



Producer

James Turk



Director JP Martinez



Project Abstract

Our project focuses on the analysis of Chinese and American movie reviews. We extracted information such as movie names, ratings, release years, and reviews of war movies from China's Douban and the United States' Rotten Tomatoes websites. Our analysis covers four aspects to examine the differing reactions to war movies between China and the U.S.

Firstly, we clean the data to remove characters that could interfere with the analysis. Then, we conduct sentiment analysis on the reviews and compare the sentiment scores with the actual ratings on the websites to assess the authenticity and effectiveness of the ratings. Next, we analyze the reviews of Chinese and American movies using word cloud visuals. Furthermore, we divide the timeline into three periods to compare the ratings of movies from different times between China and the U.S. Lastly, we perform LDA (Latent Dirichlet Allocation) topic analysis on the reviews and calculate the number of topic words for different websites using the average cosine rate.

Given the linguistic differences between Chinese and English, we processed the data separately but aimed to present our findings in a directly comparable manner, highlighting the nuanced perceptions of war movies across cultures.



Name of group members and CNetIDs

Fuyuki Tani—ftani

Qihui Zhang—qihuiz

Zirong Li—liz3

Zhen Wang—zhenzhenwang



Film Crew Positions

Code Responsibilities for Each Member

- Data_Scraper — Qihui Zhang & Zhen Wang
- Data_Clean — Zirong Li & Fuyuki Tani & Qihui Zhang & Zhen Wang
- Word Cloud — Zirong Li
- Graph_Analysis — Zhen Wang & Zirong Li
- Analysis_LDA — Qihui Zhang & Zhen Wang
- Sentiment_Analysis — Zirong Li & Qihui Zhang
- Dashboard — Fuyuki Tani
- Paper and writing process - Zhen Wang & Qihui Zhang & Fuyuki Tani



Movie Guide

Application Interaction and Produces

Clone the Repository

```
git clone https://github.com/ZhenWwvv/30122-Film-Insight.git
```

Navigate to the Repository.

```
cd [your path]
```

Install Dependencies

```
poetry install
```

Activate the virtual environment.

```
poetry shell
```

Launch the App

```
python3 -m film_insight
```

Options of the App.

- (1) Launch Dashboard
- (2) Scrape Data
- (3) Clean Data
- (4) Plot from Data
- (5) Quit



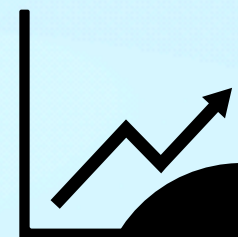
Story Structure

Structure and Connection



data_scraper

Contains the functions needed to scrape the movies' data from the website Douban and rottentomatoes to get the scores and reviews.



data_process

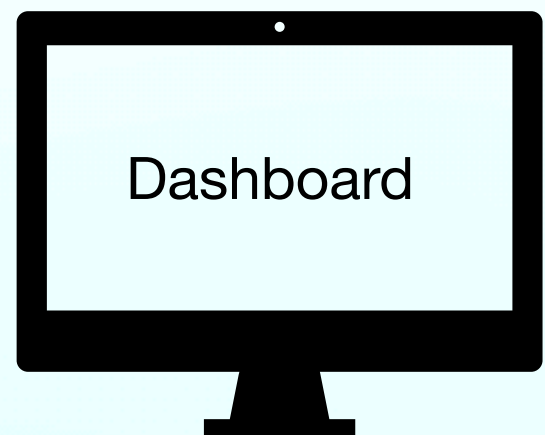
Contains all the data analysis process: -data_clean helps clean the raw data we got from the 2 websites.

-graph_analysis reflects the number of reviewers with different scores in different periods and the average rating distribution within the timeline, in order to observe people's evaluation tendencies and preferences for war movies in different periods.

-wordcloud_analysis illustrates the key elements that audiences pay attention to in war movies, highlighting recurring critical words as core elements. It reflects the interest and evaluation tendencies of audiences in different countries towards different angles of the film, different historical events and performance quality.

-sentiment_analysis uses jieba for word segmentation and CountVectorizer to convert text into vectors. Subsequently, a multinomial naive Bayes model is trained for sentiment classification. Assess the model's performance by generating a confusion matrix, which is then visualized as a heatmap to clearly show the accuracy and misclassification of the classification results.

-LDA_analysis reads the preprocessed text data and creates a dictionary using corpora.Dictionary. Then use doc2bow to create a vector of word occurrences. Gensim's LdaModel tests topic numbers from 2 to 10 to find the optimal count. Differences between subjects were measured using cosine similarity, with lower values indicating more pronounced differences. A number of different topics are selected accordingly.



Dashboard

This module is responsible for data visualization.

-dash_main.py creates a dashboard using Dash package. In addition to the graphs created by data_process, the dashboard also displays the list of movies and comments we analyzed.

-dash_prep.py makes various preparations for the dashboard. Specifically, it scrapes the web for URLs, copies image files, and performs other tasks that complement the analysis part.



Oscar Acceptance Speech

Goals to Accomplish and actuality accomplished

Our project was initially affected by several wars that broke out in 2022 and 2023. We tried to understand the discussions and emotions of people in different countries on war-related topics. We comprehensively considered the interest, feasibility, audience, etc., Among the many plans, we finally decided to analyze war-themed movies to deeply explore the differences in preferences, evaluation content, emotional tendencies, and final recognition scores between Chinese and people from other countries regarding war movies. Analysis and cause mining, and the results of the analysis and the estimation of its underlying causes will be our final conclusion.

We plan to use a web crawler to obtain data, and then conduct word cloud analysis, chart analysis, LDA analysis, etc. to find out the correlations and differences, and conduct further cause analysis.

We basically followed our original plan during the actual process, and added some mappings and regression analysis based on the original plan, added Bayesian analysis and designed the dashboard and website. Finally, we passed this application The content in our plan was more than 100% perfectly presented!