INFORMATION RETRIEVAL(CSE508)

ASSIGNMENT 3

Group Member 1: Adarsh Singh Kushwah **RollNo:** MT21111
Group Member 2: Charisha Phirani **RollNo:** MT21117

Note: The supporting code is present in .pynb file

LIBRARIES USED:

1. **os**: for importing dataset folder from directory.
2. **networkx**: python library for studying graph networks.
3. **pandas**: for implementing dataFrames.
4. **tqdm**: used for creating progress bars
5. **prettytable**: creates relational tables in python.

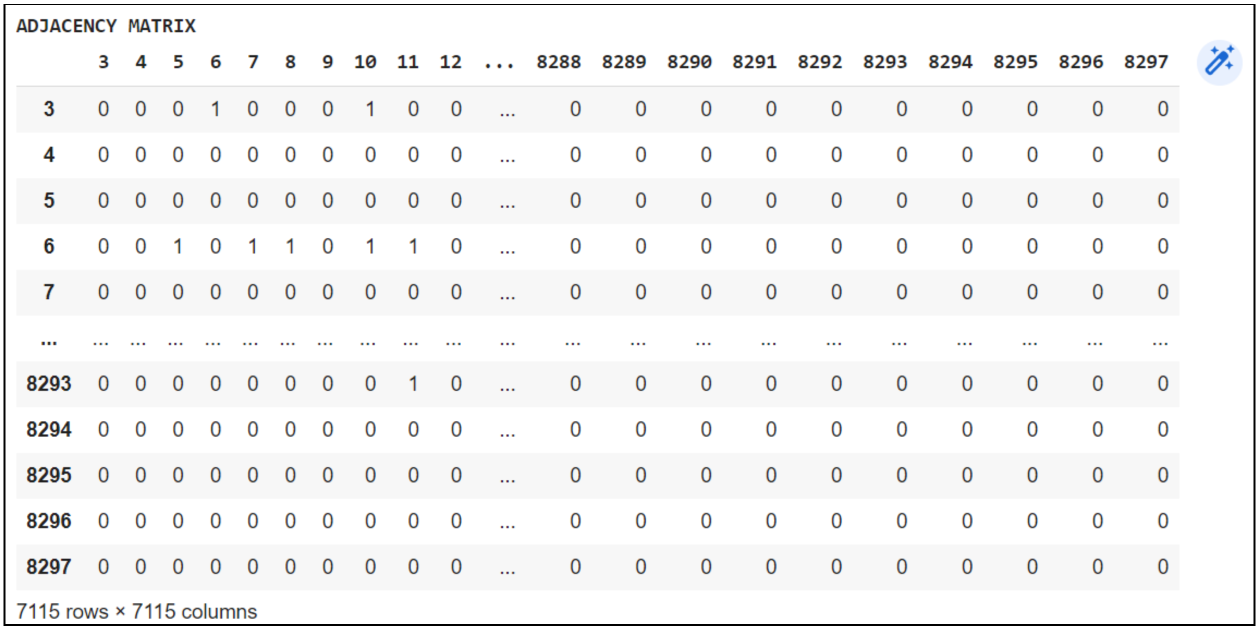
Ans 1a:

Dataset chosen: Wikipedia Vote Network

Network representation in the form of adjacency matrix.

For an adjacency matrix, if there is **an edge from node 1 to 2, then the creatematrix[1][2]=1 else it will be 0**.

