What ‘Plot’ Makes a Movie a Great One?

Teammates: Mengyue Huang, Mallika Pallipuram

Updated Motivation and Audiences:

Same as our initial proposal, we aim to use Exploratory Data Analysis (EDA) and Natural Language Processing (NLP) to explore the conversation data from movie scripts, but now with a clearer clue on the scope of the dataset we plan to use and the methodologies to implement this. In our project, we assess a movie’s success by its IMDb ratings and Twitter ratings, as the definition of ratings in our movie metadata dataset (MovieTweetings & Kaggle data). We also plan to restrict the year range of our movies to only as early as 2000, and gain a more recent understanding of the public’s tastes in films.

Our audiences may include movie industry professionals such as directors, producers, and scriptwriters to understand what the market, or the public, perceive as a box office movie, which usually also tends to be a great movie. With this project’s text analysis on the script conversations, the professionals will be able to understand what plot (words) are the most frequent for these selling movies, and see how the top words change from year to year.

Besides, in terms of Film Studies, the project contributes by providing insights into what makes a movie great and an exploration on how audience preferences and definitions of meaningful films change over time. In terms of Computing in the Humanities, the project could serve as a computational methods to understand the public’s taste for films across years, providing an aspect into our culture and art in this society.

Description:

We will first explore the basic traits of films such as ratings, year, and genres of each movie in the dataset with their review scores and box office, in order to gain some basic expectation and understanding why it is a great film. After that, we would conduct some Language Processing from the scripts and conversations from the movie dataset and try to gain some insights from their shared similarities or differences on topics and plots, obtaining some comprehension on what makes a great movie in terms of the plot.

Most of the results will be represented with plots and graphs, with interpretation markdowns alongside explaining and discussing any discoveries from them.

Goals and Methods:

The project has two goals, one focused on the metadata of movies and one focused on the scripts.

1. EDA on metadata: we stick to the same plan to study the basic metadata information of each movie such as the box office, year of release, and ratings with functions such as .describe and plots such as barplots, scatterplots, and graphs to explore the dataset and how factors such as ratings change over time. This can help us gain a better insight as to what factors contribute to higher ratings.
2. NLP on scripts: we plan to use Word Frequency Analysis and Text Similarity Analysis to dive deeper into the scripts. The frequency analysis examines the most frequently used words or phrases to identify common elements in successful movie scripts, and the similarity analysis measures the similarity between different movie scripts to find patterns or trends.
3. In order to obtain a whole dataset containing both the basic traits and its scripts of selected movies, potential dataset merging based on indexes such as title and movie code may be needed.

Citations and References:

Franklin, S. J. (2021, December 13). Sentiment Analysis of Movie Reviews in NLTK Python - S Joel Franklin - Medium. *Medium*. Retrieved March 19, 2024, from https://medium.com/@joel\_34096/sentiment-analysis-of-movie-reviews-in-nltk-python-4af4b76a6f3

Explanation: an exploration on an NLP library nltk to analyze text, supplementary materials for potential text manipulation while doing scripts exploration.

Lavin, M. J. (2019, May 13). *Analyzing Documents with TF-IDF*. Programming Historian. Retrieved March 19, 2024, from https://programminghistorian.org/en/lessons/analyzing-documents-with-tfidf

Explanation: tutorial on TF-IDF, a method to identify the terms frequent in document but rare in the whole corpus.

Pabalan, C. (2024, January 21). Evaluating Cinematic Dialogue — Which Syntactic and Semantic Features Are Predictive of Genre? *Medium*. Retrieved March 19, 2024, from https://towardsdatascience.com/evaluating-cinematic-dialogue-which-syntactic-and-semantic-features-are-predictive-of-genre-2c69a71af6e2

Explanation: an exploration on an NLP library spaCy to analyze text, supplementary materials for potential text manipulation while doing scripts exploration.

Pal, S. (2022, February 11). What is Text Similarity and How to Implement it?  | MLSAKIIT. *Medium*. Retrieved March 19, 2024, from <https://medium.com/msackiit/what-is-text-similarity-and-how-to-implement-it-c74c8b641883>

Explanation: a discussion on how to measure document similarity with a brief introduction on TF-IDF.

Stratis, K. (2023, December 1). *Use sentiment analysis with Python to classify movie reviews*. Real Python. Retrieved March 19, 2024, from <https://realpython.com/sentiment-analysis-python/>

Explanation: tutorial on how to preprocess text data such as stop word removal and normalization with spaCy.

