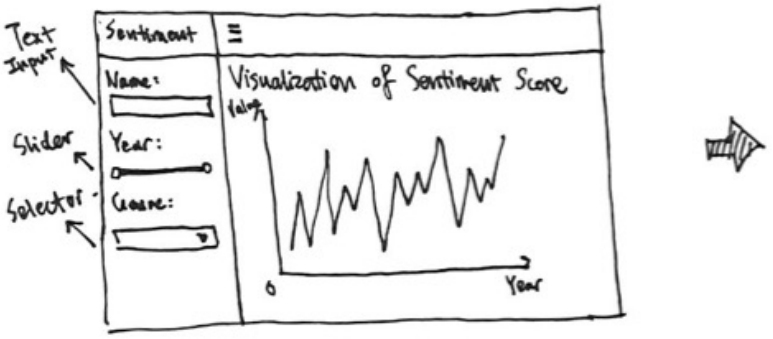
# Movie Scipt Sentiment Analysis

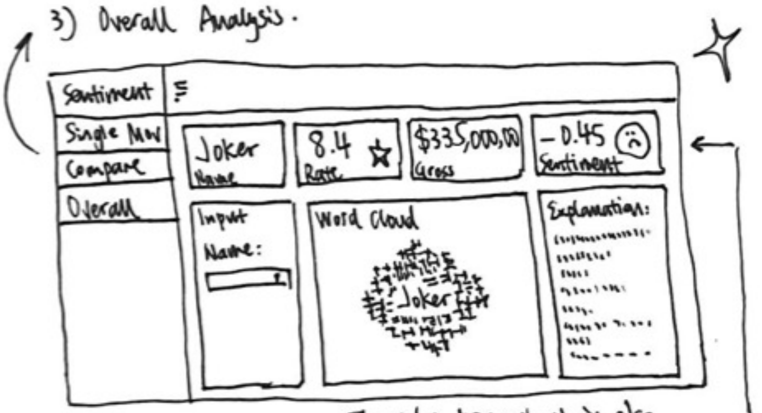
The aim of the project is to analyze the sentiment tone of movie scipts and try to compare the sentiment tone among the best movies on IMDB Top 250 chart. The thought is to design a dashboard with inputs & plot outputs to visualize interesting features within movie scipts.

## I. The Basic Layout of the Dashboard:

① The initial thought is to put the input in the sidebar, so that the plot can be bigger and more analyses can be shown on the body.



② However, after a thorough analysis on the dataset, we decide to put the input on the body of the page, and let the sidebar be a tab to select from different themes: 1) Single movie analysis 2) Comparison 3) Overall Analysis.



The value box output is also included to present the simple facts of the movie (Rating, Sentiment, Gross)...

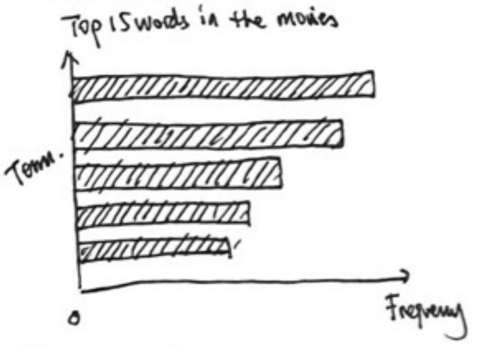
## II. Single Movie Analysis:

### Word Cloud

Word Cloud is a visually satisfying representation of the overall feature of the scipt. Therefore, it is consider to be one method for analysis.

### Word-frequency BarPlot.

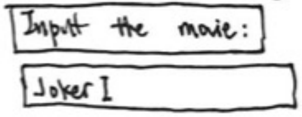
The word-frequency plot shows the top 15 most frequent words in the scipt.



Barplot is ~~chosen~~ chosen as the main plot here because barplot is easy to read and understand. The frequencies are also more clear using the barplot.

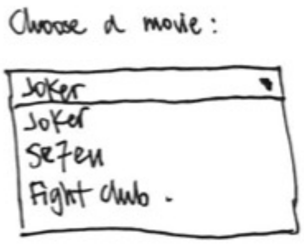
### The Input

The input used to be a textinput where the movies names are input by text.



However, it's cumbersome to type in the movies that have very long movie title. Also, users are not aware of the movies in the dataset. Another issue is that when there is no input in the textInput, there will be an error message in the plotoutput area.

Therefore, we choose to substitute the textinput as a selectorinput where the movie titles are easier to select, and more user-friendly.



## III. Movie Comparison & Sentiment Arc.

One of the central topics that we want to analyze is the change of Sentiment tone within the script of a movie. In other wod, the emotional arc is what we want to create. However, emotional arc is hard to measure, thus we decide to use Sentiment Srore as the value to measure the change In the script tone.