The value for **creatematrix[1][2] is 1 if page 1 links with page 2.**

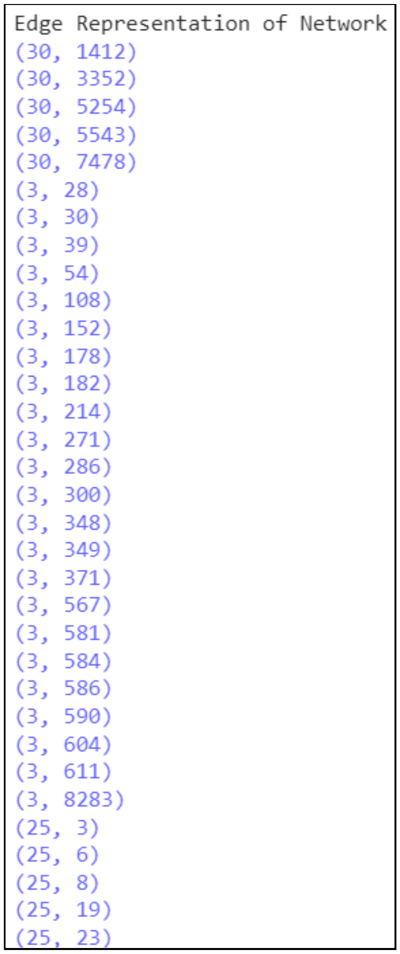


ADJACENCY MATRIX

	3	4	5	6	7	8	9	10	11	12	...	8288	8289	8290	8291	8292	8293	8294	8295	8296	8297
3	0	0	0	1	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
6	0	0	1	0	1	1	0	1	1	0	...	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8293	0	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
8294	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
8295	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
8296	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
8297	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

7115 rows × 7115 columns

Network representation in the form of **adjacency list.**



```
Edge Representation of Network
(30, 1412)
(30, 3352)
(30, 5254)
(30, 5543)
(30, 7478)
(3, 28)
(3, 30)
(3, 39)
(3, 54)
(3, 108)
(3, 152)
(3, 178)
(3, 182)
(3, 214)
(3, 271)
(3, 286)
(3, 300)
(3, 348)
(3, 349)
(3, 371)
(3, 567)
(3, 581)
(3, 584)
(3, 586)
(3, 590)
(3, 604)
(3, 611)
(3, 8283)
(25, 3)
(25, 6)
(25, 8)
(25, 19)
(25, 23)
```

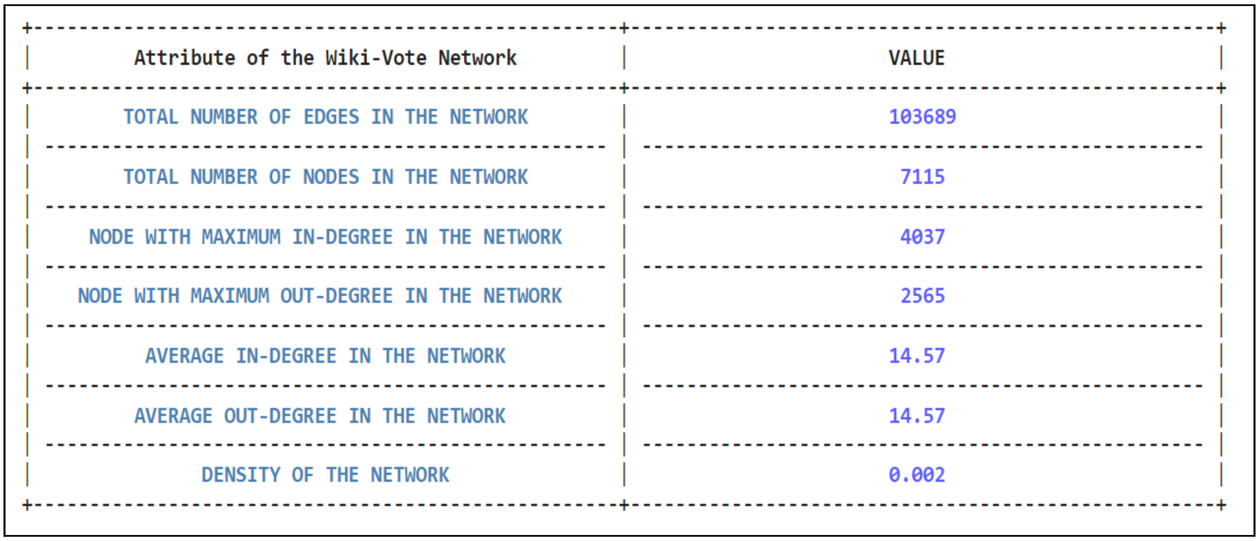
Ans 1b:

Dataset chosen: Wikipedia Vote Network

This dataset contains wikipedia voting data till January 2008.

