STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

BIA 660 Web Analytics Project

Problem Statement

Compare and contrast what makes a particular city different by mining yelp reviews

Dataset & Cleansing

- Datasets:
  - Business JSON file
  - Review JSON File

  Columns /attributes :

  - Business File : business_id, categories, city
  - Review file : business_id, text

  Cleaning:

  - We focused on US cities (Glendale, Boulder city, Harrisburgh, Peoria, Madison, Tempe, Pittsburgh, Charlotte, Phoenix, Las Vegas)
  - Businesses with more than 100 reviews were considered
  - Restaurants were considered among the different business categories

Methodology

- We imported json files(business and review) into **mongodb** (scalable)

  Code: mongoimport --db yelp --collection review--type json --file "C:Web_Analytics\Project**yelp_academic_dataset_review.json**"

  mongoimport --db yelp --collection business --type json --file "C:Web_Analytics\Project**yelp_academic_dataset_business.json**"

- Yelp database created with collections:

  Review and business

- Created new collection "yelp_review" using Yelpdataset.py

  - As discussed in the cleaning steps

  - Output: 8 cities (Glendale, Boulder city, Peoria, Madison, Tempe, Pittsburgh, Charlotte, Las Vegas)

Methodology Contd..

- Ran yelpreviewcity.py(input parameter : city) to create unbiased dataset collection "yelpclassification" with the input city as label:1 and rest 6 cities as label:0 with

  - Output: each label having 6000 records (Total 12000)

  - Boluder city was ignored as we had limited each city should have min. 6000 reviews for "train test method"

- Ran Yelpcityprediction.py for prediction using random forest classifier

- Visualized the main features of cities using wordcloud package.

