# Restaurant review analysis with NLP and recommender system construction

## Yelp capstone project report

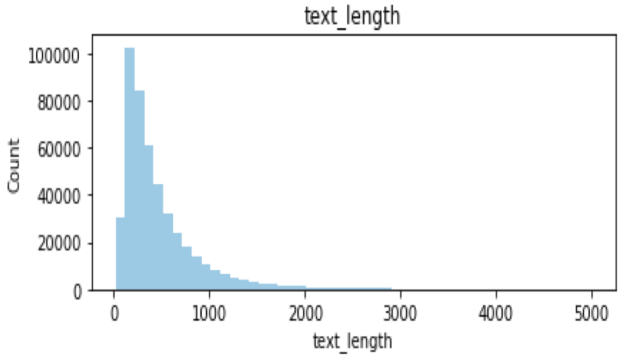
**Fan Yang**

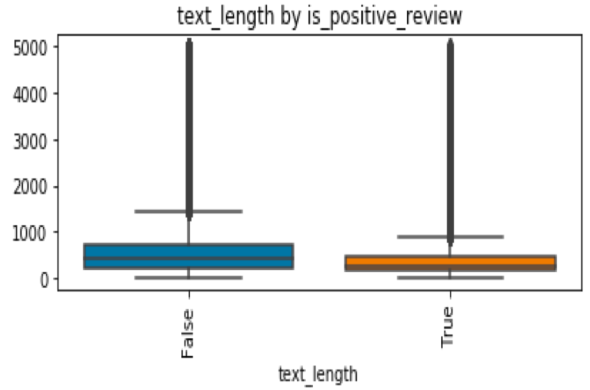
# Project summary

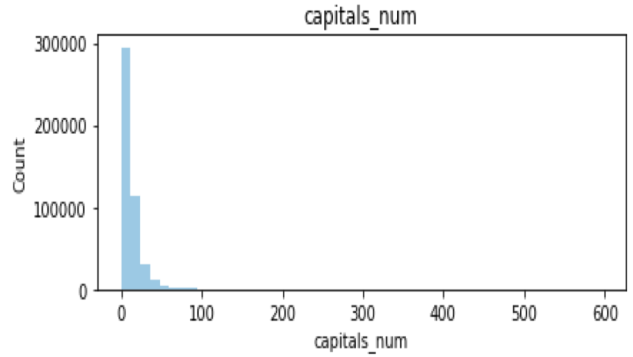
- Yelp is a crowd-sourced local business review networking site.
- Yelp dataset contains about 6 million reviews on 200k business across several major cities.
- Data were downloaded from Yelp dataset in JSON files, including business, review, checkin, tip and photo documents.
- I select review text about all restaurants localized in Arizona between 2016 and 2018.
- Natural Language Processing (NLP) and sentiment analysis were performed, resulting in sentiment classifier modeling.
- I also build restaurant recommender systems based on user rating.

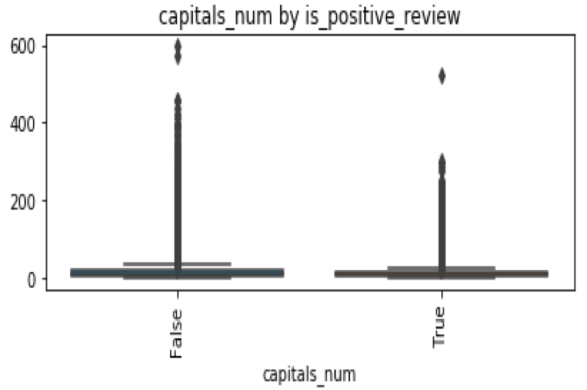
# Generate new features from unstructured review text and test their distribution between positive and negative review

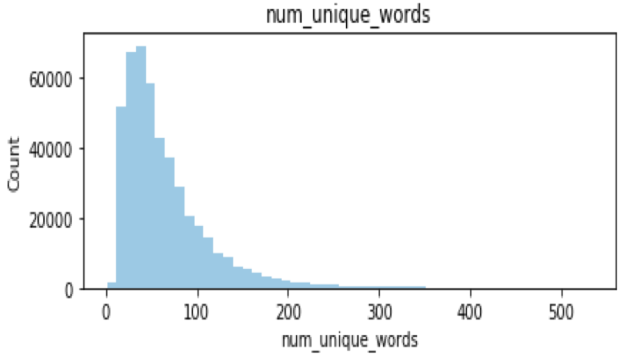
- Positive review when ‘stars’ given by review user is above 4.0.
- Detect text length, frequency of unique word, special symbol or marks.

