Date: 15/09/2022

Assignment No. 08

Title: Web Mining task

PRN: 2019BTECS00080

Batch: B7

Objective/Aim:

1. Implement the Crawler ( DFS & BFS).

The program should able to take seed input as any URLs/page and display all the links obtained by crawler.

2. Implement the PageRank algorithm to calculate the rank of each page in the file. The output should be the 10 pages with the highest rank, together with their rank values. a. Tabulate the results containing adjacency matrix and rank of pages.

3. Implement the HITS algorithm to calculate the hub and the authority weight of each web page in the data set. The output should be the 10 most authoritative pages and 10 most hubby pages. a. Tabulate the results containing adjacency matrix and rank of pages.

**Introduction:**

Web Mining is the process of Data Mining techniques to automatically discover and extract information from Web documents and services. The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns.

Applications of Web Mining:

1. Web mining helps to improve the power of web search engine by classifying the web documents and identifying the web pages.
2. It is used for Web Searching e.g., Google, Yahoo etc and Vertical Searching e.g., FatLens, Become etc.
3. Web mining is used to predict user behavior.
4. Web mining is very useful of a particular Website and e-service e.g., landing page optimization.

Web mining can be broadly divided into three different types of techniques of mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. These are explained as following below.

1. Web Content Mining: Web content mining is the application of extracting useful information from the content of the web documents. Web content consist of several types of data – text, image, audio, video etc. Content data is the group of facts that a web page is designed. It can provide effective and interesting patterns about user needs. Text documents are related to text mining, machine learning and natural language processing. This mining is also known as text mining. This type of mining performs scanning and mining of the text, images and groups of web pages according to the content of the input.
2. Web Structure Mining: Web structure mining is the application of discovering structure information from the web. The structure of the web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Structure mining basically shows the structured summary of a particular website. It identifies relationship between web pages linked by information or direct link connection. To determine the connection between two commercial websites, Web structure mining can be very useful.
3. Web Usage Mining: Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets. And these patterns enable you to understand the user behaviors or something like that. In web usage mining, user access data on the web and collect data in form of logs. So, Web usage mining is also called log mining.

**Theory/Algorithm:**

1. BFS:

Follow the below method to implement BFS traversal.

* Declare a queue and insert the starting vertex.
* Initialize a visited array and mark the starting vertex as visited.
* Follow the below process till the queue becomes empty:
* Remove the first vertex of the queue.
* Mark that vertex as visited.
* Insert all the unvisited neighbours of the vertex into the queue.

1. DFS:

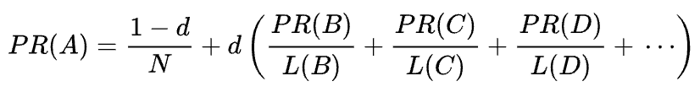
Follow the below steps to solve the problem:

1. Create a recursive function that takes the index of the node and a visited array.
2. Mark the current node as visited and print the node.
3. Traverse all the adjacent and unmarked nodes and call the recursive function with the index of the adjacent node.

3. PageRank algorithm:

Let’s have a look at how PageRank works. Each link from one page (A) to another (B) casts a so-called vote, the weight of which depends on the collective weight of all the pages that link to page A. And we can't know their weight till we calculate it, so the process goes in cycles.

The mathematical formula of the original PageRank is the following:

[](https://www.link-assistant.com/upload/news/posts/422/2.png)

Where *A*, *B*, *C*, and *D* are some pages, *L* is the number of links going out from each of them, and *N* is the total number of pages in the collection

4. HITS algorithm:

* Initialize the hub and authority of each node with a value of 1
* For each iteration, update the hub and authority of every node in the graph
* The new authority is the sum of the hub of its parents
* The new hub is the sum of the authority of its children
* Normalize the new authority and hub