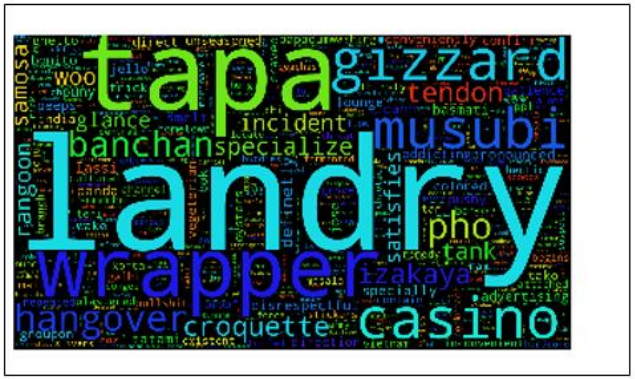
Code Demo

- Demo for the project

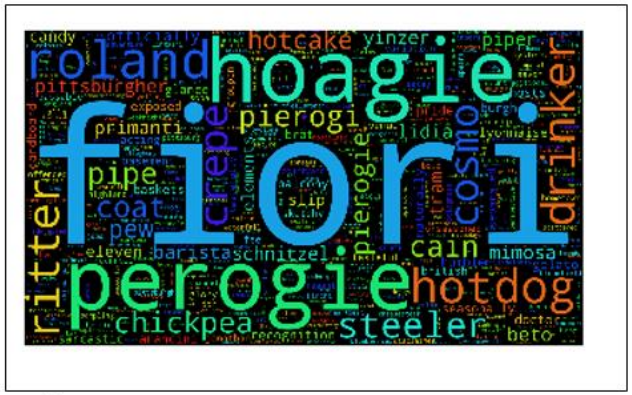
Output

Below are the features that we identified for different cities

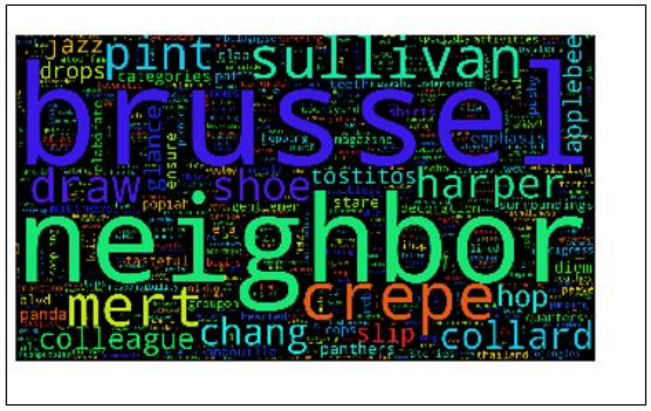
Las Vegas



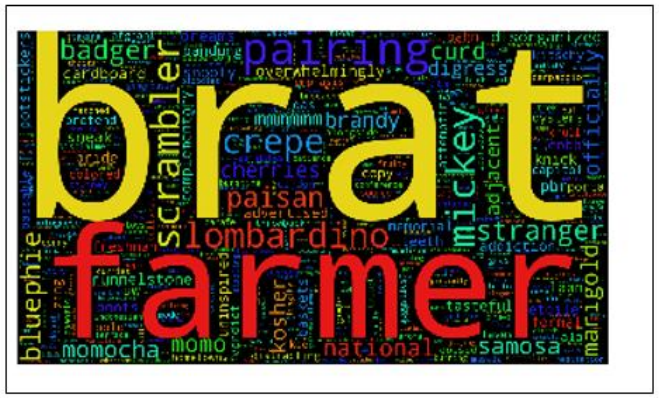
Pittsburgh



Charlotte



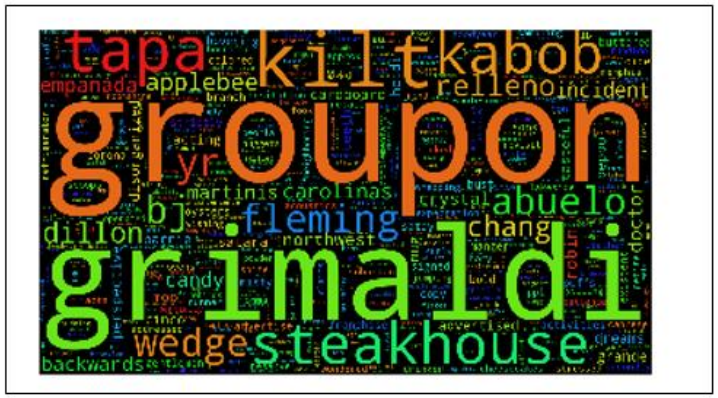
Madison



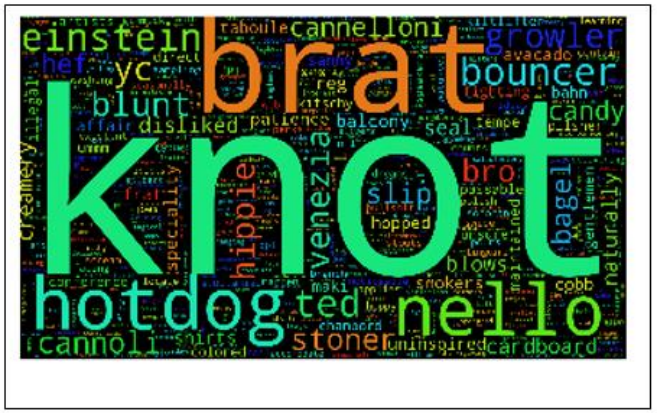
Output

Below are the features that we identified for different cities

Peoria



Tempe



More exploration

- Larger dataset which includes more cities different countries and wide range of categories can be taken.
- To get more insights we can use bigrams, trigrams or four-grams on textual reviews

References

- https://github.com/wendykan/DeepLearningMovies/blob/master/BagOfWords.py #bag of words
- https://www.kaggle.com/c/word2vec-nlp-tutorial/details/part-1-for-beginners-bag-of-words
- http://pandas-docs.github.io/pandas-docs-travis/basics.html #iterrows panda basics and row iteration
- http://blog.rizauddin.com/2009/08/python-remove-items-from-list-that_2943.html #removing elements of list from other list
- https://github.com/amueller/word_cloud # Word Cloud
- http://scikit-learn.org/stable/modules/ensemble.html #feature-importance
- https://www.dataquest.io/blog/machine-learning-python/ #train test split cross validation

Thank You!