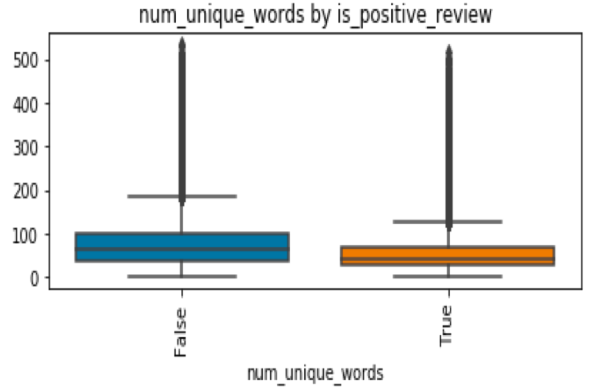


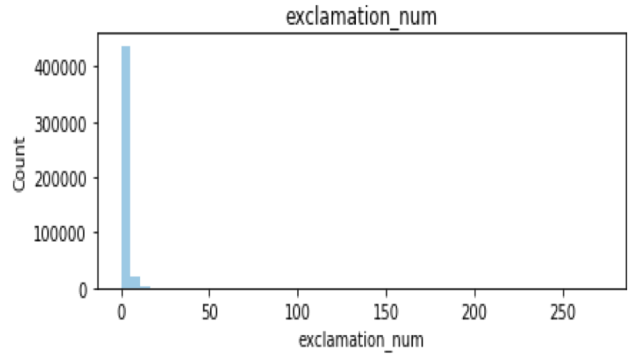



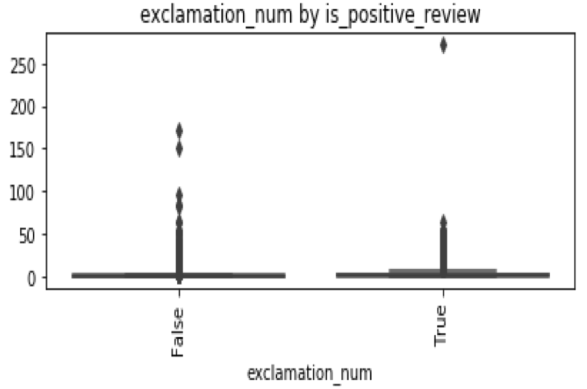



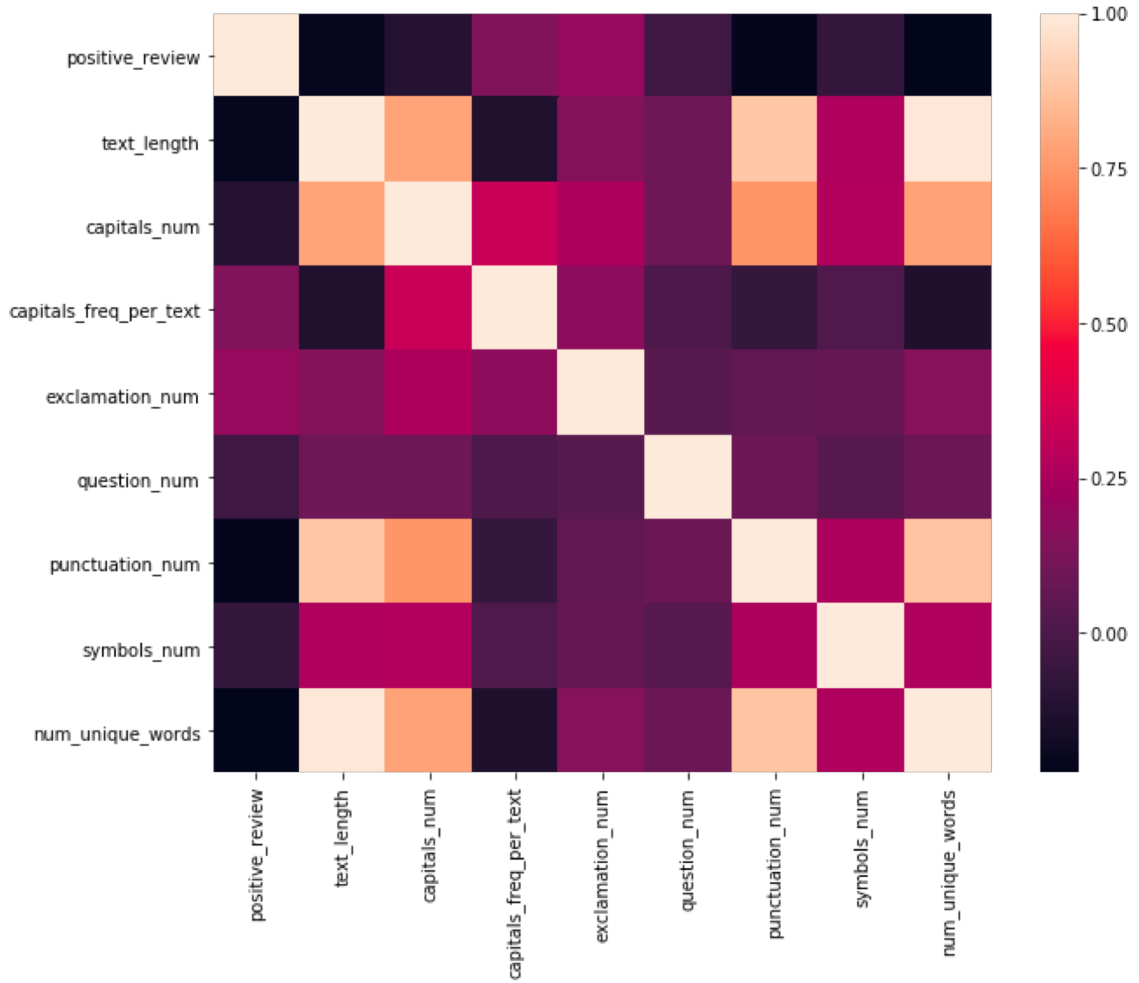



# Visualizing correlations between positive review label and text features




| | index | corr_value |
|---|---|---|
| 0 | positive_review | 1.000000 |
| 1 | exclamation_num | 0.197207 |
| 2 | capitals_freq_per_text | 0.141876 |
| 3 | question_num | -0.038420 |
| 4 | symbols_num | -0.072393 |
| 5 | capitals_num | -0.112887 |
| 6 | text_length | -0.210500 |
| 7 | punctuation_num | -0.212845 |
| 8 | num_unique_words | -0.226370 |