**Procedure:**

Import Data file

Select Algorithm

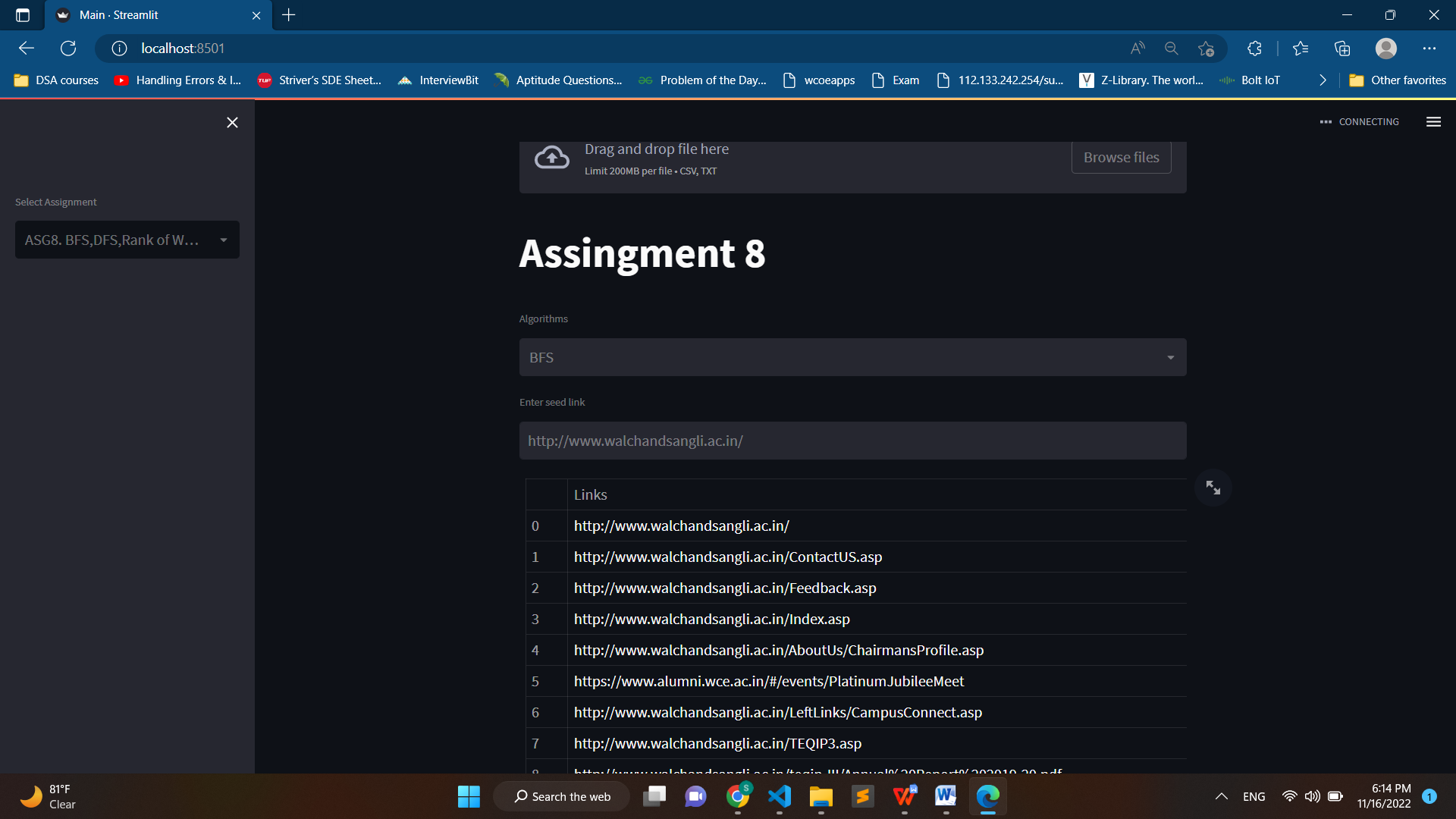
(BFS, DFS, PAGE RANK, HITS)

