text, as variations in word forms can carry stylistic or nuanced meanings. Therefore, skipping lemmatization ensured the preservation of such literary subtleties while maintaining a focus on raw word distributions.

The cleaned text was then transformed into a Bag of Words (BoW) representation, which forms the foundation for generating topic distributions through LDA. BoW was chosen due to its simplicity and effectiveness in capturing the raw frequency of words, making it particularly suitable for the unsupervised nature of LDA. Unlike Term Frequency-Inverse Document Frequency (TF-IDF), which assigns weights to words based on their uniqueness across documents, BoW treats all words equally, focusing purely on their occurrence. This aligns seamlessly with the probabilistic assumptions of LDA, which models topics based on word co-occurrence patterns rather than the relative importance of words.

By prioritizing raw word counts, BoW ensures that the model can capture thematic structures without introducing biases from weighting schemes, such as downplaying common but contextually significant words. This is especially important for literary texts like *One Hundred Years of Solitude*, where frequently occurring words often contribute to core themes and motifs.

4. Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model designed to identify groups of words that frequently appear together, representing distinct topics within a corpus. The implementation process began with extensive data preprocessing, where the text underwent cleaning to remove extraneous characters, tokenization to split the text into individual words or tokens, and vectorization. CountVectorizer was employed to convert the text into numerical representations suitable for modeling.

The model training was conducted using Scikit-learn's *LatentDirichletAllocation* class. Several key parameters were carefully configured for optimal results. The number of topics was determined iteratively to balance coherence and interpretability, ensuring the model captured the nuanced thematic structures within the text without overfitting or oversimplifying. The batch learning method was selected to process the entire dataset in one iteration, providing higher accuracy compared to the alternative online learning method, which processes data incrementally. To ensure reproducibility of results, a fixed random state was used, maintaining consistency across multiple runs of the model.

Upon training, the LDA model produced two critical outputs: the document-to-topic matrix and the word-to-topic matrix. The document-to-topic matrix provided the probability distribution of topics across each chapter of the novel, offering insights into the thematic emphasis of different sections. Meanwhile, the word-to-topic matrix highlighted the most representative words for each identified topic, serving as a basis for interpreting and naming the topics.

Finally, the topics were evaluated through coherence scores, which measure the semantic consistency of words within each topic, and manual inspection to ensure thematic relevance. This interpretative step was crucial in connecting the computational output of the model to the broader themes of *One Hundred Years of Solitude*, demonstrating the utility of LDA in uncovering latent structures within a literary text.

5. Results

The results of the topic modeling reveal the thematic diversity and complexity of *One Hundred Years of Solitude*. Following the extraction of terms, the term-topic matrix was normalized to ensure that the contribution of each word to a topic was expressed as a proportional value. This normalization step allowed for a clearer and more standardized representation of the relationship between words and topics, facilitating better interpretability and consistency across topics.

Once the topics were normalized, they were manually labeled based on the semantic coherence of their representative words. For example, a topic characterized by terms such as "Melquíades," "alchemy," “gympsied”, “artifacts”, and "magic" was labeled as Magical Realism and Discovery, reflecting the fantastical and exploratory themes in the novel.

The labeling process provided thematic clarity, linking the computational results of the model to meaningful interpretations of the novel's narrative. Each chapter of the book was then assigned a dominant topic based on the highest probability from the document-to-topic matrix. This structured approach ensured a coherent understanding of how different themes evolved across the narrative, offering a quantitative perspective on the literary depth of the book.

5.1. Topics

- Magical Realism and Discovery: These words relate to the fantastical and exploratory themes in the novel, particularly the gypsy Melquíades, his alchemical experiments, and the magical elements surrounding Macondo.
- Art, Travel, and Struggle: This topic seems to represent artistic creation, travel, and the struggle for existence and survival. It aligns with the more abstract, emotional, and itinerant aspects of the story.
- Family and Relationships: This topic highlights themes of family dynamics, romantic relationships, childbearing, and personal connections in the Buendía lineage.
- Political Unrest and Conflict: These words strongly suggest themes of political uprisings, social unrest, and class conflict, which are central to the revolutionary and war periods depicted in the novel.

- Politics and Corruptio: This topic reflects themes of politics, manipulation, and corruption, as well as elections and political figures, which are recurring in the novel's depiction of Macondo's development.
- Grief and Emotional Struggles: This topic represents themes of personal grief, emotional distress, and struggles with loss and hardship, which resonate with several characters' arcs.
- Outsiders and Modernization: This topic captures themes of outsiders entering Macondo, modernization, and the influence of foreign cultures and industries, particularly during the banana plantation period.
- Revolution and Leadership: This topic focuses on revolution, leadership, and societal change, which are central to the rise of Colonel Aureliano Buendía and the political revolutions in the novel.

Figure 1. Dominant topics across chapters.

