Data Processing

WHAT WORDS ARE MOST CHARACTERISTIC OF MY GOODREADS LIBRARY?

Ayush Mudunuru | Presentation

Introduction

Objectives

- Identify important words that characterize my library.
- Group similar books into clusters based on descriptions.

Purpose

- Applying Spacy to real life text for tokenization.
- Applying the K-Means algorithm on text to extract insights regarding books.
- Extracting meaningful insights from my library for future recommendations.

Methodology

Text Retrieval

- Scraped my web profile using my user id and GoodReads scraper package.
- Used json libraries to parse book jsons and return descriptions.

Description Tokenization

- Used Spacy to generate lemmas of descriptions.
- Vectorized text with sklearn and finally generated data file containing top 15 characteristic words for each book.
- Plot of most characteristic words of each book.

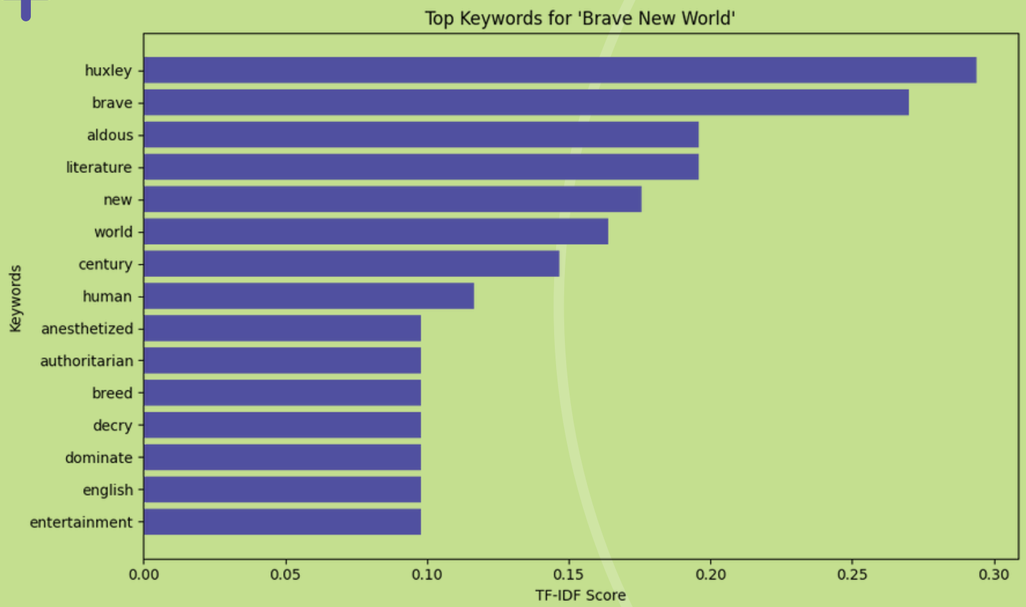
Methodology

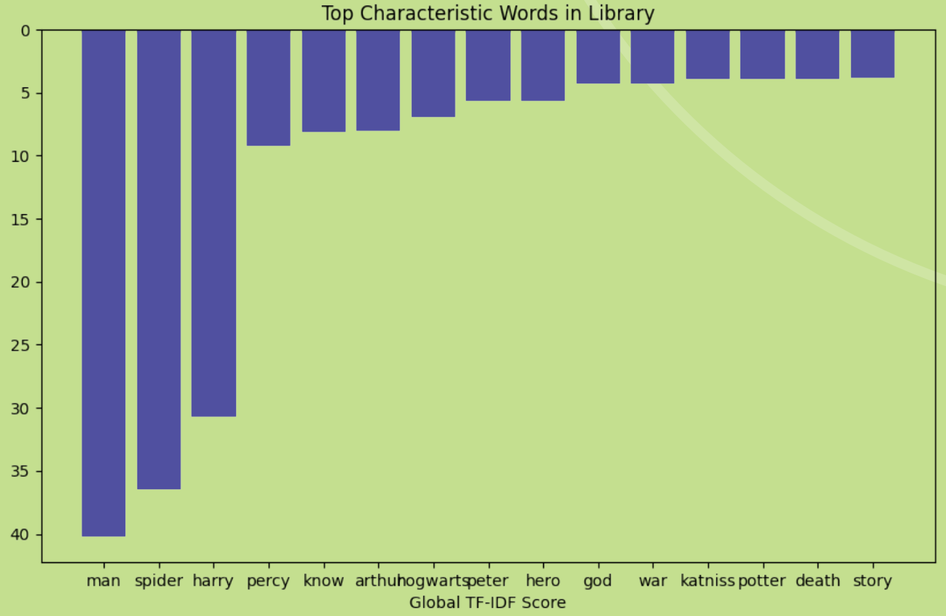
Aggregation and Plot

- Aggregated Scores(multiplied by document frequency) across all books to generate final plot.

