Project Proposal:

Topic Modeling to Identify “Niche” Book Sub-genres

Priyana Aragula
Cornell Tech
New York, NY
pa369@cornell.edu

Angky William
Cornell Tech
New York, NY
aw776@cornell.edu

Srijeeta Biswas
Cornell Tech
New York, NY
sb2555@cornell.edu

1 Motivation

In literature, readers frequently encounter challenges in discovering books that truly resonate with their unique tastes. Many genres exist, and within those broad categories lie subgenres that offer distinct reading experiences. However, the traditional genre categorizations often fall short in distinguishing these niche sub-genres, leading readers to rely more on serendipity than informed choices.

Our project aims to solve this by applying Topic Modeling, an unsupervised machine learning model that uses sentiment analysis and Natural Language Processing (NLP) techniques, to automatically cluster books into sub-genres. By doing so, we aspire to provide readers with a more refined and granular classification system, allowing them to dive deep into sub-genres like high-fantasy, horror-fiction, paranormal-romance, contemporary-comedy, and more.

2 Method

Topic Modeling[1]

Topic Modeling aims to divide a set of documents into two major categories:

1. A list of topics extracted from the documents in the given dataset

2. Groups of documents grouped by the list of topics they cover

This unsupervised machine learning algorithm works with the underlying assumption that all documents consist of a statistical combination of topics that can be computed by analyzing the distribution of the topics in the dataset along with the strength of the presence of each topic in a given document.

Using this assumption, this machine learning technique clusters the documents as per the identified topics.

The topic modeling methods that we will explore (but not limited to) are listed as follows:

1. *Latent Semantic Analysis (LSA)*

2. *Latent Dirichlet Allocation (LDA)*

Our approach is rooted in leveraging the vast textual data available about books—both in their content and associated reviews. To extract meaningful patterns from this textual data, we'll employ a combination of sentiment analysis and NLP techniques.

Sentiment analysis will allow us to gauge the mood and tone of the reviews, which is often indicative of the subgenre (e.g., darker sentiments for horror-fiction). Concurrently, NLP techniques will extract key terms, themes, and motifs that can shed light on a book's themes.

3 Experiments

Our experiments[2] will focus on the effectiveness and granularity of our topic modeling approach.

1. *Dataset*
   We will begin by sourcing a comprehensive dataset of books and their associated reviews. This dataset will serve as the foundation for all our subsequent analyses.

2. *Data Pre-Processing and NLP Techniques*
   We will perform standard pre-processing techniques such as tokenization, lemmatization, removing stopwords, etc. We will use other NLP techniques such as tf-idf, sentiment analysis, keyword extraction, etc.

3. *Dimension Reduction and Running Topic Modeling Algorithms*
   Once this preprocessing is done, our next machine learning technique will be dimension reduction, which will transform our high-dimensional data into a more manageable form. Upon extracting themes and sentiments from the data, we will run our unsupervised algorithm using a bag of words, tf-idf, etc. on our dataset.

4. *Performance Evaluation and Scoring*[3]
   The key performance aspects will be the coherence and relevance of the topics within each cluster along with the distinction between clusters. We expect each cluster to provide a clear indication of a niche sub-genre. We will use scoring metrics such as C_V, U_Mass, topic coherence pipeline, etc. Further, to validate our results, we will perform qualitative evaluations and present our clusters to avid readers and seek feedback on the accuracy and utility of our clustering.

References

[1]https://monkeylearn.com/blog/introduction-to-topic-modeling/

[2]
https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24

[3]https://www.linkedin.com/advice/0/how-do-you-evaluate-quality-relevance-your-1e#:~:text=There%20are%20two%20main%20aspects,their%20semantic%20similarity%20or%20frequency.