**Final Year B. Tech. (CSE) – I : 2021-22**

**4CS462 : PE2 - Data Mining Lab**

**Assignment No. 9**

**Group id: DM21G12**

**Group members:**

**Abhishek More(2018BTECS00037)**

**Sushil Wagh(2018BTECS00031)**

**Title :** Pagerank and HITS algorithm

**Objective/Aim :**

1. 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.

2. 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.

3. Tabulate the results containing adjacency matrix and rank of pages.

**Introduction:**

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.

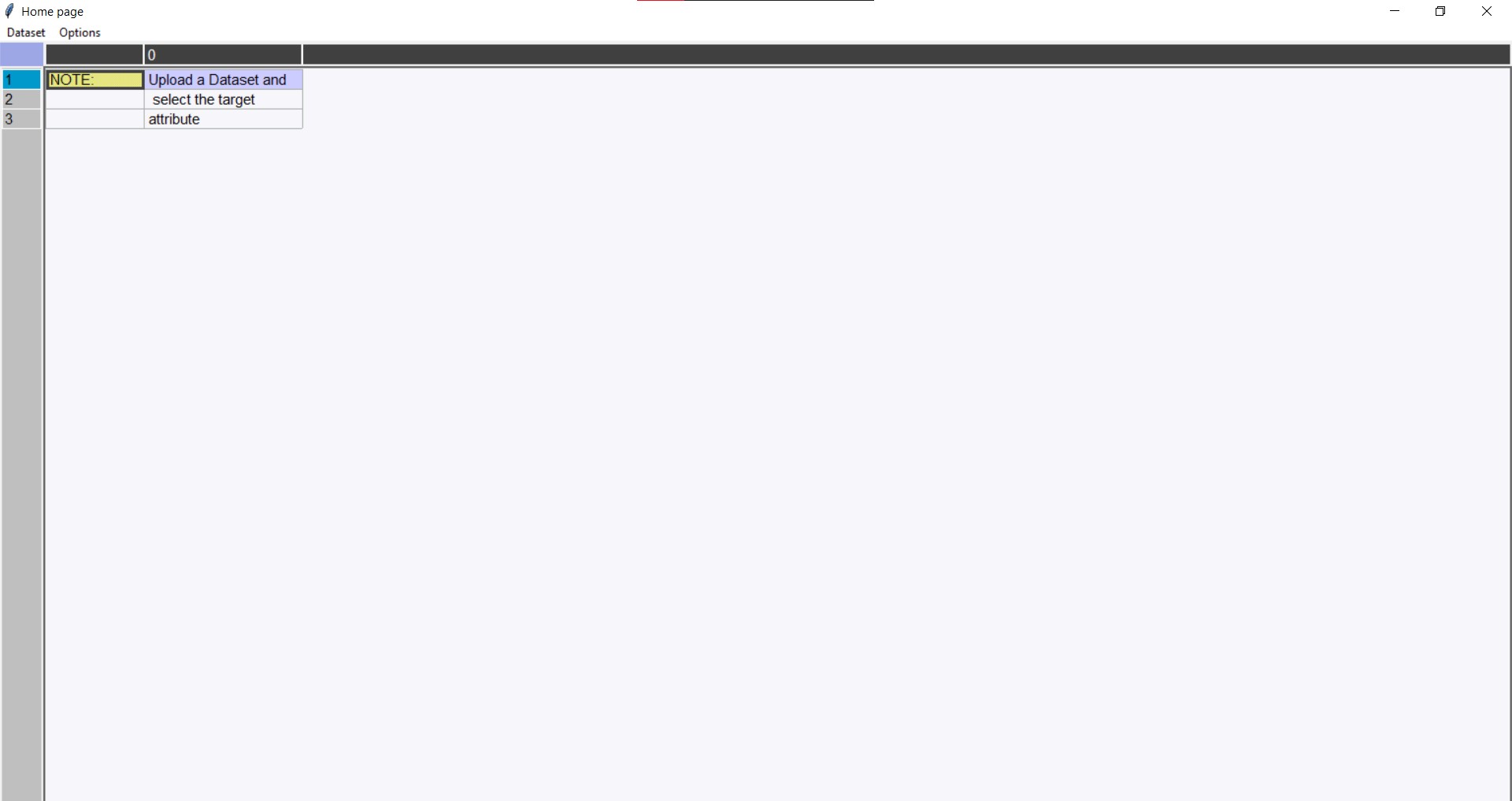
PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