Nodes in the network: **wikipedia users**

Edges in the network: **user i voted on user j**



```
+----------------------------------------------------+----------------------------------------------------+
|          Attribute of the Wiki-Vote Network          |                        VALUE                       |
+----------------------------------------------------+----------------------------------------------------+
|        TOTAL NUMBER OF EDGES IN THE NETWORK         |                       103689                       |
| -------------------------------------------------- | -------------------------------------------------- |
|        TOTAL NUMBER OF NODES IN THE NETWORK         |                        7115                        |
| -------------------------------------------------- | -------------------------------------------------- |
|     NODE WITH MAXIMUM IN-DEGREE IN THE NETWORK      |                        4037                        |
| -------------------------------------------------- | -------------------------------------------------- |
|     NODE WITH MAXIMUM OUT-DEGREE IN THE NETWORK     |                        2565                        |
| -------------------------------------------------- | -------------------------------------------------- |
|          AVERAGE IN-DEGREE IN THE NETWORK           |                       14.57                        |
| -------------------------------------------------- | -------------------------------------------------- |
|          AVERAGE OUT-DEGREE IN THE NETWORK          |                       14.57                        |
| -------------------------------------------------- | -------------------------------------------------- |
|               DENSITY OF THE NETWORK                |                       0.002                        |
+----------------------------------------------------+----------------------------------------------------+
```

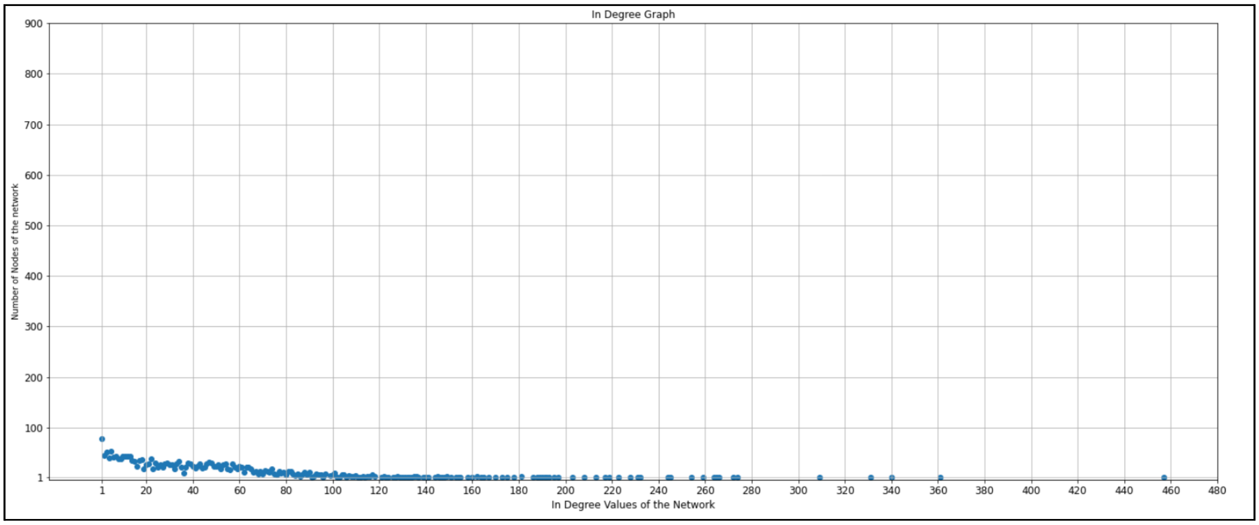
Since the graph is the directed graph, we have plotted the in-degree and out-degree distribution.

The **average indegree and outdegree will be the same** because the nodes with indegree will get balanced by the nodes with one outdegree.

