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.

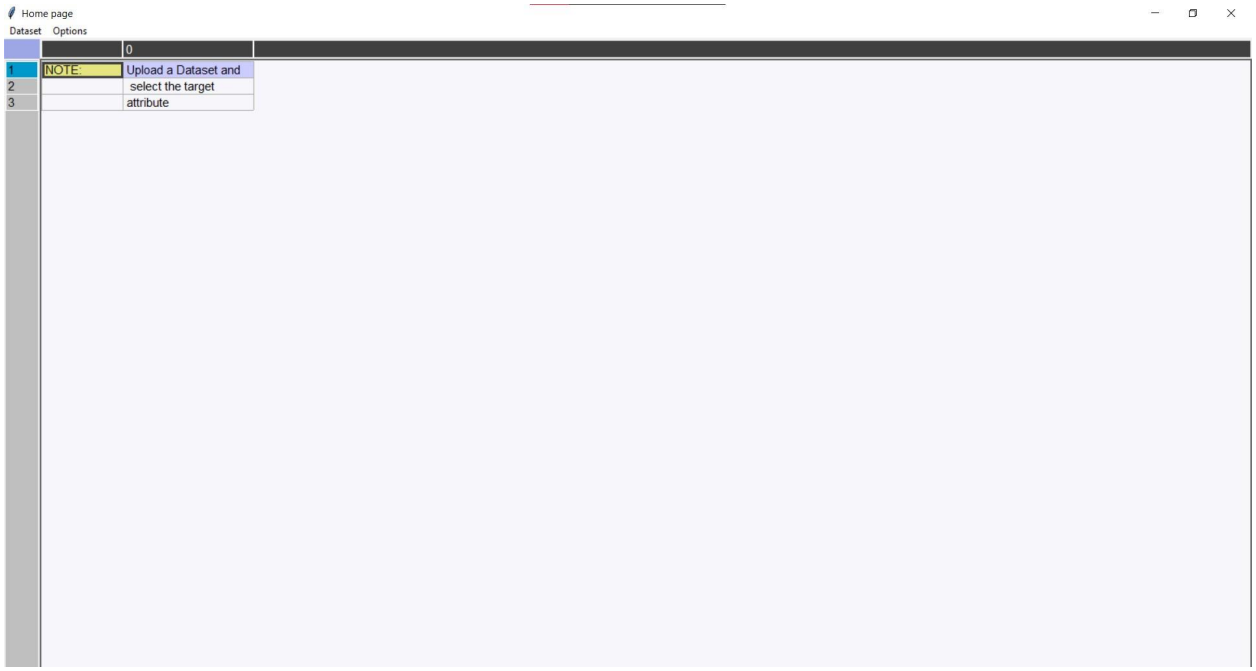
PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

