2023 CAPSTONE PROJECT: FINAL REPORT

ReelWhisperer:

Story-Driven Movie Recommendations

Soohyun Ahn
Data Science Diploma Program @ BrainStation

Problem Statement

How can we use natural language processing (NLP) and unsupervised machine learning techniques to build a personalized movie recommendation system that utilizes users' specific plot preferences and narrative elements, and provides more accurate and relevant movie suggestions?

Background

In recent years, recommendation systems have become increasingly important as they provide a way to effectively filter and recommend personalized content to users based on their preferences. Content-based and collaborative filtering techniques are commonly used in recommendation systems, but these approaches have limitations such as over-specialization and cold start problems. Hybrid techniques, which combine multiple approaches, have shown promising results in addressing these limitations. In addition, NLP and unsupervised machine learning techniques have become more advanced, offering new opportunities for personalized recommendation systems.

Value Added

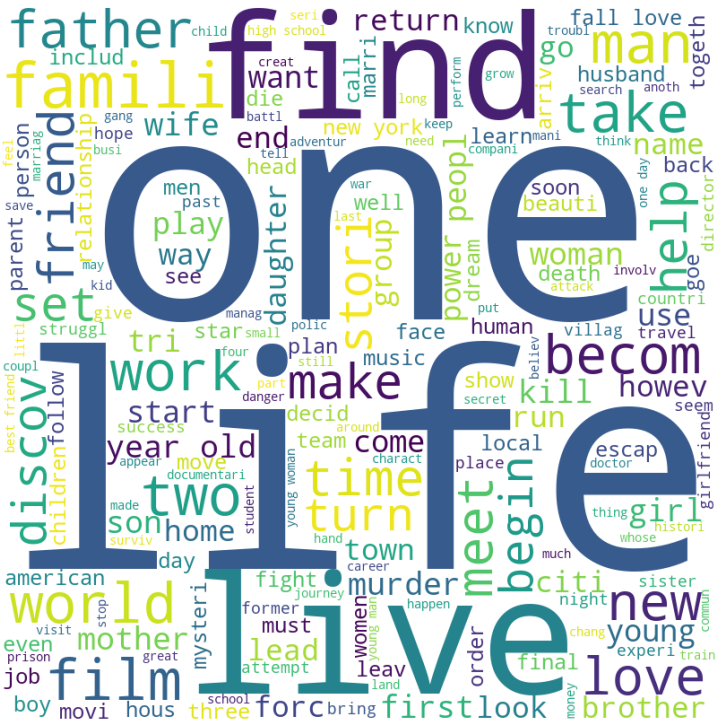
ReelWhisperer aims to provide a unique, user-centric, and engaging movie recommendation experience by incorporating state-of-the-art natural language processing and unsupervised machine learning techniques. The recommendation system utilizes OpenAI's advanced embedding model, which allows for comprehensive analysis of movie plot summaries and other textual data. ReelWhisperer has the potential to offer personalized and meaningful movie recommendations that cater to specific plot preferences and narrative elements, which may enhance user engagement and satisfaction.

Potential Use Cases

- **Personalized movie recommendations**: Users could input their desired storylines to receive movie recommendations tailored to their tastes. This could help users discover new movies that they might not have otherwise found and could enhance their overall movie-watching experience.
- **Themed movie nights**: Users could search for movies that share common themes or narrative elements, making it easier to plan themed movie nights.
- **Screenwriting inspiration:** Screenwriters could use ReelWhisperer to search for movies with similar plotlines as their own ideas, providing inspiration and guidance for their own projects.
- **Content recommendation for streaming platforms**: Streaming platforms could integrate and fully develop ReelWhisperer to complement their recommendation systems, offering users movie suggestions based on specific plot preferences or narrative elements. This could enhance user satisfaction and engagement, and ultimately drive more views and revenue for the platform.

About the Data

In this project, I used two publicly available datasets: one from the project called "MPST: Movie Plot Synopses with Tags" and the other from "The Movies Dataset" on Kaggle. Initially, the first dataset was created to automatically generate tags for movies. For my project, I used it to create a movie recommendation system, driven by user-input storylines for a more personalized experience.



To improve my movie recommendation system, I plan to incorporate supplementary data sources in the future. One option is to gather user feedback on recommended movies and use it to improve the model. Genre, actor, and director data can also be used to tailor recommendations These additions have the potential to improve the accuracy and relevance of my recommendations, enhancing the overall user experience.

Figure 1] A wordcloud created from movie synopses. The size of words represents their relative frequency and importance.

Data Cleaning, EDA, & Text Preprocessing

During the process of data cleaning, I successfully scraped a significant number of English titles for movies with non-English titles from the IMDb website. However, I also encountered a challenge when attempting to access the website, as I was not aware of the potential issues that could arise from web scraping, such as triggering website protections against bots. Despite this challenge, I was able to identify and rectify my mistake and learned the important lesson of doing web scraping responsibly.