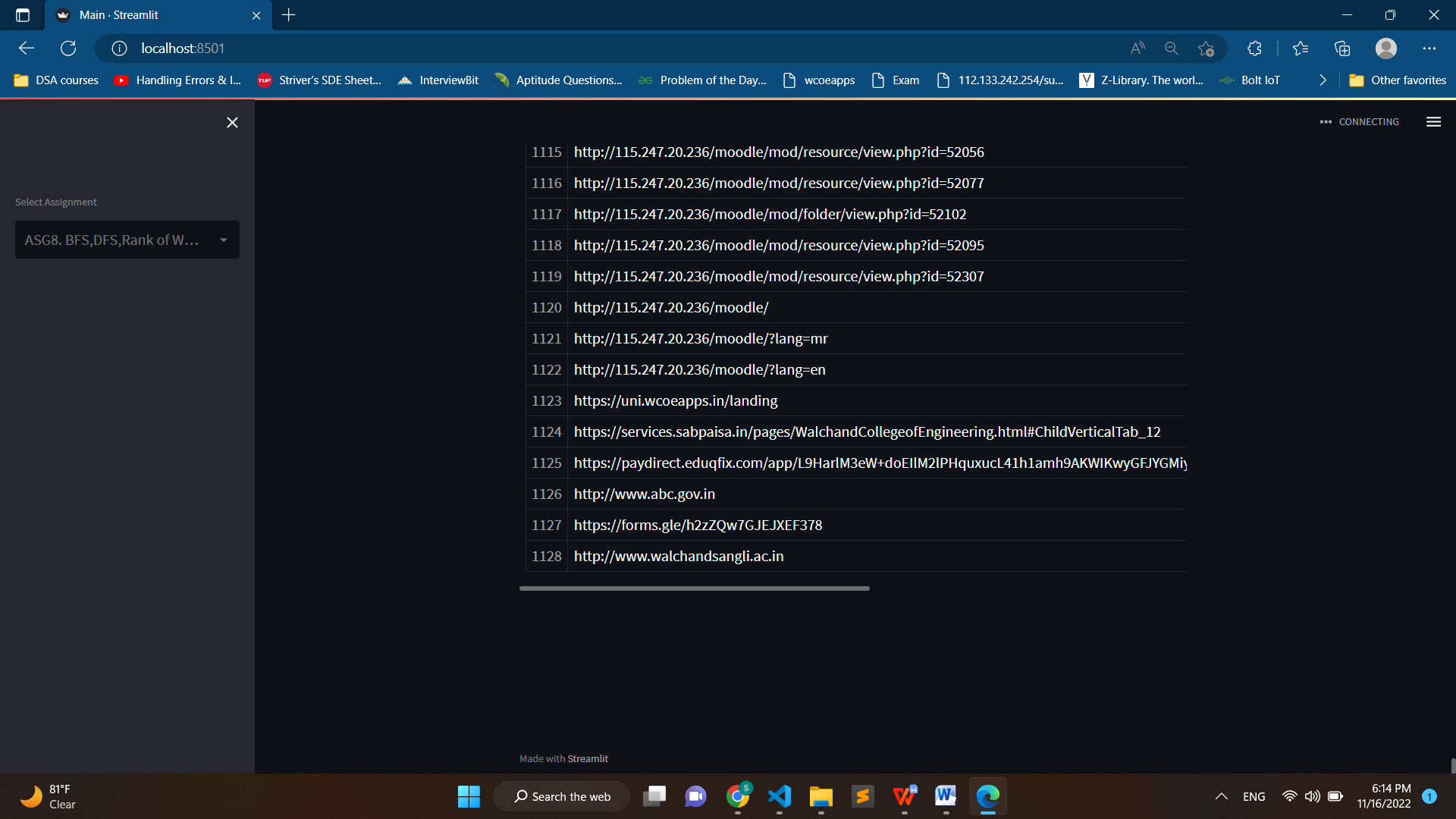
INSERT LINK

(Only for BFS, DFS)

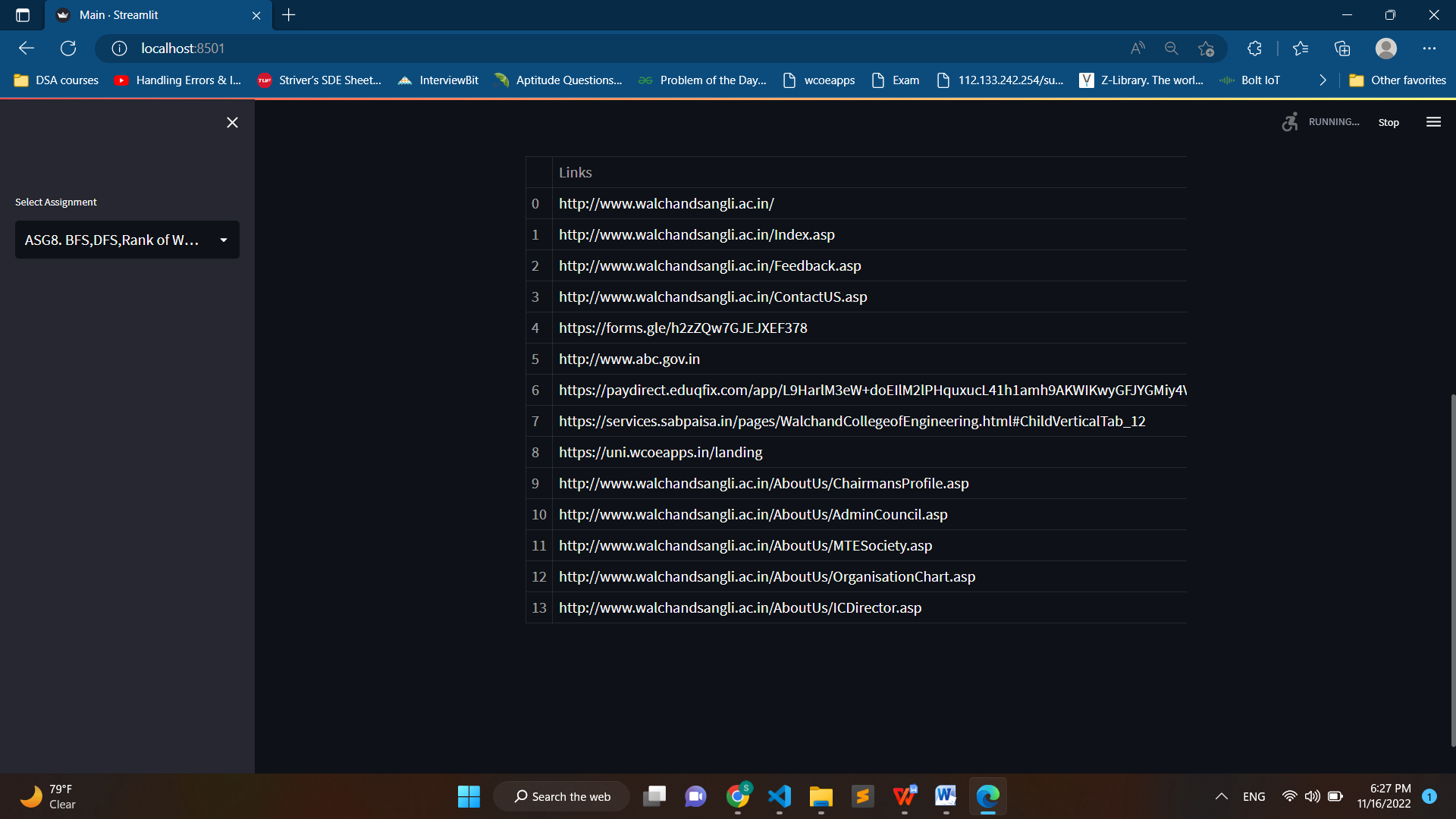
**Actual Experimentation/ simulation/ result/ Observation:**

