The Goodbooks Dataset Clustering

Matej Kulháň
ČVUT - FIT
kulhama8@fit.cvut.cz

December 31, 2024

1 Introduction

The **Goodbooks-10k** dataset [5] is a collection of the 10,000 most rated books from the Goodreads website. The dataset contains information about the books, such as the title, author, and rating. The goal of this project is to cluster the books in the dataset based on their content, such as the rating, publication year, etc. The clustering will be done using the **k-medoids** algorithm, which is a variation of the k-means algorithm. Lastly, a method to name each cluster based on the books that belong to it will be proposed.

2 Data

2.1 Data Description

The Goodbooks-10k dataset contains multiple files, of which the following are used in this project:

- **books.csv** - Contains information about the books, such as the title, author, publication year, and rating.
- **ratings.csv** - Contains information about the ratings of the books.
- **tags.csv** - Contains information about the tags of the books.
- **book_tags.csv** - Contains information about the tags of the books.

In addition, there also exists an extended version of the dataset [3] that contains additional information about the books, such as the **book description** and the **number of pages**. These additional features will be used in the clustering process.

2.2 Data Preprocessing

2.2.1 Genres

The combination of the **tags.csv** and **book_tags.csv** files is used to determine the genres of the books. Firstly, the most common tags are determined by counting the number of occurrences of each tag as one of the top tags for each book. From these tags, only the tags that represent genres are selected. The genres are then assigned to the books based on the tags that are associated with them. The genres follow the one-hot encoding scheme, where each genre is its own feature, and the value of the feature is 1 if the book belongs to the genre, and 0 otherwise. One book can belong to multiple genres.

The following genres are used in this project: *Young Adult, Fantasy, Nonfiction, Romance, Adult, Science Fiction, Contemporary, Mystery, Classics, Historical Fiction.*

2.2.2 Numerical Features

The numerical features of the books are extracted from the **books.csv** file. The features used in this project are the *average rating, number of ratings, publication year and the number of pages.* All of these features are normalized.

2.2.3 Book Description

The book descriptions are a short summary of the book content. The descriptions are preprocessed by removing any special characters, numbers, and stopwords. Lemmatization is then applied to the descriptions to reduce the words to their base form. The language of each description is determined using the **langid** library. Only books with English descriptions are used in the clustering process. This removes approximately 250 books from the dataset.

3 Methodology

3.1 Handling Mixed Data Types

The dataset contains both numerical and categorical data. This prevents the use of traditional clustering algorithms, such as k-means, which require numerical data. To handle this, the **Gower distance** [4] is used, which is a distance metric that can handle mixed data types. This creates a distance matrix that can be used in the **k-medoids** [1] algorithm.

3.1.1 Adding Additional Semantic Information

The numerical and categorical features aren't enough to accurately cluster the books. For example, the nonfiction genre contains a wide variety of books, such as biographies, self-help books, and cookbooks.

To add additional semantic information to the clustering process, the **Sentence-BERT** [2] model is used to create embeddings of the book descriptions. From these embeddings, a similarity matrix is created, which is the cosine similarity between the embeddings of the book descriptions. This matrix is inverted to create a distance matrix, which is then combined with the Gower distance matrix.

Another way to add semantic information is to use the **ratings.csv** file to determine the similarity between books based on the sets of users that rated each book. The **Jaccard similarity** is used to determine the similarity between the sets of users that rated each book. This similarity matrix is also inverted to create a distance matrix, and then combined with the Gower and Sentence-BERT distance matrices.

This results in a combined distance matrix that contains the numerical, categorical, and semantic information of the books.

3.2 Clustering

The initial number of clusters is determined using the **elbow method** and later refined through human evaluation and visual inspection of the clusters projected into a 2D space using the **UMAP** algorithm.