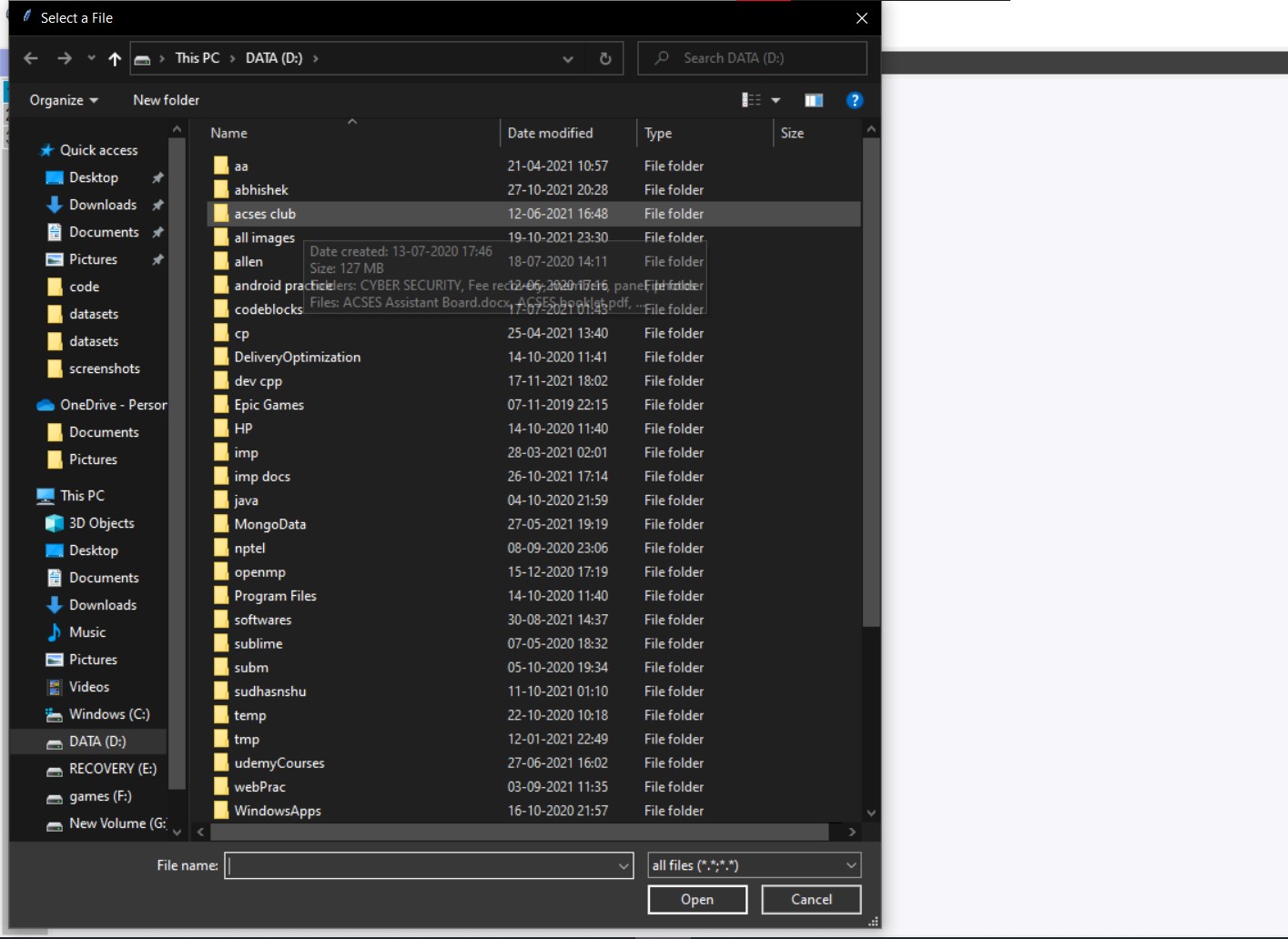
Hyperlink Induced Topic Search (HITS) Algorithm is a Link Analysis Algorithm that rates webpages, developed by Jon Kleinberg. This algorithm is used to the web link-structures to discover and rank the webpages relevant for a particular search.

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.

