Cecilia Bell 6.25.2021

NLP Project Write-Up

Content-Based Movie Recommendation System

Abstract

The goal of this project was to create a content-based movie recommendation system involving natural language processing and an unsupervised learning model. I worked with a dataset from Kaggle which included over 30,000 Wikipedia movie plots. After cleaning and preprocessing the text data, I used the TF-IDF vectorizer and the NMF model to create topics based on the corpus of the movie plots. With these topics, I used the cosine similarity between movie topic profiles to create a content-based recommendation system, which included an origin/country filter.

Design

Providing customers with recommendations can improve the customer experience and in turn increase revenue. Recommendation systems are therefore widely used by SVOD, AVOD and TVOD streaming companies. For this project, I created a content-based movie recommendation system, which would be particularly useful to a streaming company that lacked data on their customers. Another advantage of my recommendation system is its origin/country filter, which allows someone to find a movie like *The Godfather* in Chinese or *The Notebook* in Telugu.

Data

The data used is from Kaggle and can be accessed here. The raw dataset contains entries for 34,886 movies, with each row representing a movie and with 8 columns representing various attributes of the movie. After cleaning the data, 33,931 movies remained.

Algorithms

- Cleaning: plots with fewer than 20 words were dropped as most were improperly scraped and contained no information for topic modeling. Null values and duplicate movies were also dropped.
- Preprocessing: numbers and punctuation were removed from the text, which was then tokenized. After tagging the tokenized text with parts of speech, proper nouns were removed and the remaining words were lemmatized.
- Topic modeling: different vectorizers and models were tested with different parameters including stopwords. The TF-IDF vectorizer and NMF model were eventually selected, with 30 topics and custom stopwords combined with English stopwords.
- Recommendation system: the recommendation system was built using the cosine similarity between movie topic profiles. A filter for origin/country was also included.

Tools

- Numpy and Pandas for data manipulation
- NLTK for text processing
- Scikit-learn for preprocessing and modeling
- Matplotlib and Plotly for plotting

Communication

An explanation of the recommendation system, as well as visualizations displaying insights into the textual data, are displayed in a PowerPoint, which will be presented to the Metis class and shared on GitHub.