Exploring Ayush Mudunuru's GoodReads Libraries' Descriptions Using TF-IDF Analysis

Introduction

The primary goal of this project is to analyze book descriptions to uncover key insights about the themes and characteristics of books within the library of the author. By applying Natural Language Processing (NLP) techniques, we can determine the most distinctive words that define each book, group similar books into clusters, and explore patterns within the library's collection. The approach combines tokenization, lemmatization, and Term Frequency-Inverse Document Frequency (TF-IDF) analysis to compute the importance of words in book descriptions. The results provide meaningful insights that can inform book classification, thematic organization, and potential applications such as recommendation systems.

Research Questions

What words are most characteristic of Ayush Mudunuru's goodreads Library?

What distinct sub-groups emerge from clustering of Ayush Mudunuru's Goodreads library?

Methodology

Preprocessing

The first step involves retrieving and preparing the book descriptions for analysis:

1. **Scraping:** The author's goodreads library was first scraped using the goodreads_scraper package. Its documentation was instrumental in collecting the necessary data.
2. **Tokenization**: Each book description is split into individual words using the SpaCy library.
3. **Lemmatization**: Words are reduced to their base forms (e.g., "running" becomes "run") for consistency.
4. **Stopword Removal**: Commonly used words with low information content (e.g., "and," "the") are removed.
5. **Filtering Non-Alphabetic Words**: Words containing numbers, punctuation, or symbols are excluded.

After preprocessing, each book description is represented as a list of meaningful, lowercase lemmas.

TF-IDF Calculation

TF-IDF is used to measure the importance of each word within a book's description relative to the entire collection. TF-IDF scores highlight words that are frequent within a single book but rare across others, making them ideal for identifying characteristic terms.

Aggregation and Clustering

1. **Top Words per Book**:
    - For each book, the top 15 words with the highest TF-IDF scores are extracted. These words represent the most distinctive aspects of the book's description.
2. **Library-Wide Analysis**:
    - TF-IDF scores are aggregated by multiplying with document frequency across the entire collection to identify globally important words that characterize the library.
3. **Clustering**:
    - Using K-Means clustering, books are grouped based on their TF-IDF vectors. The Elbow Method is applied to determine the optimal number of clusters(4 is chosen). This grouping helps identify thematic patterns and similarities between books.
4. **Visualization**:
    - Clustered Heat Maps highlight relationships between books based on their descriptions.

Results

Key Findings

1. **Top Characteristic Words**:
    - Each book's top 15 words provides a concise summary of its themes. For example, a fantasy book might include words like "magic," "wizard," and "castle," while a science fiction book might feature "alien," "future," and "technology."
2. **Library-Level Insights**:
    - The aggregated TF-IDF analysis reveals common themes such as genres or popular topics across the library.
3. **Clusters of Books**:
    - Clustering results show distinct groups of books based on their descriptions. For instance, books in one cluster appear to only be Harry Potter books.
4. **Visualization**:

- The cluster scatter plot demonstrates thematic similarities that were found based on the PCA Decomposition of the tokenized descriptions.
- Clustered Heat Maps show similarities between books in the same cluster.

Challenges and Limitations

1. **Short/Invalid Descriptions**:
   - Books with very brief descriptions or books in a non-English language may not yield enough distinctive words for meaningful analysis.
2. **Arbitrary Optimal Number of Clusters:**
   - The Elbow method didn't have a very clear answer.
3. **Huge Scope for adding other relevant information from scraped data**

Conclusion

This project demonstrates the potential of TF-IDF and clustering to analyze book descriptions effectively. By extracting key terms, identifying themes, and grouping books into clusters, we gain a deeper understanding of the library's collection. These insights can be applied to improve book discovery, enhance reader engagement, and support various educational and marketing initiatives.

Future directions include incorporating metadata like genres, ratings, or publication dates, and extending the analysis to multilingual book descriptions for a broader scope.

Python Packages and technologies that the author learnt through this process:

- Goodreads Scraping with Python Package
- Spacy Tokenization
- Using Heatmaps to derive insights from clusters
- Elbow Method to define optimal number of clusters

References

- Source Code: https://github.com/CrownCrafter/goodreads-library-keywords
- Goodreads Scraper: https://github.com/maria-antoniak/goodreads-scraper
- Spacy Docs: https://spacy.io/usage/spacy-101#lightning-tour
- Heatmaps used in final project: https://seaborn.pydata.org/generated/seaborn.heatmap.html