I faced challenges with removing duplicate rows from my dataset, leading to issues in my model's top 5 movie recommendations. I realized that some movies had different tags from different sources, causing them to be unrecognized as duplicates. To resolve this, I identified duplicates under the columns of `title` and `imdb_id` during the cleaning process to ensure accurate removal of duplicate movies.

During text preprocessing, some rows had null values due to short text that was removed, resulting in empty strings. The standard method for identifying missing values failed to catch this problem because empty strings aren't considered missing values. I resolved this by using `str.strip()` to remove leading/trailing white spaces and checking for empty strings, ensuring consistency in data analysis.

Feature Extraction, Feature Engineering, and Analysis

In my project, I utilized various feature extraction and engineering techniques to transform raw data into features suitable for machine learning algorithms. To convert textual data into numerical vectors, I performed text vectorization using techniques such as bag-of-words, CountVectorizer, and word embeddings. Additionally, clustering techniques such as k-means clustering with t-SNE and DBSCAN were used to identify groups of similar movies in the dataset.

Creating word/document embeddings is an important part of preparing text data for feature engineering. In my analysis, I used Latent Dirichlet Allocation (LDA), a probabilistic topic modeling technique commonly used in NLP, to identify underlying themes among the movies in the dataset. To visualize the results of the LDA analysis, I used PyLDAvis to create interactive visualizations of the identified topics.

