Cecilia Bell 6.16.2021

NLP Project Proposal

Content-Based Movie Recommendation System

Question

What topics can be extracted from movie plots? Can these topics form the basis of a recommendation system? For this project, I plan to perform natural language processing on a large collection of Wikipedia movie plots to build an unsupervised learning model that creates movie profiles with topic modelling, which forms the basis of a content-based recommendation system. This type of recommendation system will be particularly helpful to movie watchers who want recommendations based on topics that they are currently interested in, rather than their viewing history.

Data

The data I plan to use 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, most of which will be dropped. An individual unit of analysis will be the plot for one movie. I expect that themes related to genres will figure as topics in the model.

Tools

- Numpy and Pandas for data manipulation
- spaCy for text processing
- Scikit-learn for preprocessing, modelling and statistics
- Matplotlib and Seaborn for plotting
- Tableau for presentation visualizations

MVP Goal

My MVP goal is to create a basic NLP pipeline to process the movie plot text data and build an unsupervised learning model that gives a preliminary set of topics and movie profiles.