INFORMATION RETRIEVAL
Assignment 3

Group Members :
Shambhavi Sharma (MT22066)
Rohit Kesarwani (MT22059)
Sushil Kumar(MT22077)

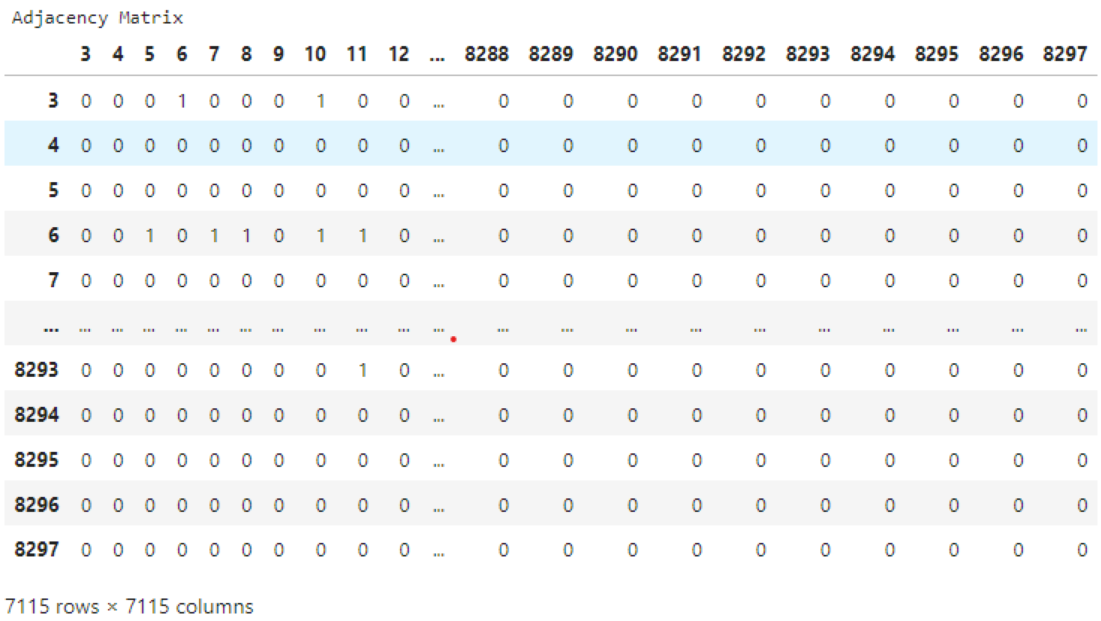
LIBRARIES USED

- os : os is used for importing dataset folders from directory.
- pandas : It provides tools for data cleaning, transformation, aggregation, and visualization, making it an essential tool for data scientists and analysts.
- Matplotlib : Matplotlib is a popular plotting library in Python.
- Prettytable : It is useful for displaying tabular data in a readable and organized format.
- tqdm : It is useful for tracking the progress of a long-running operation, such as a loop, and provides an easy way to visualize how much of the operation has been completed.
- networkx : python library for studying graph networks.

Ans 1:

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 adj_matrix [1][2]=1 else it will be 0. The value for adj_matrix [1][2] is 1 if page 1 links with page 2.