Dataset Bibliography

We looked at four datasets on movie data and primarily searched for datasets containing information about movie metadata, box office rating, the release year, and the movie script. We looked at the Humanities Data website as well as Kaggle to find these datasets.

https://humanitiesdata.com/resources/160

This dataset was interesting to look at because it compiled data on movie ratings from Twitter. This dataset aims to provide more relevant data on how movies are truly rated, as public movie rating sites tend to be outdated. The earliest data is from February of 2013 and was last updated in 2021. This dataset contains each movie’s title, year of release, and genre in addition to the rating from 0-10, user id, and rating timestamp. Although this dataset is unconventional, looking at it in addition to other sources can provide a better understanding of how people truly feel about certain movies. I believe it can be really interesting to look at social media data, as people tend to be more transparent about their opinions.

https://www.kaggle.com/code/ruchi798/movies-data-collection-eda-using-tableau

The dataset is a combination of scraped data from a comprehensive list of movies available on various streaming platforms as well as the IMDb dataset. The dataset contains information on the movie’s title, year of production, viewing audience age, IMDb rating, genre, country of production, languages, runtime, streaming platforms it is available on, and Rotten Tomatoes % rating. The methodology involved in creating this dataset was data scraping from multiple sources using BeautifulSoup4 and IMDbPy, which is a Python package for retrieving and managing data of the IMDb movie database. This dataset has interesting details about each movie such as audience age and streaming platforms it is available on, which provides a more comprehensive look at what influences the ratings.

https://www.cs.cornell.edu/~cristian/Cornell\_Movie-Dialogs\_Corpus.html

This dataset was created by researchers at Cornell University’s Information Science Department. It contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts. It has over 220,000 conversations between fictional characters, which vary in length and complexity. This dataset contains information on movie metadata such as genre, year of release, IMDb rating and number of votes, and also contains information on character metadata such as gender and place in credits. The dataset contains detailed movie metadata such as character and utterance metadata, which can show us how these specific factors present impact movie ratings. Looking into character metadata in specific conversations will also show us how conversations and better developed characters play into how people assess movies.

Instructor Feedback

I’m glad to see that you’ve started making some progress on this project, and that you’ve identified some relevant datasets and readings, as well as narrowing your focus temporally – all of which seems appropriate. However, I’m a bit concerned about the unclear progress regarding actually working with the data here or in thinking thoughtfully about you’re actually going to define these categories.

As I stated in your proposal, how data about movies is collected is not always obvious or consistent, and furthermore what makes a movie “great” is very subjective. Currently, you seem to be implicitly making an argument for conflating greatness with gross box office earnings, but you provide no rationale, nor do you cite any scholarship to corroborate this assumption. Indeed, most film scholars would categorize great films by awards, not box office earnings, so for your project to be successful, you will need to explain how and why you are defining movies as “great.” You might even consider thinking about the idea of audience engagement as creating a different conception of *greatness* that is more akin to going “viral” or reception of movies, which may or may not correspond with box office success or artistic merit.

In terms of working with the data, you provide some good details on what’s in each dataset, but don’t include any information about missingness or how well these datasets can be combined. It is imperative that you assess whether you will have to include only a small subset of this data because of limited overlap or if there is a way to combine them without losing data. For example, if you plan on using movie title to join these datasets, how will you resolve duplicated titles that refer to different movies? Ideally, you would have some shared metadata to join these files and perhaps there is some imdb related metadata across all the datasets, but even then, those ids can change, so it’s worth working on this connecting of these datasets sooner than later. Hopefully you have already started this process, but it remains a bit unclear in this update. I would also recommend trying to find at least an article or two by a scholar working with data to study films. You might try searching for “digital humanities” or “computational social science” along with cinema or film studies to find relevant materials. Hopefully finding more relevant research will help you contextualize your findings and get at some of the issues I’ve raised over thinking critically about things like audience or greatness.

Finally, it is still unclear how you plan to divide this work, which is important element of working collaboratively. I would highly recommend making a more concrete outline of how you plan to tackle the remainder of the work for this project, and be sure to divide the work in a way that enables each of you to contribute to the project. I do think your planned use of computational methods is promising, though again wondering how natural language processing methods of scripts will relate to your social media data, or how using dialogue data differs from using the full script. I hope these comments help you finalize your next steps and looking forward to seeing your final analysis!

Grade: B+