Homework 7

Homework 7

In this assignment you will first collect Yelp data by webscraping Yelp pages. You will need to use regular expressions to interpret the data on Yelp pages.

Go to https://pip.pypa.io/en/latest/installing.html to install "`pip`", a tool supported on Windows, Linux, and MacOS that helps in installing Python packages. (For Windows users, `pip` should be installed at "`...\Python27\Scripts`" [usually `C:\Python27\Scripts`]. Make sure you can find `pip.exe` at this path.)

For any package that you want to install in Python, use the command "`pip install`<package_name>".

Install the BeautifulSoup package using the command "`pip install beautifulsoup4`". (For Windows users, if your Python is located at `C:\Python27`, the command should be "`C:\Python27\Scripts\pip install beautifulsoup4`.")

Review the BeautifulSoup documentation located athttp://www.crummy.com/software/BeautifulSoup/bs4/doc/

Download the starter script `hw7.py` from the course Piazza Resources page. Copy the starter script to a new file `hw7.`<lastname>`.py`

Open the link

http://www.yelp.com/search?find_desc=restaurants&find_loc=San%20Francisco%2C+CA&sortby=rating&start=0#

This is the first page of search results of San Francisco restaurants on Yelp, sorted in descending order of ratings. Browse the page source using View Page Source (usually this can be accessed by right-clicking anywhere on the page) to become familiar with its structure and available information, and inspect the relevant elements that render the search results.

Your script will get the forty highest rated restaurants in San Francisco from the search results (pages 1 through 4) of Yelp. (You can figure out the URLs for pages 2 through four of the search results from the buttons on the first page or from the pattern in the URL.) For each restaurant, your script will figure out how many reviews it has.

In the starter code, we give an example (four lines) that reads a page into a string and calls the`preprocess_yelp_page` function to preprocess the string before proceeding to BeautifulSoup. Feel free to modify these four lines of code or write your own code, but do preprocess the page content for every web page your read. Otherwise, there might be issues when you try to find the html tag containing relevant information.

You may find it helpful to download the four web pages once and save them into html files on your local drive for the purposes of debugging your BeautifulSoup code.

Your script will create a text file named `restaurants.`<lastname>`.txt` and write in this file each of the top 40 restaurant names followed by a comma followed by the number of reviews from the four results page, one line for each restaurant, in sorted order based on the number of reviews for each restaurant. For example:

```
Ike's Place,6381
Gary Danko,3945
```

Note that a search result page may contained advertised results at the top of the page. Perform a sanity check that confirms the number of restaurants you are counting is ten. If you do encounter an advertised restaurant, figure out how to separate the advertised restaurants from the "real restaurant" results.

When your program is complete, upload the `hw7.`<lastname>`.py` and `restaurants.`<lastname>`.txt`files using the file upload tool available at https://www.ischool.berkeley.edu/uploader/?s=i206 Login with your ISchool userid

and password and follow the directions.

Extra Credit:

For extra credit, you can collect the same information in a different way, using the Yelp APIs.

Download the `hw7xtracredit.py` file from the Piazza Resources page, and rename it`hw7xtracredit.<lastname>.py`

(Note that the extra credit script uses `urllib2` rather than `urllib` for compatibility with the `oath2`package).

Run the command `pip install oauth2`.

Go to http://www.yelp.com/developers and create an account. Go tohttp://www.yelp.com/developers/manage_api_keys to generate your app key/secret and a token by providing a website URL (such as the ISchool website or a dummy URL) and giving the reason to use the APIs (homework assignment). Copy the "Consumer Key", "Consumer Secret", "Token", and "Token Secret" into the relevant portions of the extra credit script.

Go to http://www.yelp.com/developers/documentation and learn how to build the URLs to use the Yelp Search and Business APIs.

You will need to form a URL and send it to the `yelp_req` function in the script we gave you to get the API response.

(Hints: Look at the parameters `limit`, `offset`, and `sort` in the Yelp API documentation. Perform a Google search to find out how to percent-encode the parameters using `urllib2.urlencode`. Many additional parameters are needed to be appended to the URL you form for the purpose of authentication. Try printing the full URL in the HTTP request for yourself and see what parameters are included.)

The HTTP responses will be JSON strings. Your program should go through them and produce the file`restaurants2.<lastname>.txt` in the same format as `restaurants.<lastname>.txt`

Upload your `hw7xtracredit.<lastname>.py` and `restaurants2.<lastname>.txt` using the usual process.