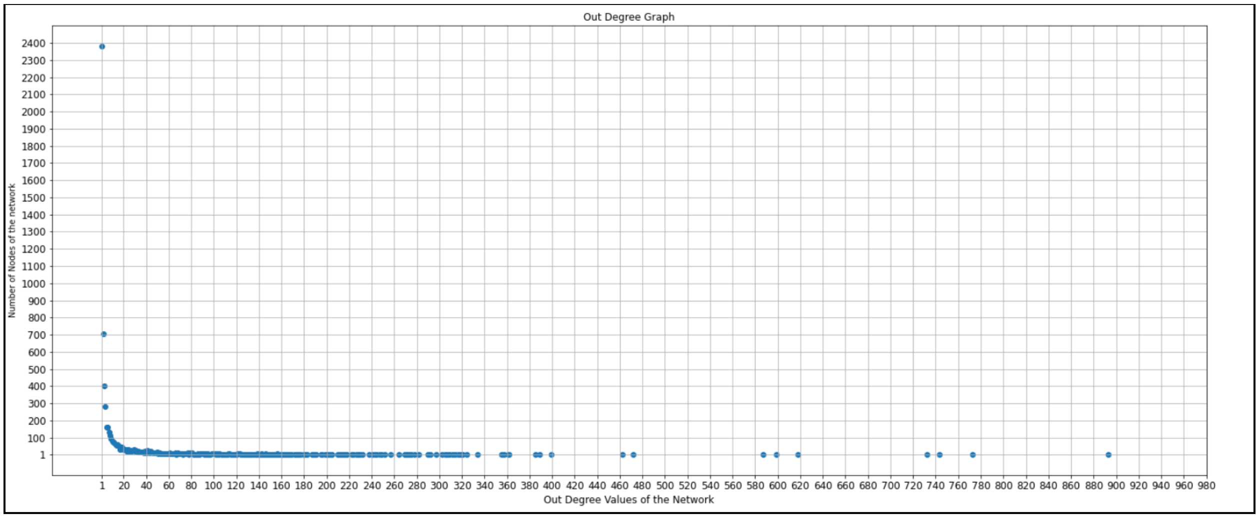
The network density tells us that if the **density is 0**, then the network has no edges and if the **density is 1**,then the network is a complete graph.

Network density is calculated as : **total count of edges / (n)*(n-1)** ; n = total count of nodes in the network (for directed graph).

1. **In degree distribution:**



2. **Out-degree distribution**



Ans 1c

Clustering coefficient of each node.

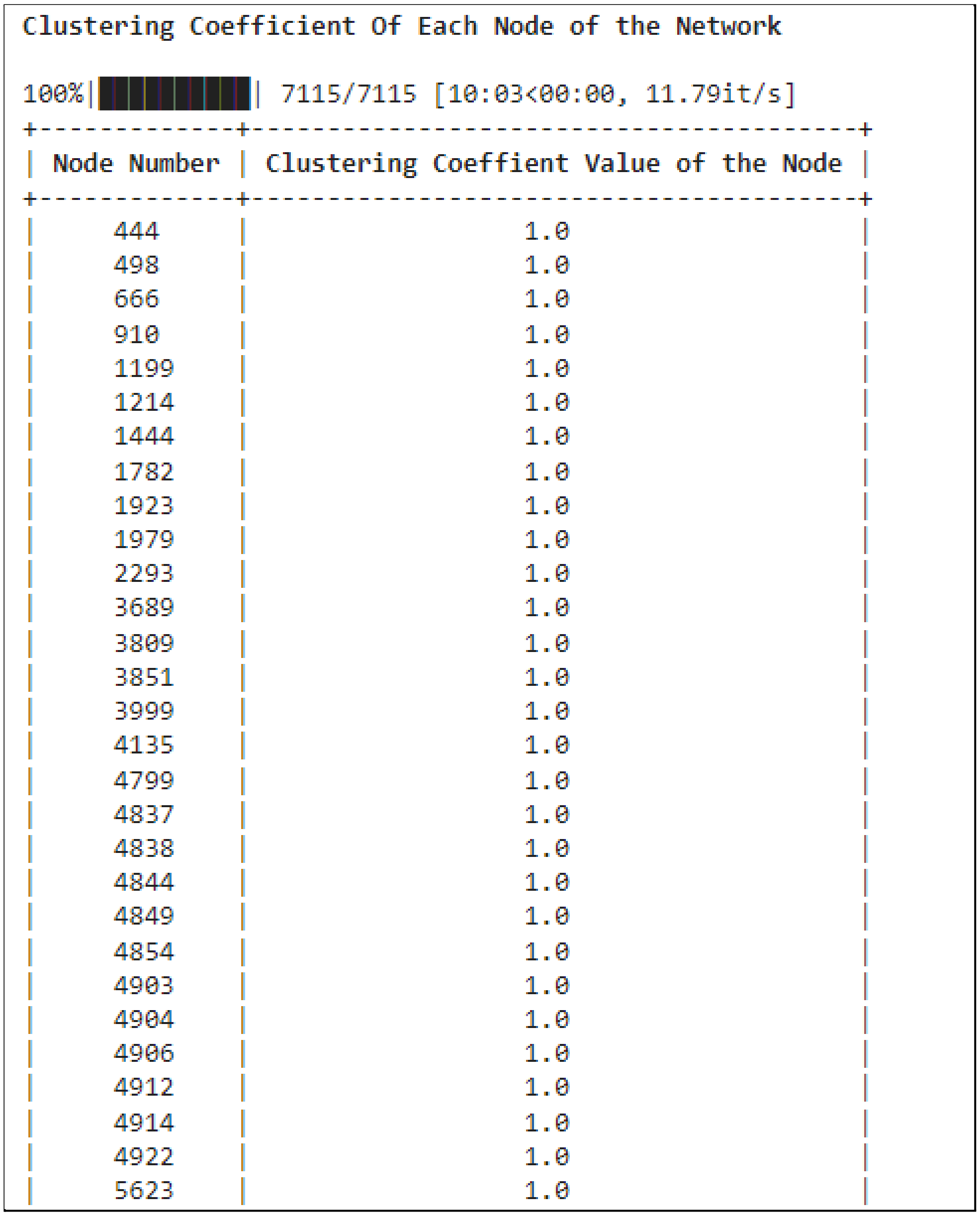
Clustering coefficient lies between 0 and 1. The clustering coefficient more skewed towards 1 gives higher certainty.

