Title: IMDb 2024 Movie Recommendation System using NLP

Submitted by:

Prasanth S – 2022506018

Hemanth K S – 2022506043

Divakar S - 2022506036

Course: NATURAL LANGUAGE PROCESSING (NLP)

Institution: Madras Institute of Technology , Anna University

Date: 21/5/2025

Abstract:

This project aims to build a content-based movie recommendation system using Natural Language Processing (NLP) techniques. By scraping IMDb's 2024 movie listings and processing their storylines, the system provides movie suggestions based on textual similarity. It leverages TF-IDF vectorization and Cosine Similarity to recommend movies either by selecting an existing title or entering a custom storyline. The solution is deployed via a user-friendly Streamlit web application.

Introduction:

IMDb (Internet Movie Database) is a widely-used platform that provides comprehensive information about movies, TV shows, and celebrities. With hundreds of new movies released annually, users often struggle to find movies tailored to their preferences. In 2024, the volume of movie data continues to grow, making intelligent recommendation systems highly valuable. This project introduces an NLP-driven recommendation engine focused on IMDb 2024 movies to enhance user discovery experiences.

Data Collection:

To gather movie data, Selenium was used to scrape IMDb's 2024 feature film listings. The key attributes collected were:

- Movie Name
- Storyline

IMDb Source URL:

https://www.imdb.com/search/title/?year=2024&title_type=feature

The script navigates the webpage, extracts titles and story summaries, and saves them into a structured CSV format (extracted_movies.csv).

Data Preprocessing:

Before applying machine learning techniques, the text data underwent several preprocessing steps:

- Lowercasing: Converted all text to lowercase for uniformity.
- Tokenization: Split sentences into individual words.
- Stopword Removal: Removed common, non-informative words (e.g., "the", "is").
- Cleaning: Removed punctuation and non-alphanumeric characters.

This preprocessing ensures meaningful comparison between movie storylines.

Methodology:

The recommendation logic uses two key NLP techniques:

1. TF-IDF (Term Frequency-Inverse Document Frequency):

Converts textual storylines into numerical vectors based on term importance.

2. Cosine Similarity:

Measures similarity between vectors. A higher cosine similarity indicates closer match between storylines.

By computing cosine similarity between a user input (or selected movie) and all other movie storylines, the top 5 most similar movies are recommended.

Implementation:

Languages & Libraries Used:

- Python 3
- Selenium (for web scraping)
- Pandas (data manipulation)
- NLTK (text preprocessing)
- Scikit-learn (TF-IDF and similarity)
- Streamlit (web interface)

Files:

- scrape_imdb.py – Collects movie data
- nlp_recommend.py – Preprocessing and recommendation logic
- preprocess.py – Streamlit frontend app

Frontend Design:

The front end was created using Streamlit, offering two input modes:

1. Select Movie Title – User selects a movie from the 2024 list.
2. Enter Storyline – User inputs custom movie description.

Upon submission, the app displays the top 5 recommended movies. The interface is clean, responsive, and designed for quick interaction.

Testing:

Several test cases were run to verify the recommendation system:

Example 1:

