Information Retrieval and Web Analytics

Semester 2 – 2020.

Assignment 2 – 10%

This is an Individual Assignment

Due Date – 11th September 2020

Marks will be reduced if code is copied from Internet or from a fellow student.

You are free to use existing Python Packages to do this assignment.

The written programs should be tested by proper inputs. The input and the output of the program should be clearly shown in the code.

**Write python program to do the followings.**

1. Test if a given page is found or not on the server.  (5 marks)
2. Download and display the content of robot.txt for en.wikipedia.org. (5 Marks)
3. Get 90 days of visits broken down by browser for all sites on data.gov.

(10 Marks)

1. Consider the url ‘<http://www2.sas.com/proceedings/sugi30/toc.html>’. Extract all the papers names, from 001-30 to 268-30. (10 Marks)
2. Consider the url=’<http://www.dictionary.com/browse/>’ and the words ‘handy’,’whisper’,’lovely’,’scrape’.  
   Build a data frame, where the first variable is “Word” and the second variable is “definitions”. Scrape the definitions from the url. (10 Marks)
3. Let consider any twitter account of any celebrity to do the following exercises.

(4\*10 = 40 Marks)

1. Get the number of followers of a given twitter account.
2. Get the number of posts on Twitter liked by a given account.
3. Count number of tweets by a given Twitter account.
4. Scrap number of tweets of a given Twitter account.
5. Select any popular ecommerce website such as AMAZON, FLIPCART, ALIBABA, DARAZ, etc. to do the following exercises.
   1. Scrap the selected website and collect all the product related information from the website and store them in a data frame. (10 Marks)
   2. Using the results from (a) identify and print/display the followings
      1. Top 10 most popular products based on sales (5 Marks)
      2. Top 10 most popular products based on customer reviews.

(5 Marks)