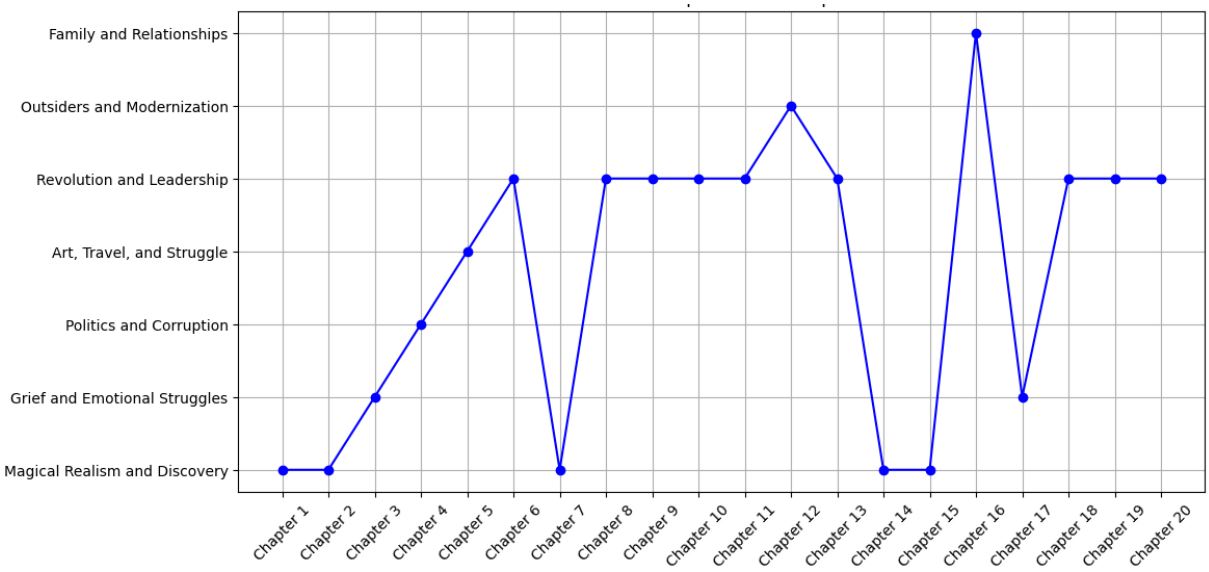

Figure 1 illustrates the dominant topics across the chapters highlighting how the thematic focus evolves throughout the novel. Each chapter is mapped to a single dominant topic, which represents the most prominent theme as determined by the LDA model.

The early chapters (1 and 2) are dominated by *Magical Realism and Discovery*, capturing the fantastical elements and the foundational exploration of Macondo's unique setting and history. Chapter 3, the focus shifts to *Grief and Emotional Struggles*, reflecting a period of personal

challenges and loss for the characters. Chapter 4 introduces *Politics and Corruption*, signaling the beginning of Macondo's sociopolitical evolution. Chapter 5 transitions to *Art, Travel, and Struggle*, highlighting themes of creative expression and existential challenges. From Chapter 6 onwards, *Revolution and Leadership* emerges as the dominant theme, underscoring the novel's exploration of Colonel Aureliano Buendía's rise and the broader revolutionary movements. This theme persists strongly through Chapters 9–13, emphasizing the centrality of political and social upheaval. In Chapter 12, *Outsiders and Modernization* takes center stage, marking the arrival of foreign influence and the banana company, which significantly impacts Macondo's trajectory. Chapters 14 and 15 return to *Magical Realism and Discovery*, reviving the novel's fantastical roots. Toward the end, Chapters 16 and 17 revisit *Grief and Emotional Struggles*, echoing the cyclical nature of tragedy in the Buendía family. The final chapters return to *Revolution and Leadership*, reinforcing the recurring themes of societal uprising and change.

Figure 2. Intertopic Distance Map (see interactive image in GitHub).