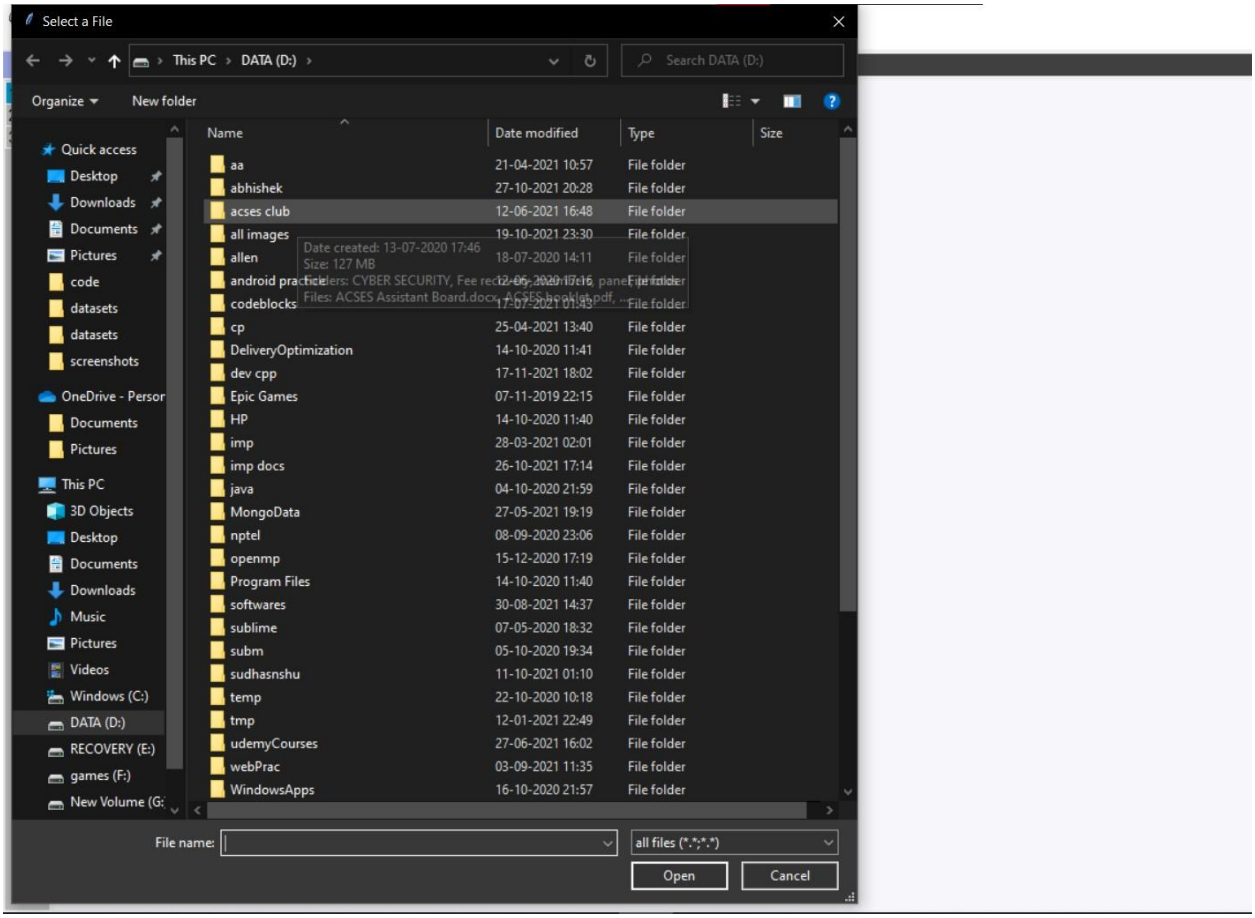
Hyperlink Induced Topic Search (HITS) Algorithm is a Link Analysis Algorithm that rates webpages, developed by Jon Kleinberg. This algorithm is used to the web link-structures to discover and rank the webpages relevant for a particular search.

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.