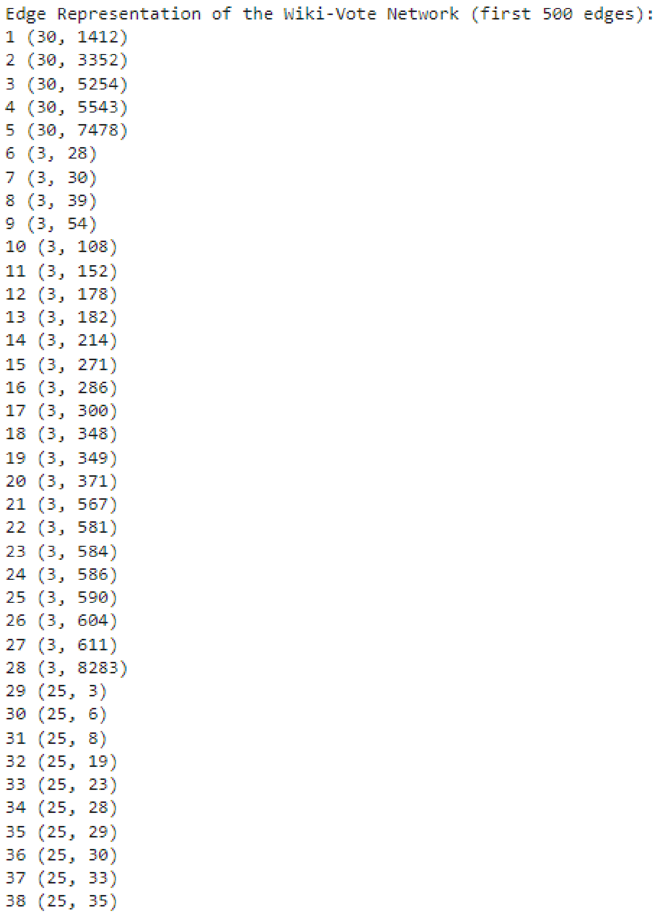


```
Adjacency Matrix
```

	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

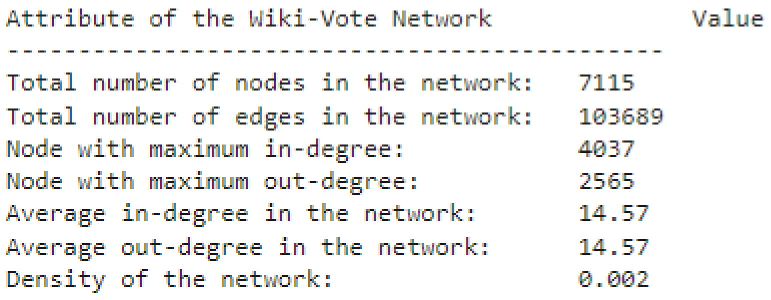
Edge Representation of the Wiki-Vote Network (first 500 edges)



```
Edge Representation of the Wiki-Vote Network (first 500 edges):
1 (30, 1412)
2 (30, 3352)
3 (30, 5254)
4 (30, 5543)
5 (30, 7478)
6 (3, 28)
7 (3, 30)
8 (3, 39)
9 (3, 54)
10 (3, 108)
11 (3, 152)
12 (3, 178)
13 (3, 182)
14 (3, 214)
15 (3, 271)
16 (3, 286)
17 (3, 300)
18 (3, 348)
19 (3, 349)
20 (3, 371)
21 (3, 567)
22 (3, 581)
23 (3, 584)
24 (3, 586)
25 (3, 590)
26 (3, 604)
27 (3, 611)
28 (3, 8283)
29 (25, 3)
30 (25, 6)
31 (25, 8)
32 (25, 19)
33 (25, 23)
34 (25, 28)
35 (25, 29)
36 (25, 30)
37 (25, 33)
38 (25, 35)
```

Ans 1.1:

Dataset chosen: Wikipedia Vote Network
Nodes in the network: wikipedia users
Edges in the network: user i voted on user j

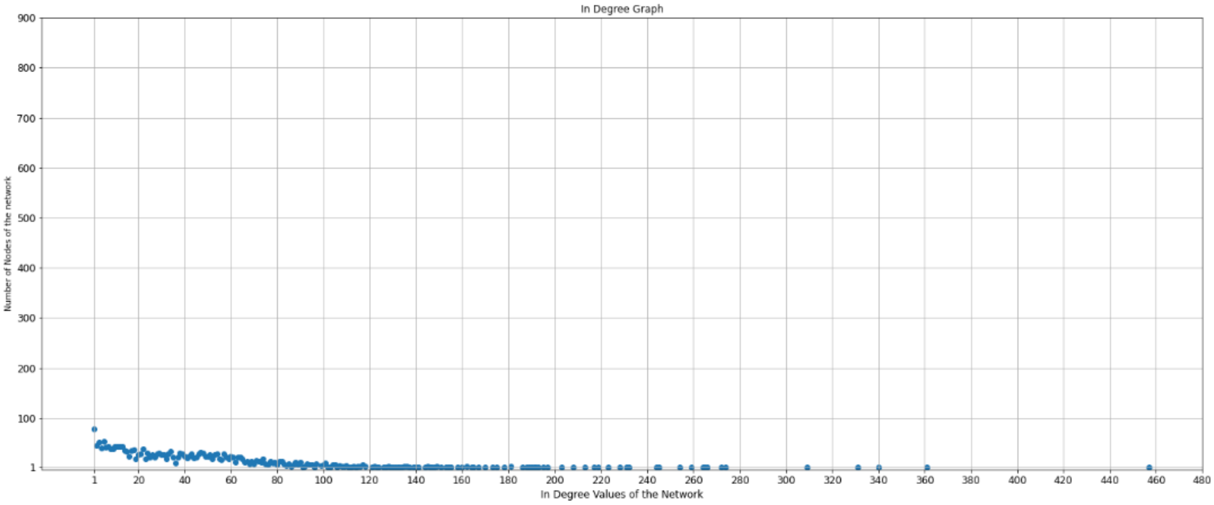


```
Attribute of the Wiki-Vote Network              Value
----------------------------------------------
Total number of nodes in the network:   7115
Total number of edges in the network:   103689
Node with maximum in-degree:            4037
Node with maximum out-degree:           2565
Average in-degree in the network:       14.57
Average out-degree in the network:      14.57
Density of the network:                 0.002
```