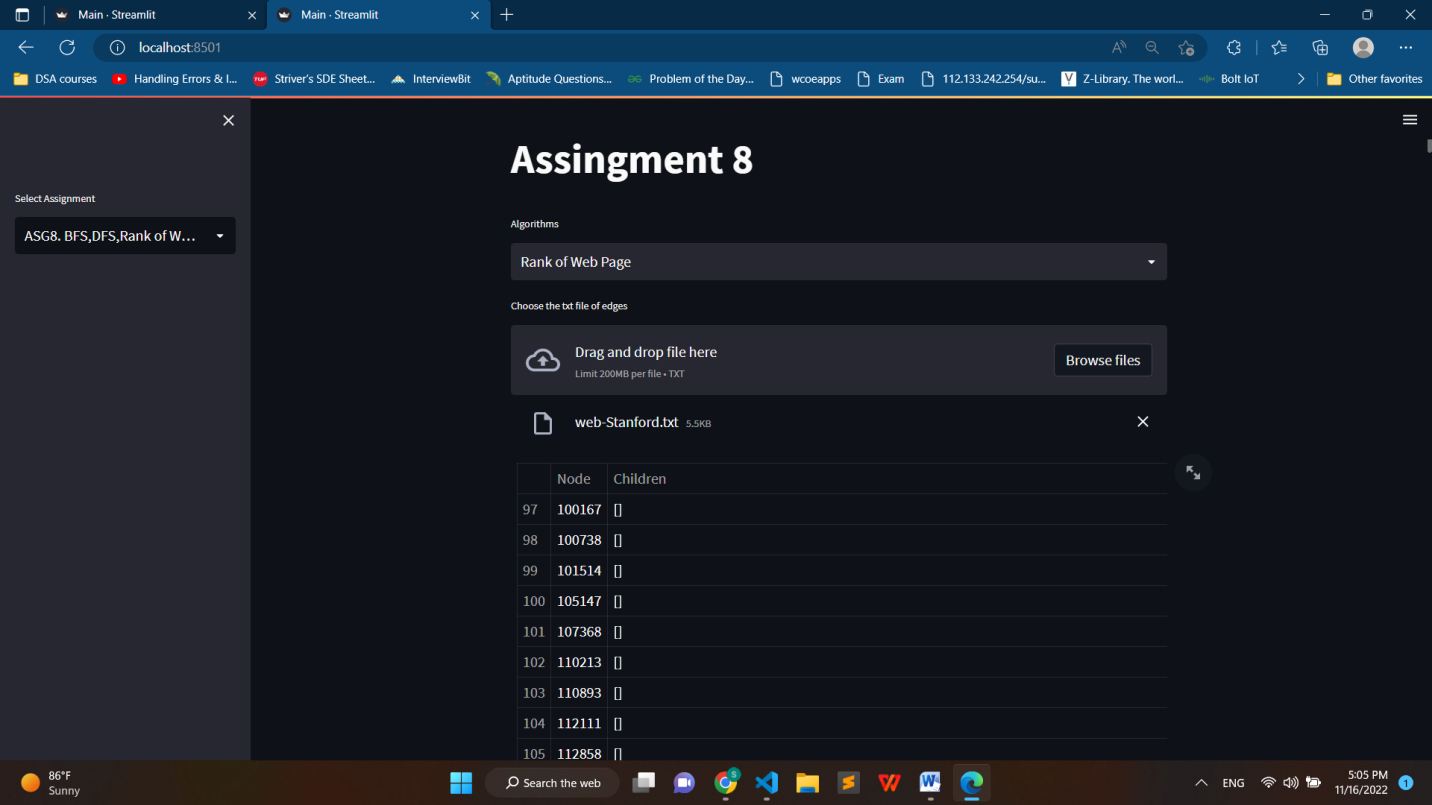
**BFS:**

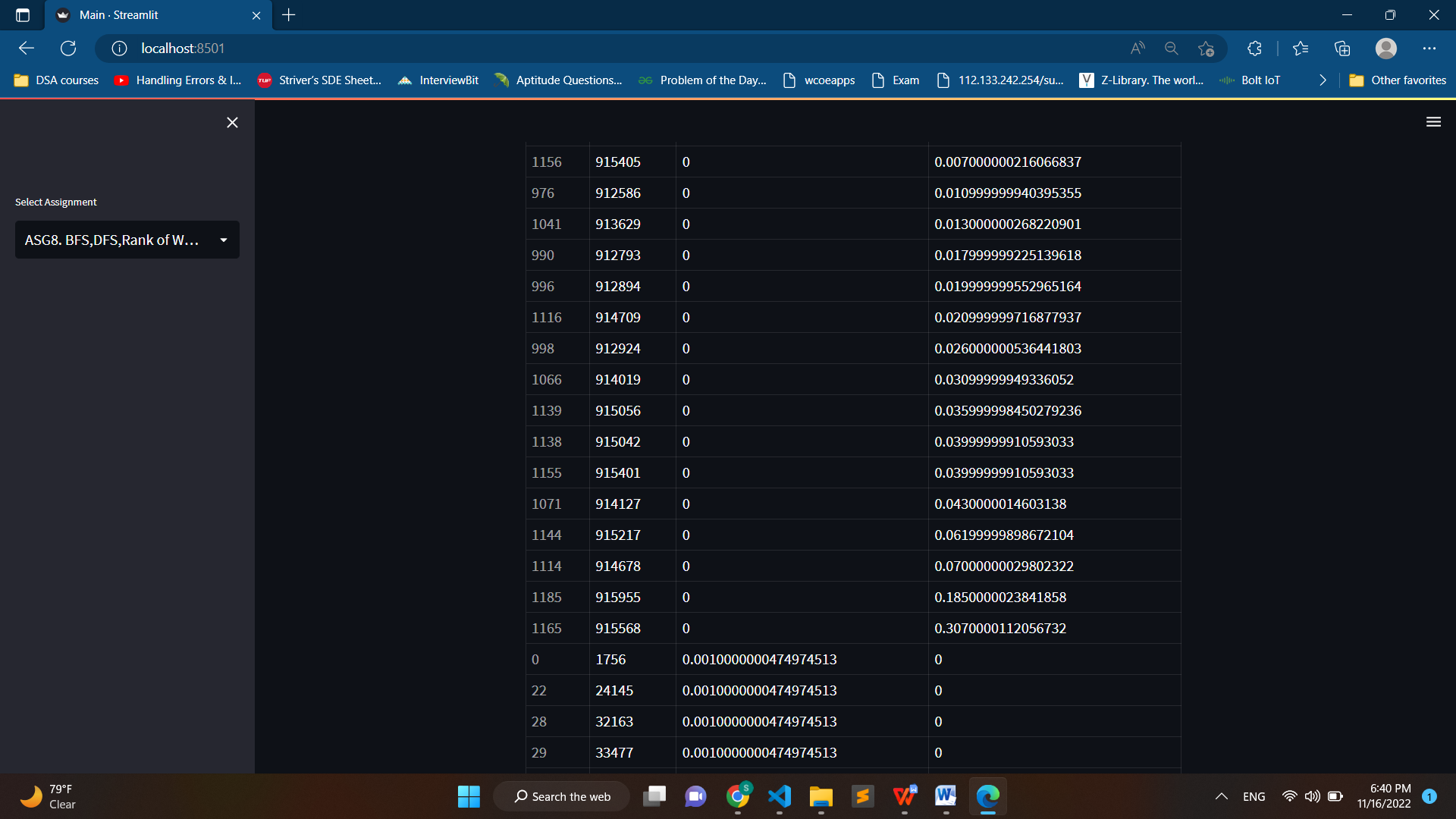