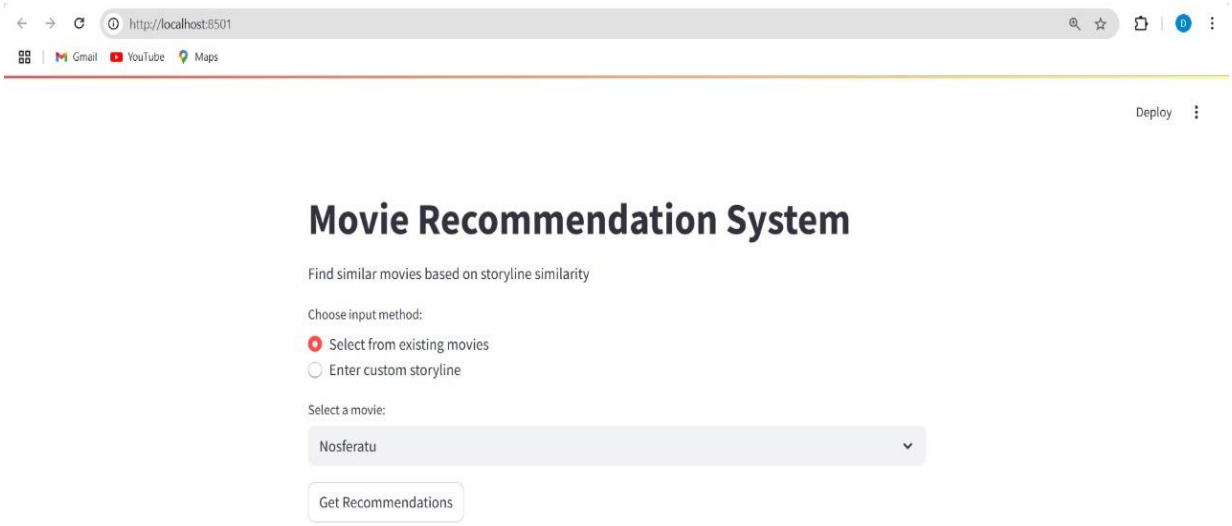
- Input: "Godzilla x Kong: The New Empire"
- Output:

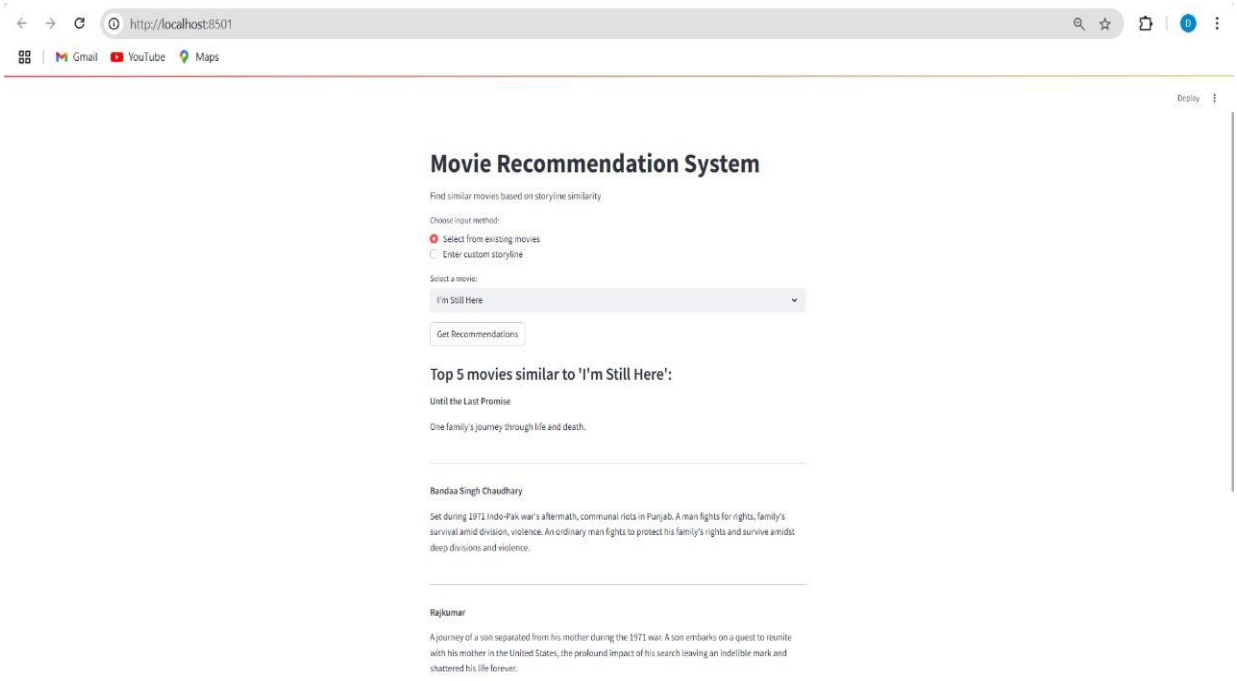
1. Dune: Part Two
2. Rebel Moon
3. Kingdom of the Planet of the Apes
4. Furiosa: A Mad Max Saga
5. The Marvels

