Project Summary - IMDb Crawler and Data Visualization

Team Name: “We Tried”   Team Member: Yaqing Peng, Muyun Ji, Ruilin Tao, Shuang Deng

1. Project Description: This project aims to collect and visualize data from IMDb, a popular online film database. Our objectives are to crawl reviews of *The Shawshank Redemption* and store that data in a CSV file, of which the code can be adapted to crawl reviews for any film on this website. Furthermore, through data analysis and visualization, our analysis dimensions and visualization methods can be applied to any other films. The project is completed by a leaderless team of four members who share a similar amount of workload.

2. User Stories: As IMDb reviewers, we aimed to understand audience preferences through movie reviews. Using the classic film "*The Shawshank Redemption*" as an example, we crawled existing reviews from IMDb, and then visualized and analyzed the data. Our data analysis allowed us to uncover:

a. The most important dimensions of the movie according to critics.
b. The high-frequency words used to rate the movie positively and negatively.
c. The distribution and trend of ratings over time.
d. Whether reviewers prefer long or short reviews, and their receptiveness to spoilers.

3. Development Circle

Stage	Task	Date Range
a. Analysis	Define project requirements (Discuss and determine the IMDb Spider project needs)	Feb 26 - Mar 1
b. Design	Design data collection methods (Design web scraping methods for movie review data)	Mar 2 - Mar 5
c. Developmen	Implement web scraping (Develop IMDb Spider for data collection)	Mar 6 - Mar 20
d. Testing	Test web scraping and data analysis (Test IMDb Spider and analyze collected data)	Mar 21 - Mar 27
e. Deployment	Visualize and present data (Create data visualizations and prepare presentations)	Mar 28 - Apr 7
f. Evaluation	Review and submit project (Review project, make necessary adjustments, and submit)	Apr 8 - Apr 18

4. Challenges and Solutions

a. Data Scraping: We collaborated to learn and develop an effective web scraping process for IMDb web pages from scratch. The team faced several challenges, such as incomplete reviews or missing page numbers, which required innovative solutions like automatically clicking loading buttons and expanding folded reviews to ensure all data was collected. Ultimately, we successfully crawled **1**0758 out of 10777 reviews, including titles, contents, spoiler warnings, ratings, dates, and votes.

b. Data Visualization: We tackled various challenges throughout the project, such as generating word clouds, handling large volumes of data, and filtering relevant information. To overcome these obstacles, we sought out an online course on web scraping to enhance our skills and learned to implement features like assigning different colors to word frequencies. Meanwhile, we addressed issues by optimizing data processing with libraries like pandas and NumPy, and applying text processing techniques. By overcoming these challenges together, we successfully analyzed movie review data, generated word clouds, and provided valuable insights, all while honing our programming skills and capacity for independent learning.

5. Conclusion and future optimization: Through data analysis and visualization, we were able to uncover deeper insights hidden within the movie reviews:

a. Actors are most discussed in the reviews, followed by writers and the director, while the production company is rarely mentioned.
b. Reviews frequently use character-related, plot-related, and praise-focused words. High-rating reviews. often use praise-focused words i.e. "Best" and "Great" etc., while low-rating reviews feature negative terms i.e. "Overrated", "Boring", etc.
c. Ratings are polarized, dominated by either near-perfect or low scores, which has remained consistent over time. A significant increase in the number of reviews can be observed between 2013 and 2019.
d. People favor writing shorter reviews less than 200 words. But data suggests they may slightly enjoy reading non-spoiler-warning reviews more.

The project also has great potential for expansion. For instance, we could crawl and analyze more movie reviews, leveraging machine learning techniques to predict the ratings for upcoming films. Additionally, we could utilize machine learning analysis to provide more personalized movie recommendations to users, eventually improving user satisfaction.

6. Technical Resource:

a. Library: Data Processing (pandas, os, csv, nltk)**,** Data Visualization (matplotlib, seaborn)**,** Web Scraping (Selenium, Webdriver, time)**,** Word Cloud Generation (Wordcloud)
b. Learning Materials & Technical Supportive: YouTube, Udemy, Bilibili, Stack Overflow, CSDN, GitHub.
c. Expertise: Prof. Ram, TA Zhengrui Lu.