Data Engineering

Mitchell Brogan

11/1/2022

**Abstract**: The goal of this project was to create a webapp capable of taking a book title as an input and recommending the top ten most similar books based on the given book’s description. This tool would be useful for avid readers and book worms looking for their next read or finding other titles with similar subject matter.

**Design**: First the data was loaded out of DB Browser for SQLite using SQL Alchemy into a pandas dataframe. The data was then cleaned by removing unwanted characters using regex and all text was converted to lowercase. A Jupyter lab session was created on a Google Cloud instance on a machine with 128 GB RAM to perform the following data processing steps. The descriptions for all 100,000 unique books were vectorized using TF-IDF vectorizer. PCA was used for dimensionality reduction while keeping an explained variance ratio above 0.90. Lastly, K-means clustering was applied on the PCA model where K = 5 clusters.

The web application was created using Streamlit and uses the pairwise cosine distances of the clustered model to compute the top 10 most similar books to a target book entered in the search bar by the user.

**Data**: 100,000 rows with 2 columns. The first column has unique book names and the second column contains descriptions of each respective book. This dataset was downloaded from the following link: **https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews?select=Books\_rating.csv**

**Algorithms**:

* Using regex to remove unwanted characters
* TF-IDF vectorizer to create 1-gram tokens
* PCA applied to vectorized grams to reduce dimensionality to 20 dimensions
* K-Means clustering on dimensionally reduced model to 5 clusters
* Pairwise distances used to compute cosine similarity between given book and other books in the same cluster
* Frontend app design using Streamlit functions

**Tools**:

* DB Browser for SQLite
* Google Cloud Services
* Streamlit
* NLTK
* SKLearn
* Pandas
* Regex

**Communication**: In addition to the powerpoint presented, the Jupyter Notebooks will be posted on the following GitHub link: <https://github.com/MitchellB9/BookBoy-Recommender-App>