Neshat Jalali Heravi
January 2022
NLP Project Proposal

Building a Content-based Movie Recommendation System

Question

What topics can be extracted from movie overviews and plots? Can these topics be used as a basis of a movie recommendation system? The objective of this project is to utilize Natural Language Processing on a movie meta-dataset and build an unsupervised learning model that creates different movie profiles using topic modeling, which can be used as a basis of recommendation system. The proposed recommendation system can be beneficial to new streaming services without sufficient user history or established streaming services' consumers who are looking for movies based on their current topics of interests rather than their user history!

Data

The dataset for this project contains information of 45,466 movies featured in the Full MovieLens dataset (movies that are released on or before July 2017) and is available on Kaggle. The dataset includes 24 column features including release dates, budget, overviews, etc. The focus of this project will be the overviews which are a short summary of the movie plots. Each row represents a movie, and the overview feature column include unprocessed text overview of each. An individual unit of analysis will be a single overview.

Tools

- Pandas and Numpy for data manipulation
- spaCy and RegEx libraries for text processing
- scikit-learn for modeling
- Matplotlib and seaborn for plotting and visualization

MVP

The MVP will include the basic NLP pipeline to process the text data and the preliminary unsupervised model that provides movie profiles based on movie overviews.