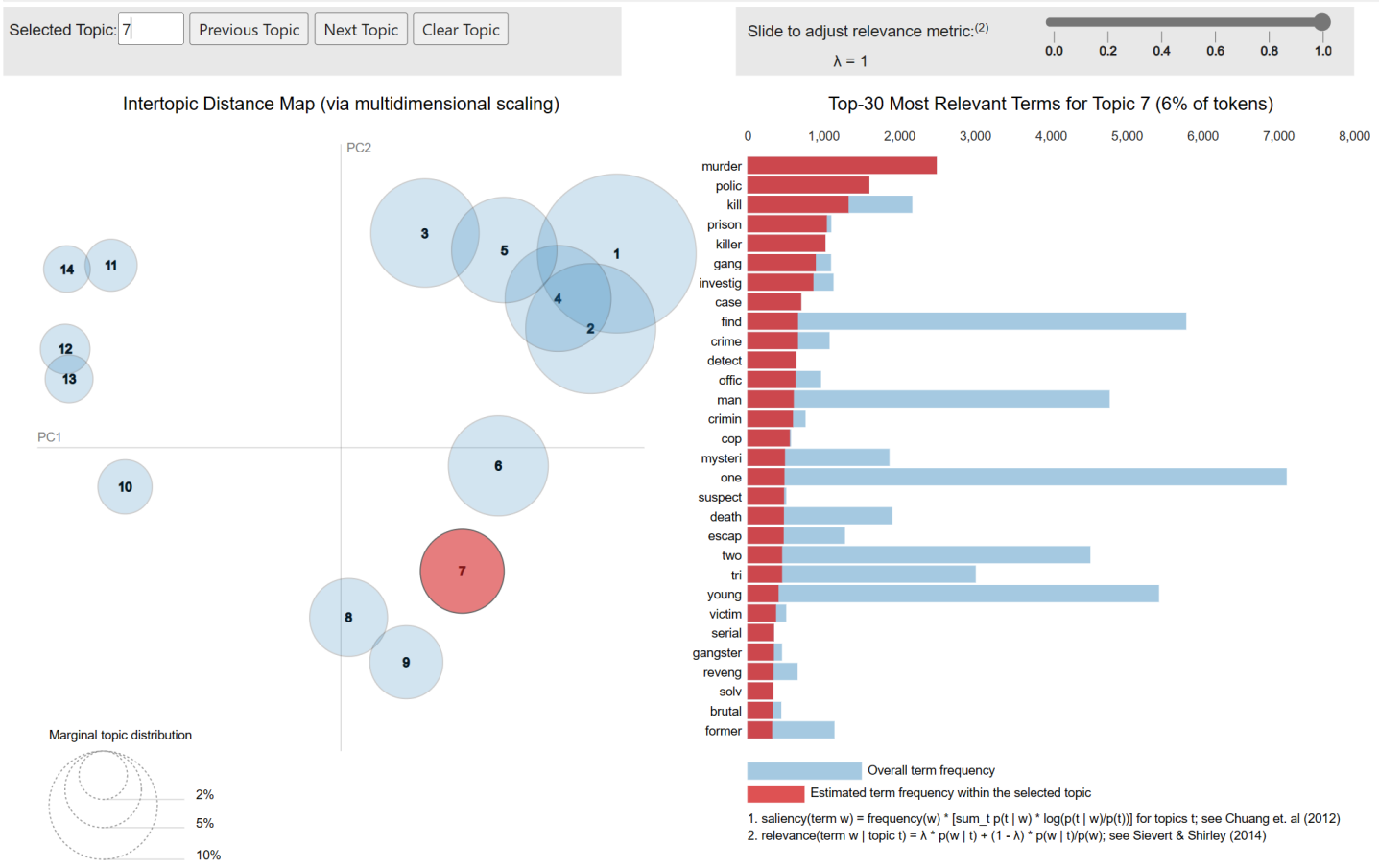


Figure 2] A PyLDAvis visualization with the number of topics set as 14. The topic 7 was identified with top relevant terms such as "murder," "police," "kill," "prison," "killer," and so on. Based on this, we can reasonably assume that topic 7 is mostly associated with themes such as "Crime," "Thriller," and "Action."

Modeling & Evaluation

In this project, I strategically examined two datasets in parallel. Throughout the development process, I focused on refining the data cleaning and text preprocessing steps to enhance the quality of the textual data. I underwent several iterations of data cleaning and text preprocessing in order to improve the accuracy and effectiveness of the recommendation system. Experimenting with various orders of preprocessing steps, such as removing people names and lowercasing, helped me identify the optimal sequence for achieving the desired results. I also added a new line of code that expands contractions to remove words such as "nt" that may have been missed by other preprocessing steps.

In terms of modeling, I explored several clustering algorithms, including k-means clustering and DBSCAN, and made significant efforts to optimize the clustering process. While my experiment did not yield the desired results, I was thorough in my efforts to identify the optimal number of clusters based on the

density of the data points. This included exploring various clustering algorithms and tuning their respective parameters.

To evaluate the performance of my recommendation system, I utilized metrics, such as silhouette scores, to assess the quality of the clustering results. While not all experiments were successful, this iteration process motivates me to further develop my movie recommendation system in the future.

One of the refinement steps I took was to combine the embeddings-based method with a keyword-based method to improve the model. One of the functions I defined extracts relevant keywords from the user input and calculate the Jaccard similarity between the keywords of the user input and those of the movie descriptions in the dataset. With the process of adjusting the weights for cosine similarity and Jaccard similarity, the model becomes more reliable.

While my movie recommendation system represents a promising approach to personalized movie recommendations, there is still room for improvement. Although I have prioritized thoroughness and optimization in creating the model, it is not perfect and could benefit from additional fine-tuning and complementary features (e.g., genre information, user feedback) to enhance its accuracy and user satisfaction. Despite these challenges, the model has already provided valuable insights into the patterns and themes in the movie dataset, and with continued refinement, it has the potential to offer a more accurate and satisfying viewing experience for users.

Streamlit Application

To demonstrate my movie recommendation system, I created a simple yet functional application using Streamlit. I incorporated a function that fetches poster images from The Movie Database (TMDb) API. TMDb provides API keys for developers upon signing up and submitting project information.

Based on my experience with the app, I recommend users include genre-specific keywords in their input. Although my model doesn't explicitly filter movies by genre, movie descriptions frequently contain genre-related terms, and text embeddings can capture the relationships between these words and their contexts. Consequently, mentioning genre-related keywords can still help users obtain more relevant recommendations.

The model isn't perfect and may occasionally suggest seemingly unrelated movies. However, during this experimental stage, I hope users can appreciate and enjoy even the unexpected and surprising movie recommendations!

ReelWhisperer

Tell me about the movie plot you're in the mood for!

A thrilling science fiction movie with a futuristic setting, advanced technology, and maybe some space travel or exploration.

Discover Movies

Top 5 Movie Recommendations:

Gattaca - Science fiction drama about a future society in the era of indefinite eugenics where humans are set on a life course depending on their DNA. The young Vincent Freeman is born with a condition that would prevent him from space travel, yet he is determined to infiltrate the GATTACA space program.

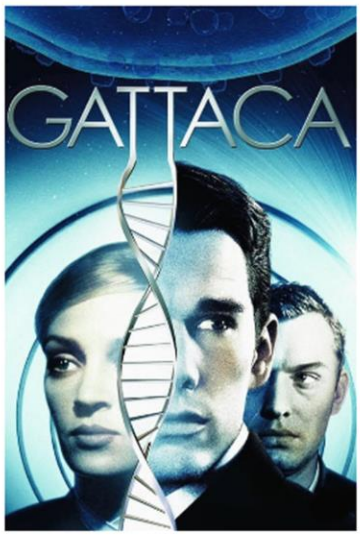


Figure 3] My Streamlit app recommends a classic Sci-Fi movie, ***Gattaca*** (1997) as the top choice to the user input of "a thrilling science fiction movie with a futuristic setting, advanced technology, and maybe some space travel or exploration".