# KMeans divide review text into clusters and reveal top features from each cluster

Transform review text into tf-idf by TfidfVectorizer

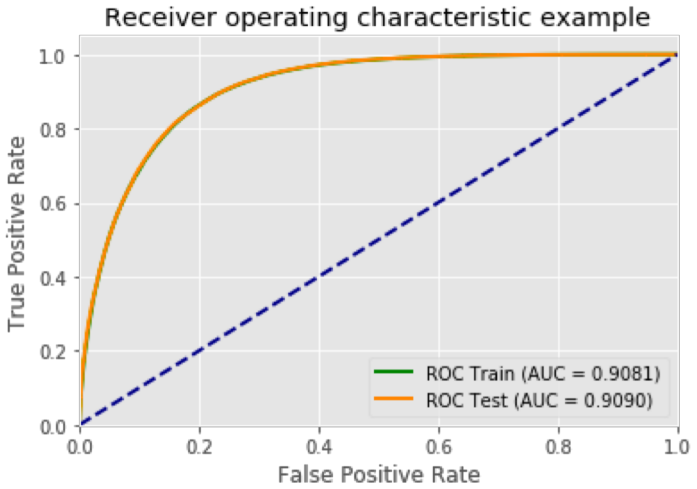
5 cluster

```
top features for each cluster:
0: chicken, good, food, fried, place, rice, ordered, salad, great, like
1: order, food, time, minutes, service, just, came, asked, got, didn
2: good, food, place, best, delicious, service, love, like, amazing, really
3: pizza, crust, good, great, place, wings, service, best, cheese, love
4: great, food, service, place, good, friendly, love, atmosphere, amazing, staff
```

