Project 3: Web APIs & NLP

by Casey Liu

Problem Statement

The goal of this project is to classify posts from two different subreddits.

- Data Collection
- Data Cleaning & EDA
- Preprocessing and Modeling
- Evaluation
- Conclusion and Recommendations

Data Collection

- Total of 1785 posts: 944 from Boston and 841 from Seattle

r/boston

- 421k members
- A reddit focused on the city of Boston, MA and the Greater Boston Area.

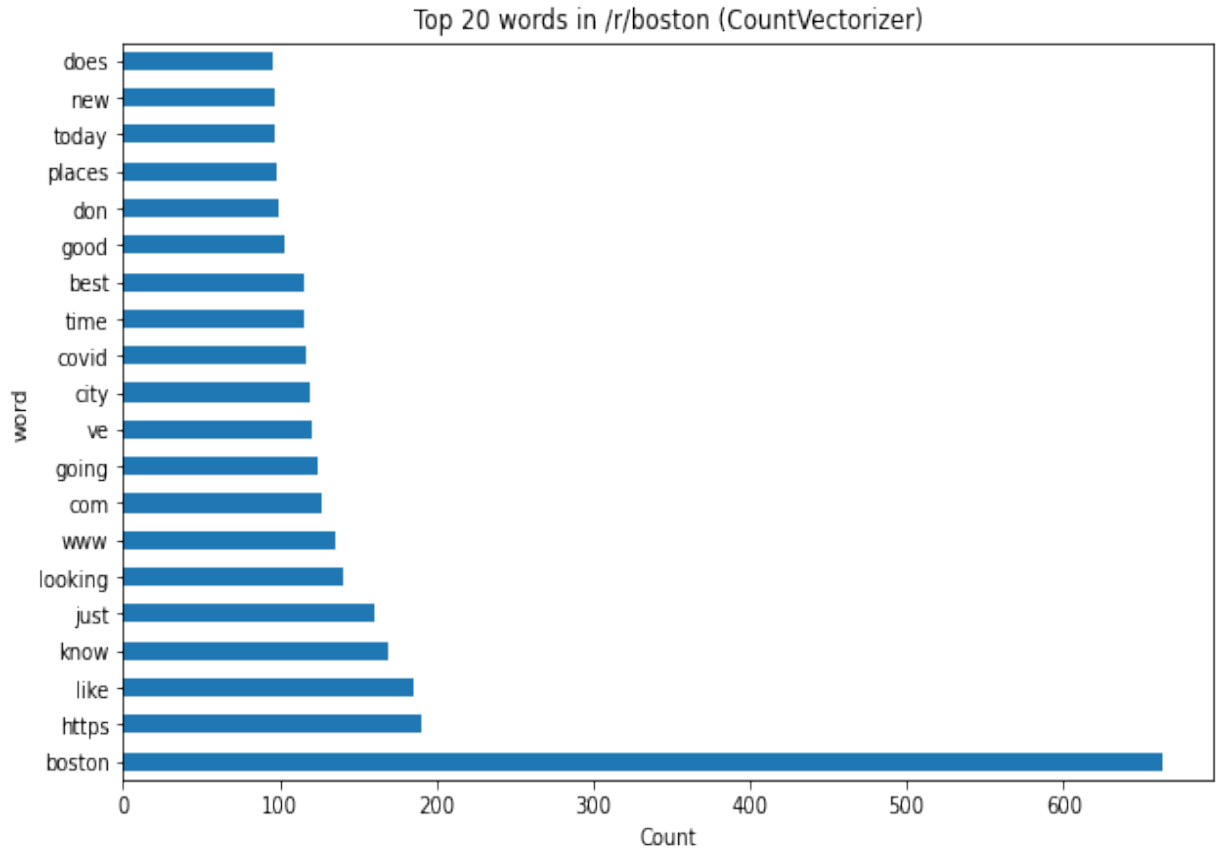
r/seattle