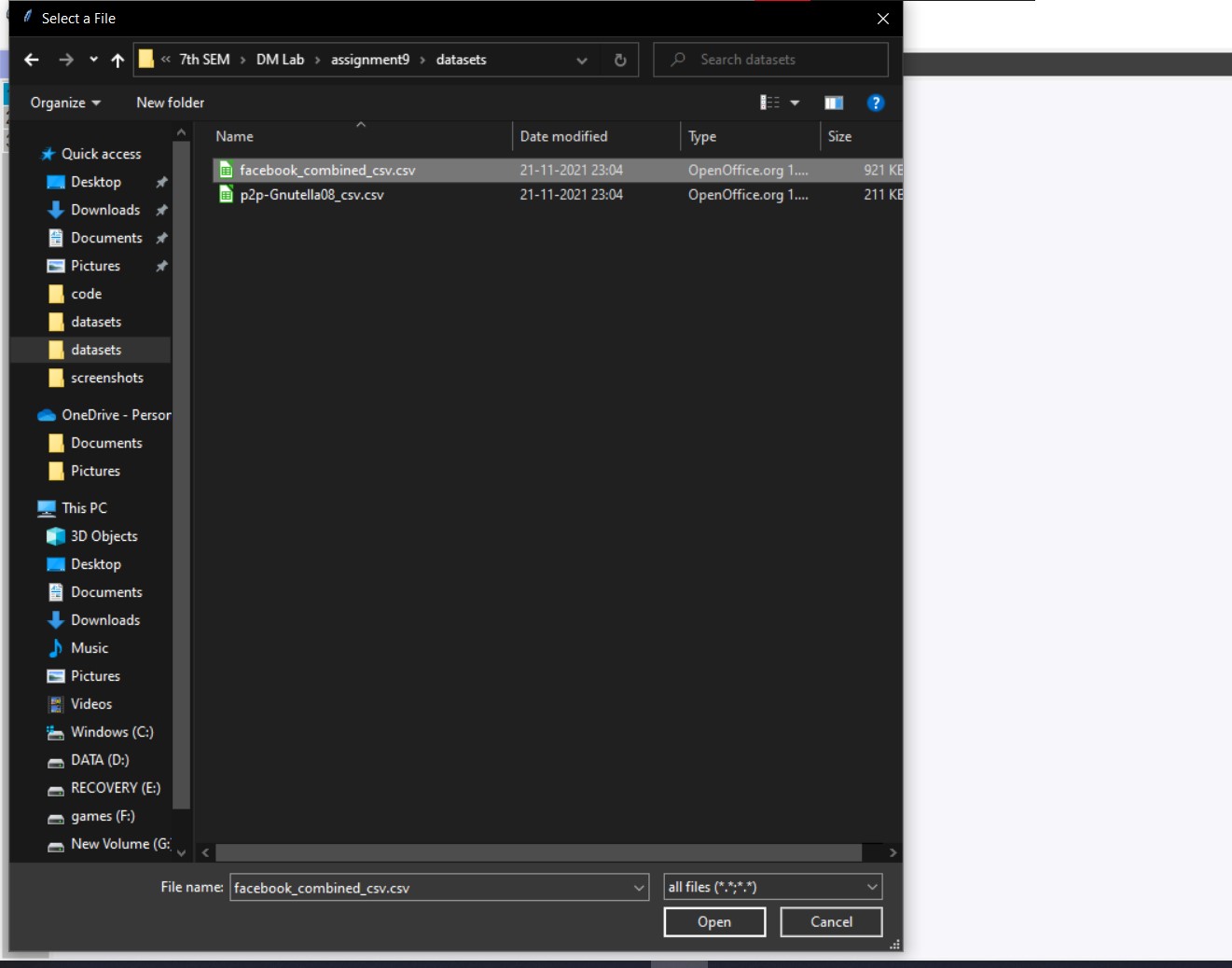
* Given a query to a Search Engine, the set of highly relevant web pages are called Roots. They are potential Authorities.
* Pages that are not very relevant but point to pages in the Root are called Hubs. Thus, an Authority is a page that many hubs link to whereas a Hub is a page that links to many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