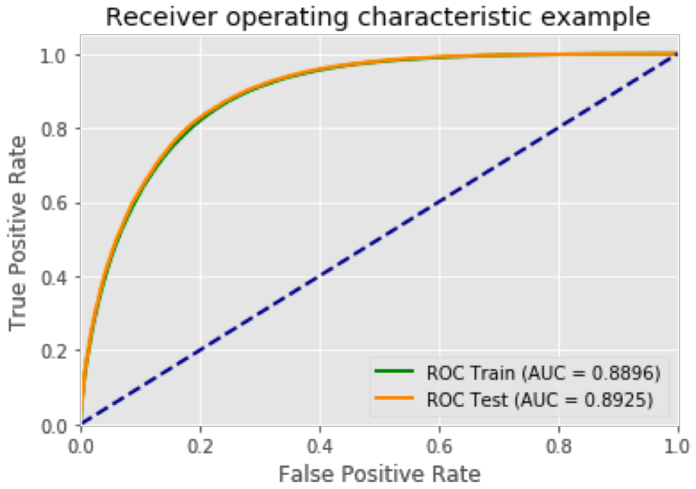
2 cluster

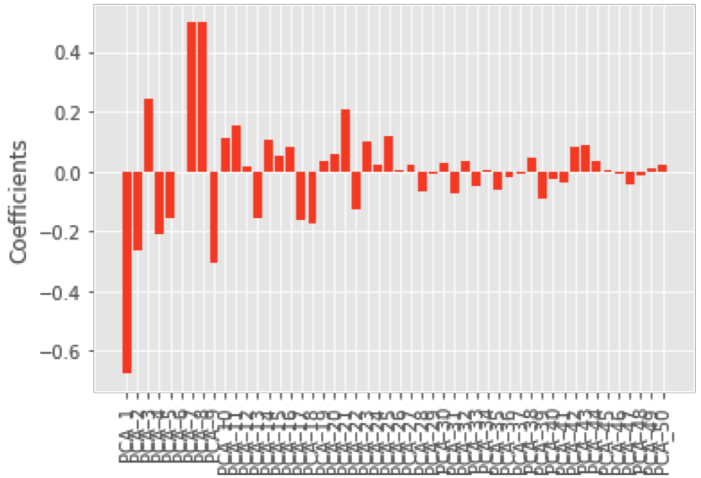
```
top features for each cluster:
0: good, food, place, just, like, time, order, ordered, service, chicken
1: great, food, place, service, love, amazing, friendly, good, staff, delicious
```

