Clustering the Goodbooks-10k Dataset

Matej Kulháň

January 6, 2025

Data

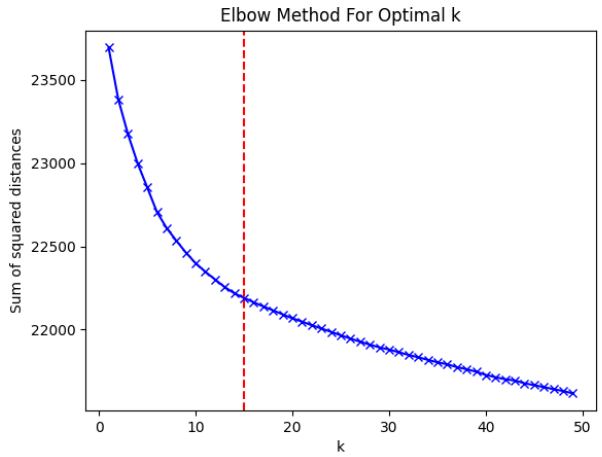
- **Dataset**:
  - `books_enriched.csv`
  - `tags.csv`
  - `book_tags.csv`
  - `ratings.csv`
- **Data Preprocessing**:
  - `Genre Extraction`
  - `Description Cleaning`
  - `Removing Non-English Books`

Distance Matrix

- **Handling Mixed Data Types**:
  - Calculated the distance matrix using the `Gower` distance
- **Adding Text Embeddings**:
  - Used `SBERT` embeddings to capture the semantic context of book descriptions
  - Created a distance matrix using the `cosine` distance
- **Adding User Interactions**:
  - Used the `ratings` dataset to capture user interactions with books
  - Created a distance matrix using the `Jaccard` distance for sets of user ratings
- **Combining Distance Matrices**:
  - Summed the distance matrices from numerical and binary features, text embeddings, and user interactions into a single distance matrix

Clustering

- **K-Medoids Clustering**:
  - Chose the number of clusters to be 15
  - Used the `K-Medoids` algorithm on the combined distance matrix



Cluster Naming Process

- **Cluster Metadata**:
  - Extracted keywords from book descriptions using `TF-IDF`
  - Used the cluster's mean values for numerical features and genre distributions
- **Generating Cluster Names:**
  - Used a pretrained `LLM` model to generate cluster names based on the extracted metadata

Cluster Example

- **Cluster Metadata:**
  - *Average Rating:* 13 (Above Average Rating)
  - *Original Publication Year:* 12 (Slightly Newer Books)
  - *Pages:* 11 (Average Length)
  - *Ratings Count:* 5 (Not The Most Popular)
  - *Dominant Genres:*
    - Fantasy: 0.84
    - Romance: 0.97
    - Adult: 0.59
    - Historical Fiction: 0.14
- **Keywords:** kresley, vamp, bridgerton, paranormal, brotherhood, fae, anita, crawfield, shapeshifter, year vampire
- **Generated Cluster Name:** *Vampire Lore and Fae Mythology: Dark Romantic Fantasies*