Crowdsourcing to Predict Current Popular Movies

CHRIS GLYNN, JOSH ALPHONSE, ABHILASH MANDADI

Our Motivation

- this application will predict the current popular movie utilizing crowdsourcing.
- Potential users will be general population who would like to know the popular movie to go see. Future uses could include movie critics to use in their work. Movie creators could use the features extracted to find out what makes a movie popular.

Major Components

- Data Collection
- Data Quality
- Data Exploration
- Classification
- Clustering

Data Collection

- Dynamic queries are used
- The queries are created using a three-step process.
- First, retrieve a list of movie titles from a geographical location
- Next, Remove stop words ,all punctuations and capitalizations
- Add the movie key term to each query
- Each movie is queries separately

Query Example

- “Pirates of the Caribbean: Dead Men Tell No Tales ”
- “pirates caribbean dead men tell no tales movie”

Data Quality

- Query and positive attributes are added to the random sample
- Random Sample: 1,966 false tweets. 34 positive tweets

API Recall: 5.7%

Quality Precision: 1.0

Quality Recall: 5.7%

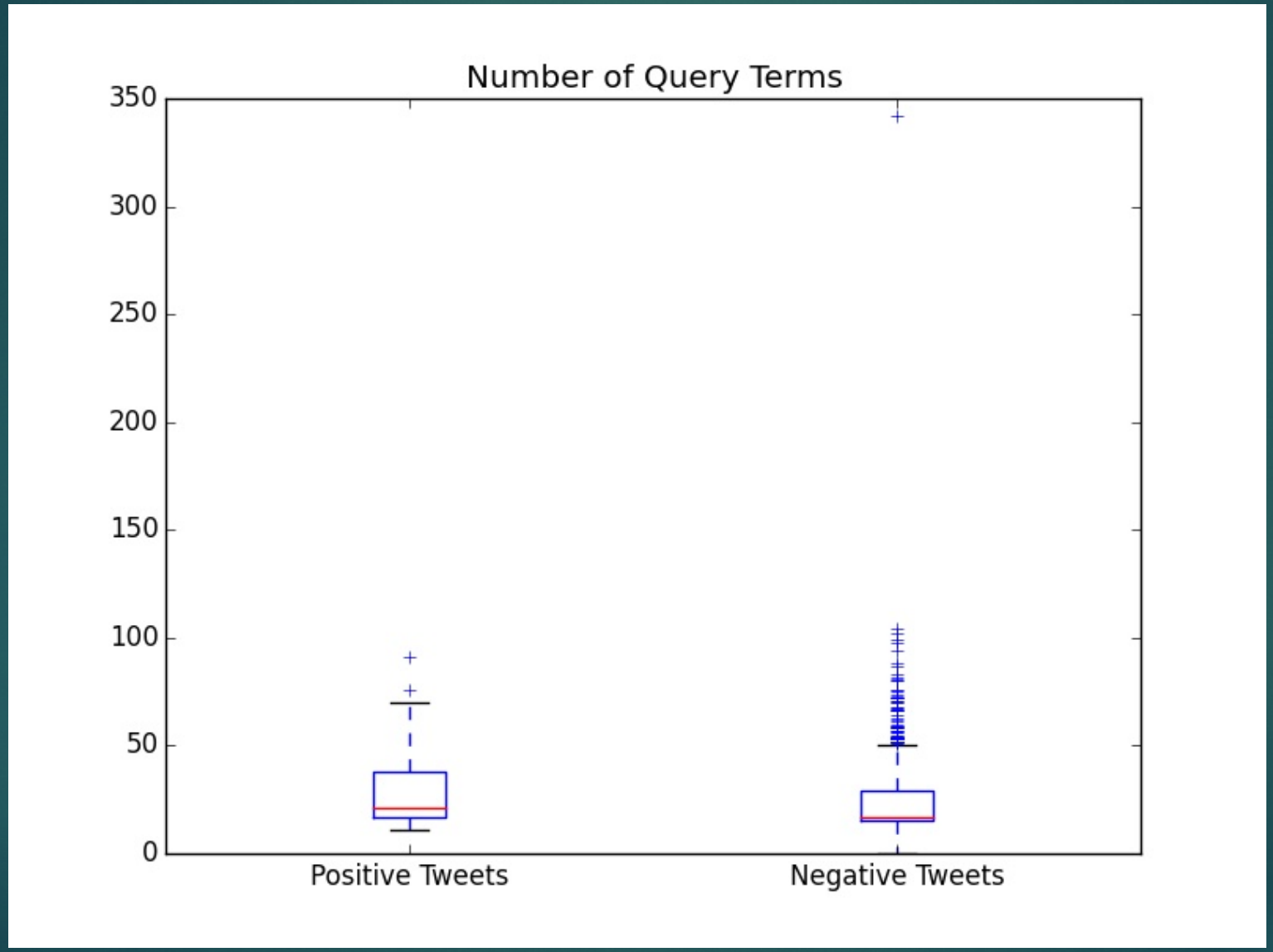
Data Exploration

- Set of collected tweets with attributes: length, number of query terms and contains query terms.
- Positive tweets have less variance on length than negative tweets.
- Positive tweets always contain query terms.
- Extract features that will be used in predicting popular movie.