# Build Logistic regression model classifying positive review




Without PCA

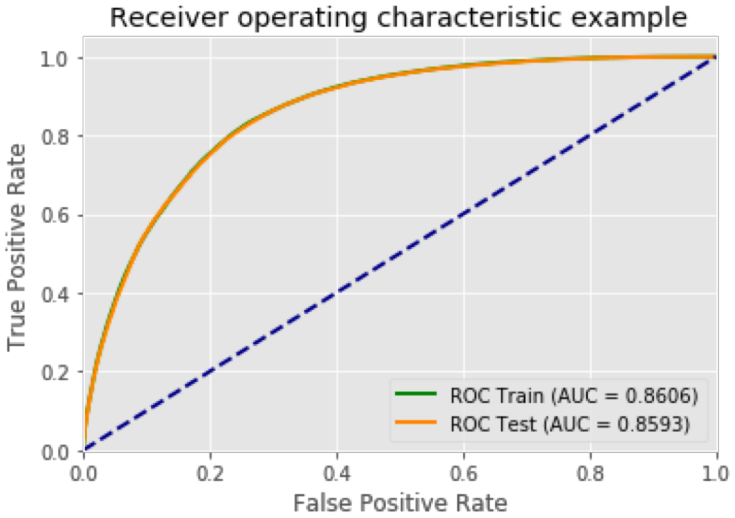


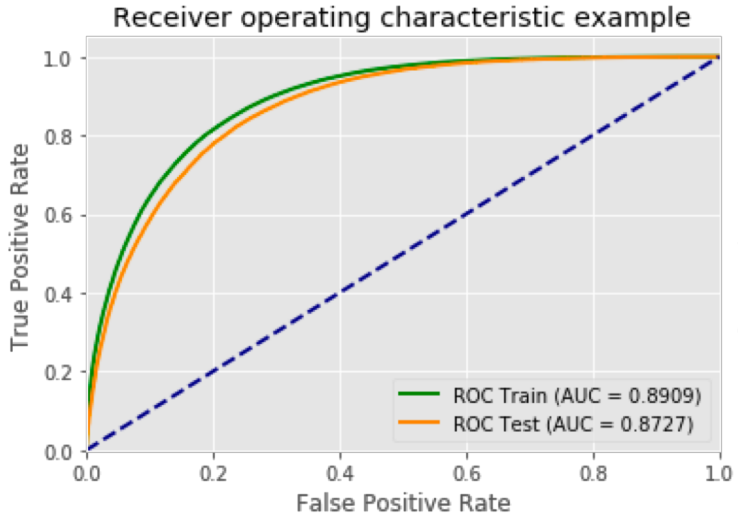

PCA transformation

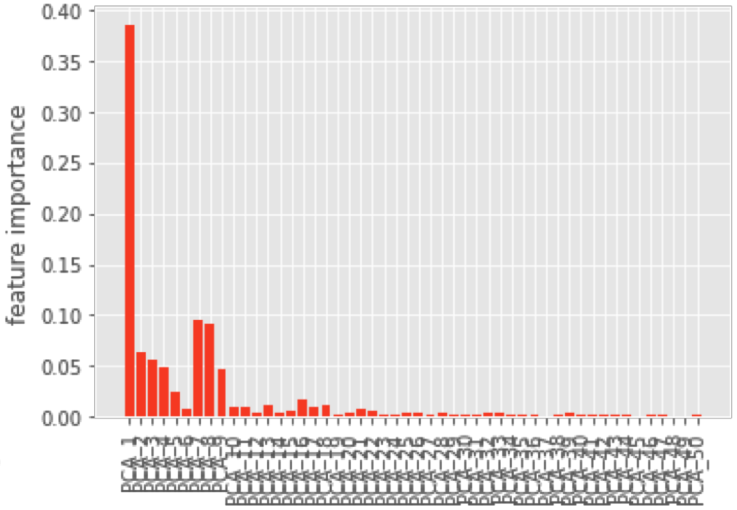
PCA reduce dimensionality and speed up model

# Build Random forest model classifying positive review




Without PCA

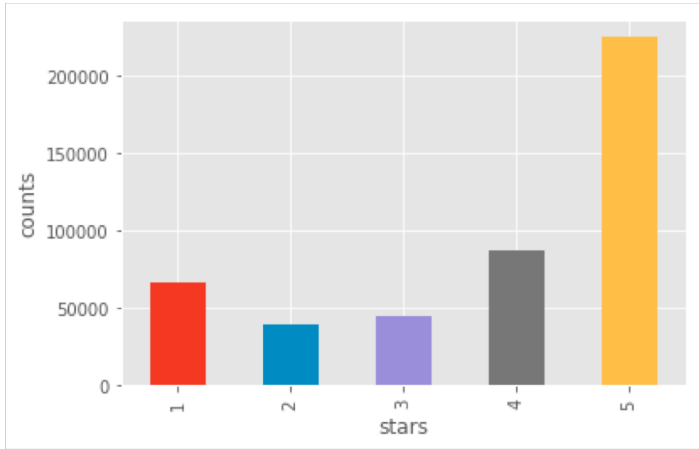



PCA transformation

PCA reduce dimensionality and speed up model

# Generating utility matrix and build restaurant recommender system




| | business_id | user_id | stars |
|---|---|---|---|
| 0 | -01XupAWZEXbdNbxNg5mEg | GbYhdXKQGYGp6D2_S3Oyfw | 1 |
| 1 | -01XupAWZEXbdNbxNg5mEg | CU_RU1o3sKSwymRotn3DUg | 2 |
| 2 | -01XupAWZEXbdNbxNg5mEg | -XoCb6sUMa7NoFayUW0FIA | 1 |

| business_id | -01XupAWZEXbdNbxNg5mEg | -092wE7j5HZOogMLAh40zA | -0WegMt6Cy966qlDKhu6jA | -0aIra_B6iALlfqAriBSYA | -0tgMGl7D9B10YjSN2ujL |
|---|---|---|---|---|---|
| user_id | | | | | |
| -0udWcFQEt2M8kM3xclofw | 0 | 0 | 0 | 0 | |
| -8rSnT5ztVk6vmTDkxTqsQ | 0 | 0 | 0 | 0 | |
| -9S_Fh-sQebyB1yhEM5zHw | 0 | 0 | 0 | 0 | |

Item-item similarity recommender

Matrix factorization recommender (NMF, UVD)

# Project conclusions

1. This project use TF-IDF to extract unstructured review text data.
2. Logistic regress and random forest models were built to classify positive and negative reviews.
3. Review text were grouped by Kmeans method by choosing 2 and 5 clusters, which identify top features from each cluster
4. Restaurant recommender systems were built based on users rating records. Performance of recommender were evaluated by MSE.