

Project Report: IMDb 2024 Movie Recommendation System

1. Introduction

The IMDb 2024 Movie Recommendation System is a content-based movie recommender that uses NLP techniques to analyze and compare movie storylines. The system scrapes IMDb for movie data, processes the storyline text, and recommends similar movies using TF-IDF and Cosine Similarity. A Streamlit-based web application provides an interactive interface for users to input a storyline and receive recommendations.

2. Project Deliverables

2.1 CSV File

- A structured dataset containing movie names and their corresponding storylines scraped from IMDb 2024.

	Movie Name	Rating	Storyline	Duration
0	1. Deadpool & Wolverine	7.6	Deadpool is offered a place in the Marvel Cine...	2h 8m
1	2. Sonic the Hedgehog 3	7.0	Sonic, Knuckles, and Tails reunite against a p...	1h 50m
2	3. Kraven the Hunter	5.4	Kraven's complex relationship with his ruthles...	2h 7m
3	4. Moana 2	6.8	After receiving an unexpected call from her wa...	1h 40m
4	5. Carry-On	6.5	A mysterious traveler blackmails a young TSA a...	1h 59m

Fig:Sample data from imdb\_movies.csv

2.2 Python Scripts

- **Scraping Script:** A Selenium-based script to extract movie names and storylines from IMDb.
- **NLP and Recommendation Script:** A script for processing the storyline text, vectorizing it using TF-IDF, and calculating similarity scores using Cosine Similarity.
- **Streamlit App Script:** A script to build an interactive UI where users can enter a storyline and get movie recommendations.

2.3 Streamlit Application

- A live web application that allows users to input a movie storyline and receive a list of similar movies.

# Movie Recommendation System

Enter a storyline, and we'll recommend movies with similar plots!


Enter a movie storyline:

After being captured by terrorists in Afghanistan, billionaire weapons manufacturer Tony Stark builds a high-tech armored suit to escape. Upon returning home, he refines the suit, becoming the superhero Iron Man. As he uncovers a conspiracy within his own company, led by Obadiah Stane, he must use his suit to protect the world and take responsibility for his inventions.

Find Similar Movies

## Top 5 Similar Movies:

 37. Ghostbusters: Frozen Empire

 When the discovery of an ancient artifact unleashes an evil force, Ghostbusters new and old must join forces to protect their home and save the world from a second ice age.

## 3. Approach Breakdown

### 3.1 Data Scraping

- Used **Selenium** to navigate IMDb's website and extract movie names and storylines.
- Processed the extracted data and stored it in a CSV file for further analysis.

### 3.2 Data Preprocessing

- Cleaned and tokenized the storylines using **Natural Language Processing (NLP)** techniques:
  - Removed punctuation and special characters.
  - Tokenized words and removed stop words.
  - Used **TF-IDF Vectorizer** to convert text into a numerical format suitable for similarity analysis.

### 3.3 Cosine Similarity Calculation

- Used **Cosine Similarity** to measure the textual similarity between movie storylines.
- Computed similarity scores and ranked movies based on their relevance to the input storyline.

### 3.4 Streamlit Interface

- Developed an interactive **Streamlit** application that:
  - Accepts a user-provided storyline.
  - Computes similarity scores using the preprocessed dataset.
  - Displays the **top 5 recommended movies** based on Cosine Similarity.
  - Provides a user-friendly UI for easy interaction.

## 4. Implementation Details

### 4.1 Technologies Used

- **Python Libraries:** Selenium, Pandas, BeautifulSoup, Scikit-learn, NLTK, Streamlit
- **Web Scraping Tool:** Selenium for dynamic content extraction
- **NLP Techniques:** Tokenization, Stopword Removal, TF-IDF Vectorization
- **Similarity Calculation:** Cosine Similarity from Scikit-learn
- **Front-end Interface:** Streamlit for an interactive user experience

### 4.2 Code Workflow

1. **Scraping Script:**
  - Navigate IMDb 2024 movies page using Selenium.
  - Extract movie names and storylines.
  - Store data in a structured CSV file.
2. **NLP and Recommendation Script:**
  - Read the CSV file and preprocess the text.
  - Convert storylines into a numerical format using TF-IDF.
  - Compute Cosine Similarity scores.
  - Generate recommendations based on similarity rankings.
3. **Streamlit App:**
  - Load preprocessed data and similarity model.
  - Accept user storyline input.
  - Display top 5 recommended movies based on input.

## 5. Results and Insights

- Successfully extracted and processed IMDb 2024 movie data.
- Built a functional recommendation system using text-based similarity.
- Streamlit app provided an intuitive user experience for exploring similar movies.
- The TF-IDF + Cosine Similarity approach worked effectively for content-based filtering.