Count of nodes with clustering coefficient 0:

Count of nodes with clustering coefficient 1:

Overall clustering coefficient of the network:

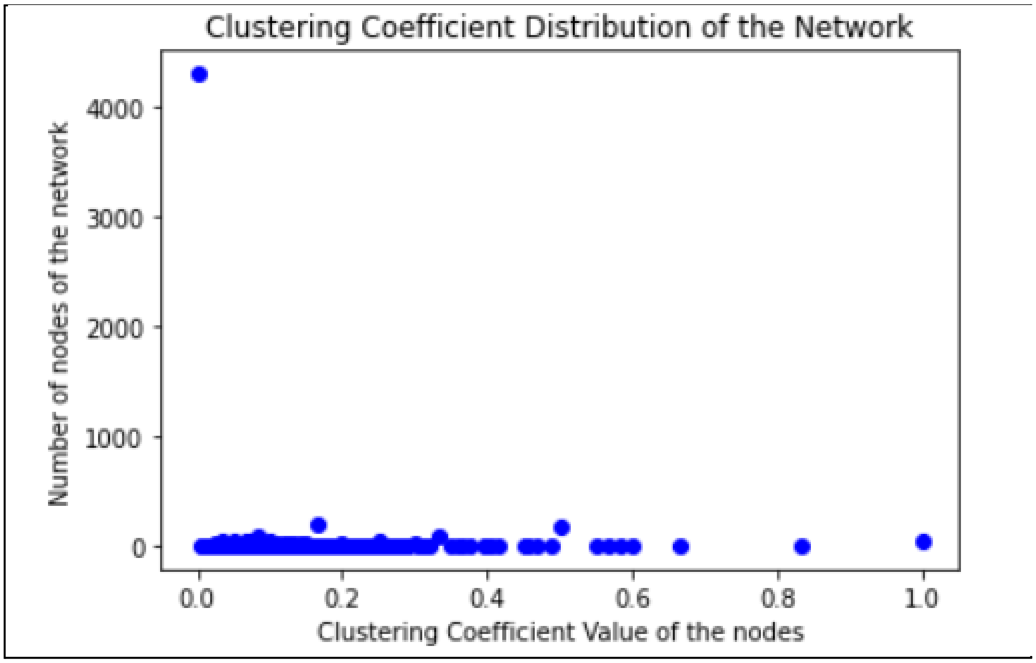
Formula for calculating the clustering coefficient of the network for directed graph: **N/n*(n-1)** here n = total node neighbors, N = number of edges among n neighbors of that node in the network.



```
Clustering Coefficient Of Each Node of the Network

100%|██████████| 7115/7115 [10:03<00:00, 11.79it/s]
+-------------+---------------------------------------+
| Node Number | Clustering Coeffient Value of the Node |
+-------------+---------------------------------------+
|     444     |                  1.0                  |
|     498     |                  1.0                  |
|     666     |                  1.0                  |
|     910     |                  1.0                  |
|     1199    |                  1.0                  |
|     1214    |                  1.0                  |
|     1444    |                  1.0                  |
|     1782    |                  1.0                  |
|     1923    |                  1.0                  |
|     1979    |                  1.0                  |
|     2293    |                  1.0                  |
|     3689    |                  1.0                  |
|     3809    |                  1.0                  |
|     3851    |                  1.0                  |
|     3999    |                  1.0                  |
|     4135    |                  1.0                  |
|     4799    |                  1.0                  |
|     4837    |                  1.0                  |
|     4838    |                  1.0                  |
|     4844    |                  1.0                  |
|     4849    |                  1.0                  |
|     4854    |                  1.0                  |
|     4903    |                  1.0                  |
|     4904    |                  1.0                  |
|     4906    |                  1.0                  |
|     4912    |                  1.0                  |
|     4914    |                  1.0                  |
|     4922    |                  1.0                  |
|     5623    |                  1.0                  |
```