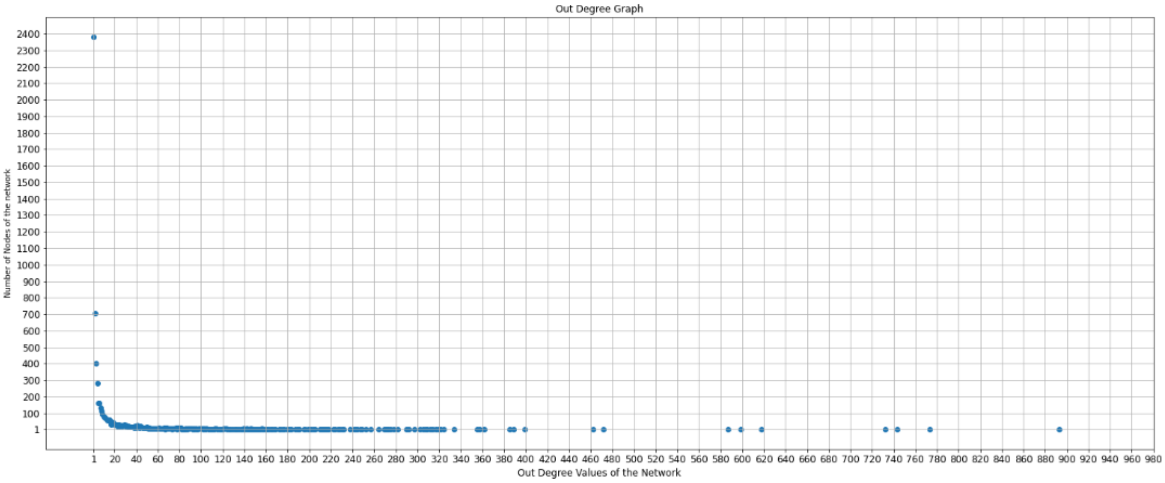
we have a directed graph, we can analyze its in-degree and out-degree distribution. The average in-degree and out-degree will be equal because nodes with a high in-degree will be balanced out by nodes with a high out-degree.

Network density is a measure of how connected a graph is. If the density is 0, then there are no edges in the network, while a density of 1 indicates a complete graph. To calculate network density for a directed graph, we use the formula: total number of edges divided by n times n-1, where n is the total number of nodes in the network.

