Authors: Jasmine Kim, Shirin Haji Amin Shirazi, Bailey Herms
Fall 2017 CS 235
Final Project: YTrend

Dataset
Categories.json
Retrieved from: https://www.yelp.com/developers/documentation/v3/all_category_list
Business.json, reviews.json, user.json
Retrieved from: https://www.yelp.com/dataset/challenge

Necessary Tools
PyPlot, Pandas, Beautiful Soup, Fast DTW, SciPy, WordCloud, NLTK

Files
test_script.py
This is the main file that performs the data mining. It calls 4 tests (run_test_1 to run_test_4) that were run to see what we could find.

preprocessing.py
This file contains all the library functions to pre-process and clean up the JSON files into a CSV format that we need to data mine.

get_csv.py
This file contains all the library functions that does the creation of CSV files. Examples such as creating a CSV given a category, data frame, or business name.

ts_lib.py
This file contains all library functions that pertain to time series data mining. Examples such as pre-processing using a rolling average, time normalization using quartiles, z-normalization, DTW, and pattern finder.

test_lib.py
This file contains all of the test functions that are called in test_script.py. It loops through different category and business name combinations.

review_plotter.py
This file contains a plotter function.

google_v_google.py
This file contains the code to compare Google Trends. This file is standalone.

textMining.py
This file contains the code used to text mine and text visualizations. This file is standalone.

How to Run

Before running, please make sure you have all of the dataset downloaded and necessary tools listed above.

In order to run the the data mining portion, please run test_script.py. Please set the first_time_flag to 1 if this if your very first time running it in order to clean up the JSON dataset. This will do the following:

1. Clean up JSON dataset and create all necessary CSVs with needed information to data mine.
2. Run test 1: This compares all restaurant categories to each other and scores it based off DTW.
3. Run test 2: This is the same as run test 1 but it Z-normalizes and time normalizes using quartiles.
4. Run test 3: This runs a comparison of Chipotle Mexican Grill against all other restaurant categories.
5. Run test 4: This runs a comparison of Chipotle Mexican Grill against all other restaurants that fall under the “Mexican” category.