**Instruction for compiling and running Web Crawler**

* Files: Zip file comes with following files
  1. *Crawler.py:* python file implemented on python 3, which is implementation of web crawler.
  2. *unfocused\_crawling\_URL\_list*: URLs retrieved while running crawler without any keyphrase.
  3. *focused\_crawling\_URL\_list*: Relevant URLs retrieved while running crawler with keyphrase “Concordance”
* Steps to run Crawler.py

1. Unfocused Crawler: (Without keyphrase)

>python3 Crawler.py “http://en.wikipedia.org/wiki/Hugh\_of\_Saint-Cher”

Run Time: ~2000 seconds (

1. Focused Crawler: (With Keyphrase “Concordance”)

>python3 Crawler.py “http://en.wikipedia.org/wiki/Hugh\_of\_Saint-Cher” “Concordance”

Run Time: ~1600 seconds

* Pre-requisites
  1. python version 3.x (preferable 3.5.0)
  2. BeautfulSoup4 module in python3

>python3 -m pip install BeautifulSoup4

* Proportion of focused crawling with the keyphrase “Concordance”:
  + Total relevant pages found with keyphrase “Concordance” (n)= 56
  + Total pages crawled (N) = 1000
  + Proportion = (n/N) = 56/1000 = 0.056

***References***:

* Python Tutorials: <http://www.tutorialspoint.com/python/>
* BeautifulSoup4: <http://www.crummy.com/software/BeautifulSoup/bs4/doc/>

Collaborated with: Aniket to Handle connection errors in python.

Note:

1. I have considered all the links on wikipedia to start with “http://” and not “https://”. So the same links starting with http:// and https:// will be considered as different links as answered by instructor on piazza.
2. As discussed with Prof. Nada I’m crawling 1000 pages which is a stopping condition in case of focused crawling and out of which I got 56 relevant pages with keyword “Concordance”.