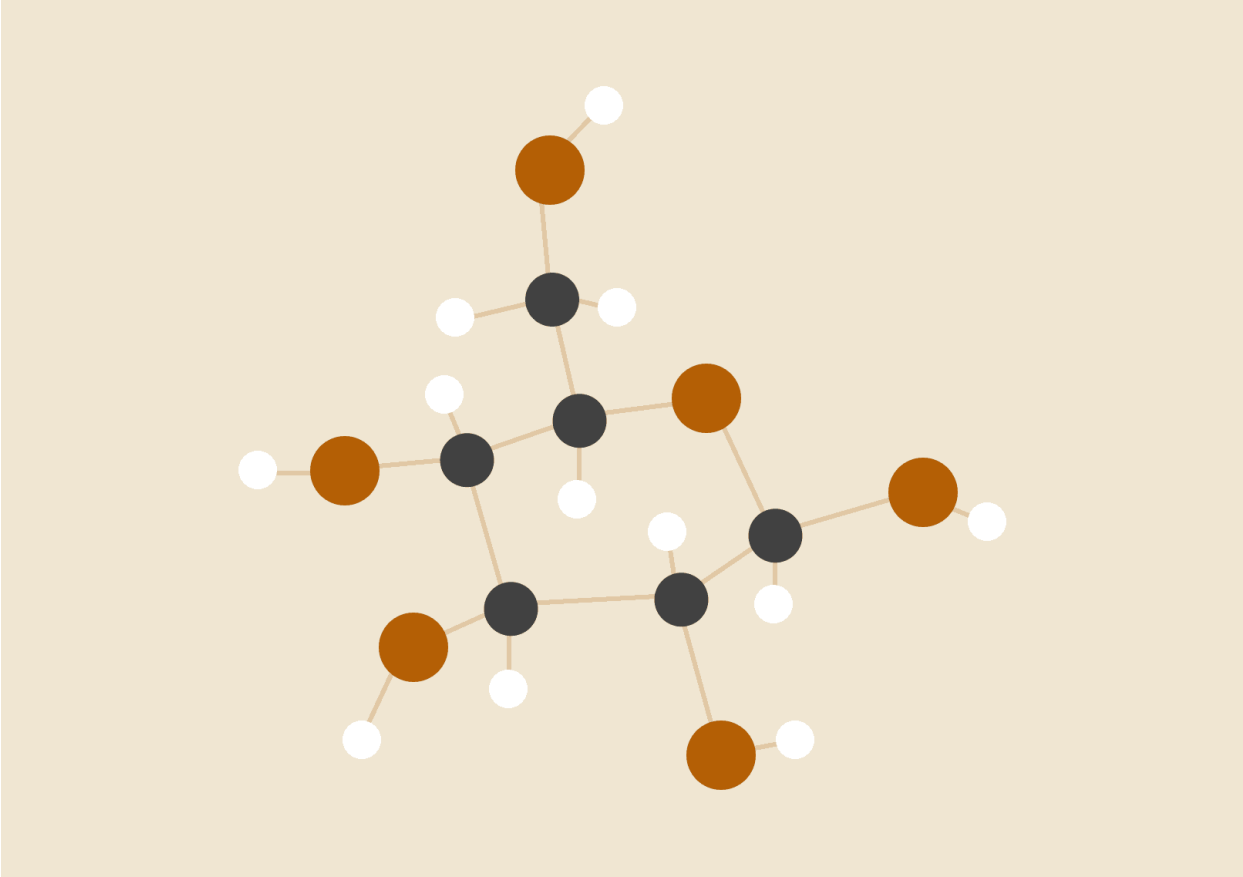
Analysis of Variety in Inter and Intra-Genre Writing Styles

Project Abstract



Jason Curtis, Yanzheng Liu, Xiaoyang Fei

NOV 25, 2024
DS 5230

INTRODUCTION

Text written by different authors is typically grouped into a genre-subgenre schema to help give readers an idea of what to expect in a book prior to selection. Genres such as Fantasy can be further specified into subgenres such as Sci-fi, High Fantasy, Fairy Tales, and so on. However, each writer typically has their own independent style which may not be captured by these broader designations. We intend to apply unsupervised machine learning techniques to learn more about what differentiates similarly grouped writers and learn more about how these styles develop.