### Sentiment Analysis Method:

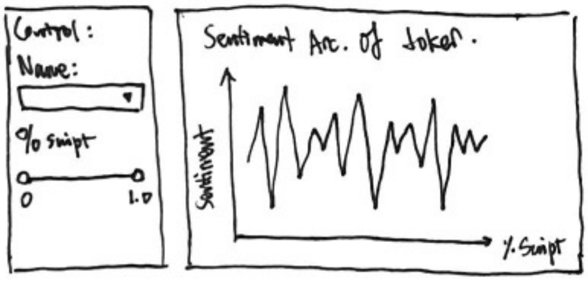
The sentiment Analysis method utilized is the Hu & Liu Sentiment dictionary. The value of sentiment score is calculated by the following formula:

$$\frac{\text{positive.count} - \text{negative.count}}{\text{positive.count} + \text{negative.count}} = \text{Sentiment.Score.}$$

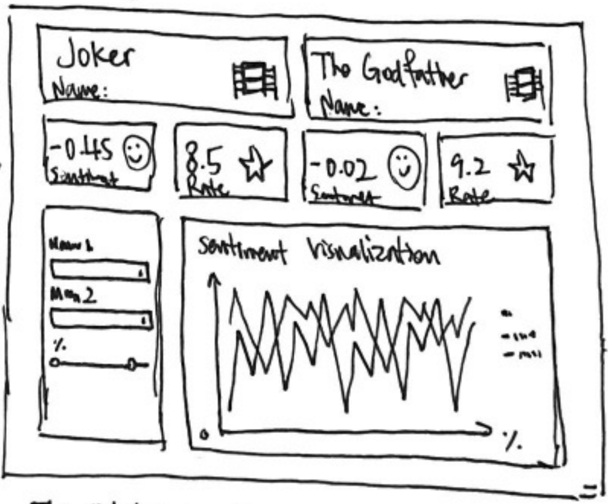
* The idea comes from Nayomi Chibana's interactive visualization of emotional arcs of movie scripts.

### Implementation

The idea is very simple. The input is the movie name and a Sliderinput for the % of scripts included in the plot.



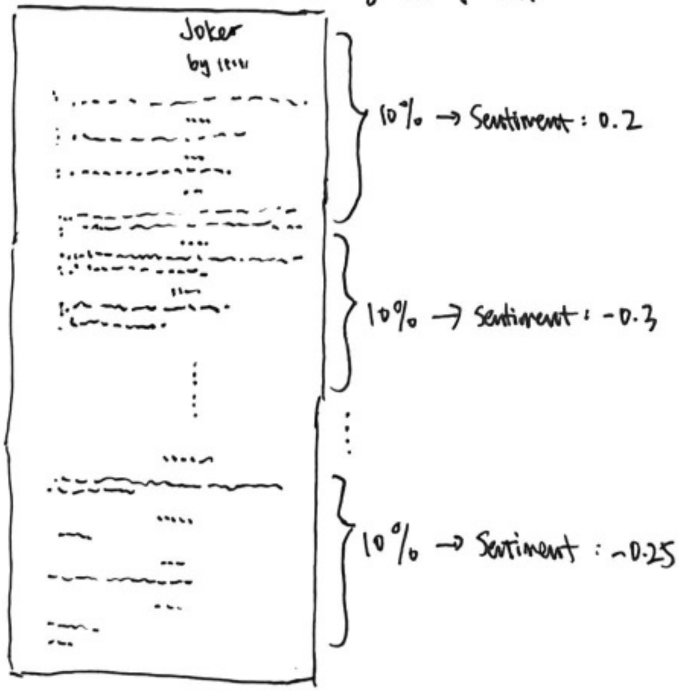
However, the visualization result of one movie is too simple, We decide to add another movie as input to make a comparison on the two movies. With valuebox to indicate the movies sentiment sure and rating.



The valuebox are illustating the rating and overall sentiment score, comparing between two movies.

### The sentiment score measurement:

The movie script is divided by percentage into 20 equal portions. and calculate the sentiment score of each portion.

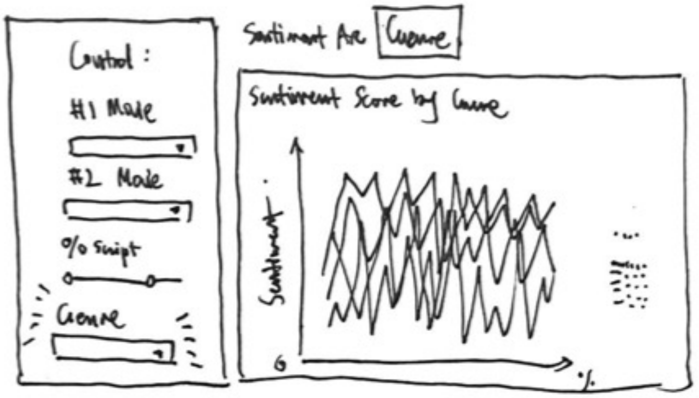


The % of script input can only take certain values: (0, 0.05, 0.1 .... 0.95, 1.0) Since the scripts are divided into 20 parts.