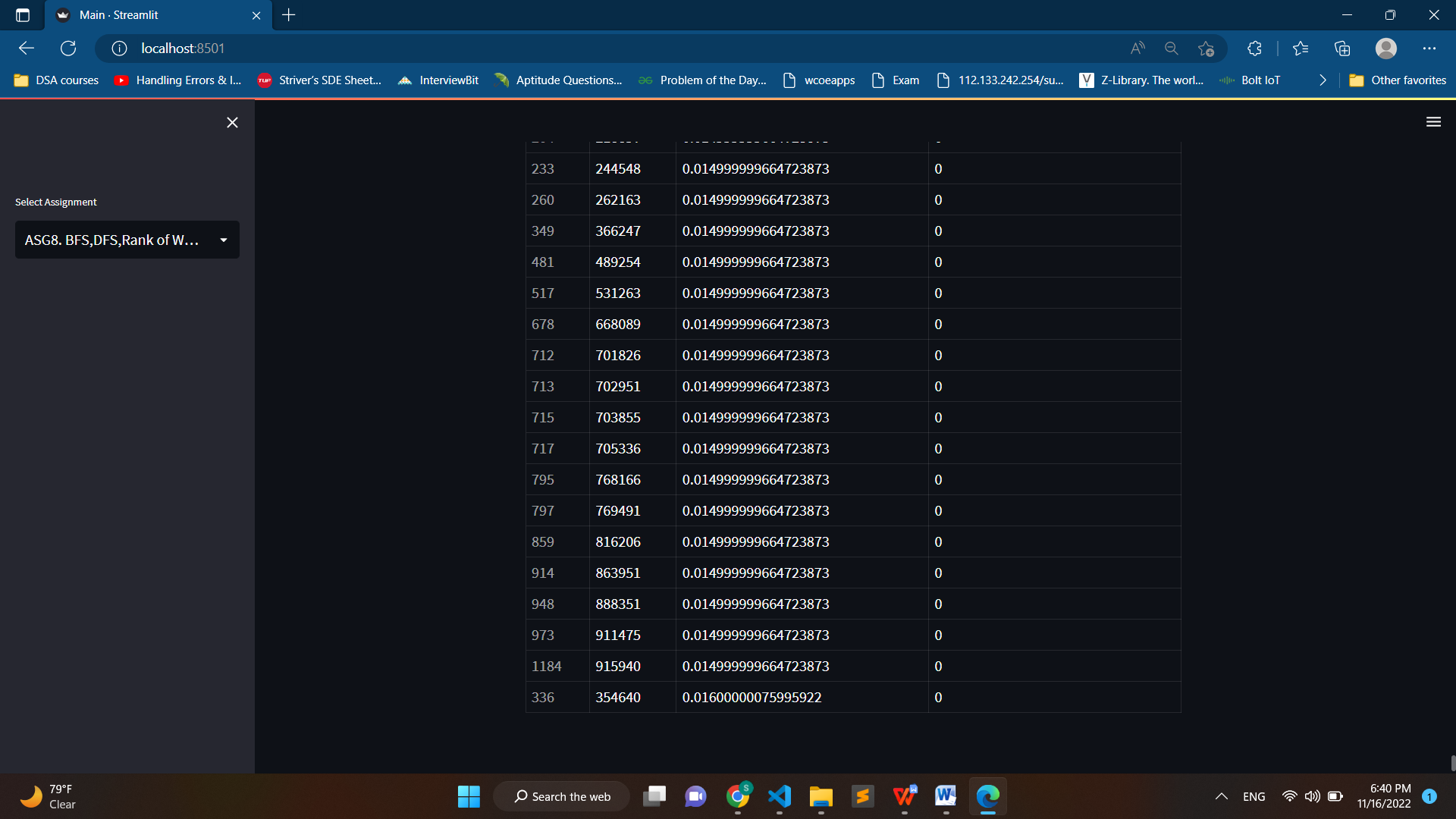
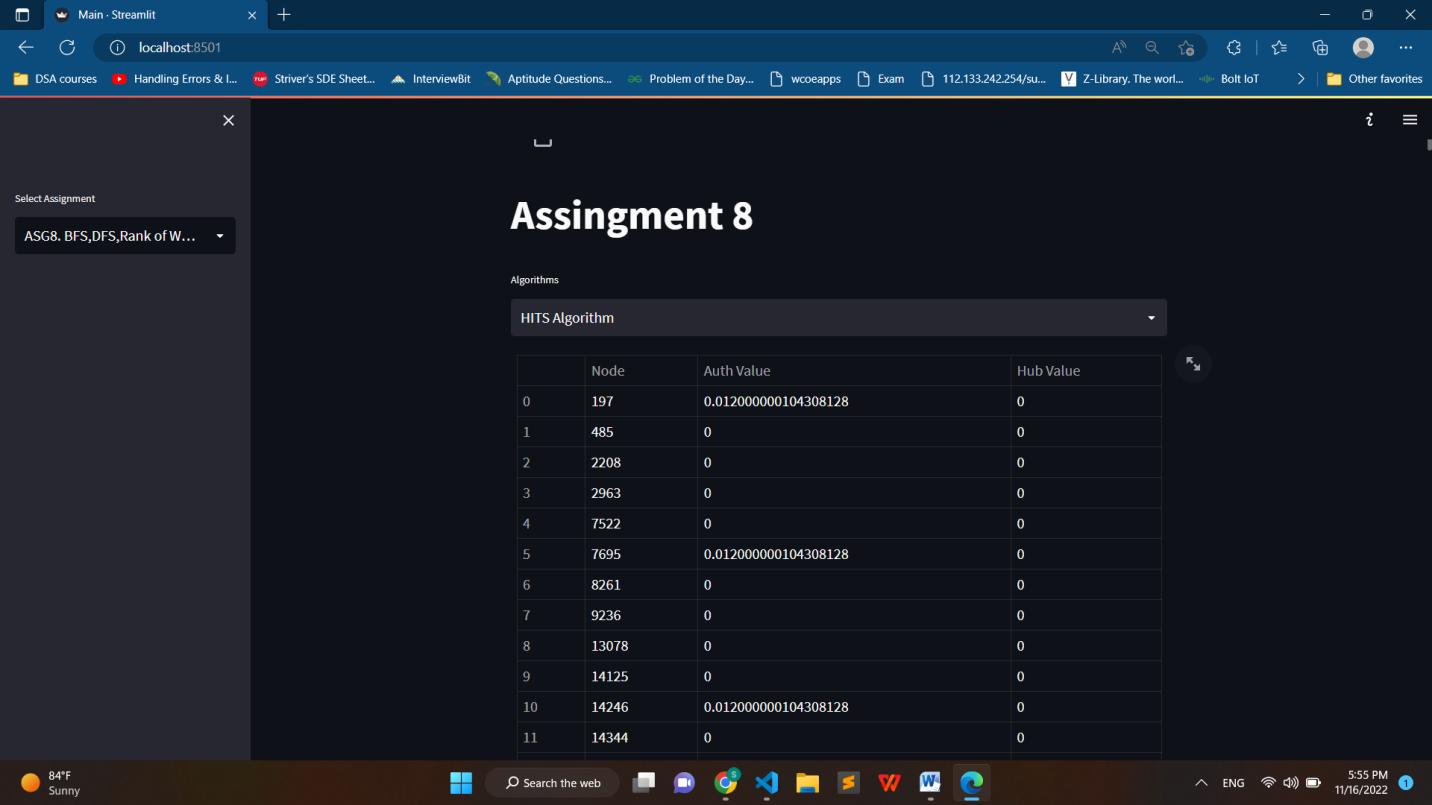
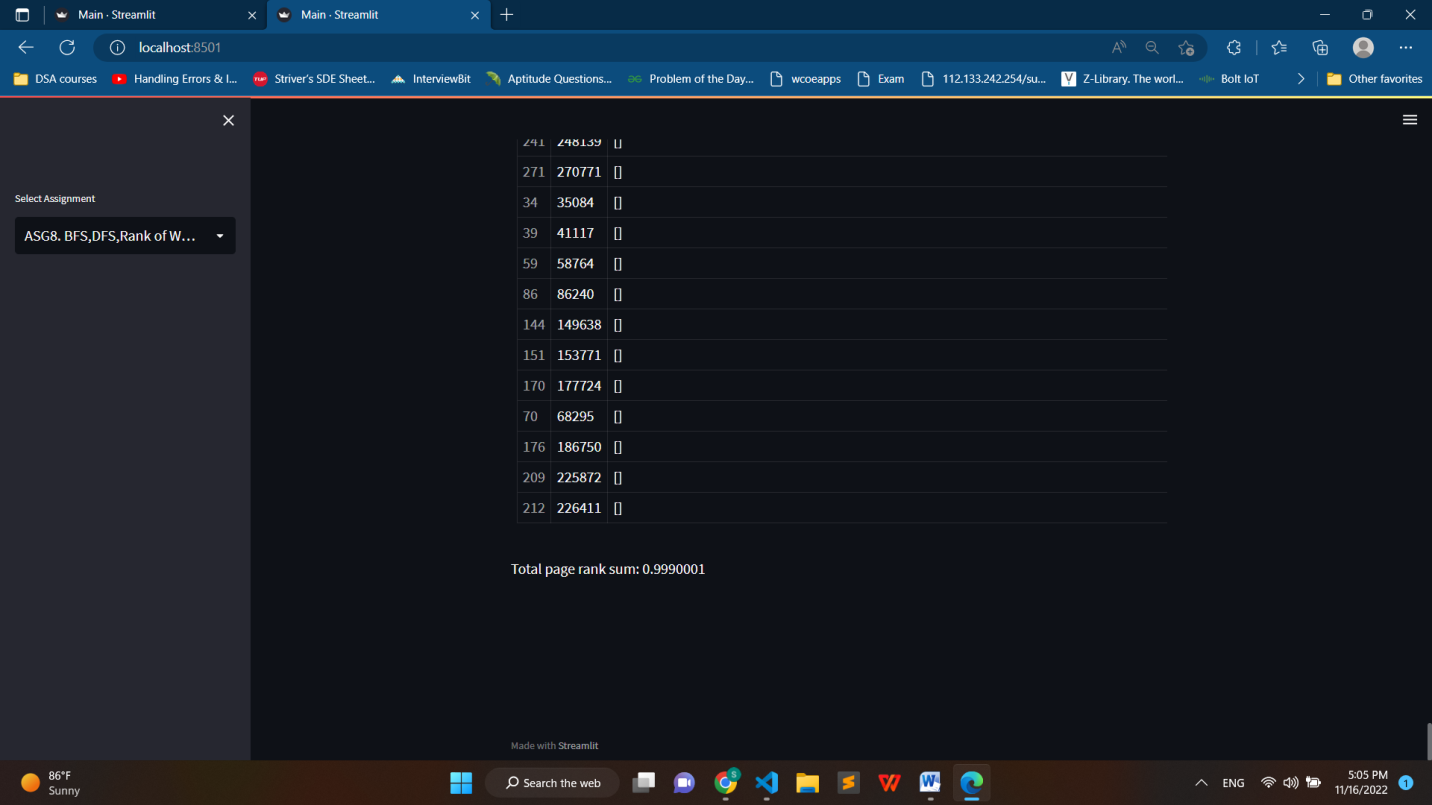