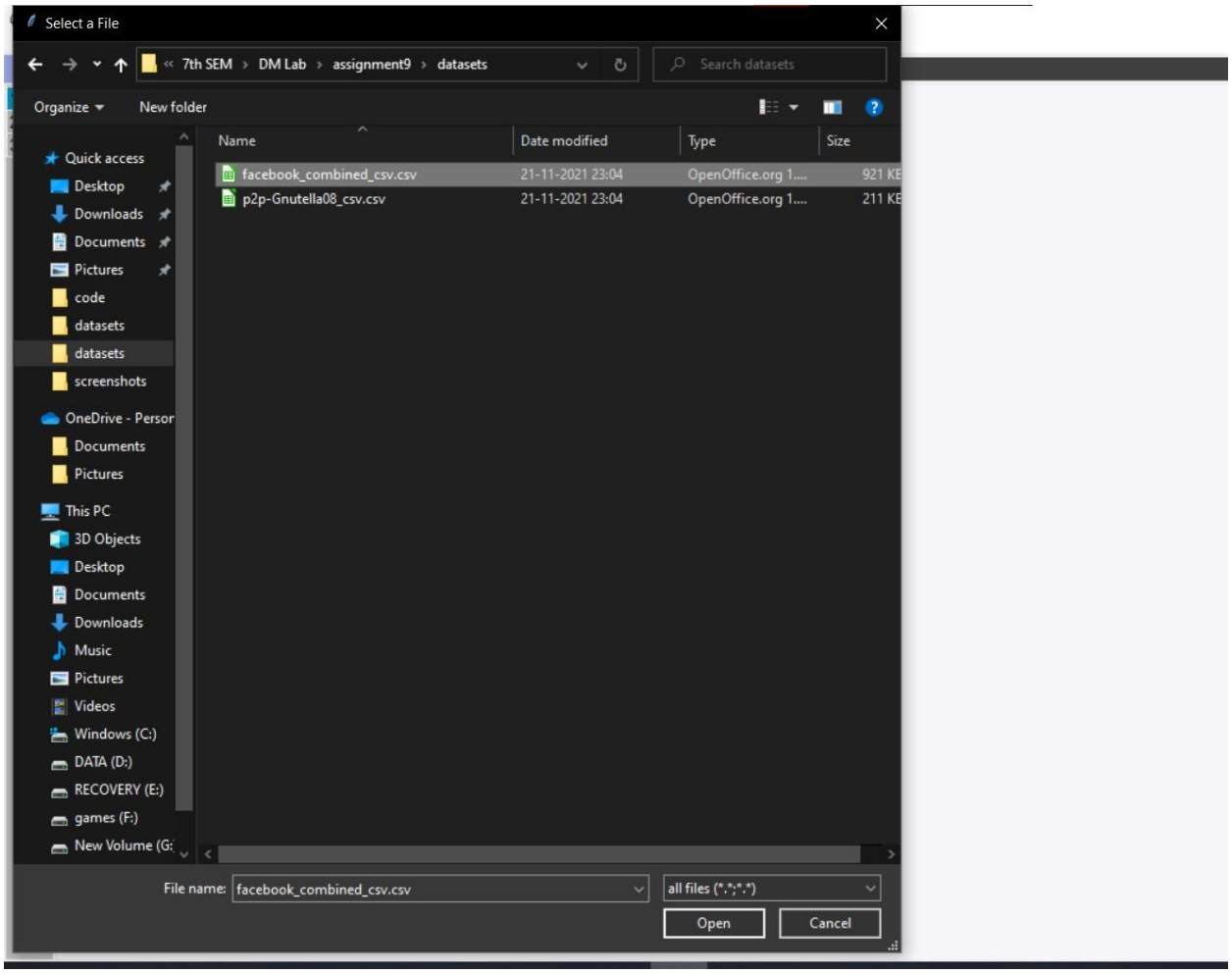
- Given a query to a Search Engine, the set of highly relevant web pages are called Roots. They are potential Authorities.
- Pages that are not very relevant but point to pages in the Root are called Hubs. Thus, an Authority is a page that many hubs link to whereas a Hub is a page that links to many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

Result/Observations/Screenshots:

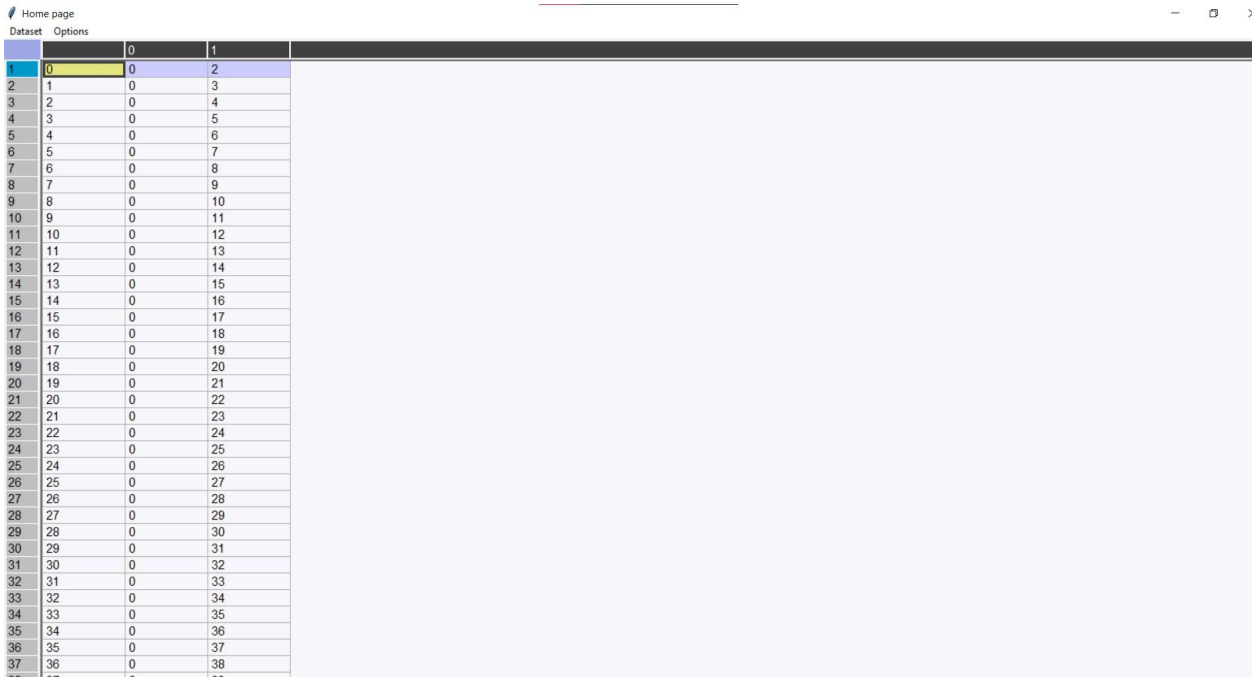



Select a File
This PC › DATA (D:) ›
Search DATA (D:)
Organize
New folder
Quick access
Desktop
Downloads
Documents
Pictures
code
datasets
datasets
screenshots
OneDrive - Persor
Documents
Pictures
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
Windows (C:)
DATA (D:)
RECOVERY (E:)
games (F:)
New Volume (G:
Name
Date modified
Type
Size
aa 21-04-2021 10:57 File folder
abhishek 27-10-2021 20:28 File folder
acses club 12-06-2021 16:48 File folder
all images 19-10-2021 23:30 File folder
allen 18-07-2020 14:11 File folder
android practicle 12-06-2020 17:16 File folder
codeblocks 17-07-2021 01:43 File folder
cp 25-04-2021 13:40 File folder
DeliveryOptimization 14-10-2020 11:41 File folder
dev cpp 17-11-2021 18:02 File folder
Epic Games 07-11-2019 22:15 File folder
HP 14-10-2020 11:40 File folder
imp 28-03-2021 02:01 File folder
imp docs 26-10-2021 17:14 File folder
java 04-10-2020 21:59 File folder
MongoData 27-05-2021 19:19 File folder
nptel 08-09-2020 23:06 File folder
openmp 15-12-2020 17:19 File folder
Program Files 14-10-2020 11:40 File folder
softwares 30-08-2021 14:37 File folder
sublime 07-05-2020 18:32 File folder
subm 05-10-2020 19:34 File folder
sudhasnshu 11-10-2021 01:10 File folder
temp 22-10-2020 10:18 File folder
tmp 12-01-2021 22:49 File folder
udemyCourses 27-06-2021 16:02 File folder
webPrac 03-09-2021 11:35 File folder
WindowsApps 16-10-2020 21:57 File folder
Date created: 13-07-2020 17:46
Size: 127 MB
Files: ACSES Assistant Board.docx, ACSES booklet.pdf, ...
File name:
all files (*.*;*.*)
Open
Cancel

