Letterboxd Sentimental Analysis

Data Sheet

Motivation

I have always considered myself a cinephile. Normally on the weekend, I will pick an intriguing film to enjoy with my friends as a weekly ritual. Although my friend and I all have similar film taste, we still spend an incredible amount of time in choosing the film that we ended up watching.

There are various ways to measure if a selected film is worth our time or not. For us, we enjoy referencing film reviews by both critics and normal audiences from difference sources.

No matter how popular and well-received a film is, it is certain that there will be voice of objections at any given point in time. The mix of perspectives about films really fascinates me. That is why I decided to leverage knowledge that I learned from the course to conduct an in-depth sentimental analysis on film reviews on *Letterboxd*.

Composition

Corpus Link: https://github.com/zejiachen9912/letterboxd-sentimental-analysis/tree/main/finalProjectDataset

Since Nolan is such a controversial director while his films range from sci-fi to suspense, I thought that it will be interesting to conduct the analysis based on his works.

As a result, my current dataset is a corpus that contains reviews of Christopher Nolan' films that I scrapped from Letterboxd. All the documents are now in `.csv` format, but I will adjust the file format accordingly depending on the future testing and modeling.

Films that I selected from Nolan's career are as followed:

Film Title	**Year**	**Number of Reviews**
Inception	2010	331
The Dark Knight	2008	333
Interstellar	2014	282
Dunkirk	2017	298
Tenet	2020	245
Memento	2000	351
The Prestige	2006	373
Insomnia	2002	333
Following	1998	316

Because the project's main object is to access the sentiment toward these films, there is only one column of data, film review (String), available in all documents.

There are on average 313 reviews in each csv file, and 2862 review across the entire corpus. (Do you think this is a good size to start with? I can certainly scrap more if needed)

There are other attributes (film ratings, reviewer) on the website that might be useful to my analysis. I will consider adding the information if needed.

Collection Process

Scrapper Link: https://github.com/zejiachen9912/letterboxd-sentimental-analysis/blob/main/Notebook/Scrapper.ipynb

I collected all my data from Letterboxd, a social platform focused on sharing opinions about, and love of, film. Thanks to the clean formatted website URL, I was able to scrap all the reviews I need with `Request` and `BeautifulSoup`. To ensure the quality of reviews, I decided to scrap the review in an order of "Review Activity." (amount of likes, comments received) To store the data locally, I first put the review in a data frame, and subsequently stored the data frame as a local csv file.

Preprocessing/Cleaning/Labeling

The current data is still, what I considered, raw data. Because there is only one column of data entry, there is not too much that I need to do in terms of cleaning or labeling.

Nevertheless, I did preprocess the data while I was scrapping it form *Letterboxd*. After examining some pages of reviews, I have come to an understanding that there are non-English reviews on Letterboxd. As a result, I took this into account while I was collecting the review from the website. I designed a function `isEnglish()` that takes an addition step to check if a review is consisting of English characters, ensuring the corpus is free from reviews that consist with non-English characters or emojis.

Use

The dataset was contrived by me. Therefore, it has not been used for any analysis yet. I am planning on testing the corpus next week to see if it can possibly fit into my analysis.

The current state of the data can be used for conducting sentimental analysis using models like VADER or BERT. I might as well convert documents inside the corpus to text files to fit into the topic modeling analysis that I wish to kickstart later.

Distribution

Distribution wise, I have pushed my entire final project workflow to GitHub for version control and showcasing purposes. The repository contains hitherto everything I did for the final project, and I will continue update it as the project progress.

Maintenance

I will oversee the maintenance of the corpus. Although the corpus is planned be used solely for the final project, if there are any other use cases in the future, I will also consider updating and maintain the integrity of the dataset.