In-degree Distribution of Network



Out-degree Distribution of Network



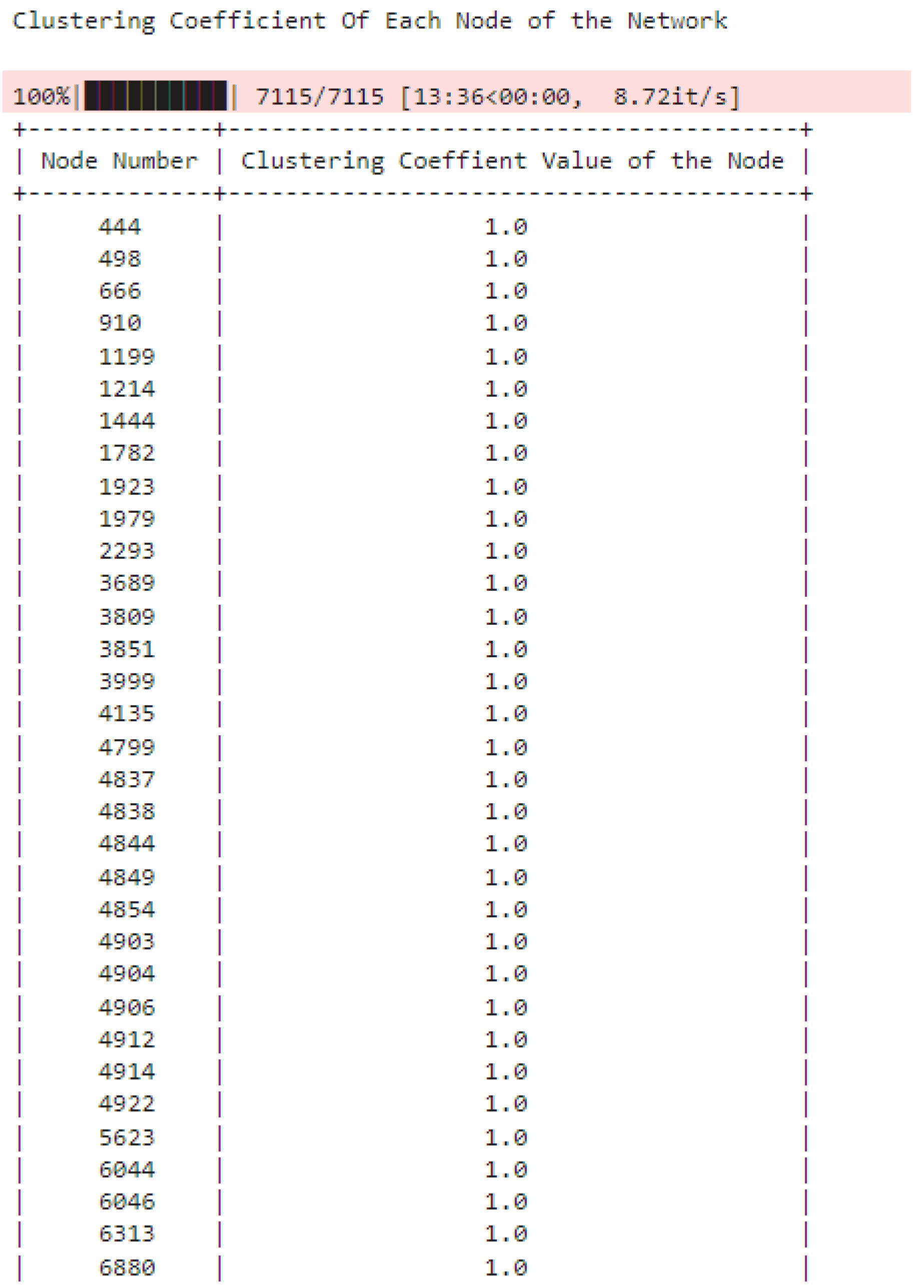
Ans 1.2

We can calculate the clustering coefficient of each node in a graph. The clustering coefficient ranges between 0 and 1, with values closer to 1 indicating a higher level of certainty.

We can also count the number of nodes with a clustering coefficient of 0 and 1 in the graph.

The overall clustering coefficient of the network can be determined by calculating the average clustering coefficient across all nodes.