Select dataset

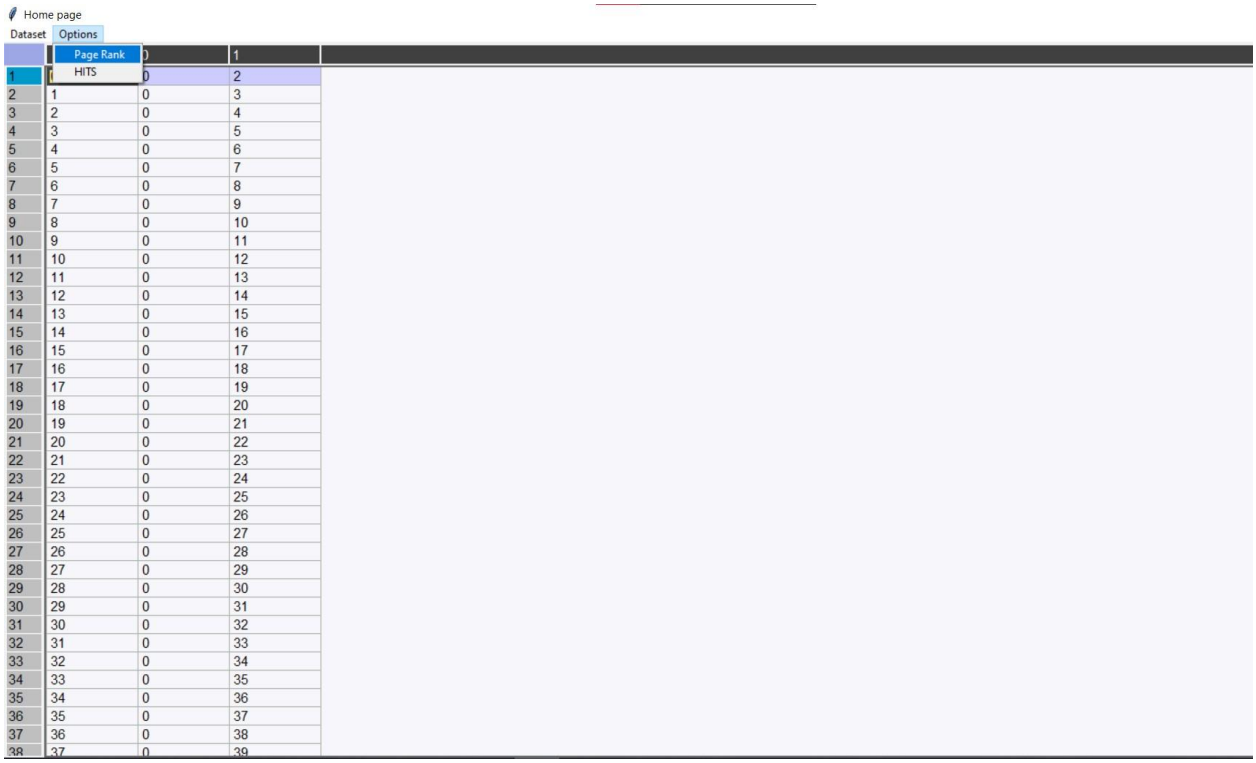


Displaying dataset



		0	1
1	0	0	2
2	1	0	3
3	2	0	4
4	3	0	5
5	4	0	6
6	5	0	7
7	6	0	8
8	7	0	9
9	8	0	10
10	9	0	11
11	10	0	12
12	11	0	13
13	12	0	14
14	13	0	15
15	14	0	16
16	15	0	17
17	16	0	18
18	17	0	19
19	18	0	20
20	19	0	21
21	20	0	22
22	21	0	23
23	22	0	24
24	23	0	25
25	24	0	26
26	25	0	27
27	26	0	28
28	27	0	29
29	28	0	30
30	29	0	31
31	30	0	32
32	31	0	33
33	32	0	34
34	33	0	35
35	34	0	36
36	35	0	37
37	36	0	38

Choosing PageRank



		0	1
1	0	0	2
2	1	0	3
3	2	0	4
4	3	0	5
5	4	0	6
6	5	0	7
7	6	0	8
8	7	0	9
9	8	0	10
10	9	0	11
11	10	0	12
12	11	0	13
13	12	0	14
14	13	0	15
15	14	0	16
16	15	0	17
17	16	0	18
18	17	0	19
19	18	0	20
20	19	0	21
21	20	0	22
22	21	0	23
23	22	0	24
24	23	0	25
25	24	0	26
26	25	0	27
27	26	0	28
28	27	0	29
29	28	0	30
30	29	0	31
31	30	0	32
32	31	0	33
33	32	0	34
34	33	0	35
35	34	0	36
36	35	0	37
37	36	0	38
38	37	0	39

