**Final Year B. Tech. (CSE) – I: 2022-23**

**5CS462: PE5 - Data Mining Lab**

**Assignment No. 8**

**PRN: 2019BTECS00077 Date:02 oct 2022**

**Full name: Biradar Avinash Vishnu**

**----------------------------------------------------------------------------------------------------------**

**Title:** Design the data analysis tools (GUI) to perform the following task.

1. Implement web crawler (DFS/BFS)
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.
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.

**Objective/Aim:**

1. To implement data analysis tool using python programming language.
2. To design and implement web crawler
3. To design and implement PageRank algorithm
4. To design and implement HITS algorithm.

**Introduction:**

A web crawler, spider, or search engine [bot](https://www.cloudflare.com/learning/bots/what-is-a-bot/) downloads and indexes content from all over the Internet. The goal of such a bot is to learn what (almost) every webpage on the web is about, so that the information can be retrieved when it's needed. They're called "web crawlers" because crawling is the technical term for automatically accessing a website and obtaining data via a software program.

**Theory:**

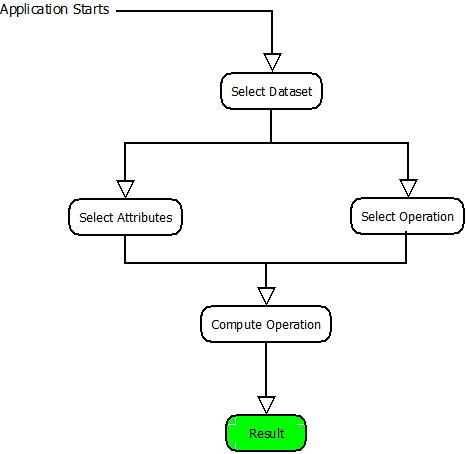
PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages.

**Algorithm**   
The PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank can be calculated for collections of documents of any size. It is assumed in several research papers that the distribution is evenly divided among all documents in the collection at the beginning of the computational process. The PageRank computations require several passes, called “iterations”, through the collection to adjust approximate PageRank values to more closely reflect the theoretical true value.

Hyperlink Induced Topic Search (HITS) Algorithm is a Link Analysis Algorithm that rates webpages, developed by Jon Kleinberg. This algorithm is used to the web link-structures to discover and rank the webpages relevant for a particular search.   
HITS uses hubs and authorities to define a recursive relationship between webpages. Before understanding the HITS Algorithm, we first need to know about Hubs and Authorities.

* Given a query to a Search Engine, the set of highly relevant web pages are called Roots. They are potential Authorities.
* Pages that are not very relevant but point to pages in the Root are called Hubs. Thus, an Authority is a page that many hubs link to whereas a Hub is a page that links to many authorities.