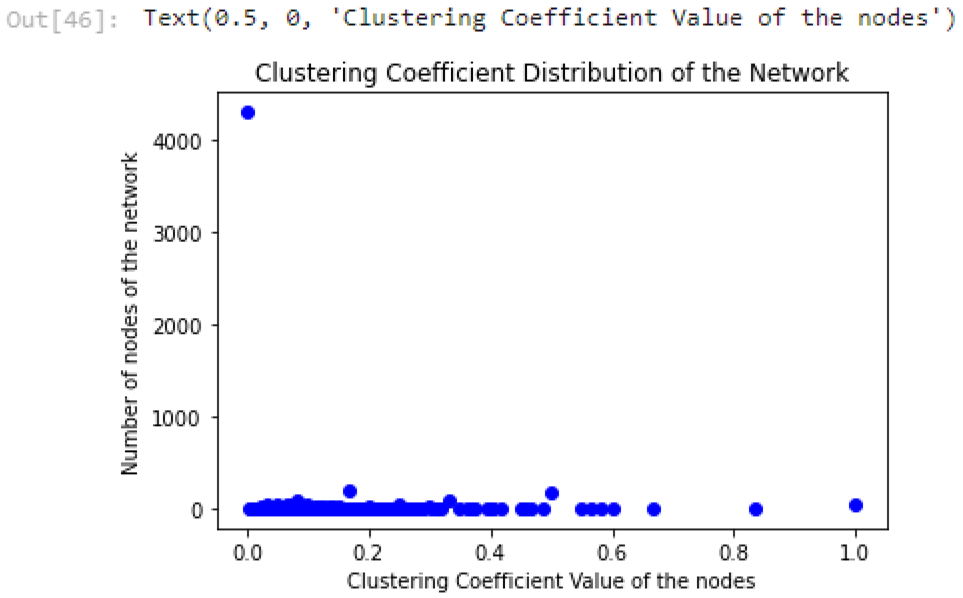
To calculate the clustering coefficient of a directed graph, we use the formula: N divided by n times (n-1), where n is the total number of neighbors for the node, and N is the number of edges among those neighbors in the network.



```
Clustering Coefficient Of Each Node of the Network

100%|██████████| 7115/7115 [13:36<00:00,  8.72it/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                  |
|     6044    |                  1.0                  |
|     6046    |                  1.0                  |
|     6313    |                  1.0                  |
|     6880    |                  1.0                  |
```

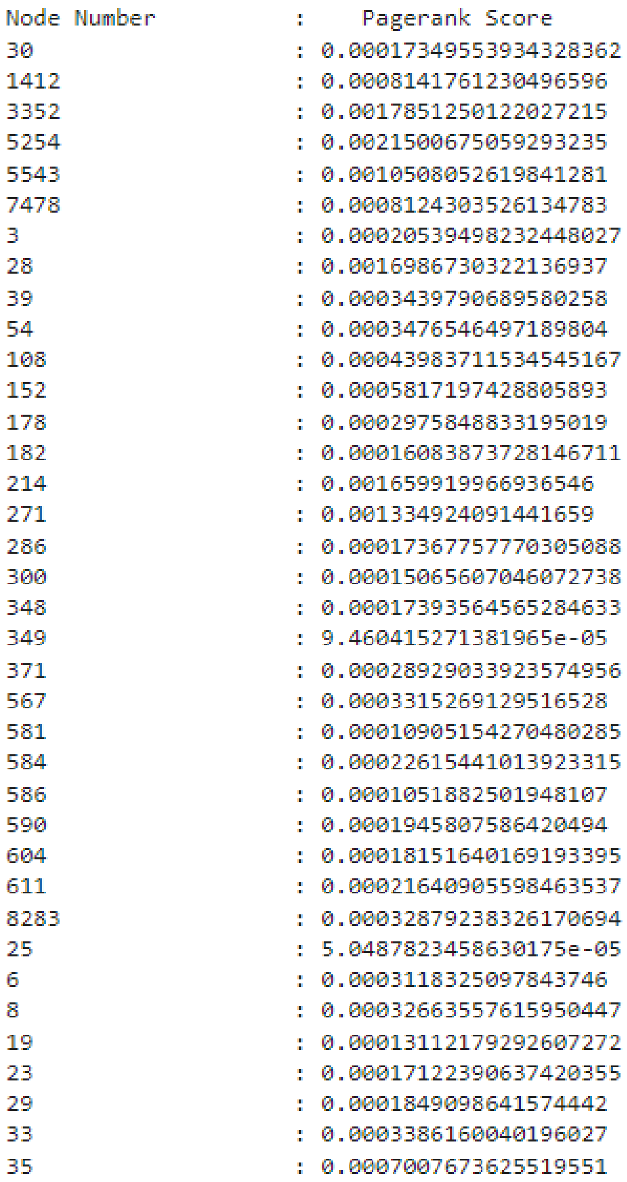
Graph of clustering coefficient distribution



```
Out[46]:  Text(0.5, 0, 'Clustering Coefficient Value of the nodes')
```