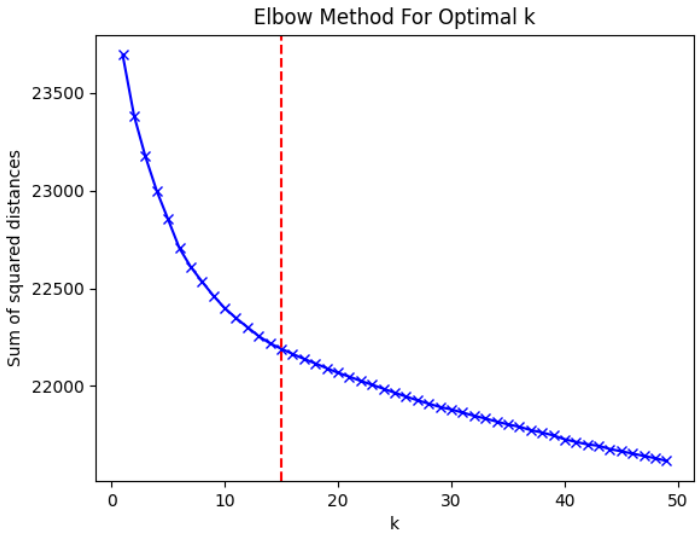

Figure 1: Elbow Method

The optimal number of clusters is determined to be 15, which is slightly higher than what the elbow method suggests. This is due to the complexity of the dataset and the need to capture the diversity of the books.

The **k-medoids** algorithm is used to cluster the books. This algorithm is a variation of the k-means algorithm that allows for the use of a precomputed distance matrix. The algorithm is run with the distance matrix, and uses the **FasterPAM** algorithm to find the medoids of the clusters.

3.3 Naming the Clusters

3.3.1 Keyword Extraction

To identify the thematic essence of each cluster, the descriptions of all books within a cluster are concatenated into a single document. From these documents, keywords are extracted using the **Tf-idf** algorithm, which highlights terms that are distinctive for the cluster relative to the entire dataset.

3.3.2 Cluster Metadata

The numerical and categorical features of the books within each cluster are aggregated to generate a comprehensive metadata profile. Key metadata elements include the most common genres, average rating, publication year, and number of pages. Genres are represented as continuous features ranging from 0 to 1, indicating the percentage of books in the cluster belonging to each genre. Other metadata features are normalized on a scale from 1 to 20, where higher values denote a stronger presence of a feature across clusters.

3.3.3 Cluster Naming Process

The extracted keywords and aggregated metadata are used to generate descriptive and concise names for the clusters. An external language model is tasked with naming the clusters, following specific guidelines to ensure meaningful, creative, and human-readable names. These guidelines emphasize thematic interpretation of keywords, integration of dominant genres, and creative use of metadata to reflect the unique characteristics of each cluster. Nonfiction clusters are explicitly labeled as such if nonfiction exceeds 50

4 Results

4.1 Without Semantic Information

The clustering without the use of semantic information results in clusters, that are almost entirely based on the genres of the books. This results in clusters that are too broad and don't capture the diversity of the books.

