We have used 3 code files in our project. Here you can find description of what each file does.

1. **AIT590\_DataCollection.ipynb**:

The yelp dataset has 4 different files https://www.kaggle.com/yelp-dataset/yelp-dataset. But for this project we used business and review files. In this file we have converted the json files to csv.

1. Converting business and reviews JSON data to CSV

2. Extracting only Restaurants information

3. Extracting information of Restaurants in Phoenix, Cleveland, Charlotte, Vegas, Pittsburgh

4. Join the business and reviews data on business\_id

5. Collect the Top 100 restaurants in each city on condition if they have atleast 20 reviews from 20 different users.

6.Save the above data frame. That will be the final file we will work on entire project.

1. **AIT590\_FeatureSelection.ipynb**

Here we used the csv file created from DataCollection.ipynb to extract features.

* 1. Data Pre-processing
  2. Lemmatization
  3. Extract top 1000 frequently used words
  4. Implement word2vec model
  5. Binarize the words in the reviews based on features selected
  6. Aggregate the features for each restaurant
  7. Apply Tf-IDF
  8. Save the csv file

1. **AIT590\_Models.ipynb**

From the features selected from AIT590\_FeatureSelection. We have generated models.

1. Subset the data for each city

2. Generate user-Rating Matrix

3. Apply K-Means clustering

4. Apply Extra Tree Classifier to extract feature important

1. I didn’t attach the main yelp file from which we extracted AIT590\_Data.csv because it is huge file of 4GB. You can find it in <https://www.kaggle.com/yelp-dataset/yelp-dataset>.