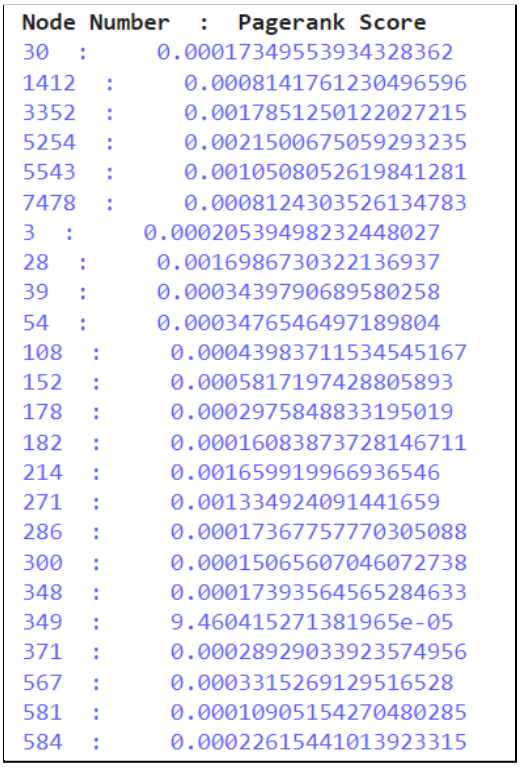
Graph of clustering coefficient distribution



Ans 2a:

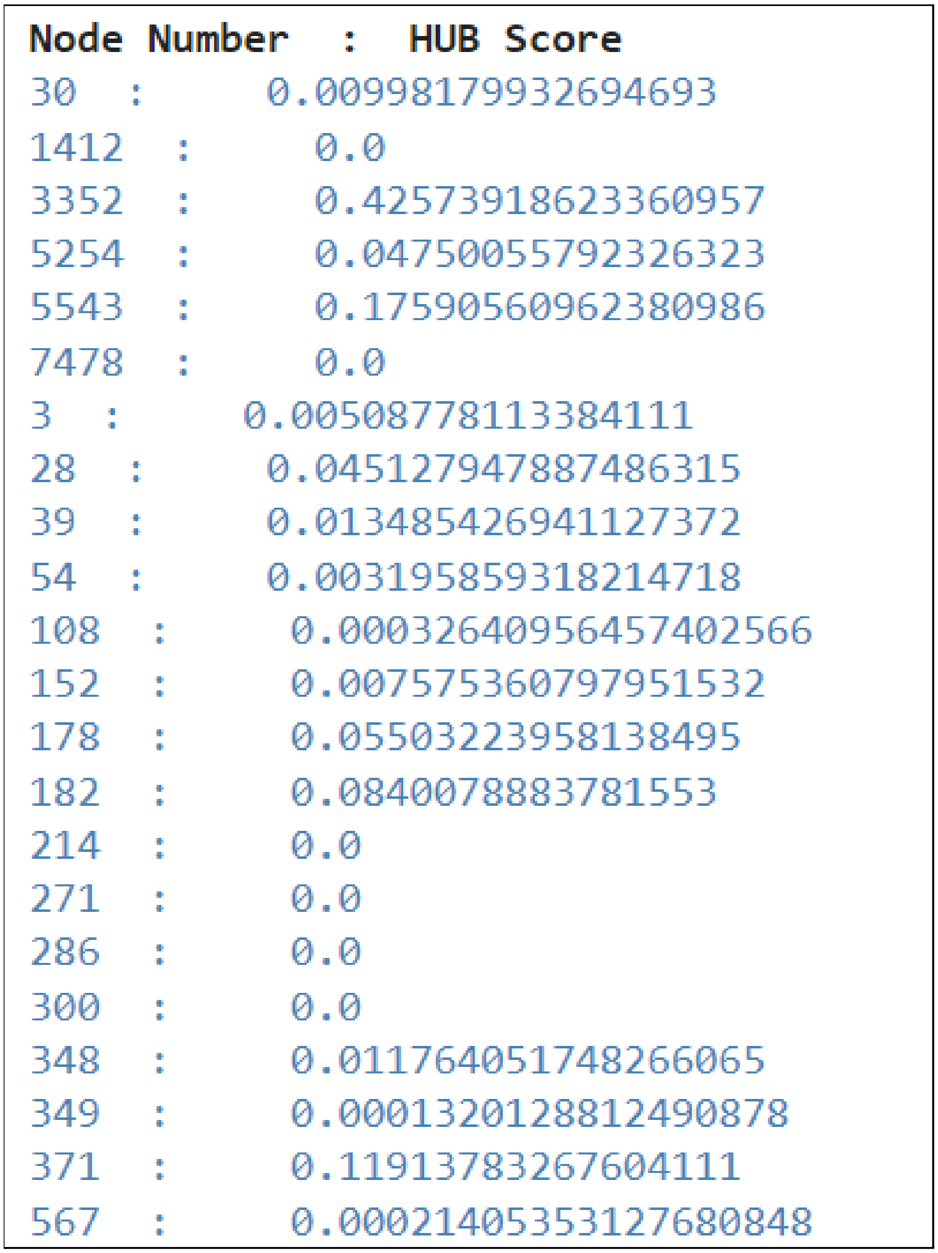
Page rank scores:

Page rank ranks the web pages and returns them in the order of relevance. For the nodes with higher incoming edges, high page rank is assigned.

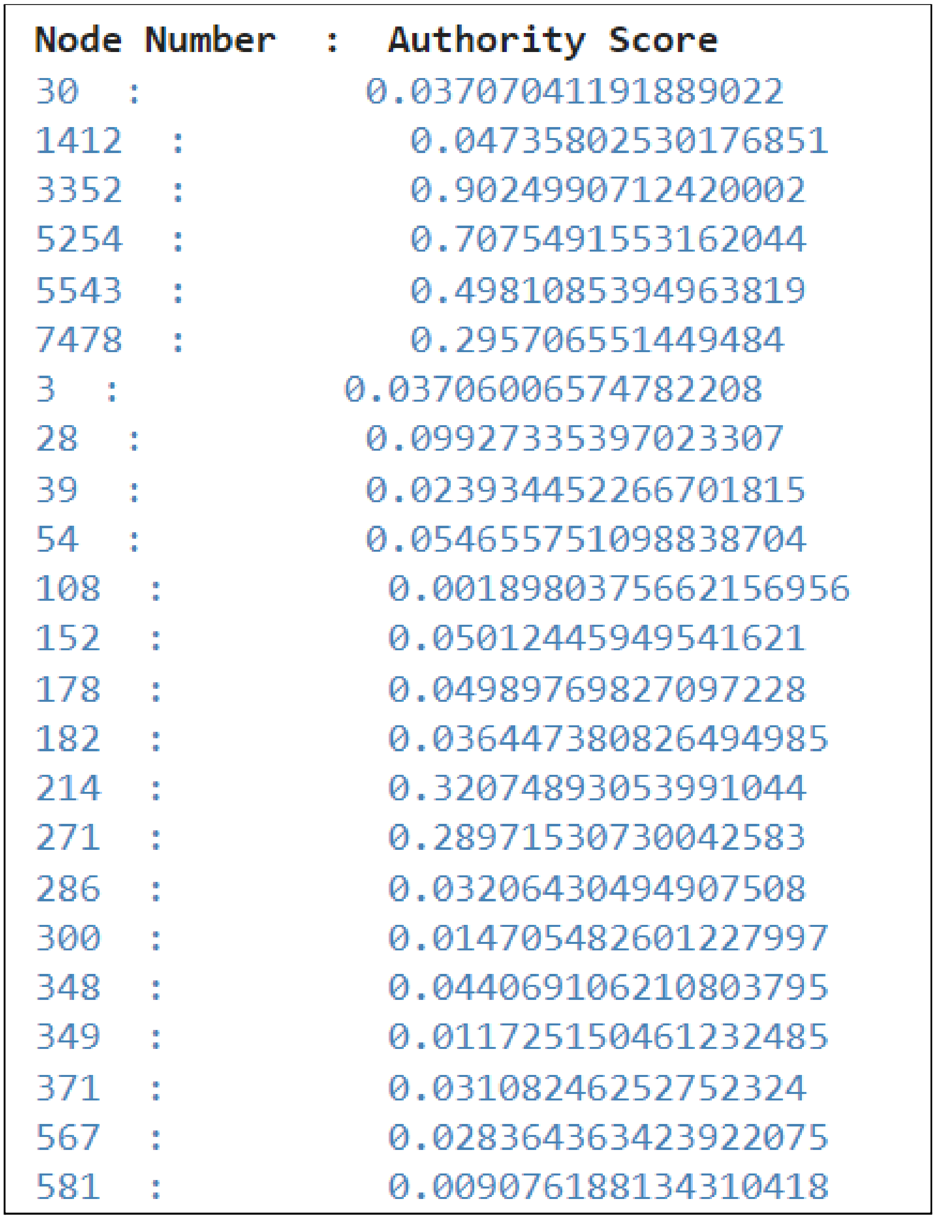


```
Node Number  :  Pagerank Score
30   :     0.00017349553934328362
1412   :     0.0008141761230496596
3352   :     0.0017851250122027215
5254   :     0.0021500675059293235
5543   :     0.0010508052619841281
7478   :     0.0008124303526134783
3   :     0.00020539498232448027
28   :     0.0016986730322136937
39   :     0.0003439790689580258
54   :     0.0003476546497189804
108   :     0.00043983711534545167
152   :     0.0005817197428805893
178   :     0.0002975848833195019
182   :     0.00016083873728146711
214   :     0.001659919966936546
271   :     0.001334924091441659
286   :     0.00017367757770305088
300   :     0.00015065607046072738
348   :     0.00017393564565284633
349   :     9.460415271381965e-05
371   :     0.00028929033923574956
567   :     0.0003315269129516528
581   :     0.00010905154270480285
584   :     0.00022615441013923315
```