Description Tokenization

- Used Spacy to generate lemmas of descriptions.
- Vectorized text with sklearn and finally generated data file containing top 15 characteristic words for each book.
- Plot of most characteristic words of each book.

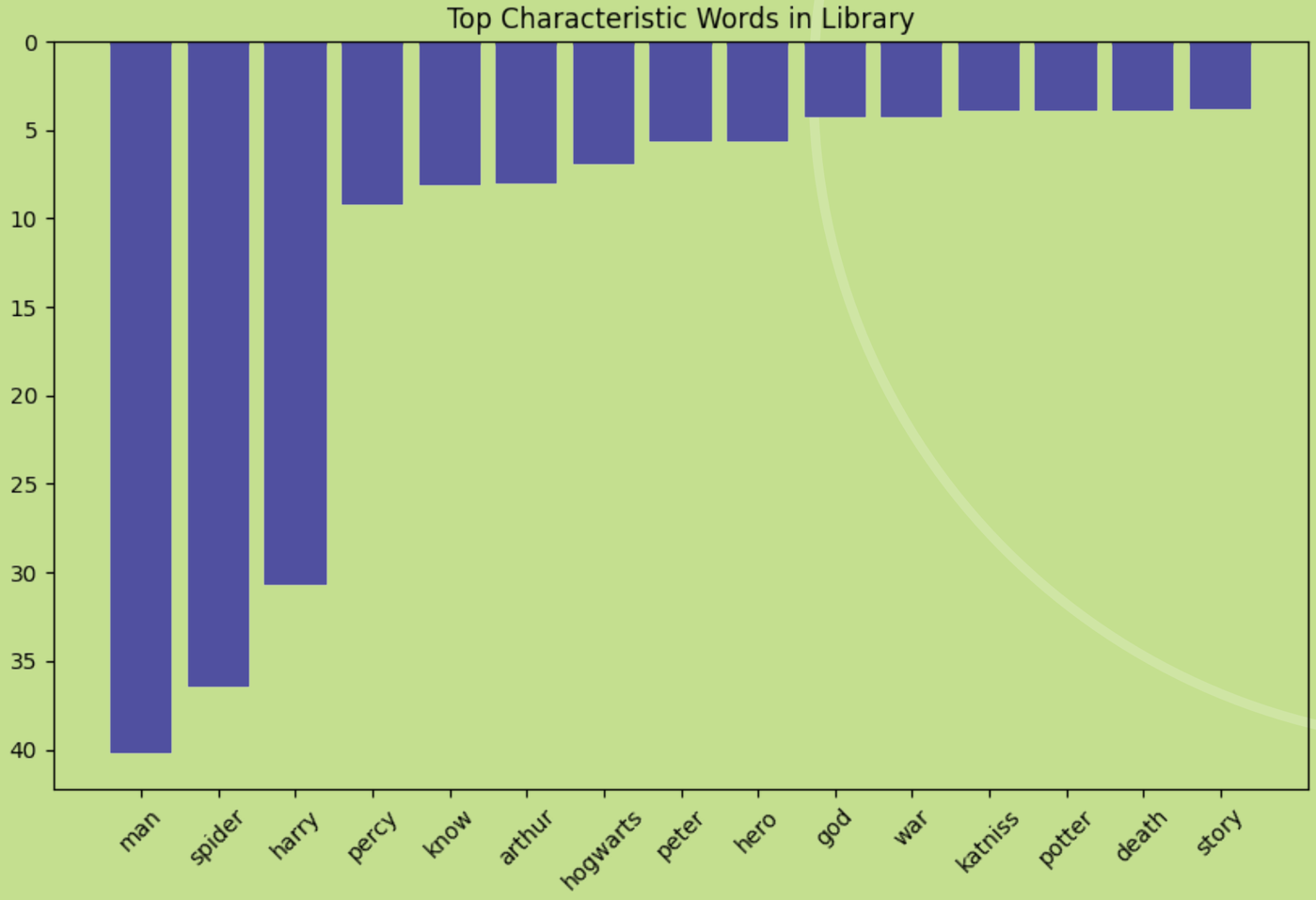




Results

Most Important Keywords in Library

- Can make out franchises that I have/had interest in.



