Project 4 – Medium Recommendation System

Bianca Orozco

Project Design

Medium is a popular online publishing platform with 150 million monthly readers. It currently has an article recommendation system and my goal was to create one of my own using a Kaggle dataset.

Data

The dataset included 1.4 million observations and the features I would be using include the following: Titles, Subtitles, and Claps.

Tools

Jupyter Notebook
Data Analysis:

- Pandas
- Numpy
- Matplotlib
- Sklearn

Modeling:

- CountVectorizer
- TfdifVectorizer
- TruncatedSVD
- KMeans

Methods

First, I cleaned the data and combined Titles and Subtitles to make one text. Then I created a pipeline starting with TF-IDF using unigrams and bigrams. Next in my pipeline, I used LSA for dimensionality reduction and NLP. Lastly, I applied K-Means to my pipeline, specifying 4 clusters. I was then able to create a function that allowed me to take in a user's input in the form of a string, use my pipeline to predict a cluster, and print out the top 3 related article titles within that assigned cluster.

What I Would Do Differently

Because my dataset originated from Kaggle and lacked full Medium articles, I would web scrape my own dataset with the title, subtitle, and article. I would also use cosine similarity to first find the closest articles, then the most popular. I would also use GridSearchCV to optimize my parameters.