**Result/Observations/Screenshots:**

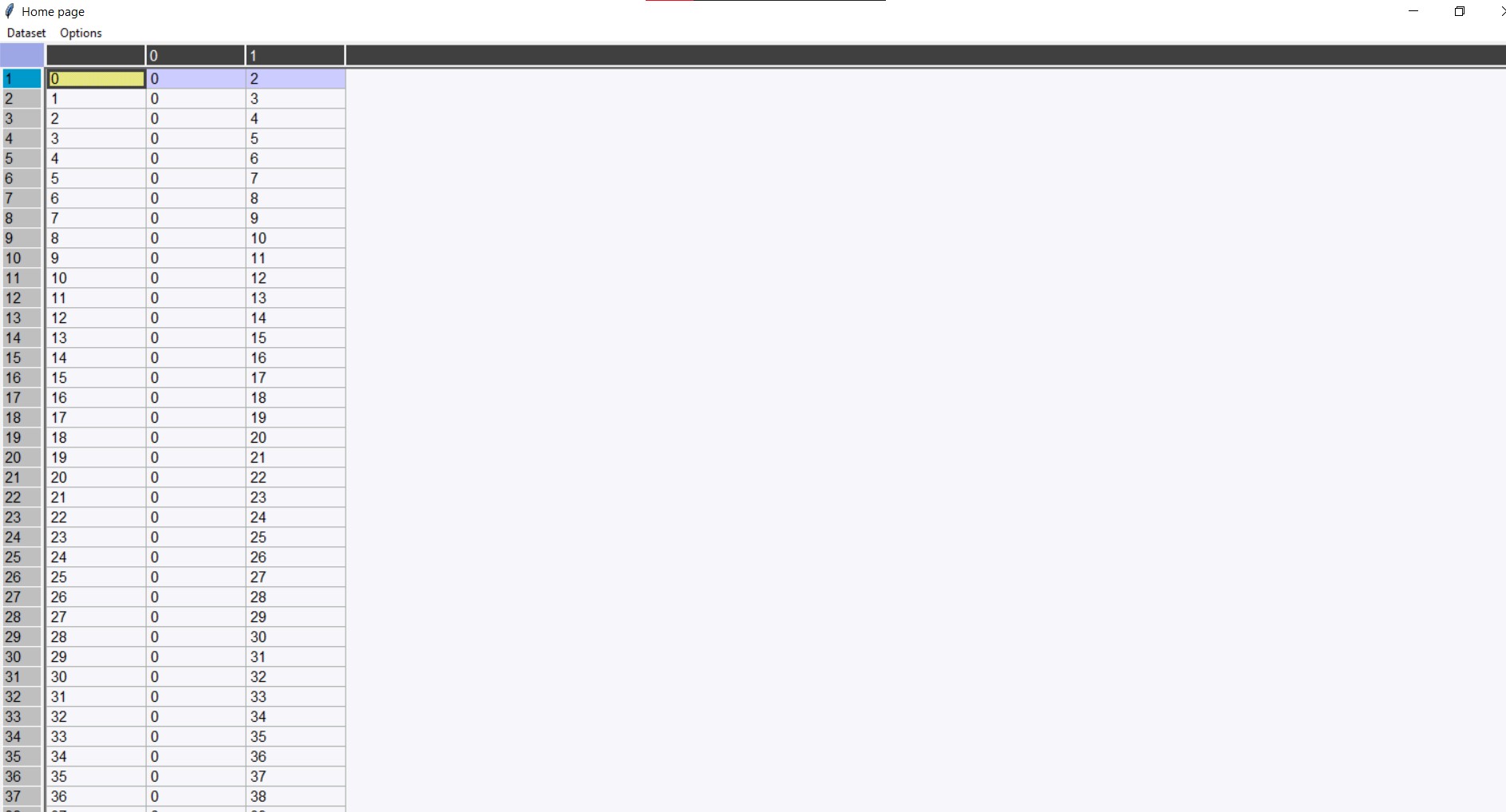
****

****