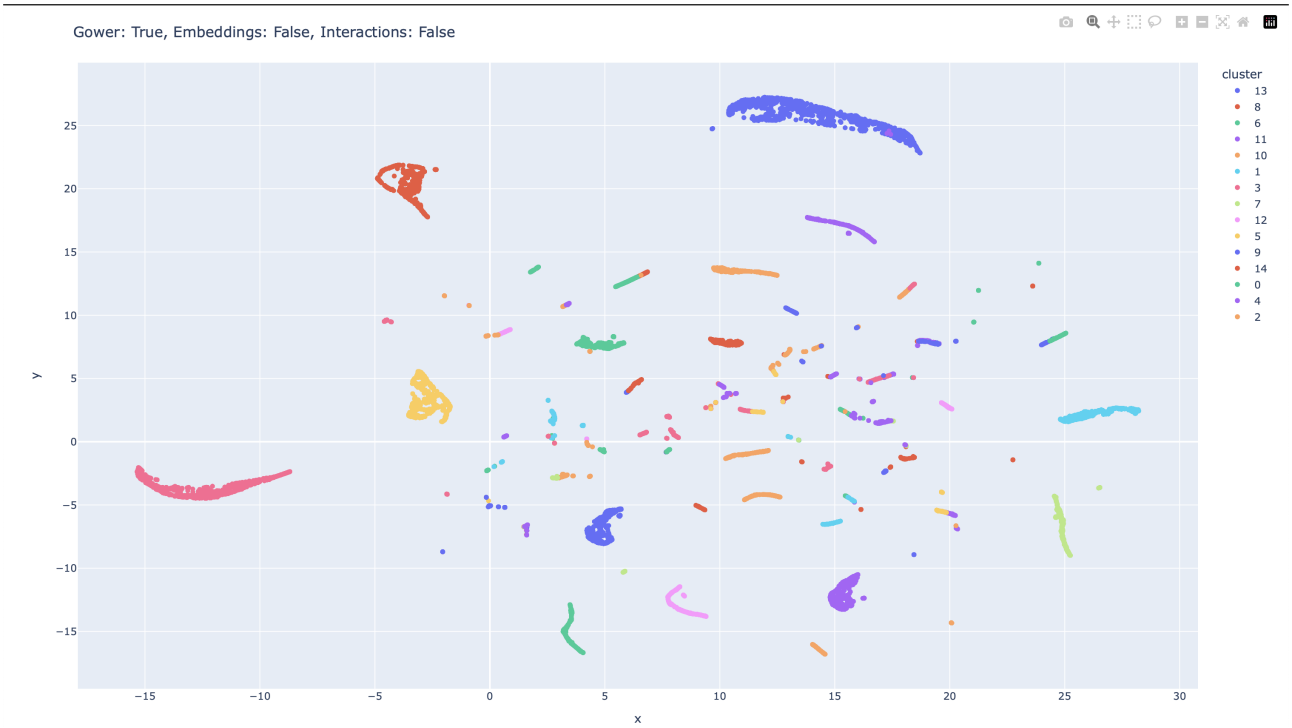

Figure 2: Clustering Without Semantic Information

As seen in Figure 2, there are a few big clusters that contain the most popular combinations of genres, and noise that contains the rest of the books. This is due to the lack of semantic information. There is only one cluster that contains nonfiction books. The cluster name is *Autobiographical Insights of Leadership and Personal Growth Through Candid Reflections*.

4.2 With Semantic Information

The clustering with the use of semantic information results in clusters that are more diverse and capture the essence of the books.

Figure 3: Clustering With Semantic Information

As seen in Figure 3, the clusters are more evenly distributed and contain a variety of genres and themes. The clusters are more diverse and capture the essence of the books. There are also two different clusters that contain nonfiction books, which is a significant improvement over the clustering without semantic information. The cluster names are:

- *Cluster 1: Business Insights of a Visionary Life*
- *Cluster 2: Inspirational Spiritual Journeys of Self-Discovery and Faith*

An example of a named cluster:

- **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*

5 Conclusion

The clustering of the Goodbooks-10k dataset was done using the k-medoids algorithm with the Gower distance matrix. The clustering was done with the use of semantic information, such as the book descriptions and the user interactions with the books. The clusters were named based on the keywords extracted from the book descriptions and the aggregated metadata of the books within each cluster. The results show that the use of semantic information improves the clustering and results in more diverse and meaningful clusters.

References

[1] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.

[2] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[3] Olivier Simard-Hanley. goodbooks-10k-extended: An extended dataset for book recommendations, 2017. GitHub repository, accessed December 31, 2024.

[4] Gulanbaier Tuerhong and Seoung Bum Kim. Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. *Expert systems with applications*, 41(4):1701–1707, 2014.

[5] Zygmunt Zając. goodbooks-10k: A dataset for book recommendations, 2017. GitHub repository, accessed December 31, 2024.