RESEARCH QUESTIONS

By developing several related research questions we intend to complete a broad research approach that investigates many aspects of the question: what makes a writing style distinct?

1. What stylistic features (e.g. word frequency, sentence length, dialogue usage) are most distinct among authors within the same subgenre?
2. Can authors within a subgenre be grouped into clusters based on stylistic features?
3. Does clustering analysis align with traditional genre/subgenre classification, or are new patterns revealed?
4. What specific linguistic features are most important in distinguishing clusters of authors?
5. What themes or topics are consistently present across different authors in the same subgenre?
6. Do topic modeling results suggest latent subgenres not captured by existing classifications?

DATA AND TECHNIQUES

Data will be drawn from publicly available sources such as Project Gutenberg. Specifically, we will curate a dataset of texts spanning multiple subgenres within an umbrella genre, such as Fantasy or Mystery.

We will employ the following unsupervised machine learning and text analysis techniques:

- **Clustering:**
  - K-means
  - DBSCAN or hierarchical clustering
- **Dimensionality Reduction:**
  - Principal Component Analysis (PCA)
  - Kernel PCA
  - t-SNE
- **Topic Modeling:**
  - Latent Dirichlet Allocation (LDA)

PROCEDURE

1. Collect a varied dataset of texts spanning multiple subgenres within a chosen umbrella genre
2. Preprocess texts to structure dataset for analysis
    a. Tokenization, lemmatization/stemming
3. Feature Engineering
    a. Lexical Features
        i. Word Frequency
        ii. Sentence Length
        iii. Vocabulary richness
    b. Syntactic Features
        i. Word embedding
        ii. Topic/themes
    c. Structural Features
        i. Sentiment analysis
        ii. Figurative language use

4. Unsupervised technique application
    a. Clustering
    b. Dimensionality Reduction
    c. Topic Modeling
    d. Network Analysis

DATA EXAMPLE

Project Gutenberg

- **Description**: A large collection of free eBooks, including many public domain works spanning multiple genres and subgenres.
- **Why Use It**: It provides a vast selection of classic texts that are ideal for cross-genre and historical comparisons.
- **URL**: https://www.gutenberg.org
- **Examples**:
    - **Fantasy**: Works by J.R.R. Tolkien, Lewis Carroll.
    - **Sci-fi**: Texts by H.G. Wells, Jules Verne.
    - **Mystery**: Agatha Christie's early works.

Author	Title	Genre	Subgenre	Text
J.R.R. Tolkien	The Hobbit	Fantasy	High Fantasy	In a hole in the ground where lived a hobbit.
Isaac Asimov	Foundation	Sci-fi	Space Opera	Hari Seldon turned to face the audience as the lights dimmed.
Agatha Christie	Murder on the Orient Express	Mystery	Detective	The train was unusually crowded that morning, as Poirot boarded.

RESULTS

The results will include:

1. Clusters of authors within subgenres based on their stylistic features.
2. Visual representations of dimensionality reduction analyses show overlaps and separations between traditional subgenre classifications.
3. Topic modeling outputs reveal latent subgenres and thematic consistencies.
4. Identification of key linguistic features that most strongly define authorial styles within subgenres.

CONCLUSION

Our analysis will shed light on the nuances of writing styles within genres and subgenres. By leveraging unsupervised learning techniques, we aim to uncover latent structures and relationships among authors that may not align with traditional genre classifications. This research can offer new perspectives on how writing styles evolve and how authors can be better understood beyond their genre labels.

REFERENCES

1. Jurafsky, D., & Martin, J. H. (2022). *Speech and Language Processing*. Pearson.
2. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
4. Available works from Project Gutenberg (Datasets)