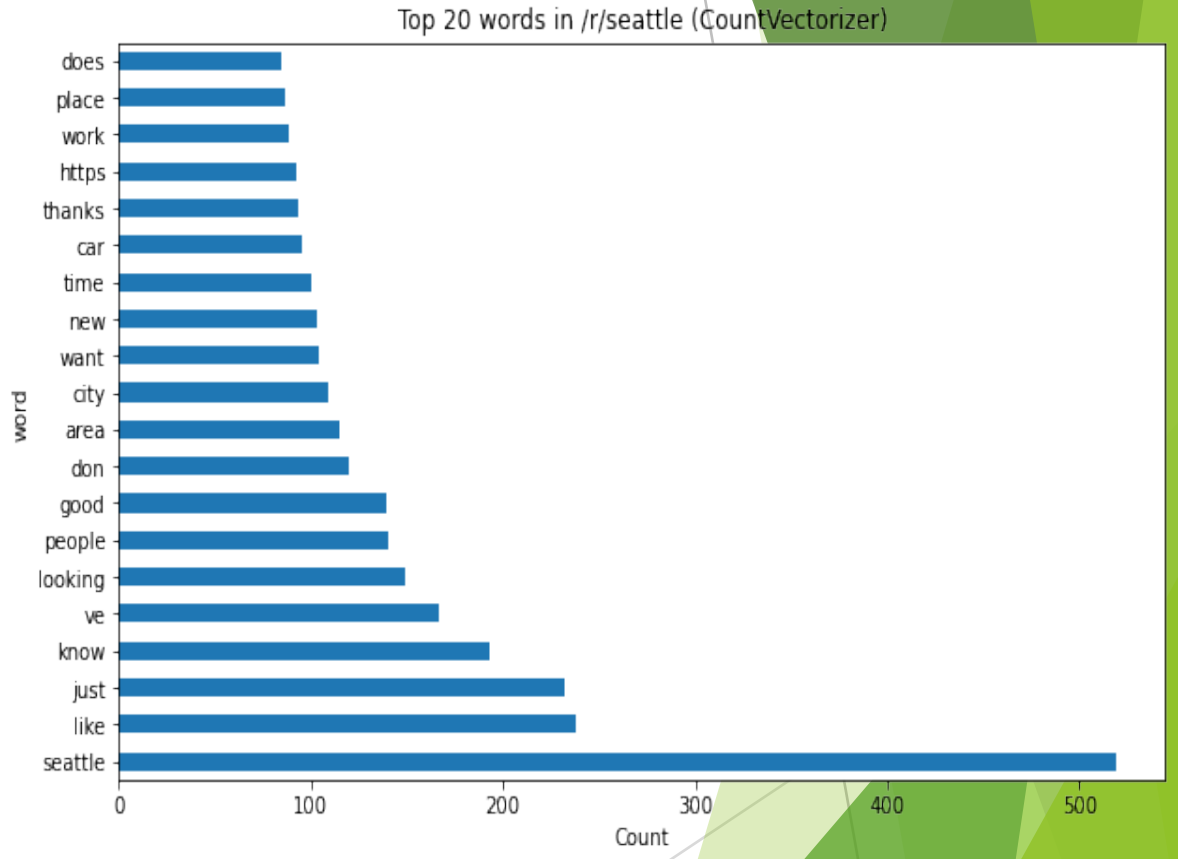
- 382k members
- News, current events in & around Seattle, Washington, USA.

Data Cleaning

- Fill NaN
- Check duplicate rows
- Combine title and selftext in one column
- Convert subreddit to 0 and 1

Frequent words with CountVectorizer





Boston:

- covid, line, thread, mass, mbta, winter, event, walk, regularly, public, south, police, sunset

Seattle:

- work, car, washington, homeless, experience, love, service, food, moved, county, ballard, lake, king