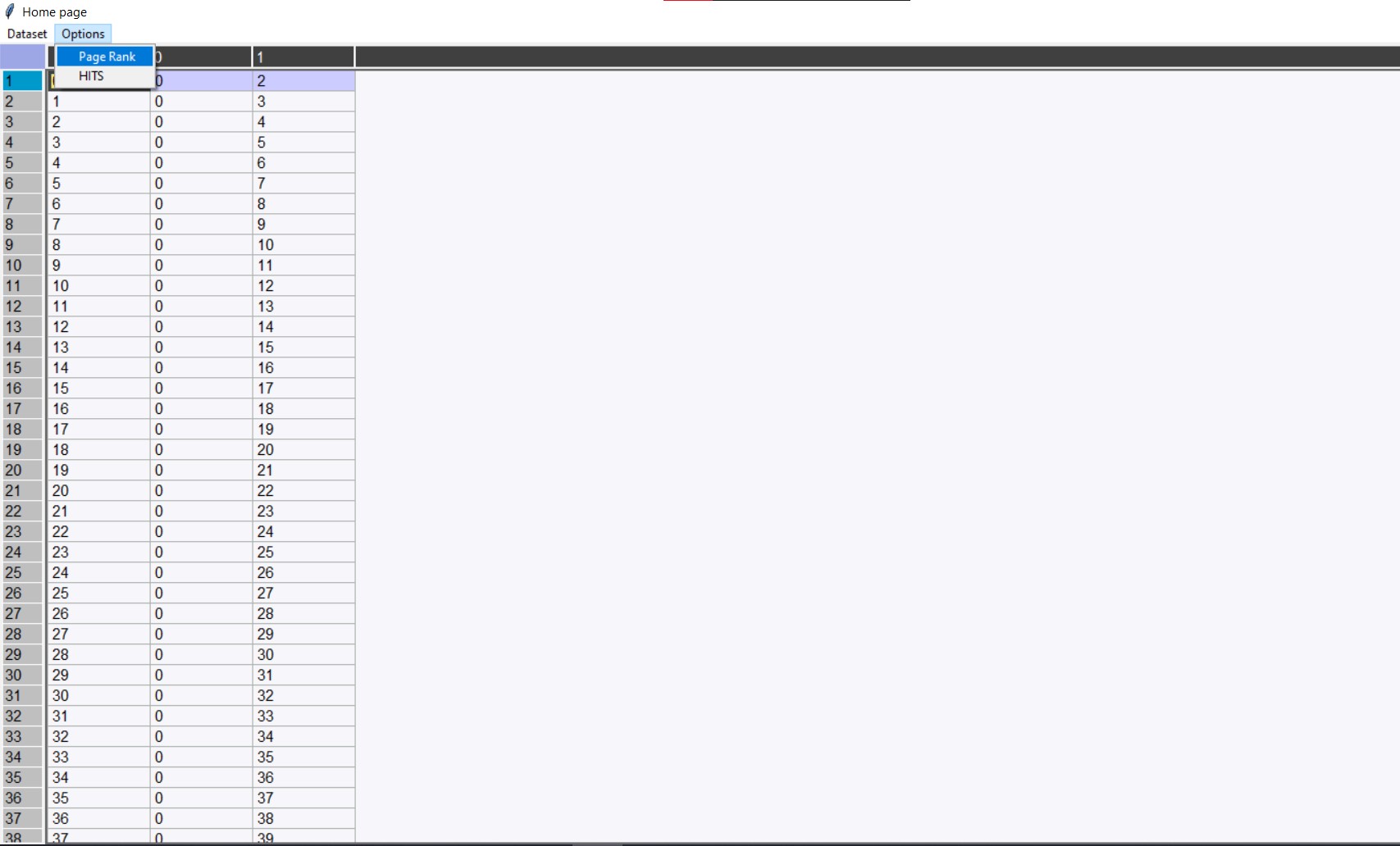
**Select dataset**

****

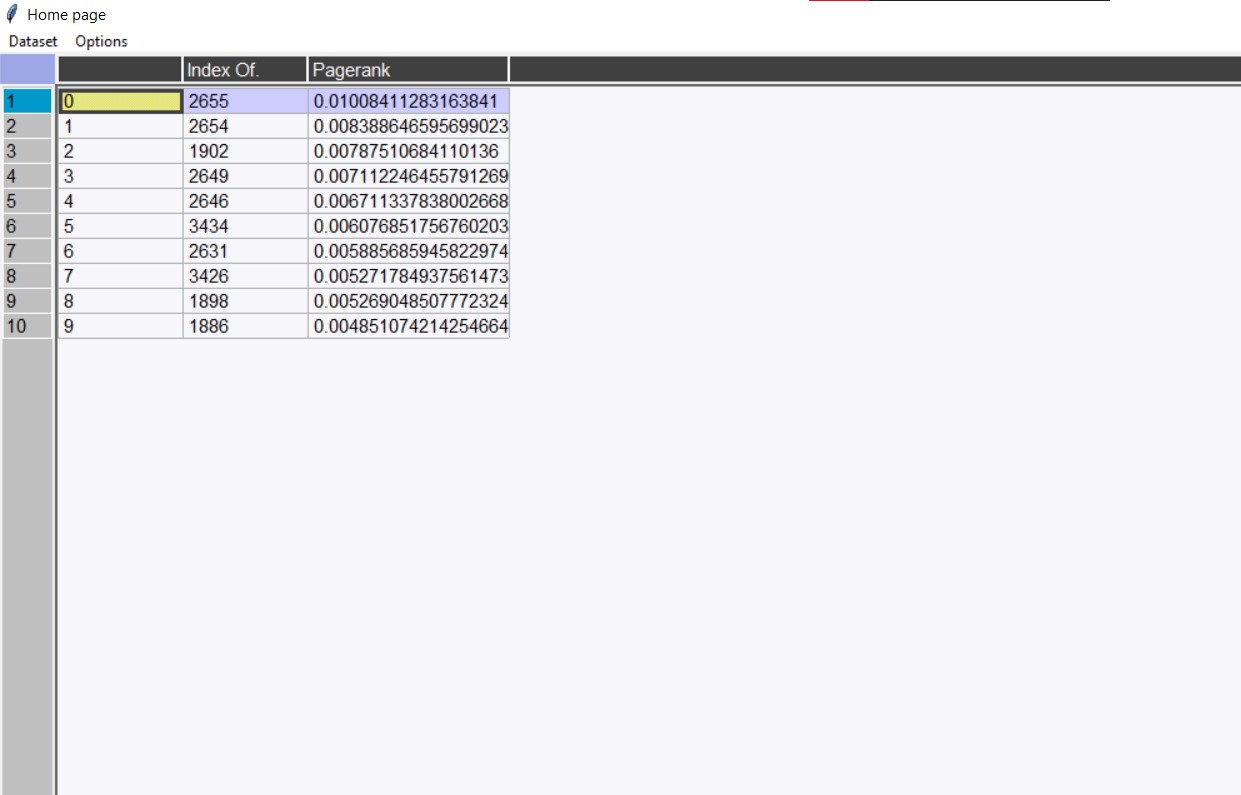
