Movies Recommender System and Topic Modeling

By: Sarah Alabdulwahab & Asma Althakafi

Introduction

Recommender systems have two main techniques: collaborative and content-based. Content-based recommender systems recommend based on data that the user likes. For this Project, we will try to find the degree of similarity between movies in order to recommend movies based on their plots and perform topic modeling on the movie plots as well. Since we are working with textual data, we will use Natural Language Processing (NLP), which is a field of study that works on making computers understand and interpret the human language.

Data Description

We obtained the dataset from Kaggle.com (from this link) and it consists of 45K movies released on or before July 2017. There are 24 features such as budget, genres, overview, revenue, release dates, languages, production companies, TMDB vote counts and vote averages, etc. However, these are the features that we are focused on:

- id: The ID of the movie in TMDB.
- imdb_id: The ID of the movie in IMDB.
- original_title: The name of the movie.
- overview: The plot of the movie from TMDB.
- genres: The genres of the movie.

In addition, we will use The Movie Database (TMDB) API to retrieve the keywords for each movie and we will scrape the plots of the movies from IMDB to increase our word count. Here are the additional features:

- keywords: The keywords that describe the movie.
- imdb_plot: The plot of the movie from IMDB.

Tools

- Pandas and Numpy for data manipulation.
- Beautiful Soup for web scraping.
- Sklearn for unsupervised learning.
- NLTK and spaCy for text manipulation.

MVP Goal

The expected outcome is a content-based movie recommender system, a topic modeling of movies based on their plots, a dataset, a report of the analysis, and finally, a presentation highlighting all the main points from beginning to end.