**Readme File - IR Assignment 1**

**Setup**

Files/Software to Download:

Python

<https://www.python.org/download/releases/3.0/>

BeautifulSoup

For Documentation:

<https://pypi.org/project/beautifulsoup4/>

pip install bs4

Requests

Reference:

<http://docs.python-requests.org/en/master/>

pip install requests

PyCharm

<https://www.jetbrains.com/pycharm/>

**Compile**

Task 1

1] To execute the file for task 1 go to the folder containing the file Task1.py

2] Open the file using PyCharm

3] Create the folders: FolderCar, FolderCarbon, FolderTime in the new project that you import the Task1.py to

4] Run the Task1.py from the terminal inside PyCharm or use the shortcut to run it

5] The Task1.py file has to be run using three see urls which can be added to the task1() function which is called from the main() method.

6] The folder links where the files are stored need to be updated when running different files to be able to correctly store the entire raw HTML pages in separate directories for the different seeds.

**Task 2**

1] Task 2 assumes correct completion of Task1

2] To execute the file for task2 go to the folder containing the file Task2.py

3] Create the folder: MergedList in the same project

4] After creating the folder run file Task2.py

**Task 3**

1] Task 3 does not require completion of Task1 or Task2

2] To execute the file for task3 go to the folder containing the file Task3.py

3] Add that file to the project that you are working in.

4] Use the terminal or the shortcut to run the file

5] You will be prompted to enter the keywords that you want to search for: Enter them and separate different keywords with a comma(,)

6] The seed url is prespecified, but can be changed from the task3() method

7] You will need to create a folder called FocusedCrawler that will contain the links that are crawled using the keywords provided

8] The final text file with all the duplicates will also be in this folder by the name finallist.txt

**Max Depth for Task 1:**

**Seed:** <https://en.wikipedia.org/wiki/Time_zone>

**Max\_ Depth :** 3 before 1000 links found

**Seed:** <https://en.wikipedia.org/wiki/Electric_car>

**Max Depth :** 3 before 1000 links found

**Seed:** <https://en.wikipedia.org/wiki/Carbon_footprint>

**Max Depth :** 3 before 1000 links found

**Max depth for Task 2:**

**Max Depth :** 3 before 1000 links found

**Max depth for Task 3:**

**Max Depth:** 6 before 1000 links found

Bibliography:

1] <https://www.pythonforbeginners.com/beautifulsoup/beautifulsoup-4-python>

2] <https://www.dataquest.io/blog/web-scraping-tutorial-python/>

3] <https://www.youtube.com/watch?v=XQgXKtPSzUI>

4] <https://www.youtube.com/watch?v=0_VZ7NpVw1Y>

**Contact Information:**

**Mihir Gandhi**

**gandhi.mi@husky.neu.edu**