**Displaying dataset**

****

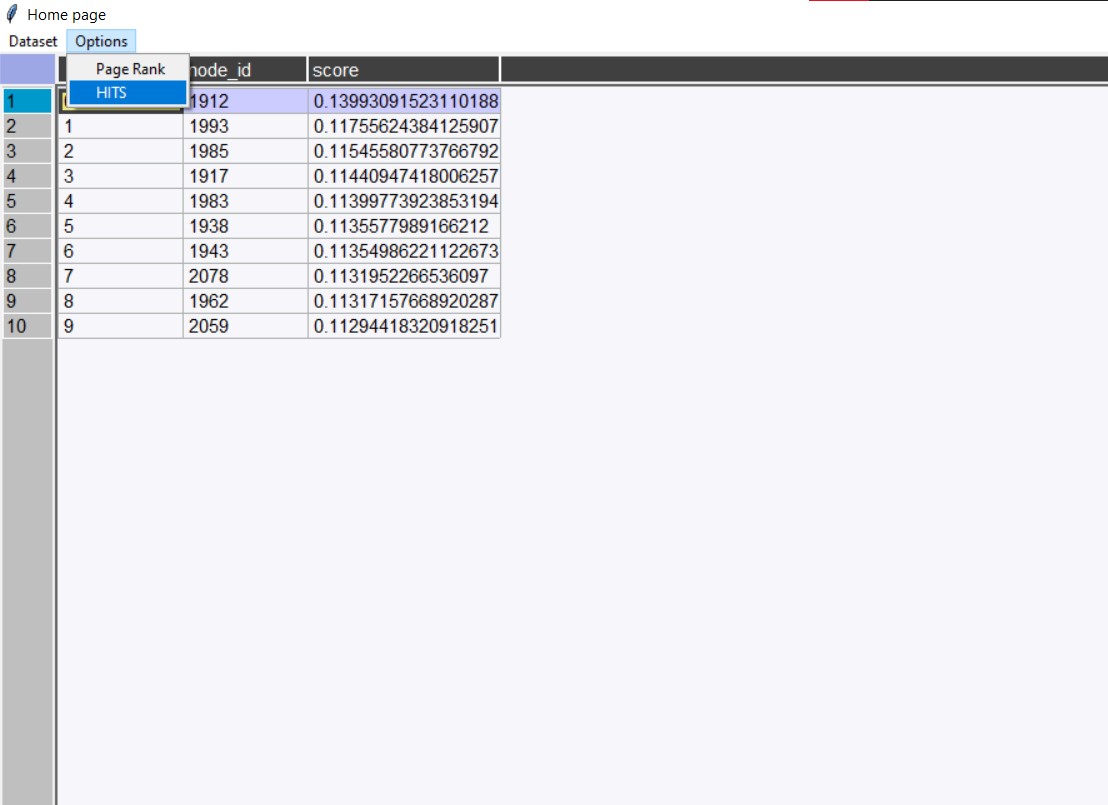
**Choosing