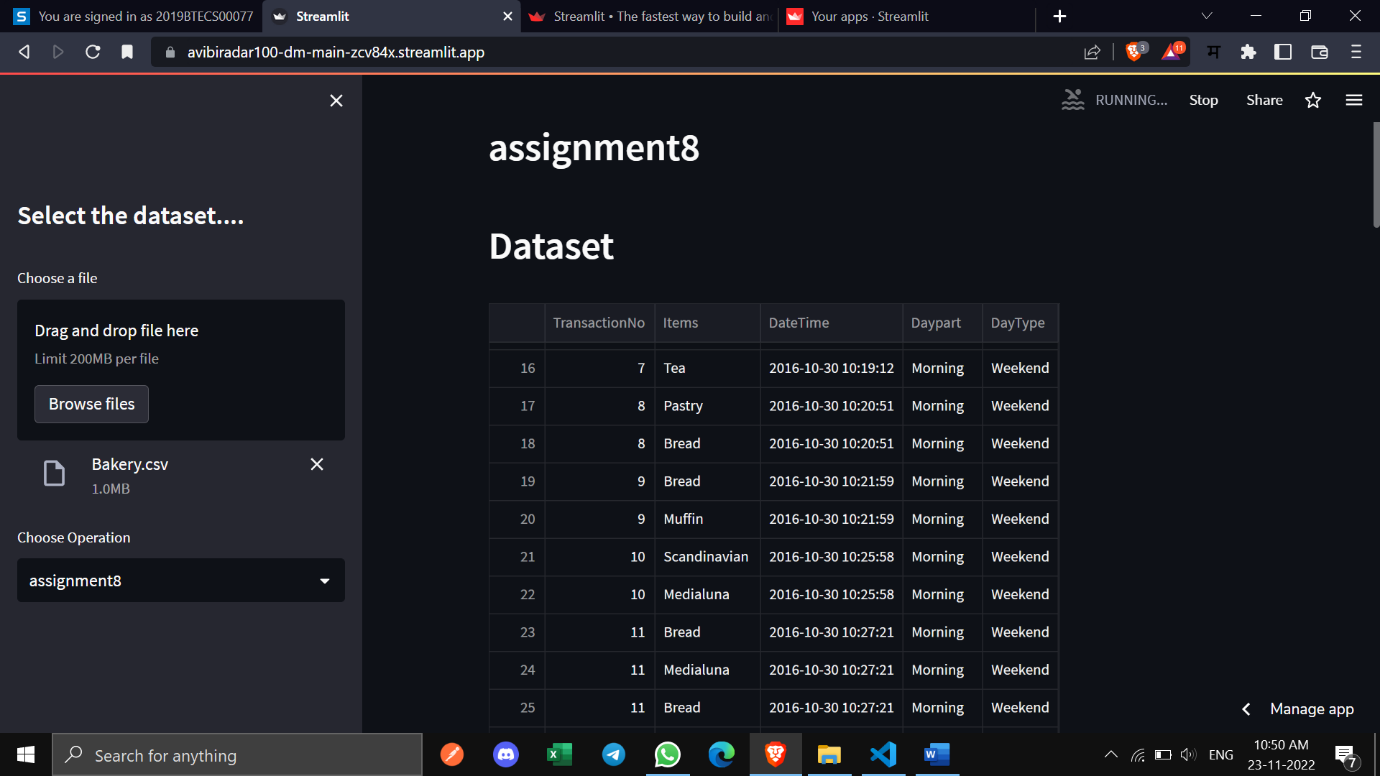
**Functional Block Diagram:**

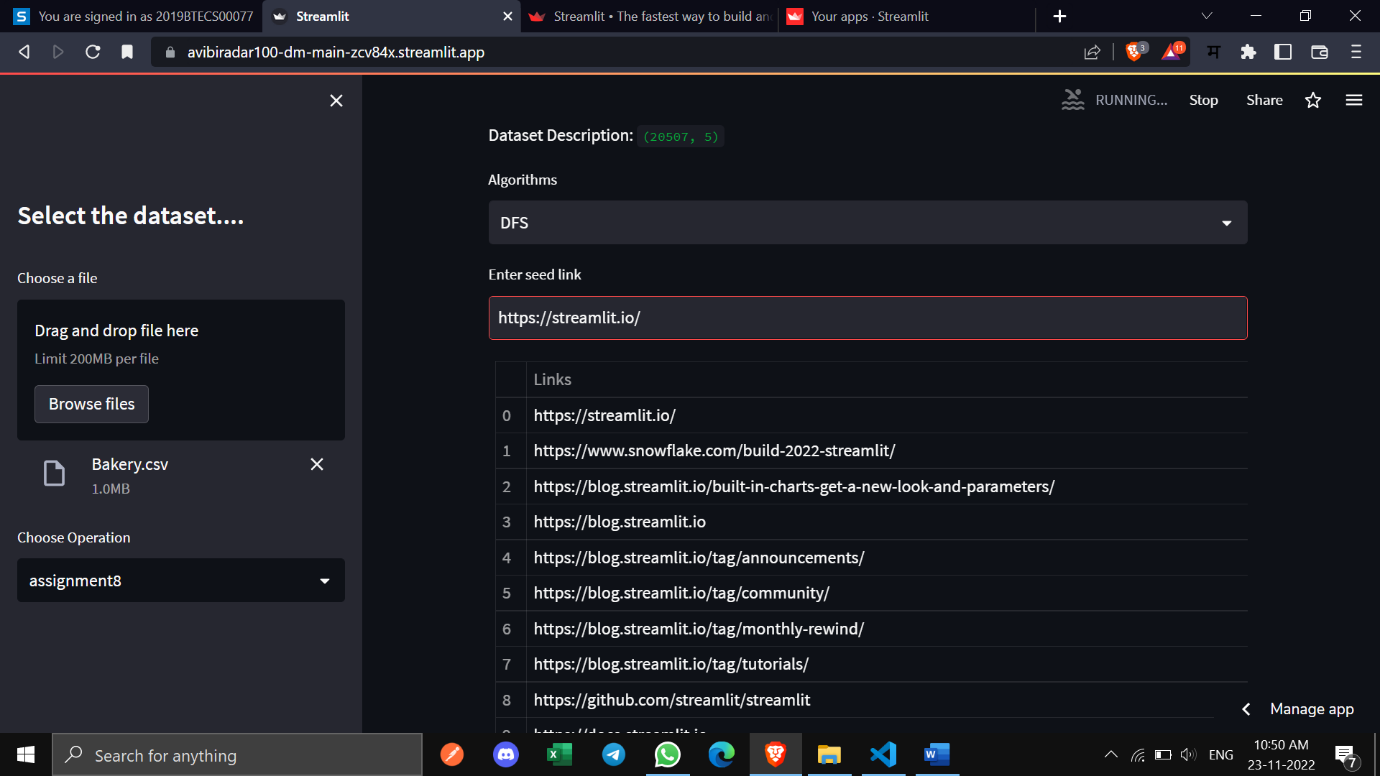


