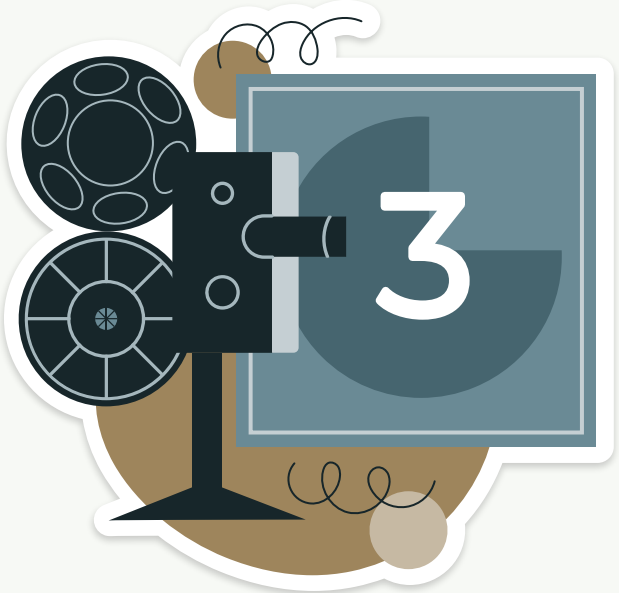
CS105 Project Proposal

Eden Fraczkiewicz, Darren Liang, Annabelle Le, Malina Martinez, Alex Tran

Summary

- Many people rely on critic reviews to determine if they will watch a movie
- Rotten Tomatoes is a company popular for their movie reviews
- For our project we wanted to see if we could analyze the text in critic reviews to see if we could predict whether they rated the movie as
    - Positive (fresh)
    - Negative (rotten)



Data Set

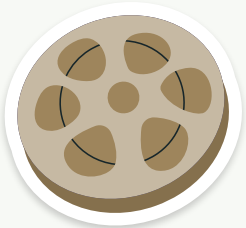
- Dataset: https://drive.google.com/file/d/1N8WCMci_jpDHwCVgSED-B9yts-q9_Bb5/view

- 480,000 critic reviews from rotten tomatoes. There are two columns across multiple movies
- 2 columns: Freshness and Review.
  - The Freshness column contains 1s and 0s, 1 meaning that the critic rated the movie “fresh” and 0 meaning that the critic rated the movie “rotten”.
  - The Review column contains strings of critic reviews from various movies.

	Freshness	Review
0	1	Manakamana doesn't answer any questions, yet ...
1	1	Wilfully offensive and powered by a chest-thu...
2	0	It would be difficult to imagine material mor...
3	0	Despite the gusto its star brings to the role...
4	0	If there was a good idea at the core of this ...

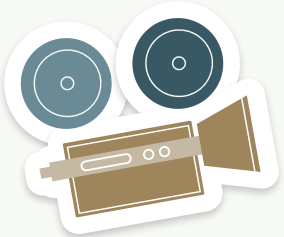
Text Mining

- **Sentiment Analysis**
  - We will be extracting sentiment from movie reviews to determine whether the tone is positive,negative or neutral
- **Bag of Words**
  - The Bag of Words will represent each review as a vector of word frequencies without considering their order. We will analyze and compare frequencies of certain words or punctuation.
- **Term Frequency - Inverse Document Frequency (TF-IDF)**:
  - TF-IDF will be used to find the normalized frequency of words with meaning and intent behind them. Words which do not show across all types of reviews, but which usually will entail either a positive or negative review.



KNN

- **KNN (supervised learning):** To determine the K-value for KNN we will choose different values of K until we find the elbow point of which will be determined to be the best K-value and which will be used on the dataset.
- **K-means (unsupervised learning):** We will test specific features such as length of the movie review, frequency of exclamation marks, and frequency of question marks to see whether the clustering via K-means will corroborate with our hypotheses.



Hypotheses

- We want to find out features that are unique between positive and negative reviews. We think that...
  - Positive reviews will use more exclamations points because they tend to be used with positive emotions
  - Negative reviews will use more question marks, questioning why choices were made or asking rhetorical/judgmental questions
  - Positive reviews will be shorter and straighter to the point as they won't suggest as much they wish was different
  - Negative reviews will be long, as people might want to explain all the reasons why they disliked the movie so much
  - Positive reviews will have positive words like "good", "fun", and "exciting"
  - Negative reviews will have negative words like "bad", "boring", and "waste"
- Our above hypotheses will guide our exploratory data analysis.