PageRank**

****

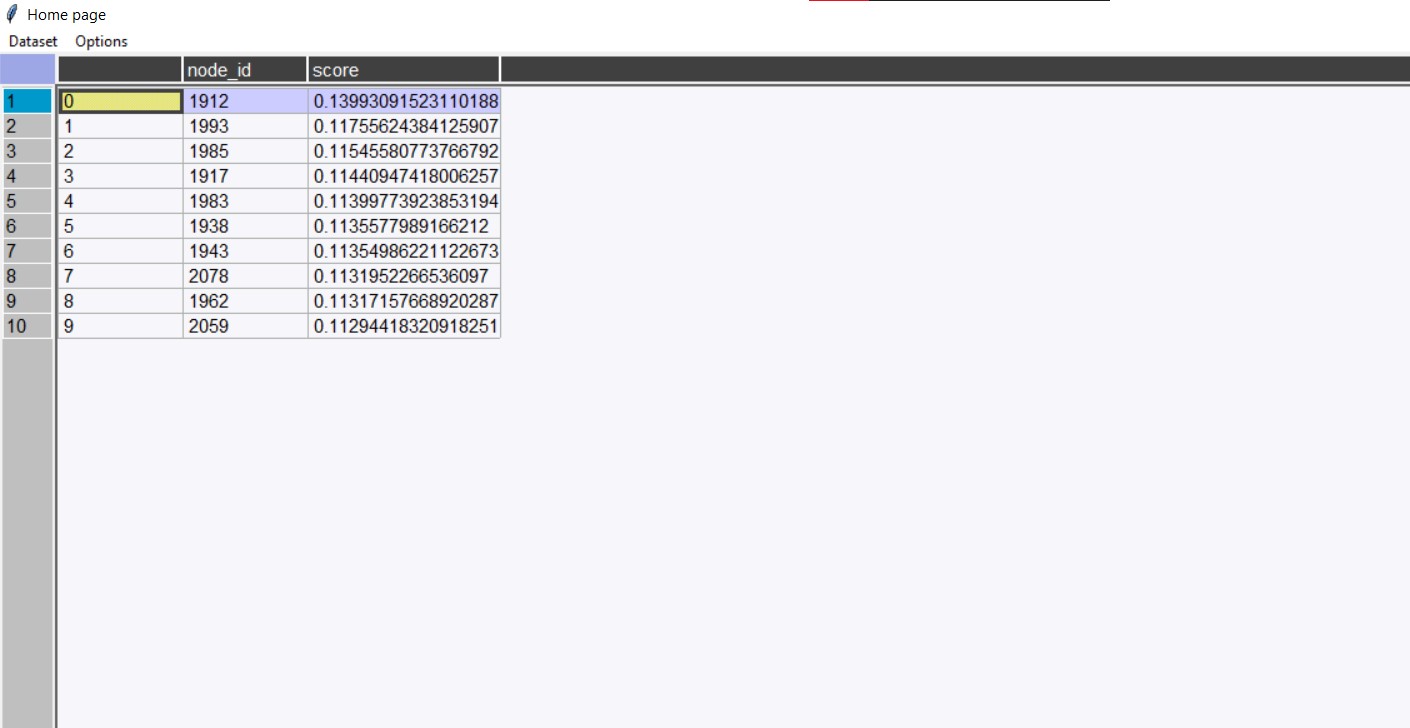
**Output for Page Rank ie.Top 10 pages**