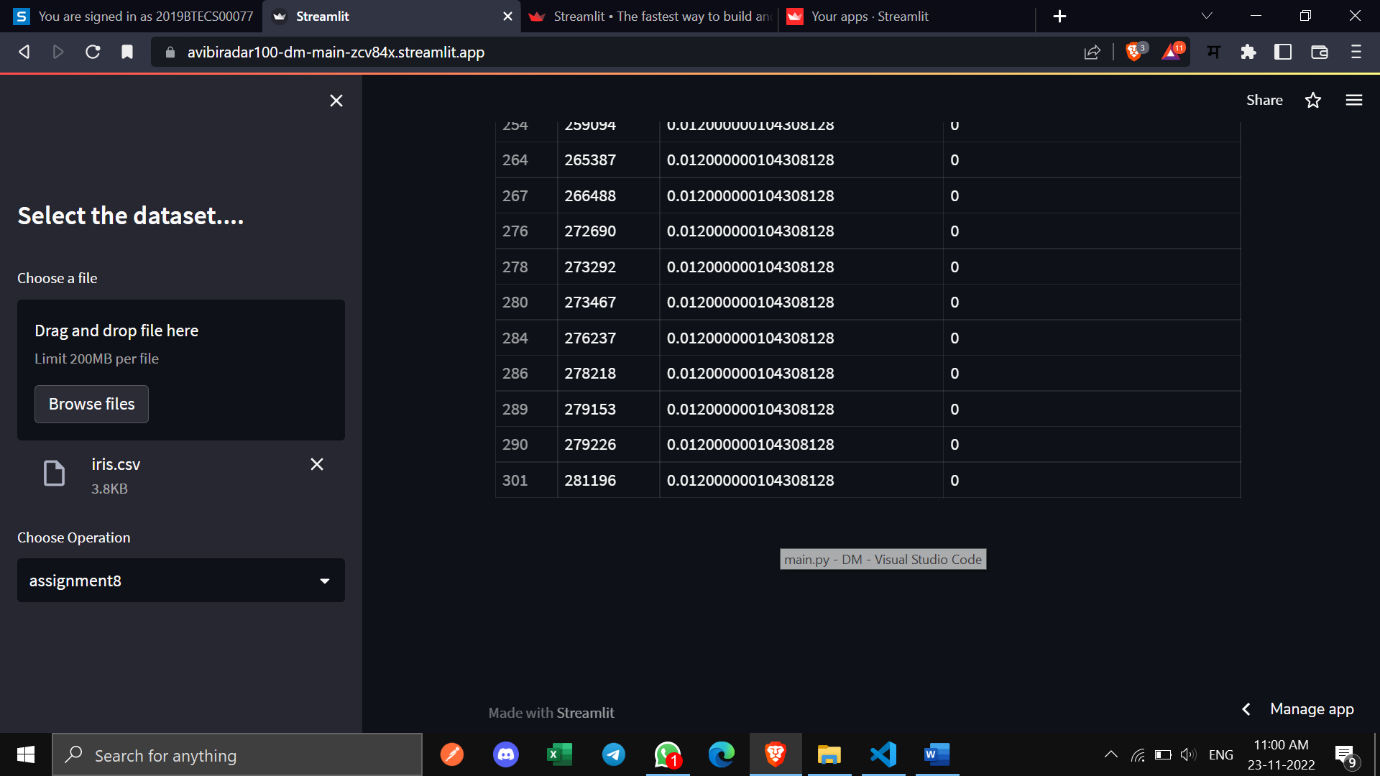
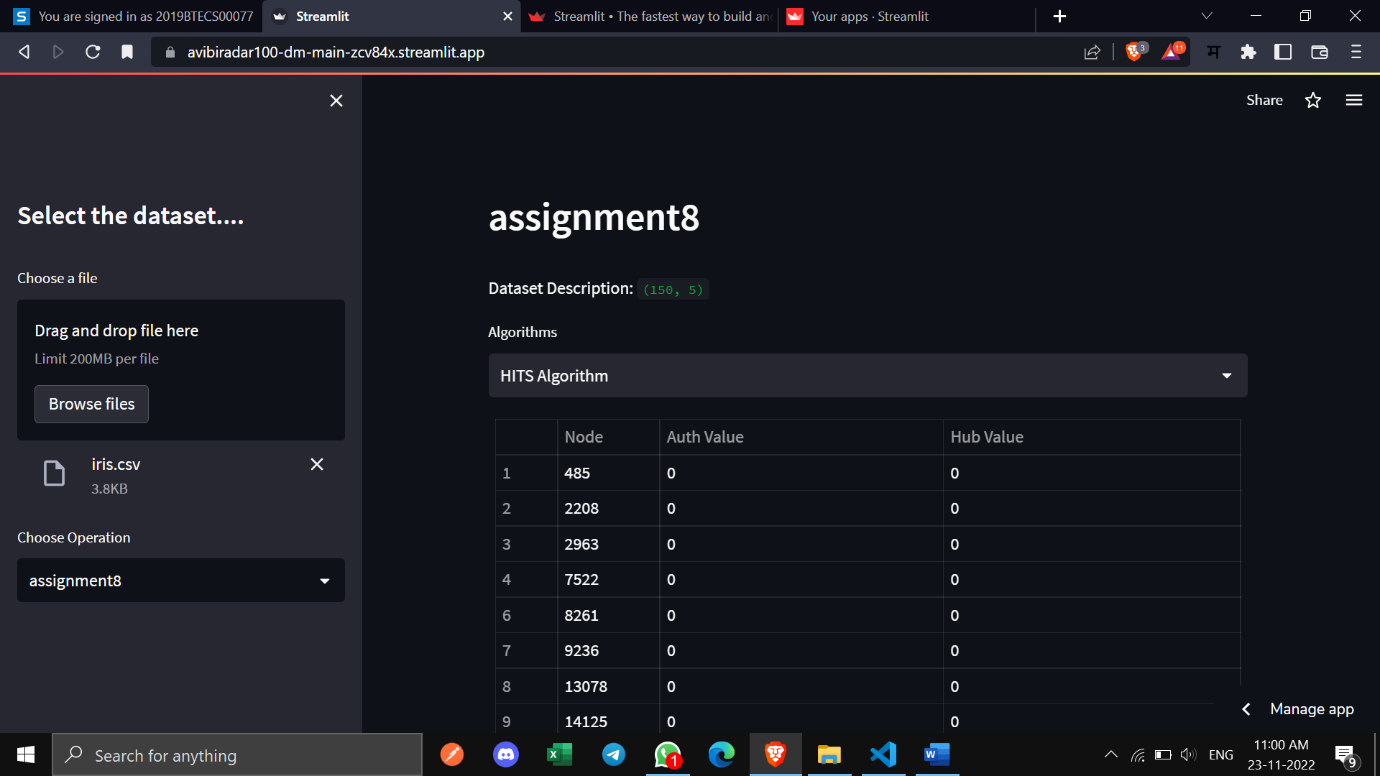
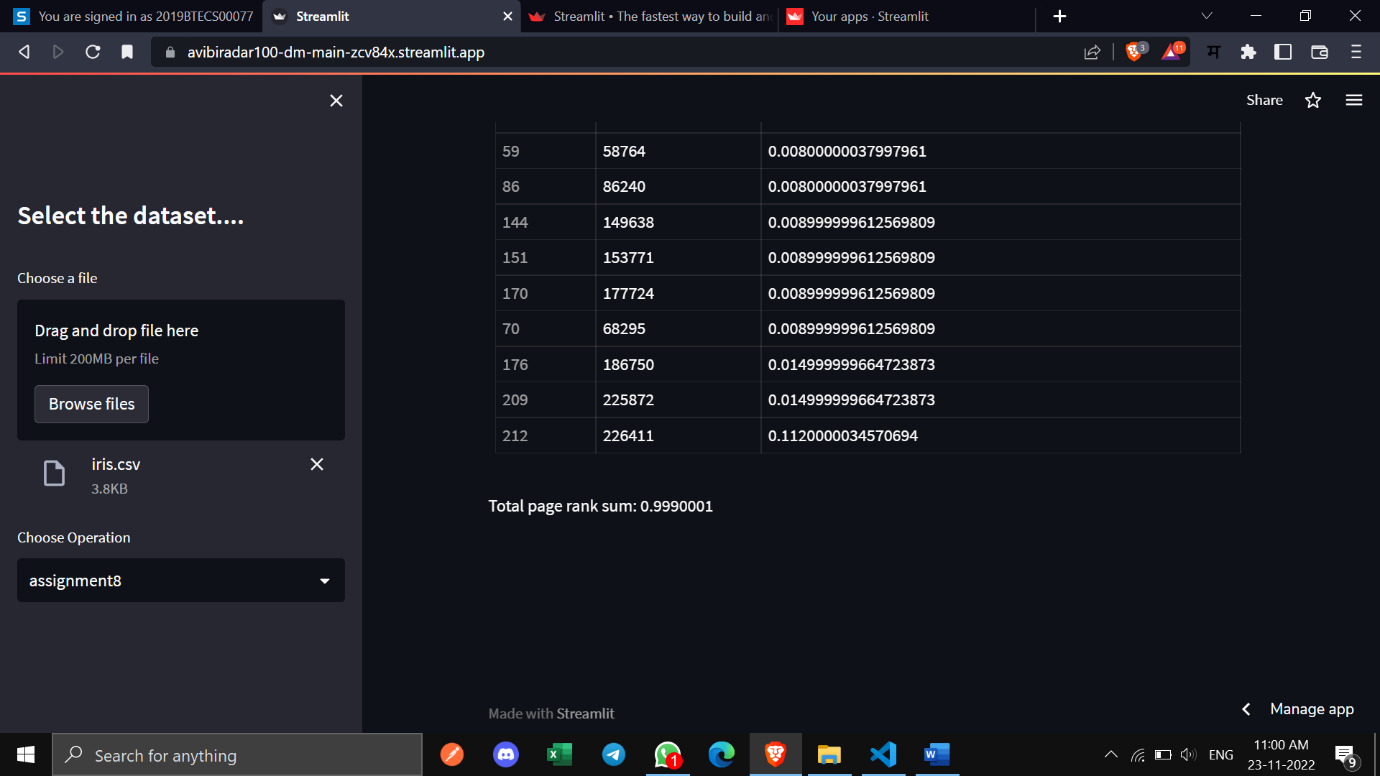
