The Goodbooks Dtaset Clustering

Matej Kulháň

December 13, 2024

Dataset Overview

- **Original Dataset**:
  - `books.csv`
  - `tags.csv`
  - `book_tags.csv`
- **Extended Dataset**:
  - `books_enriched.csv`.
  - Additional features:
    - Book descriptions
    - Pages

Data Preprocessing

- **Genre Extraction**:
  - Extracted the most frequent tags.
  - Identified tags that corresponded to actual genres.
  - Assigned a genre to books with the tag in their top 10 tags.
  - Final list of genres:
    - `young-adult`, `fantasy`, `nonfiction`, `romance`, `adult`, `science-fiction`, `contemporary`, `mystery`, `classics`, `historical-fiction`.
- **Description Cleaning**:
  - Cleaned `description` column for NLP tasks:
    - Removed special characters, stopwords, etc.
    - Removed common entities like names, locations, etc.
    - Lemmatized the text.
  - Filtered out non-English books.

Distance Matrix

- **Handling Mixed Data Types**:
  - Used the `Gower` distance, which is specifically designed to handle datasets with a mix of numerical and categorical features.
  - Allows for fair comparison across different feature types without needing to normalize all variables to the same scale.
- **Features Used**:
  - **Numerical Features**:
    - `average_rating`
    - `original_publication_year`
    - `pages`
    - `ratings_count`
    - `genre_count`
  - **Binary Features**:
    - `genres`

Clustering

- **K-Medoids Clustering**:
  - Used the `K-Medoids` algorithm.
  - Chose the number of clusters to be 12 based on a combination of the `elbow method` and human evaluation.

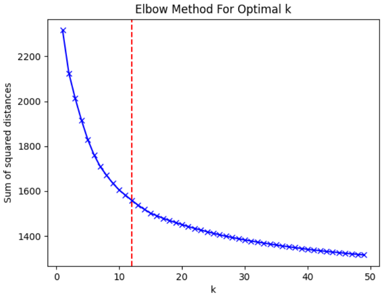

Figure: Elbow Method

Cluster Examples

Nonfic.	Fantasy	Pub. Year	Cluster Description
0.01	0.07	1911	American literary classics about the pioneer lifestyle.
0.00	0.81	1993	Science fiction about space exploration and intergalactic conflicts.
1.00	0.00	1985	Nonfiction books about leadership growth and personal development.
0.07	0.08	1850	Mystery books with a focus on crime investigation and characters.

Table: Examples of clusters with selected columns and rounded data.

Text Embeddings

- **Adding Semantic Context**:
  - Used text embeddings to capture the nuanced meaning of book descriptions.
  - Helped differentiate books within broad genres (e.g., various types of nonfiction).
- **SBERT Embeddings**:
  - Used the `Sentence-BERT` model to generate embeddings.
  - Calculated the distance matrix using the `cosine` distance.
  - Combined the Gower distance matrix (from numerical and binary features) with the distance matrix derived from text embeddings using a weighted sum approach.

Clustering

- **K-Medoids Clustering**:
  - Used the `K-Medoids` algorithm.
  - Chose the number of cluster to be 14 based on a combination of the `elbow method` and human evaluation.

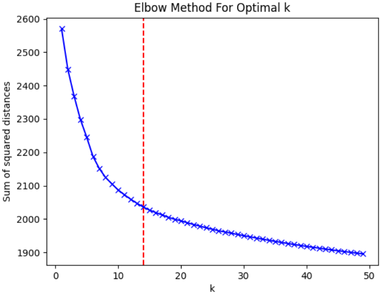

Figure: Elbow Method

Cluster Examples

Nonfic.	Fantasy	Pub. Year	Cluster Description
0.92	0.03	1972	Inspirational nonfiction about personal faith and spiritual growth.
0.00	0.71	1992	Science fiction about space exploration and technological anomalies.
0.96	0.00	1977	Intellectual nonfiction on leadership and personal development.
0.03	0.05	1985	Whimsical children's stories with humorous and relatable protagonists.

Table: Examples of clusters with selected columns and rounded data.

Cluster Naming Process

- **Extracting Keywords:**
  - Used the `Tf-Idf` algorithm to identify the most relevant keywords from book descriptions.
- **Generating Cluster Names:**
  - Used a pre-trained `Large Language Model (LLM)` for generating descriptive and human-readable cluster names.
  - The model was instructed to focus on conciseness and relevance based on extracted keywords.
- **Example**:
  - Keywords: space, galaxy, aliens, technology
  - Generated Name: `Science fiction about space exploration.`

Future Work

- **Clustering**:
  - Explore different clustering algorithms such as `HDBSCAN` and compare their performance.
- **Evaluation**:
  - Move beyond visual inspection by implementing systematic methods to evaluate cluster quality in a reproducible and scalable manner.
- **Keyword Extraction**:
  - Experiment with improved approaches for extracting meaningful keywords from book descriptions, such as:
    - Rule-based techniques like `RAKE`.
    - Leveraging the power of `Pre-trained LLMs`.