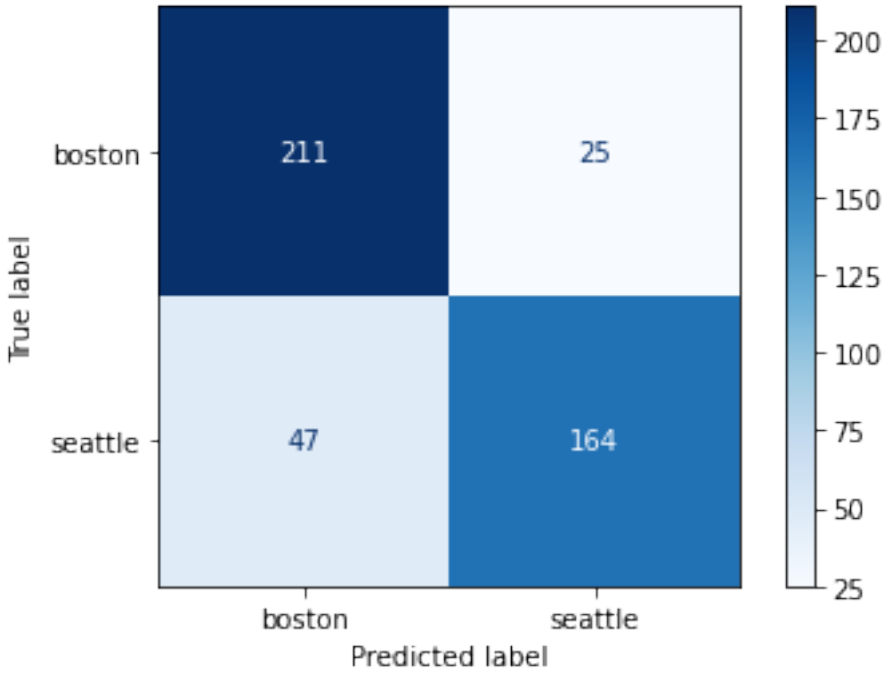
Models

- Logistic Regression with CountVectorizer:
  - GridSearch best params: max_df: 0.5, min_df: 2, ngram_range: (1,2)
  - Train/test scores: 0.9933/0.8389

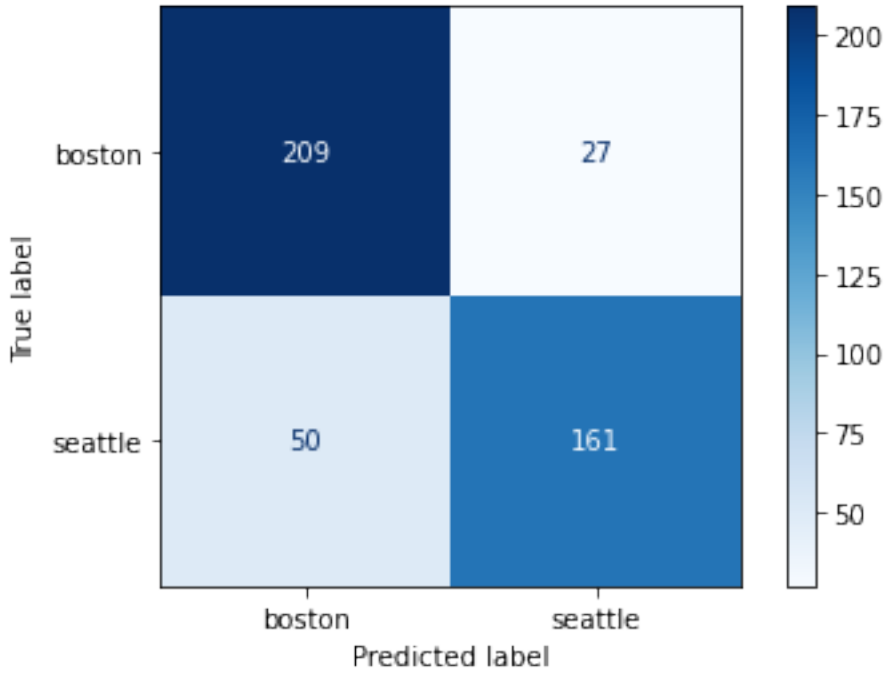
- Random forest with CountVectorizer:
  - GridSearch best params: max_df: 0.5, min_df: 2, ngram_range: (1,2)
  - Train/test scores: 0.9985/0.8277

Evaluation

logistic regression



Random forest



When we compare two models, it seems that logistic regression model performs better than random forest model.

In the logistic regression model:

- The model correctly predicts 83.39% of observations.
- From all posts that the model predicted to be in r/seattle, we have 86.67% of them correctly classified.
- From all posts that are in r/seattle, we have 77.73% of them correctly classified.
- From all posts that are in r/boston, we have 89.41% of them correctly classified.

Sentiment Analysis

Boston

- mean: 0.2617
- 53.28% posts have positive scores

Seattle

- mean: 0.2941
- 58.15% posts have positive scores

Conclusion and Recommendations

- Our logistic regression model performed well with an accuracy score of 83.39%. The random forest model works equally with score of 82.77%.
- Sentiment analysis shows that the comments in r/seattle are slightly positive overall than r/boston. However, we can't conclude which city is better before doing further research.
- Future improvements:
  - Choose significantly different subreddits may improve the model.
  - Collect more data sample. Split data in an appropriate way.
  - Include lemmatization, stemming and spell checks to have cleaned post texts.