Book Clustering

Hey, what if we tried to see which books are similar to each other to help future recommendations?

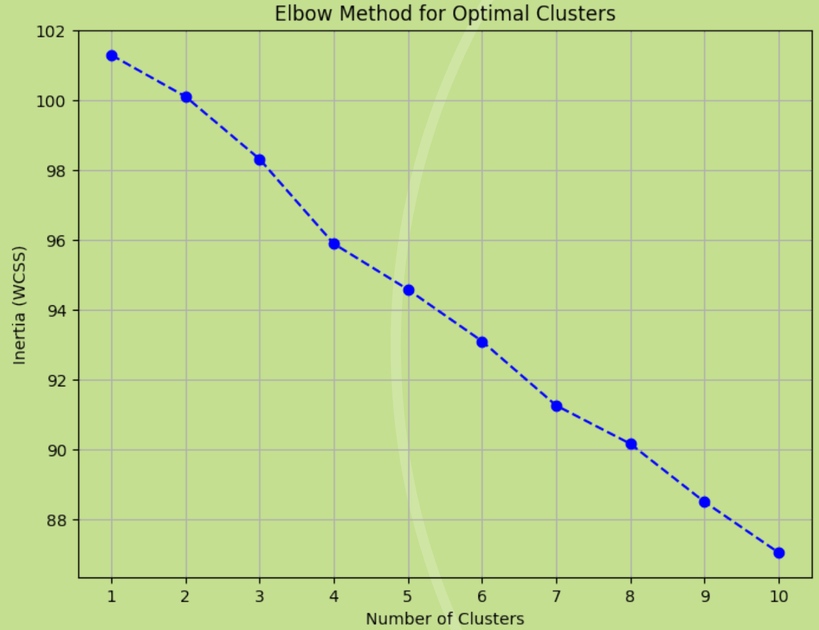
Secondary Research Question

What distinct sub-groups emerge from clustering of my Goodreads library?

Methodology

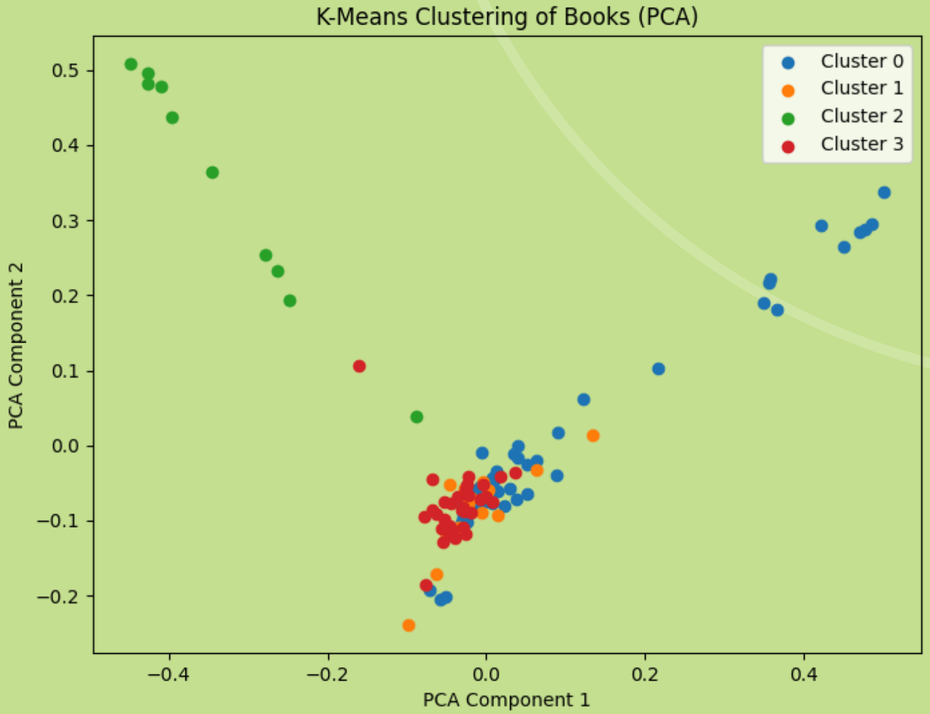
How many clusters are appropriate?1

- Elbow Method to decide.
- Picked 4.