Ans 2.1 : Page Rank

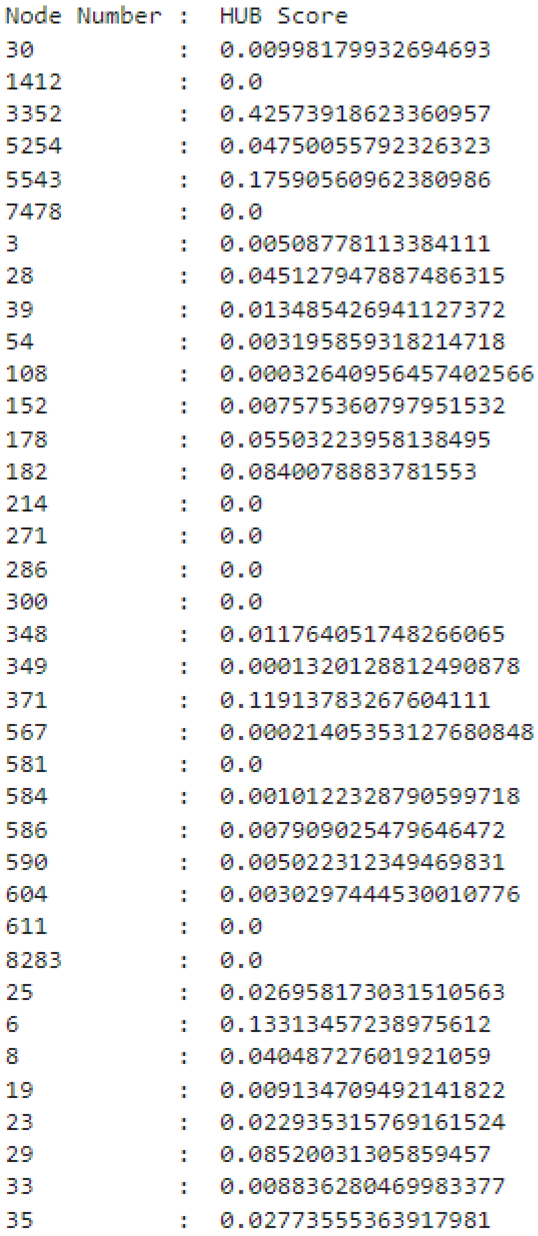
Page Rank is an algorithm that ranks web pages based on their relevance and returns them in order. Pages with more incoming edges are assigned a higher Page Rank score.



```
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
586                  : 0.0001051882501948107
590                  : 0.0001945807586420494
604                  : 0.00018151640169193395
611                  : 0.00021640905598463537
8283                 : 0.00032879238326170694
25                   : 5.0487823458630175e-05
6                    : 0.0003118325097843746
8                    : 0.00032663557615950447
19                   : 0.00013112179292607272
23                   : 0.00017122390637420355
29                   : 0.0001849098641574442
33                   : 0.0003386160040196027
35                   : 0.0007007673625519551
```

Ans 2.2 : Hubs

This method is used to measure the importance of web pages, where the root nodes are the web pages that are highly related to the provided query. Non-relevant pages that point to these root nodes are referred to as hubs. A good authority page will have many hubs pointing to it. Conversely, a page that many hubs link to is also considered important. The set of highly relevant web pages that are identified as roots are also known as potential authoritative pages.



```
Node Number :  HUB Score
30          :  0.00998179932694693
1412        :  0.0
3352        :  0.42573918623360957
5254        :  0.04750055792326323
5543        :  0.17590560962380986
7478        :  0.0
3           :  0.00508778113384111
28          :  0.045127947887486315
39          :  0.013485426941127372
54          :  0.003195859318214718
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         :  0.0001320128812490878
371         :  0.11913783267604111
567         :  0.00021405353127680848
581         :  0.0
584         :  0.0010122328790599718
586         :  0.007909025479646472
590         :  0.005022312349469831
604         :  0.0030297444530010776
611         :  0.0
8283        :  0.0
25          :  0.026958173031510563
6           :  0.13313457238975612
8           :  0.04048727601921059
19          :  0.009134709492141822
23          :  0.022935315769161524
29          :  0.08520031305859457
33          :  0.008836280469983377
35          :  0.02773555363917981
```

Compare the results obtained from both the algorithms in parts 1 and 2 based on the node scores.

The time required to evaluate scores in the HITS algorithm is typically greater than the time taken to evaluate scores using the PageRank algorithm. This is because HITS creates mutual reinforcement between authority and hub scores, whereas PageRank only considers authority. As a result, the HITS algorithm may yield less relevant results compared to PageRank.

PageRank's popularity is due to its efficiency, feasibility, and lower query time cost, among other features. These features are absent in the HITS algorithm, which may contribute to its lower popularity compared to PageRank.