****

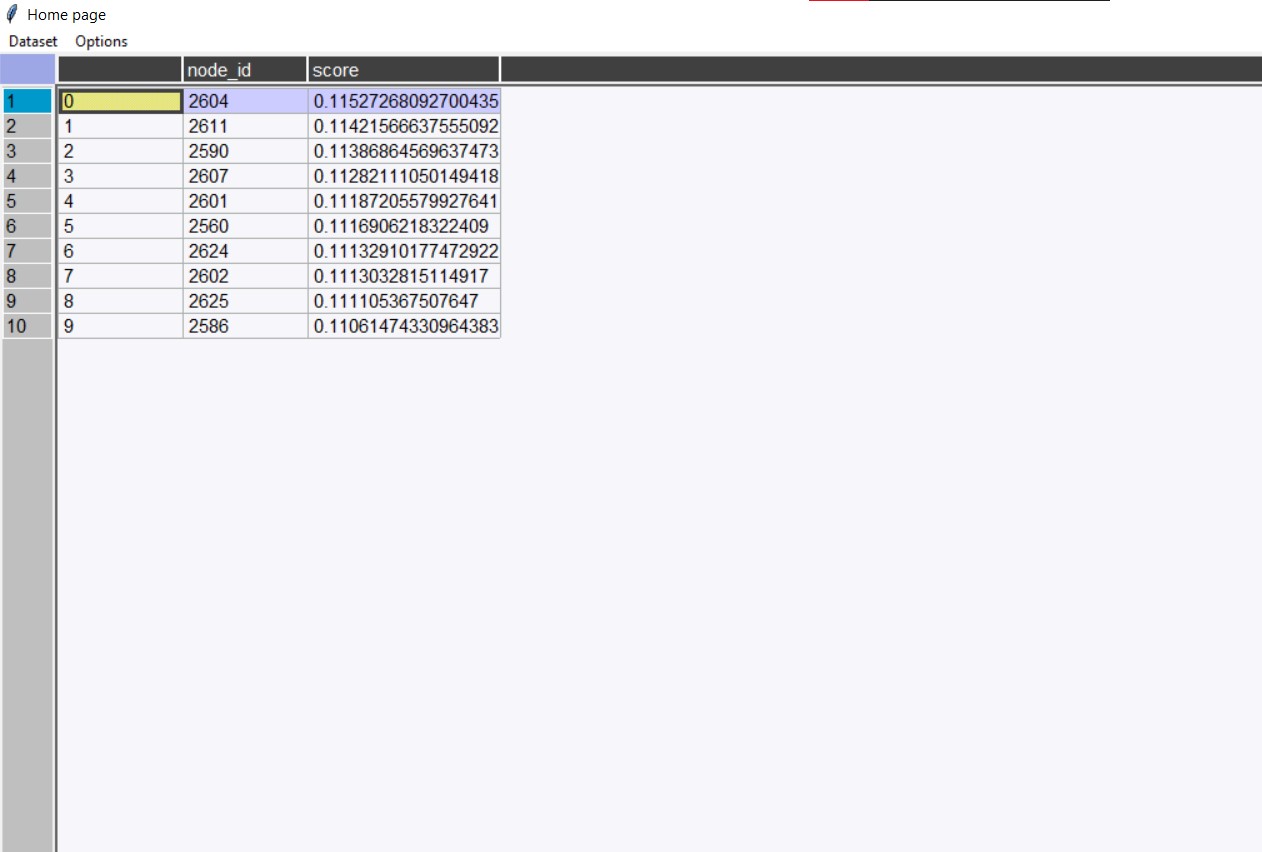
**Choosing HITS**

****

**Output 1 for HITS ie top 10 pages by hub score**

****

**Output 2 for HITS ie. top 10 pages by authoritative score**

****

**Conclusion:**

1. Implemented the PageRank algorithm to calculate the rank of each page in the file.

The output is the 10 pages with the highest rank, together with their rank

values.

2. Implemented the HITS algorithm to calculate the hub and the authority weight of each

web page in the data set. The output is the 10 most authoritative pages

and 10 most hubby pages.

3. Tabulated the results containing the rank of pages.