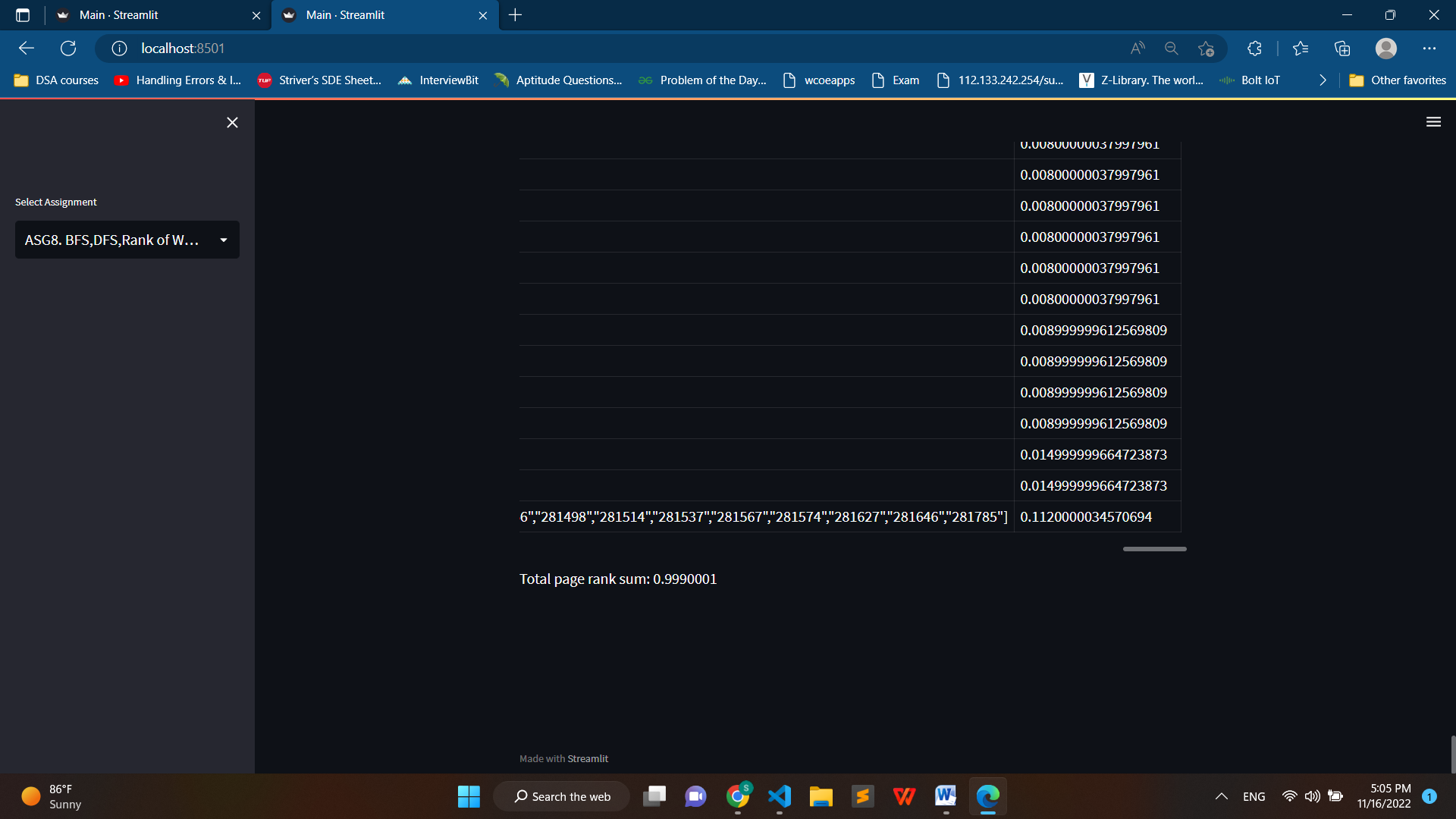




**DFS:**







**Conclusion:**

We can apply standard BFS, DFS algorithm for web crawler. Web Page ranking and HITS algorithm is used by google to rank web page after user search.

Reference:

<https://towardsdatascience.com/pagerank-3c568a7d2332>