README.md

YelpWebScraping

Harvesting Yelp Data using API and Website - www.yelp.com

Yelp Reference Links

- Locations - http://www.yelp.com/locations
- Neighbourhoods - https://www.yelp.com/developers/documentation/v2/neighborhood_list

Code Components:

- config.py
- getLocation.py
- restaurants.py
- scrapeDate.py
- scrapeReviews.py

Initial Configuration

1. Generate Yelp API key on https://www.yelp.com/developers/manage_api_keys
2. Add the values of key generated to `config.py`
3. Install required packages. Run `pip install -r requirement.txt` on terminal.
4. After successful installation of above package, download the nltk corpus. Open terminal and type `python` and press *Enter*. Do following two steps. A new window will open. Select `all` identifier in window and click on *download*

```
> import nltk
> nltk.download()
```

Data Scraping for Popular Restarants

1. **Input File format:** Sample input file can be found in input/SampleRestaurantInputFile

   ```
   Minneapolis|Downtown
   Minneapolis|University
   St Paul|Cathedral Hill
   ```

2. **restaurant.py**

   This python program is a data scrapper that will fetch 40 records for each neighborhood in city mentioned in Input file. Here, we `import getLocation` to get longitude and latitude for each neighborhood through Google Geocoding API. The logic for getting location coordinates is written in `getLocation.py`. The output file generated by the program will be in

`output/SEARCHTERM.csv` . With context of this project, the *SEARCHTERM* is **restaurants**. The file usage is as show below:

```
$ python restaurant.py [-h] -f FILENAME -s SEARCHTERM
  Scrape Data through Yelp API
  optional arguments:
  -h, --help            show this help message and exit
  -f FILENAME, --fileName FILENAME
                        File containing neighbourhoods and their respective
                        cities in pipe-delimited fashion
  -s SEARCHTERM, --searchTerm SEARCHTERM
                        Search Category for the data. eg. restaurants, bars,
```

3. **scrapeDate.py**

   This python program will fetch first review date for the restaurants from Yelp website. The program takes input a file with restaurant IDs for which you need the first review date. If there are no reviews or reviews aren't in English language, the program would insert `99999` as dummy value. The output file generates would be `output/firstReviewDate.csv` The file usage is as shown below:

```
$ python scrapeDate.py [-h] -f FILENAME

optional arguments:
-h, --help            show this help message and exit
-f FILENAME, --fileName FILENAME
                      input file with restaurantIDs
```

Data Scraping for Reviews

1. **scrapeReviews.py**

   This python program will take location, category (eg. chinese, indian, mexican, etc.) and term (eg restaurants, bars, etc.) as input, and will fetch the user reviews which will suggest improvements for the restaurants. The filtered reviews will be output in `reviewData` file. The program usage is as shown below:

```
$ python scrapeReviews.py [-h] -l LOCATION -c CATEGORY -t TERM

  Scrape Reviews through Yelp API and Web
  optional arguments:
  -h, --help            show this help message and exit
  -l LOCATION, --location LOCATION
                        Location where you want to search reviews
  -c CATEGORY, --category CATEGORY
                        Category for the reviews. eg. chinese, indian, etc.
  -t TERM, --term TERM  Search by restaurant, bar, etc
```

Datasets Created:

1. RestaurantData.xlsx - The raw dataset generated after running `restaurant.py`
2. filteredRestaurantData1.xlsx - filtered dataset after removing duplicates and noise data (no location, zipcode,etc)

3. filteredRestaurantData2.xlsx - filtered dataset after removing cities restaurants with restaurants less than 100
4. filteredRestaurantData3.xlsx - final dataset for restaurants with review_count > 100 and rating > 4.5
5. reviewData.xlsx - filtered reviews giving improvements for restaurants. This dataset is generated after text mining code `scrapeReviews.py`

Visualization Files and Codes:

1. plotting distribution of rating and review count.R
2. Tableau_Worbook.rar
3. Yelp.R - impute missing vales using mice package
4. Additionally each workbook in dataset above has a sheet called pivots which has analysis along with figures for that workbook.

Sample Files:

These files can be found in `input` folder. Place the files in same location as corresponding scripts.

1. SampleRestaurantInputFile - used with `restaurant.py`
2. sampleScrapeDateInputFile - used with `scrapeDate.py`