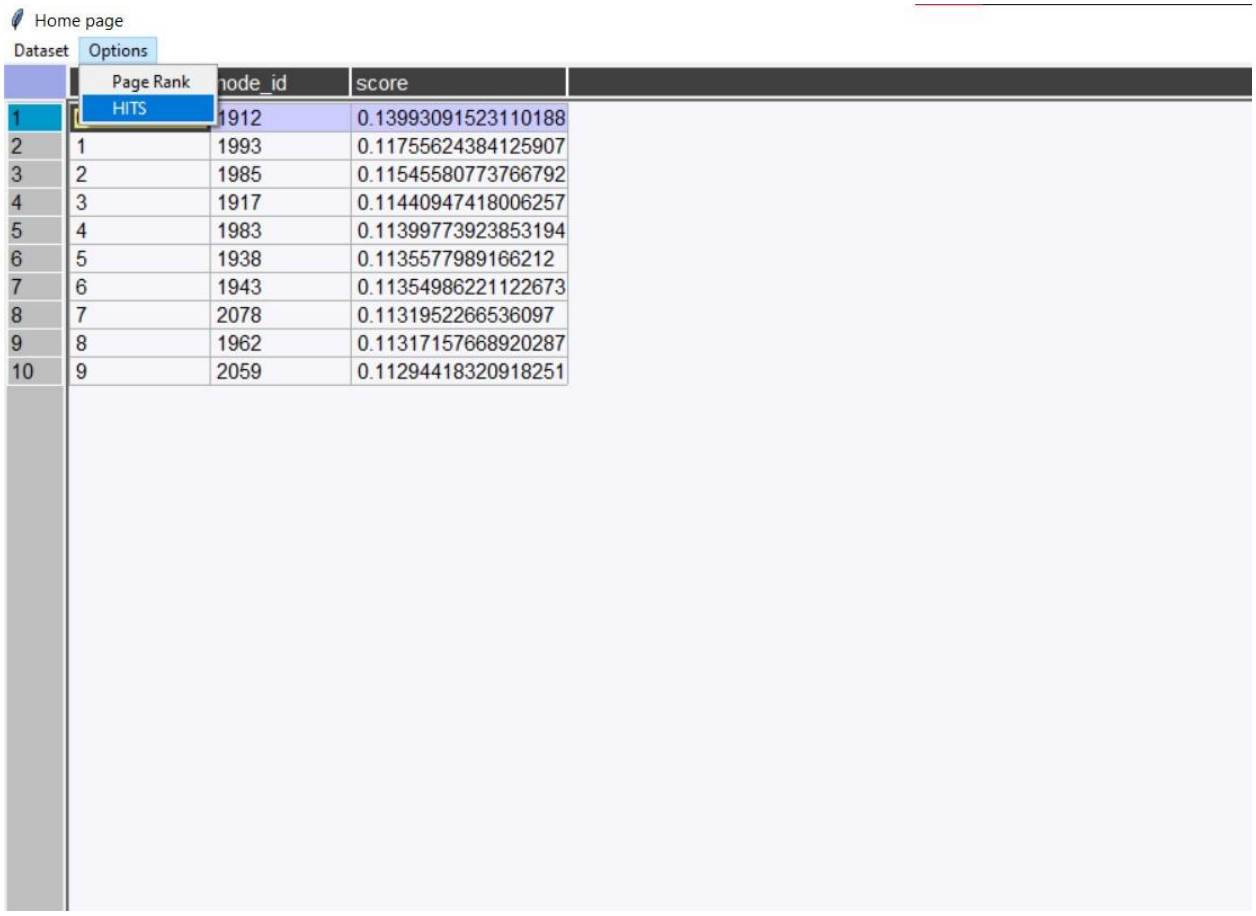
Output for Page Rank ie.Top 10 pages

Home page

Dataset Options

		Index Of.	Pagerank
1	0	2655	0.01008411283163841
2	1	2654	0.008388646595699023
3	2	1902	0.00787510684110136
4	3	2649	0.007112246455791269
5	4	2646	0.006711337838002668
6	5	3434	0.006076851756760203
7	6	2631	0.005885685945822974
8	7	3426	0.005271784937561473
9	8	1898	0.005269048507772324
10	9	1886	0.004851074214254664

Choosing HITS



Output 1 for HITS ie top 10 pages by hub score

Home page

Dataset Options

		node_id	score
1	0	1912	0.13993091523110188
2	1	1993	0.11755624384125907
3	2	1985	0.11545580773766792
4	3	1917	0.11440947418006257
5	4	1983	0.11399773923853194
6	5	1938	0.1135577989166212
7	6	1943	0.11354986221122673
8	7	2078	0.1131952266536097
9	8	1962	0.11317157668920287
10	9	2059	0.11294418320918251

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

Home page

Dataset Options

		node_id	score
1	0	2604	0.11527268092700435
2	1	2611	0.11421566637555092
3	2	2590	0.11386864569637473
4	3	2607	0.11282111050149418
5	4	2601	0.11187205579927641
6	5	2560	0.1116906218322409
7	6	2624	0.11132910177472922
8	7	2602	0.1113032815114917
9	8	2625	0.111105367507647
10	9	2586	0.11061474330964383

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.