To further anaylze the script, more portions can be created to smoothen the line.

## Genre

A new input genre is later added to the input section. The genre only include the first and the primary genre of each movie, because, otherwise, there will be too many genres and certain genres, such as "Film-Noir" only contains a small number of movies, which may not be available for comparison.



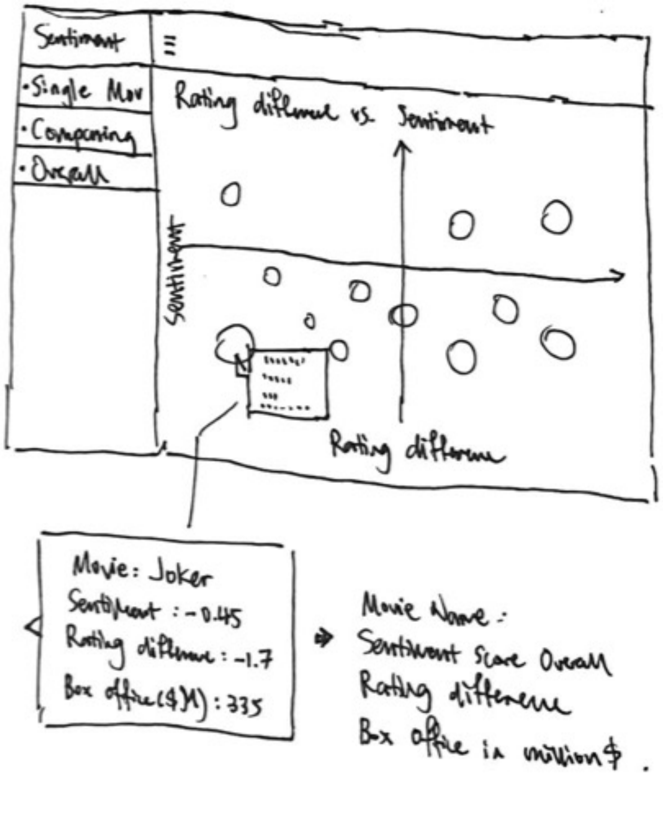
* limitation: Some genres have too many movies, making the plot too messy for analysis.

## IV. Overall Sentiment vs. Rating difference.

The third tab is a plotly interactive graph showing the sentiment score vs. the rating difference. The initial thought is to plot a sentiment score vs. rating, but the dataset only collects movie on the IMDB top 250. The plot might not be accurate or meaningful under this limitation on the data collection process. Therefore, we chose the rating difference between critic & audience as the x-axis and Sentiment Score as y-axis. The rating difference is calculated by ~~[illegible]~~ subtracting ~~IMDB~~ critic rating - ~~critic~~ IMDB rating.

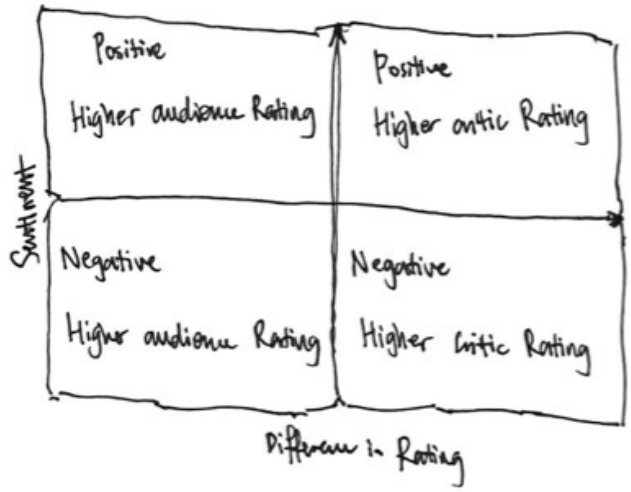
The layout would just be a graph without any control input. The reason is that we only have 46 movie scripts from the imsdb.com and thus 46 sentiment scores to visualize.

The size of the bubble plotly is the gross box office.



⇒ Movie Name:
Sentiment Score Overall
Rating difference
Box office in million$.

Layout of the plot:


Positive
Higher audience Rating
Positive
Higher critic Rating
Negative
Higher audience Rating
Negative
Higher critic Rating
Sentiment
Difference in Rating