Movie Review Sentiment Analysis Using Natural Language Processing

Johnny Dryman

IMDb

Goals

- Use Natural Language Processing (NLP) with machine learning models to classify IMDB user reviews as 'Positive' or 'Negative'
  - Best model – Logistic Regression with Term Frequency – Inverse Document Frequency (TF-IDF)
- Understand how our model distinguishes between 'Positive' and 'Negative'
- Explore alternative scoring system for movies based on sentiment analysis

Data

- Obtained 50,000 IMDB movie reviews
- Reviews are labeled 'positive' or 'negative,' and are considered "highly polar" reviews
  - 25,000 positive
  - 25,000 negative
- No associated movies or traditional 1-10 rankings were provided
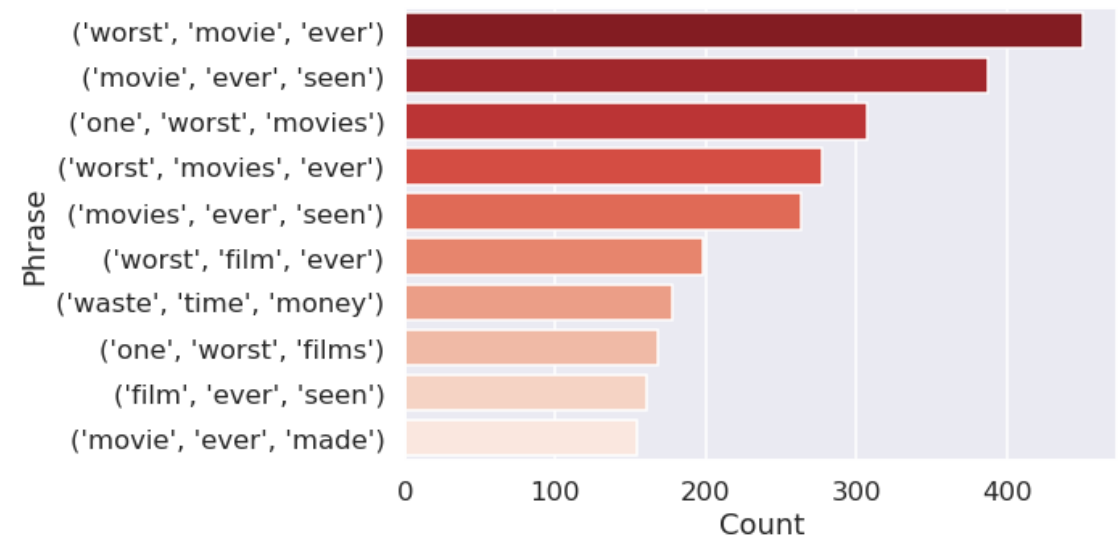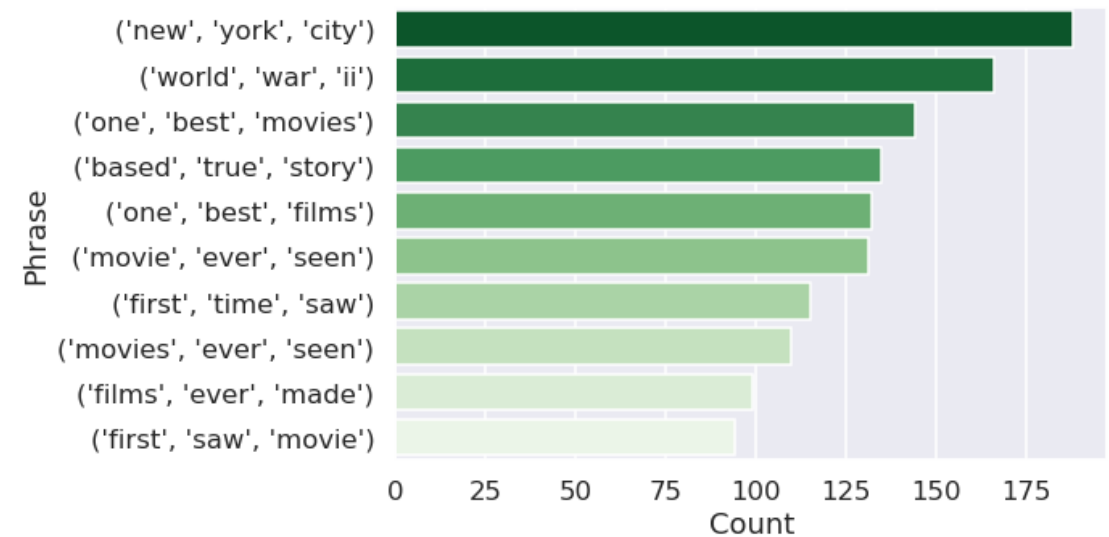- Model targeted 'positive' or 'negative' sentiment