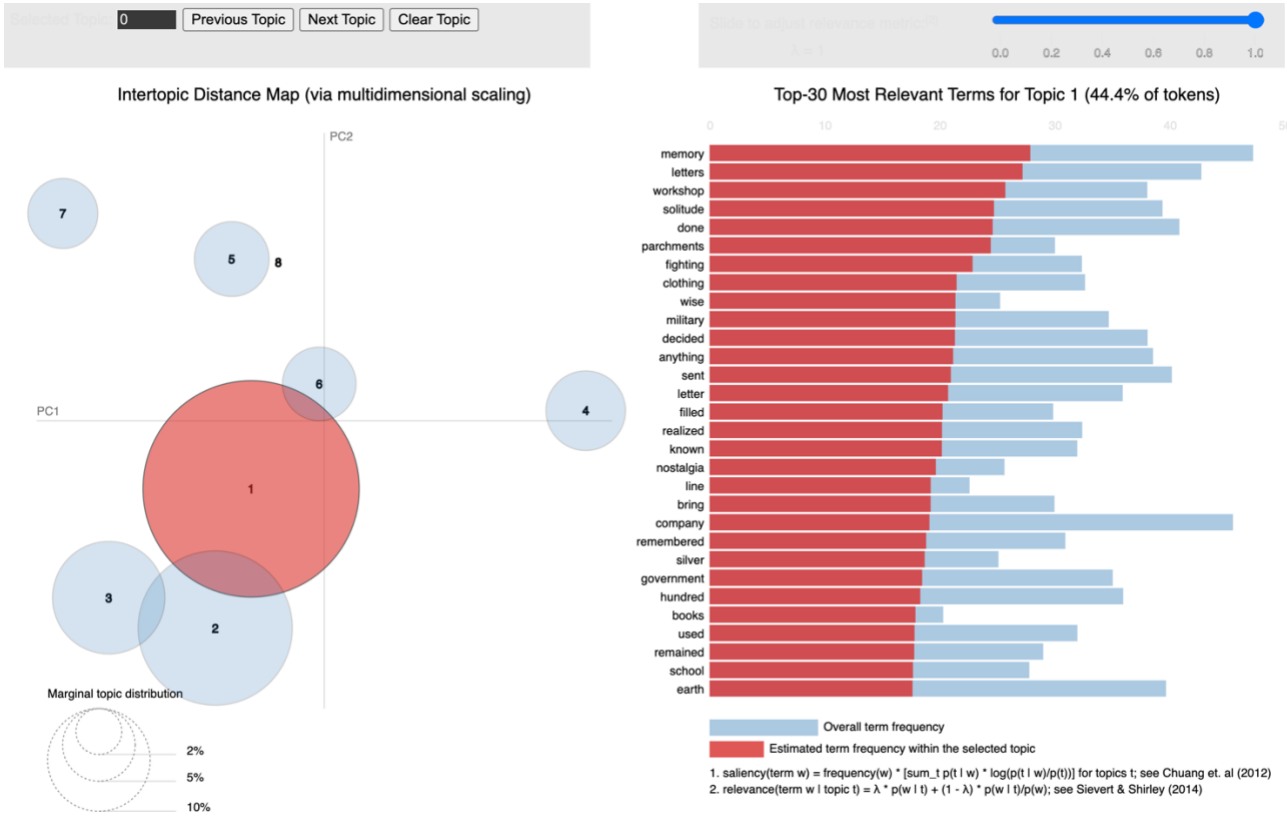

Figure 2 is a screenshot of the interactive visualization generated by *PyLDAVis*. It provides an overview of the topics identified by the LDA model, offering insights into their prevalence and distinctiveness. The left panel represents the intertopic distance map, where each circle corresponds to a topic. The size of each circle indicates the proportion of tokens in the corpus assigned to that topic. For example, Topic 1, the largest circle, accounts for 44.4% of the tokens, making it the most dominant topic. Topics positioned closer together, such as Topics 2 and 3, share more overlapping terms and are thematically similar, while topics further apart, such as Topics 1 and 6, are more distinct.

The right panel lists the top 30 most relevant terms for the selected topic (Topic 1). The relevance metric considers both the term frequency within the topic and how distinctive the term is compared to other topics. Terms such as "memory," "letters," and "workshop" are highly specific to Topic 1, reflecting its thematic focus, while others like "government" and "nostalgia" provide additional nuance. The red bars show the frequency of terms within Topic 1, while the blue bars indicate their overall frequency in the corpus. This approach differs slightly from raw probabilities because it adjusts for both topic-specific contributions and broader corpus-wide patterns, ensuring that the most distinctive and informative terms are highlighted.

This visualization is valuable for exploring the thematic structure of the corpus and understanding how topics overlap or diverge. It goes beyond raw probabilities by incorporating measures of distinctiveness, offering a more nuanced perspective on the relationship between words and topics. This distinction allows for clearer interpretation of the topics' unique characteristics and their connections within the broader narrative.

Overall, the graphs demonstrate the novel's thematic richness and the dynamic interplay of topics as the narrative unfolds, providing a structured visualization of *One Hundred Years of Solitude's* literary depth.

6. Conclusion

This study successfully applied Latent Dirichlet Allocation (LDA) to uncover latent thematic structures in Gabriel García Márquez's *One Hundred Years of Solitude*. The analysis highlighted the novel's thematic complexity, identifying eight distinct topics ranging from magical realism and discovery to political unrest and familial relationships. The results demonstrate the power of topic modeling not only to interpret literary texts but also to provide structured insights into the evolution of themes across a narrative.

Beyond literary analysis, this study underscores the broader potential of LDA in various domains. From policy-making, where it can analyze public sentiment and legislative priorities, to the private sector, where it can enhance customer insights and advertising strategies, the versatility of this method is evident. By leveraging unsupervised learning, organizations and researchers alike can derive actionable insights from textual data, enriching decision-making processes.

Future work could extend this approach to analyze thematic progression in multilingual texts or compare thematic patterns across different works of literature. The findings of this research reaffirm the interdisciplinary relevance of computational techniques like LDA in uncovering meaningful patterns in unstructured data.

Find the full analysis here:
https://github.com/isabella-ut/LDA_OneHundredYearsOfSolitude

7. References

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022.
- Raschka, S. (2024). *Applying Machine Learning to Sentiment Analysis*. Chapter 8.
- Göbel, S. (2024). *Session 7: Basic Models II*, Natural Language Processing Course.

The text was downloaded from: https://github.com/hrbn/100-years-of-solitude