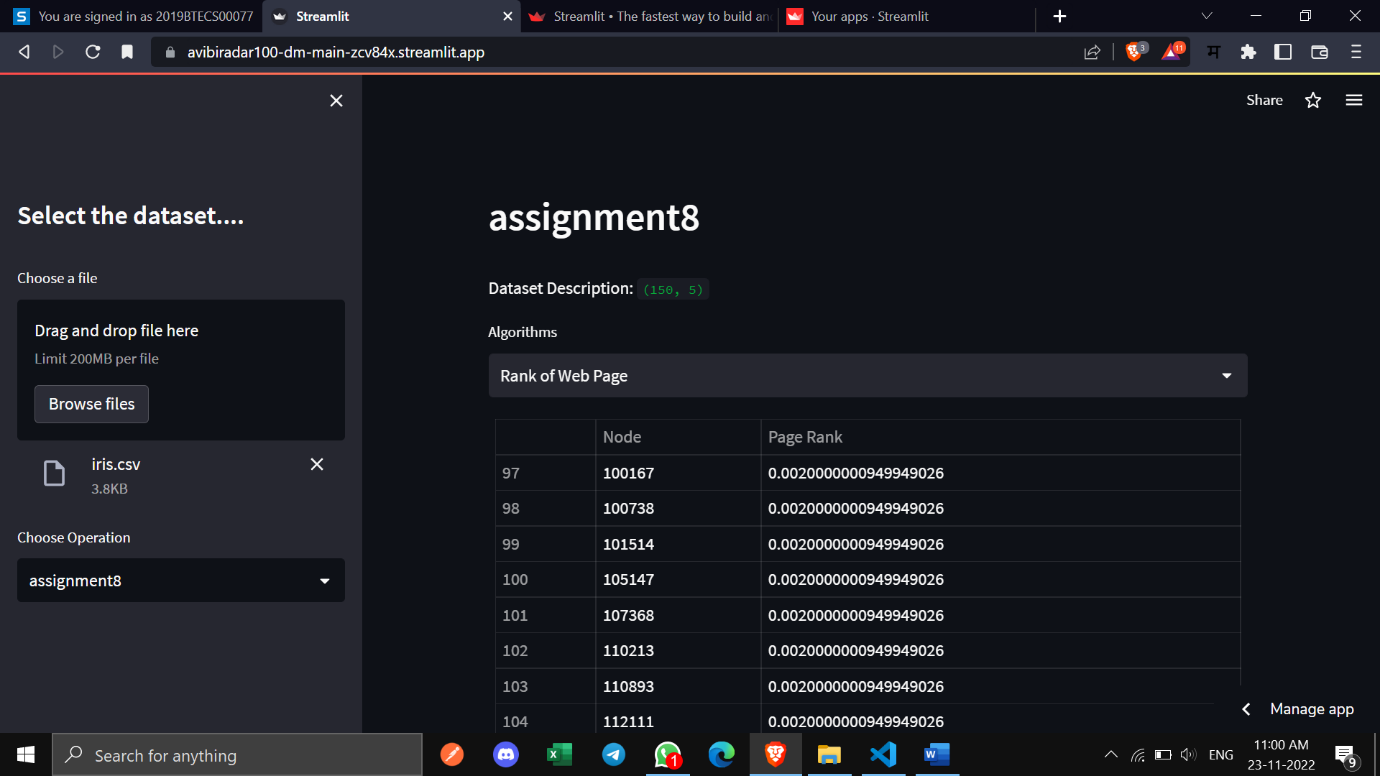
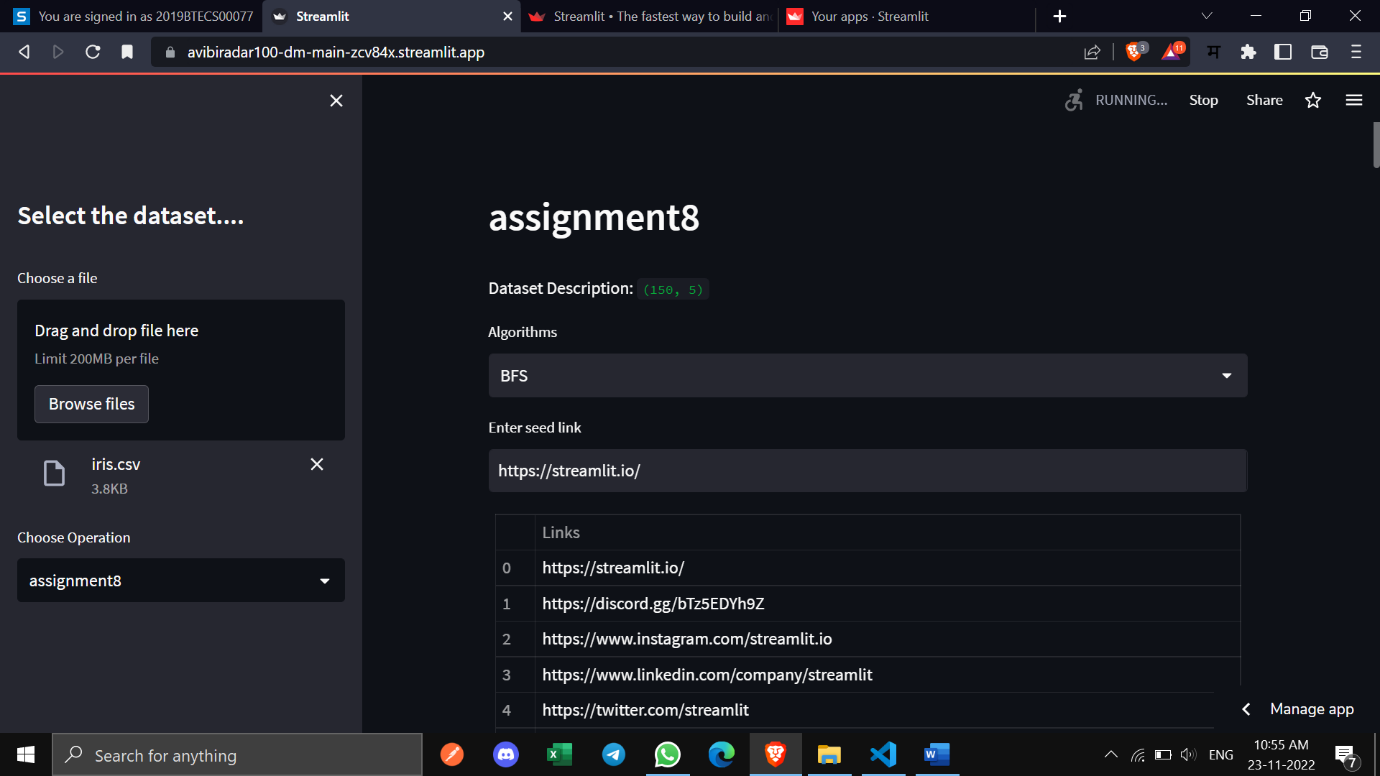
**Procedure:**

Given problem statement is solved using python programming language and specifically used tkinter module to implement GUI application and pandas module to load .csv file as dataset.

**Actual Experiments/Screenshots:**

****

****

****

**Conclusion:**

Successfully implemented data analysis tool (GUI) for implementation of web crawler algorithm.