PCA Decomposition and Clustering

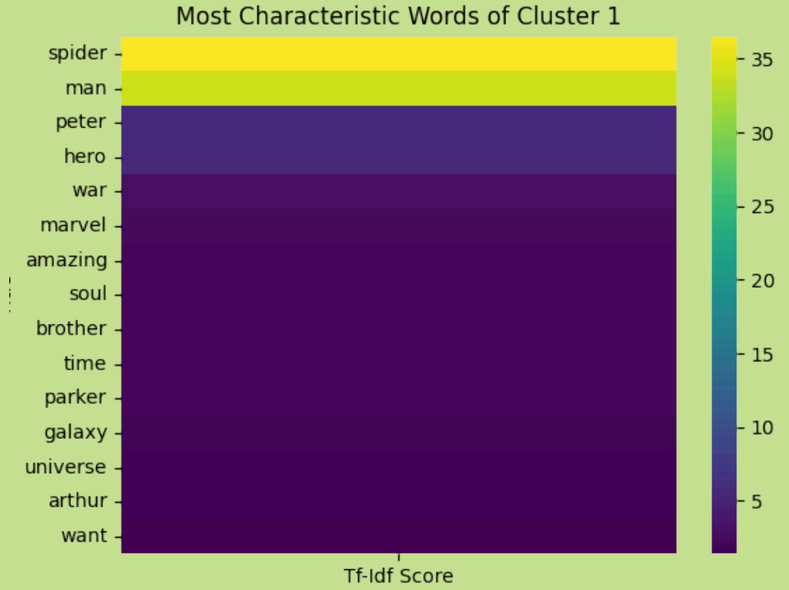
- Decomposed to 2 PCA Components.
- Returned data file with formed clusters.

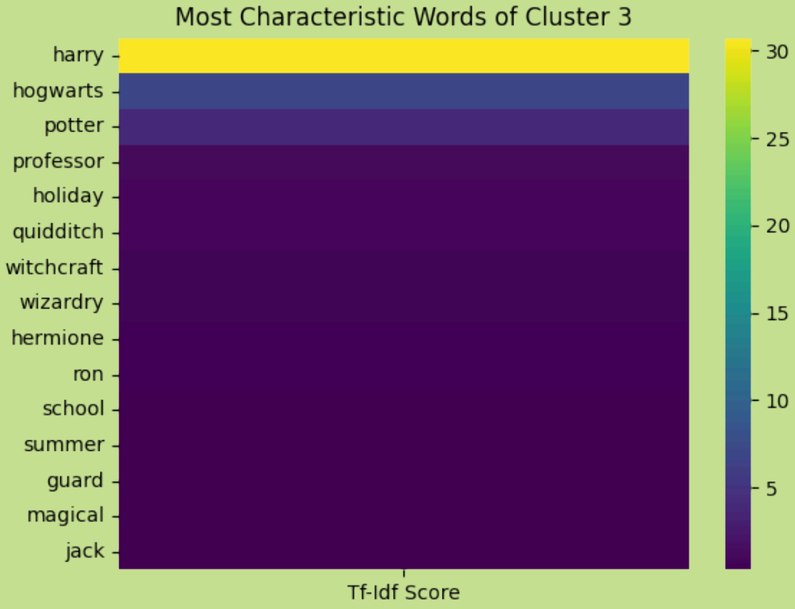


Results

Clusters

- Comics
- Sci-Fi
- Harry Potter
- Young Adult





Challenges and Limitations

- Some books have very short descriptions.
- Huge Scope for incorporating other relevant information.
- Difficult to find insights from Clusters before heatmaps
- Arbitrary optimal number of clusters

thank you