**System requirements:**

* Python 2.7.XX (It will not compile on Python 3)
* Requests module (using pip installer `pip install request`)
* Beatifulsoup4 (using pip installer `pip install BeautifulSoup4 ––upgrade`)

**How to run:**

* Download the attach package “package.zip” and unzip it. There will be 5 files inside:

1. **crawler.py:**  this will send a HTTP ‘GET’ request to the website to parse the HTML content of the webpage. Then initializing the DataCleaner object to scrape data from the HTML content.
2. **abautosDataCleaner.py:**  a class handles data from <http://www.abwautos.com/>
3. **yellowpagesDataCleaner.py:** a class handles data from <http://yellowpages.com/>
4. **stackoverflowDataCleaner.py:** a class handles data from <http://stackoverflow.com/>
5. **execute.py:** automation script to run crawler.py with different datacleaner objects

* To run it. Put these files in an empty directory. “cd” to the directory and run the following command `python execute.py` and observe the results. The program will scrape data from these three websites and generate data in a separate text files. Each output text file represents data for each page on the web. To modify number of pages (or iteration) to scrape data, open execute.py and modify the item of each similar following line

**sys.argv = ['crawler.py','http://www.abwautos.com/', r'vehicles?Page=#', 'output','abwautosDataCleaner', '7' ]**

For example, instead of getting 7 text files, it could have generated 5 text files by changing the **‘7’** with a **‘5’**. In order to change the name of output files, replace the **‘output’** with whatever name you want. Please keep the rest of items intact otherwise the program will crash.

If you don’t want to use the **“execute.py”**, you can scrape data for each website with the following format for each website with your Terminal (Mac, Unix) or Command Prompt (Windows)

* For <http://www.abwautos.com/>:

*python crawler.py "http://www.abwautos.com/" "vehicles?Page=#" [OUTPUT\_FILENAME] abwautosDataCleaner [NUMBER\_ITERATION]*

*ex:* python crawler.py "http://www.abwautos.com/" "vehicles?Page=#" output abwautosDataCleaner 7

* For <http://yellowpages.com/>:

*python crawler.py "http://www.yellowpages.com/" "search?search\_terms=restaurant&geo\_location\_terms=Waipahu%2C%20HI&page=#" [OUTPUT\_FILENAME] yellowpagesDataCleaner.py [NUMBER\_ITERATION]*

*ex:* python crawler.py "http://www.yellowpages.com/" "search?search\_terms=restaurant&geo\_location\_terms=Waipahu%2C%20HI&page=#" yellowpagesOutput yellowpagesDataCleaner.py 6

* For <http://stackoverflow.com/>

*“python crawler.py "http://stackoverflow.com/questions" "?page=#&sort=newest" [OUTPUT\_FILENAME] stackoverflowDataCleaner.py [NUMER\_ITERATION]”*

*ex:* python crawler.py "http://stackoverflow.com/questions" "?page=#&sort=newest" stackoverflowOutput stackoverflowDataCleaner.py 4

**Note:**

* Abautos.com only has 7 pages at the moment. So not to try more than 7 iteration if you don’t want to get a empty output file.
* For the command line prompt, it does not matter that your [ABC]DataCleaner has “.py” or not. The **“crawler.py”** should handles it.
* Lastly, you should use the “python execute.py” command in order to run this. Since the prompt for the command line is quite tricky.