Hubs and Authority scores:
Used to measure the importance of web pages. Root nodes are the highly related web pages for the query provided. Non relevant pages pointing to the root nodes are called hubs. A good authority has many hubs pointing to it. A page that many hubs link joined to.Set of highly relevant web pages are called Roots. They are also known as potential



```
Node Number   :   HUB Score
30   :      0.0099817993269469З
1412   :      0.0
3352   :      0.42573918623360957
5254   :      0.04750055792326323
5543   :      0.17590560962380986
7478   :      0.0
3   :      0.005087781133841111
28   :      0.04512794788748631 5
39   :      0.01348542694112737 2
54   :      0.00319585931821471 8
108   :      0.00032640956457402566
152   :      0.007575360797951532
178   :      0.05503223958138495
182   :      0.0840078883781553
214   :      0.0
271   :      0.0
286   :      0.0
300   :      0.0
348   :      0.011764051748266065
349   :      0.00013201288124908 78
371   :      0.1191378326760411 1
567   :      0.00021405353127680848
```



```
Node Number  :  Authority Score
30  :          0.03707041191889022
1412  :          0.04735802530176851
3352  :          0.9024990712420002
5254  :          0.7075491553162044
5543  :          0.4981085394963819
7478  :          0.295706551449484
3  :          0.03706006574782208
28  :          0.09927335397023307
39  :          0.023934452266701815
54  :          0.05465575109883704
108  :          0.0018980375662156956
152  :          0.05012445949541621
178  :          0.04989769827097228
182  :          0.036447380826494985
214  :          0.32074893053991044
271  :          0.28971530730042583
286  :          0.03206430494907508
300  :          0.014705482601227997
348  :          0.044069106210803795
349  :          0.011725150461232485
371  :          0.03108246252752324
567  :          0.028364363423922075
581  :          0.009076188134310418
```

Comparison between algorithm 1 and algorithm 2:

The time taken for evaluation the scores in HITS algorithm is greater than the time taken in evaluating the scores in Pagerank algorithm.

As the HITS creates mutual reinforcement between authority and hub scores and page rank just do it on the basis of authority, the HITS results are less relevance than the page rank scores.

This popularity is due to the features like efficiency, feasibility, less query time cost, etc. which are absent in HITS algorithm.