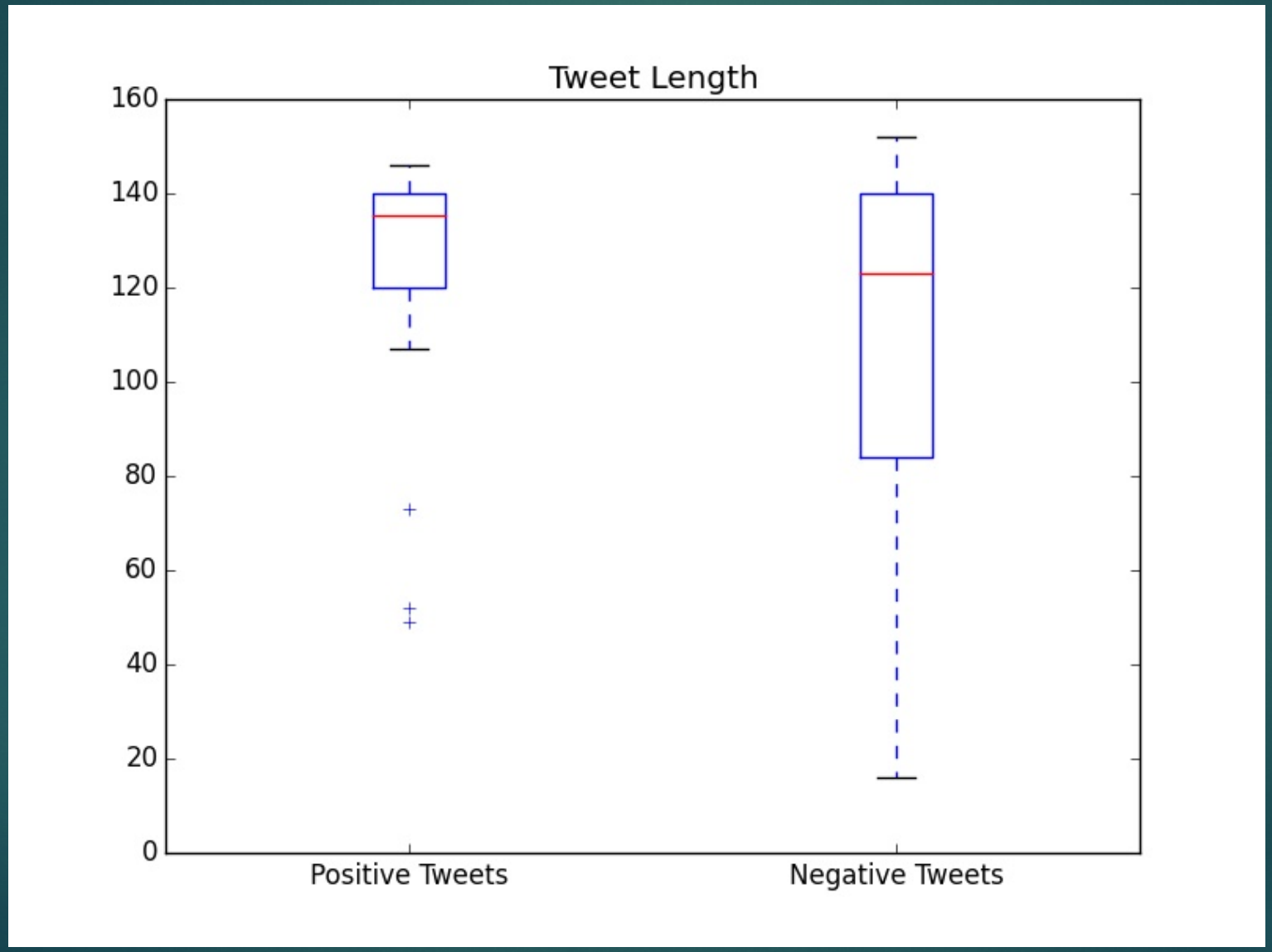
Data Exploration Statistics

- Positive Tweet Length Mean: 125.71875
- Negative Tweet Length Mean: 109.456112009
- Positive Tweet Length STD: 24.0744085792
- Negative Tweet Length STD: 34.5558219705
- Positive Tweet Length Max: 146
- Negative Tweet Length Max: 152
- Positive Tweet Length Min: 49
- Negative Tweet Length Min: 16
- Positive Tweet Length Median: 135.5
- Negative Tweet Length Median: 123.0
- MAD Positive Tweet Length : 6.0
- MAD Negative Tweet Length : 17.0
- Positive Tweet Query Count Mean: 31.75
- Negative Tweet Query Count Mean: 21.3957996769
- Positive Tweet Query Count STD: 19.9436706752
- Negative Tweet Query Count STD: 16.4554536029
- Positive Tweet Query Count Max: 91
- Negative Tweet Query Count Max: 342
- Positive Tweet Query Count Min: 11
- Negative Tweet Query Count Min: 0
- Positive Tweet Query Count Median: 21.5
- Negative Tweet Query Count Median: 17.0
- MAD Positive Tweet Query Count : 6.0
- MAD Negative Tweet Query Count : 4.0

Query Terms Calculations



Data Exploration



Data Exploration

Positive Tweet Wordle

Classification

- Train an SVM model.
- Utilize features from data exploration.
- Compute SVM best parameters
  - Kernel: rbf C: 10 gamma: .001
- Training Set:
  - 4,201 Tweets
  - 1,799 positive
  - 2,402 negative
- Classification is used to remove negative tweets.

Clustering

- Kmeans
- Number of clusters k is the number of movies
- 100 initial centroids are used
- Kmeans and agglomerative were compared
- The largest cluster represents the most popular movie

Questions?