- **Word Clouds**

- **Positive** and **Negative**
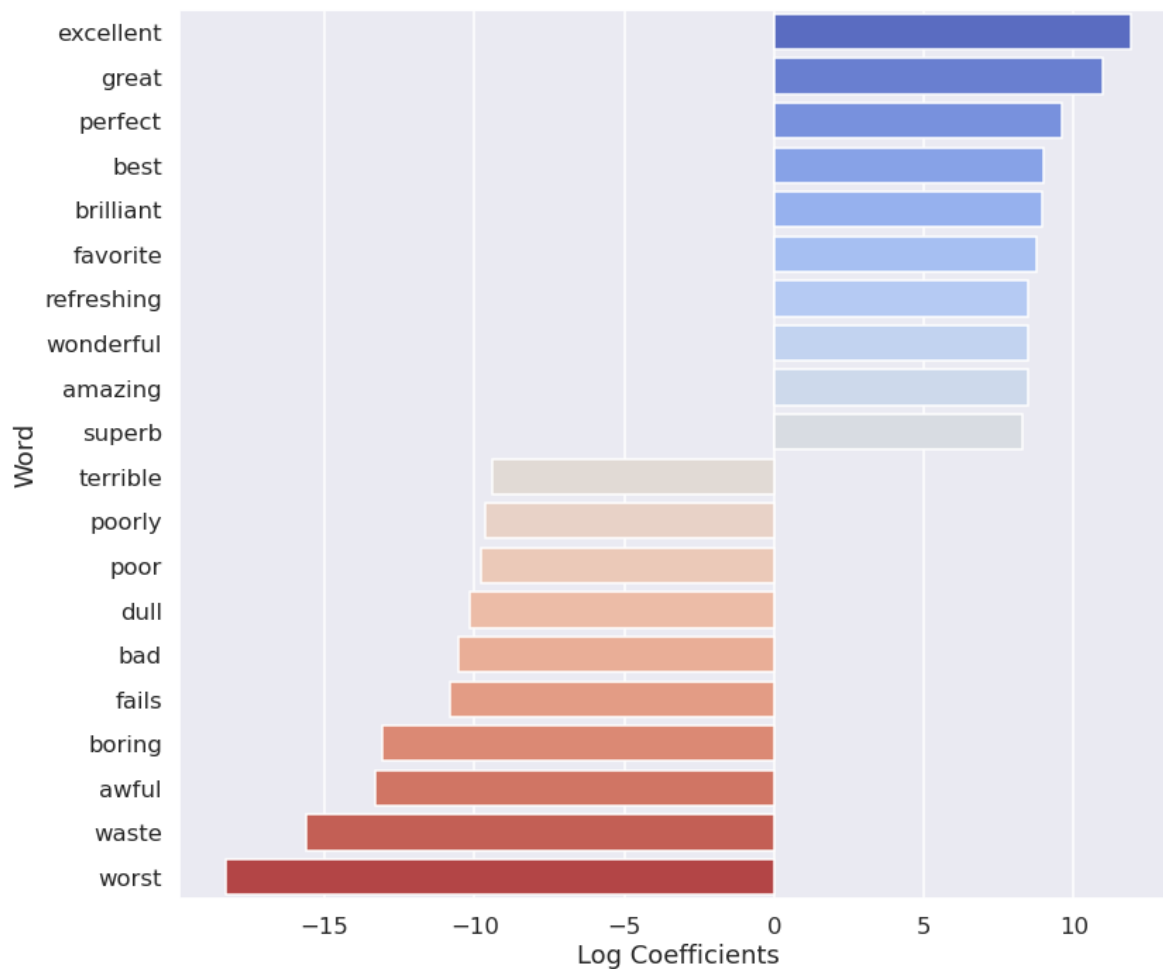- Other than color palette, difficult to distinguish

- **Trigrams**

- **Positive** and **Negative**
- Collection of three words that appear most commonly
- Generally superlative
  - "One best movies"
  - "Worst film ever"

- **Best Model: Linear Regression**

- Model Accuracy: **89.51%**

- **Coefficients:**
- **Positive** and **Negative**
- Words that influenced towards or away from classifying a review as 'positive'
- Mostly strong positive or strong negative adjectives

Conclusions

Model depended on overtly positive or overtly negative adjectives

Review sentiment can be reasonably predicted with an accuracy of 89.5%

Recommendations

- Use 'positive' and 'negative' sentiment model to calculate new score for all movies in database

- Compile a new list of top 250 movies

- Compare user scored top 250 with top 250 as calculated by sentiment analysis

- Depending on results, release to public

Next Steps

- Experiment with different tokenizers and stop word lists

- Reviews were added to dataset based on "highly polar" nature, quantify or qualify what this means

- Compile new dataset with buckets for multiclassification
  - Score of 1-3 / 10 = Negative
  - Score of 4-6 / 10 = Neutral
  - Score of 7-10 / 10 = Positive

- Explore alternative machine learning models
  - Support vector machines
  - K nearest neighbors
  - Deep learning / neural network models

Thanks for your time!

Please feel free to ask any questions.