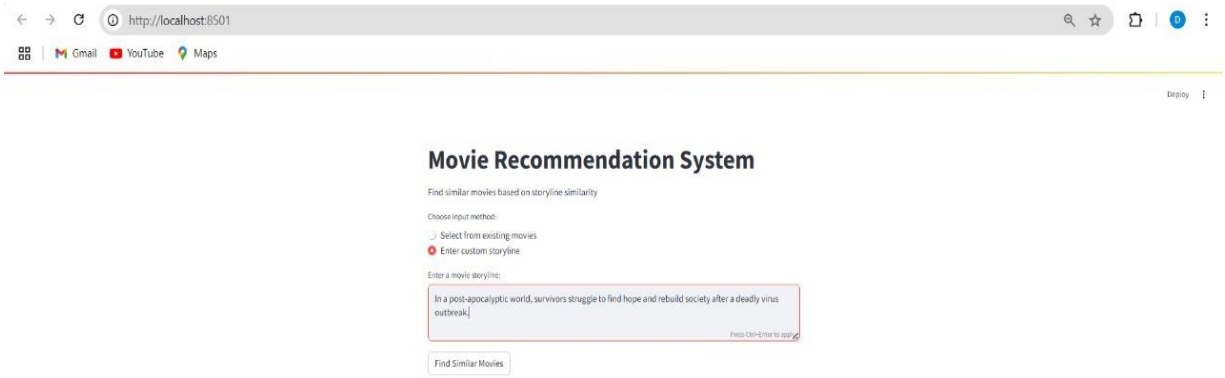
Example 2:

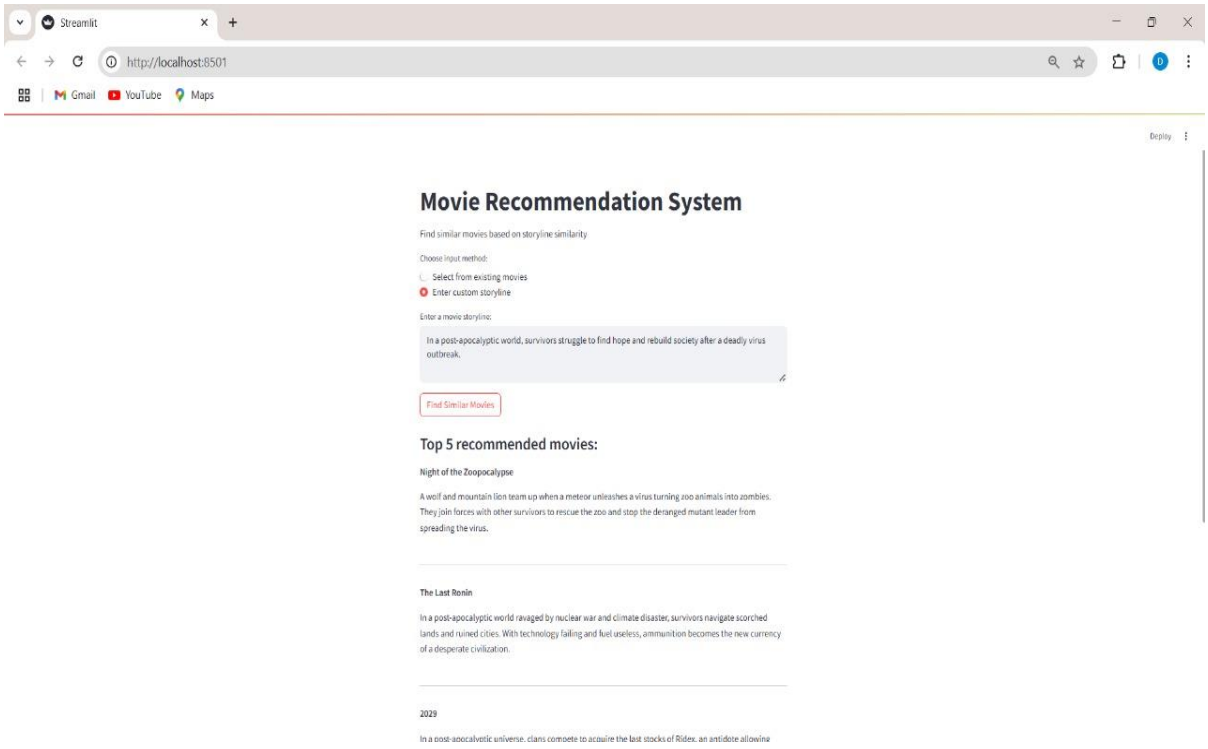
- Input: "A young prince embarks on a journey to reclaim his throne from evil sorcerers."
- Output:

1. Damsel
2. The Tiger's Apprentice
3. Kraven the Hunter
4. Rebel Moon
5. Aquaman and the Lost Kingdom


http://localhost:8501
Gmail
YouTube
Maps
Deploy
Movie Recommendation System
Find similar movies based on storyline similarity
Choose input method:
Select from existing movies
Enter custom storyline
Select a movie:
Nosferatu
Get Recommendations


http://localhost:8501
Gmail
YouTube
Maps
Deploy
Movie Recommendation System
Find similar movies based on storyline similarity
Choose input method:
Select from existing movies
Enter custom storyline
Select a movie:
I'm Still Here
Get Recommendations
Top 5 movies similar to 'I'm Still Here':
Until the Last Promise
One family's journey through life and death.
Bandaa Singh Chaudhary
Set during 1971 Indo-Pak war's aftermath, communal riots in Punjab. A man fights for rights, family's survival amid division, violence. An ordinary man fights to protect his family's rights and survive amidst deep divisions and violence.
Rajkumar
A journey of a son separated from his mother during the 1971 war. A son embarks on a quest to reunite with his mother in the United States, the profound impact of his search leaving an indelible mark and shattered his life forever.


http://localhost:8501
Gmail
YouTube
Maps
Deploy
Movie Recommendation System
Find similar movies based on storyline similarity
Choose input method:
Select from existing movies
Enter custom storyline
Enter a movie storyline:
In a post-apocalyptic world, survivors struggle to find hope and rebuild society after a deadly virus outbreak.
Press Ctrl+Enter to apply
Find Similar Movies


Streamlit
http://localhost:8501
Gmail
YouTube
Maps
Deploy
Movie Recommendation System
Find similar movies based on storyline similarity
Choose input method:
Select from existing movies
Enter custom storyline
Enter a movie storyline:
In a post-apocalyptic world, survivors struggle to find hope and rebuild society after a deadly virus outbreak.
Find Similar Movies
Top 5 recommended movies:
Night of the Zoopocalypse
A wolf and mountain lion team up when a meteor unleashes a virus turning zoo animals into zombies. They join forces with other survivors to rescue the zoo and stop the deranged mutant leader from spreading the virus.
The Last Ronin
In a post-apocalyptic world ravaged by nuclear war and climate disaster, survivors navigate scorched lands and ruined cities. With technology failing and fuel useless, ammunition becomes the new currency of a desperate civilization.
2029
In a post-apocalyptic universe, clans compete to acquire the last stocks of Ridex, an antidote allowing

Conclusion:

This project successfully demonstrates a practical use of NLP for content-based recommendations. The system provides meaningful suggestions for 2024 IMDb movies using TF-IDF and cosine similarity. It can be extended to include genres, actors, or even collaborative filtering in the future. Furthermore, the Streamlit app ensures ease of use and instant accessibility.

References:

1. IMDb Website: https://www.imdb.com
2. Scikit-learn Documentation: https://scikit-learn.org/
3. NLTK Documentation: https://www.nltk.org
4. Streamlit Docs: https://docs.streamlit.io
5. Selenium Docs: https://www.selenium.dev/documentation/