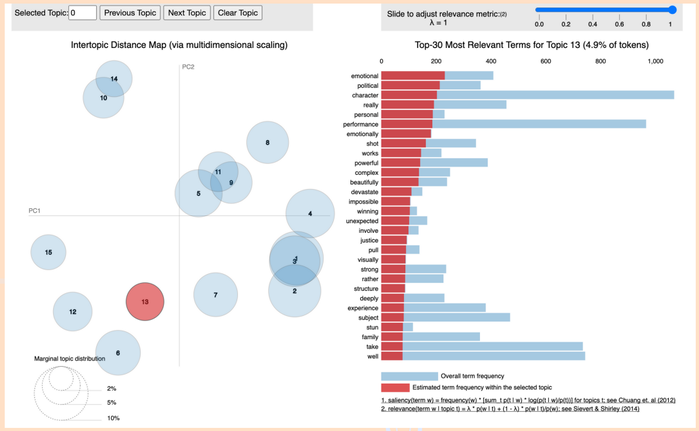
# ROTTEN TOMATOES DATA ANALYSIS

By Anya Chandorkar, Ethan Lee, Joseph Perez, and Moksha Poladi

## What do movie reviews and metadata tell us about their public reception?


### WEB SCRAPING

- Our team web scraped all of our data from the Rotten Tomatoes website
- Utilized python and Beautiful Soup.
- Retrieved the audience and critic reviews and other metadata from each movie
- Scraped the top 100 movies of each year from 2000 to 2020 to include in our dataset.

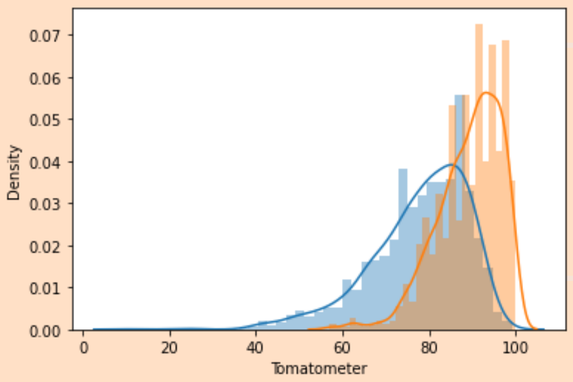



### TOPIC MODELING

- Unsupervised machine learning technique which identifies "topics" in a text corpus
- Used Latent Dirichlet Allocation to categorize our audience reviews and critic reviews into 14 topics respectively
- Given a subset of review summaries, this model would now be able to cherry pick which 'topic' that movie should belong to

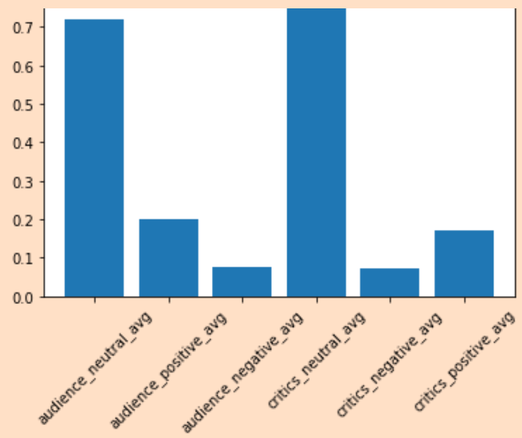
### PREDICTIVE ANALYSIS

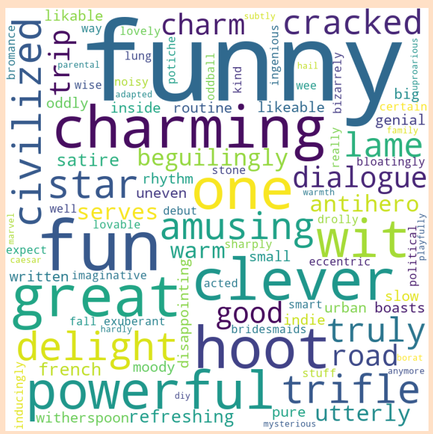
- A key component of our analysis was predicting the Rotten Tomatoes audience score of a given movie.
- We used our generated sentiment scores for critic and audience reviews as features for a RandomForest model.
- The predicted audience score based on our sentiment scores were found to be similar to RottenTomatoes published audience scores (87% accuracy).




### SENTIMENT ANALYSIS

- Sentiment analysis is a technique that uses natural language processing and machine learning to analyze text
- Determines if text is positive,  negative, or neutral
- We used sentiment analysis to analyze the reviews and to quantify the text information as a numeric value
- Utilized in our predictive analysis model




### TF-IDF

- An NLP technique that assesses the importance of a word in its document in relation to the entire corpus
- Created summaries consisting of the most important words in reviews for each movie in the dataset
- These summaries were utilized in the rest of our analysis including sentiment analysis, topic modeling, and even creating word clouds categorizing different genres (picture on left shows the Comedy genre)

```
Example:'Chicken Run is a good movie'

Temperature:0.2

Chicken Run is a good movie that's a wonderful, and it's a worthwhile, but it's a magnificently funny, a
it's a gorgeously funny, and accomplished film that's a

Temperature:0.5

Chicken Run is a good movie, swift and skillful, aftertaste that the pure pleasure to the commitment o
unfolding homecoming-of-age story.As a warm, charming and

Temperature:1.0

Chicken Run is a good movie that isn't really charmingly uprotcashed as anyone on its sensibility or
development.At times, there's a real feel-being of himself, and bancing is here

Temperature:1.2

Chicken Run is a good movie.It's a coming-of-age movie that is just about how that seems dancial or
evapices.You will fus for those marriage while watching another to countern entertainment.A
```

### TEXT GENERATION

- Our text generation model was created using the entire body of reviews that imitates a critic, generating a sort of pseudo-review as its output.
- The model was trained using tools such as GPT-2 and aitextgen.
- The text that we trained it on was partitioned based on the four most popular genres of movies in our dataset.

### OUR FINDINGS

- One of the most important factors for considering the audience scores of the movies is the Tomatometer, or the critic's ratings
- Sentiment scores of the reviews proved invaluable in evaluating the text and its position about the movie that its written about
- Using the metadata and NLP data from the reviews, we were able to achieve an accuracy of 87% in predicting the audience scores of the movies
- Utilizing the mass of text data from the reviews achieved a text generation model that can effectively generate an understandable and coherent "pseudo-review" for each movie gengre

VIEW OUR WORK! LINK:
HTTPS://SHARE.STREAMLIT.IO/JOSEPHPEREZ4/TEST
/MAIN/ROTTEN_TOMATOES_STREAMLIT.PY

**github.com/Ethan-Lee7/RottenTomatoes**