Information Retrieval

Assignment-3

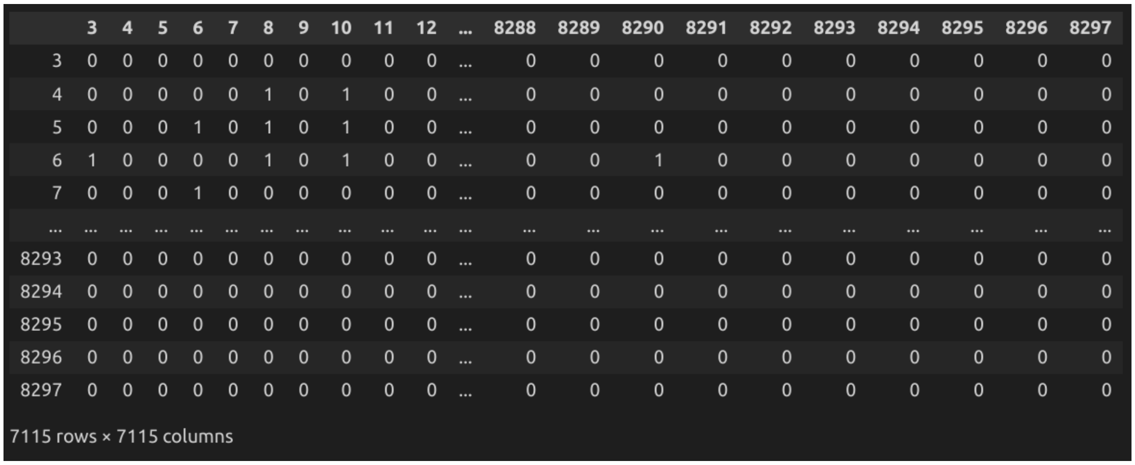
Shubham Agrawal (MT22124)   Amrita Aash (MT22011)   Ayush Agarwal (MT22095)

Question 1

The dataset chosen for question 1 is given in the following link:

https://snap.stanford.edu/data/wiki-Vote.html

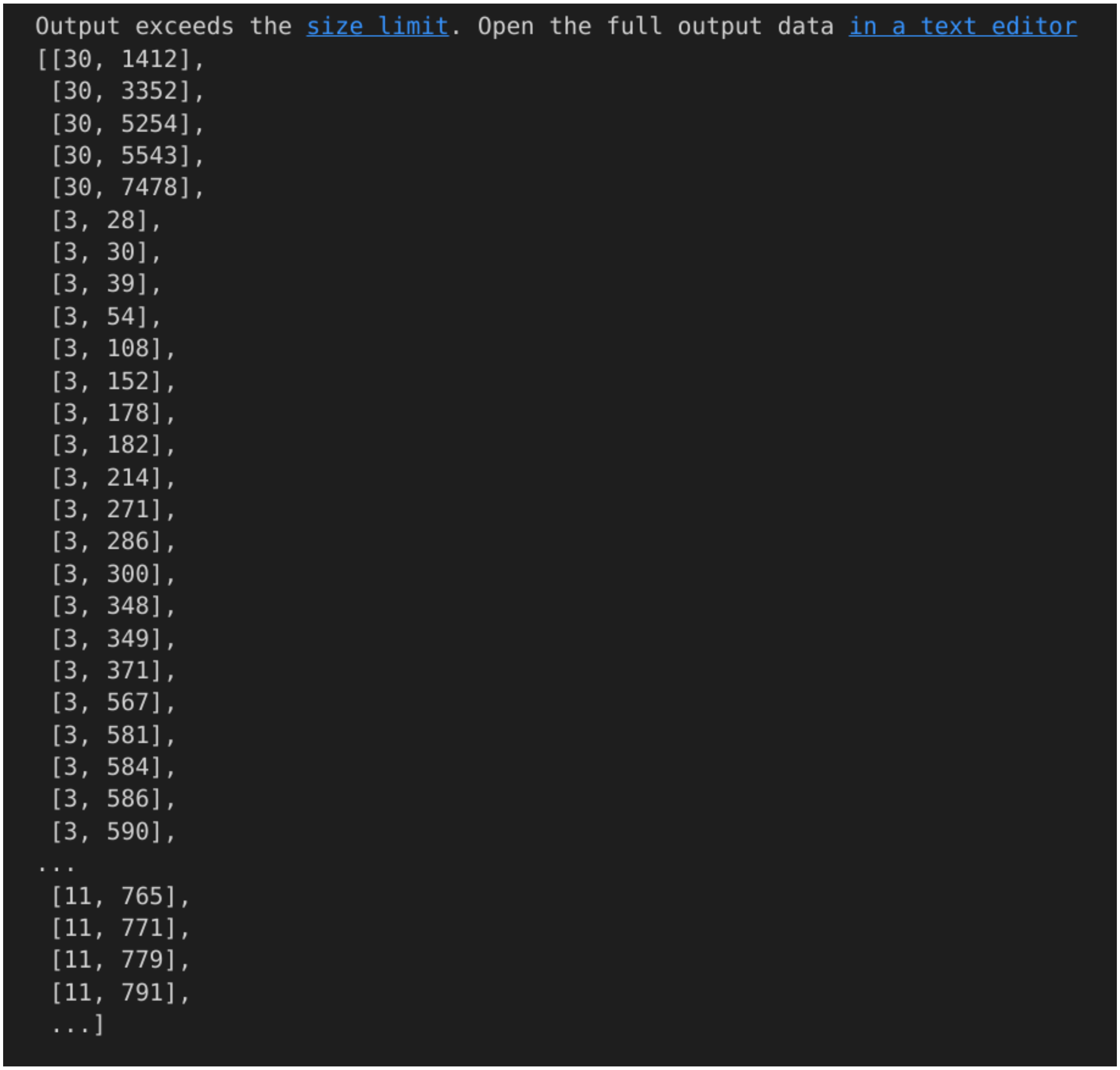
Adjacency matrix representation:



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



```
Output exceeds the size limit. Open the full output data in a text editor
[[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],
...
 [11, 765],
 [11, 771],
 [11, 779],
 [11, 791],
 ...]
```

Number of nodes:

7115

Number of edges:

103689

Avg In-degree:

Formula: Sum(in-degrees of all unique nodes)/length(all unique nodes)

14.573295853829936

Avg. Out-Degree:

Formula: Sum(out-degrees of all unique nodes)/length(all unique nodes)

14.573295853829936

Node with Max In-degree:

Node 4037 has maximum in-degree with in-degree as 457

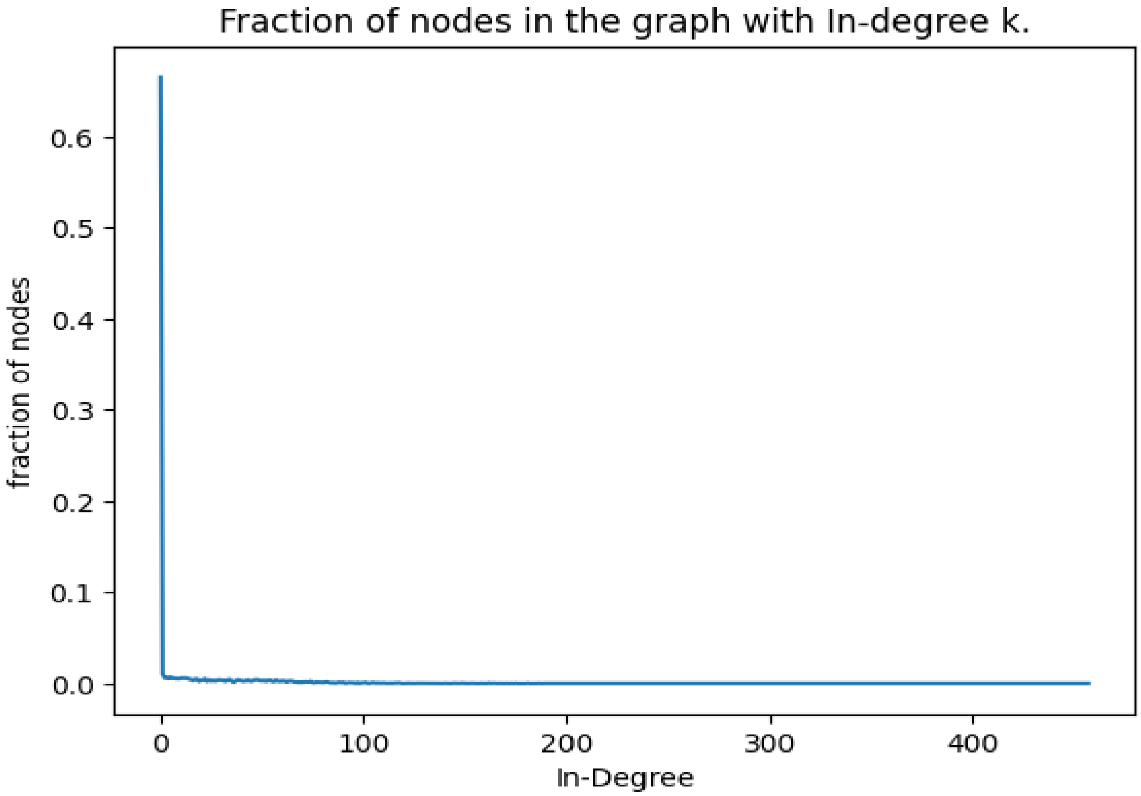
Node with Max out-degree:

Node 2565 has a maximum out-degree with out-degree as 893

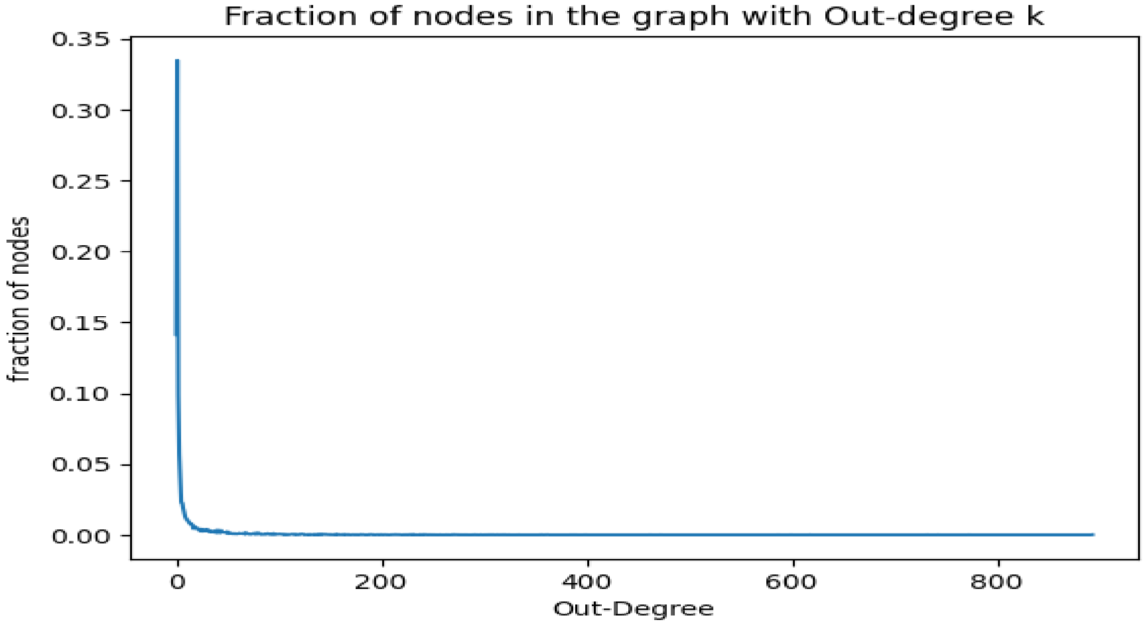
The density of the network:

0.0020485375110809584

In-degree distribution of the network:



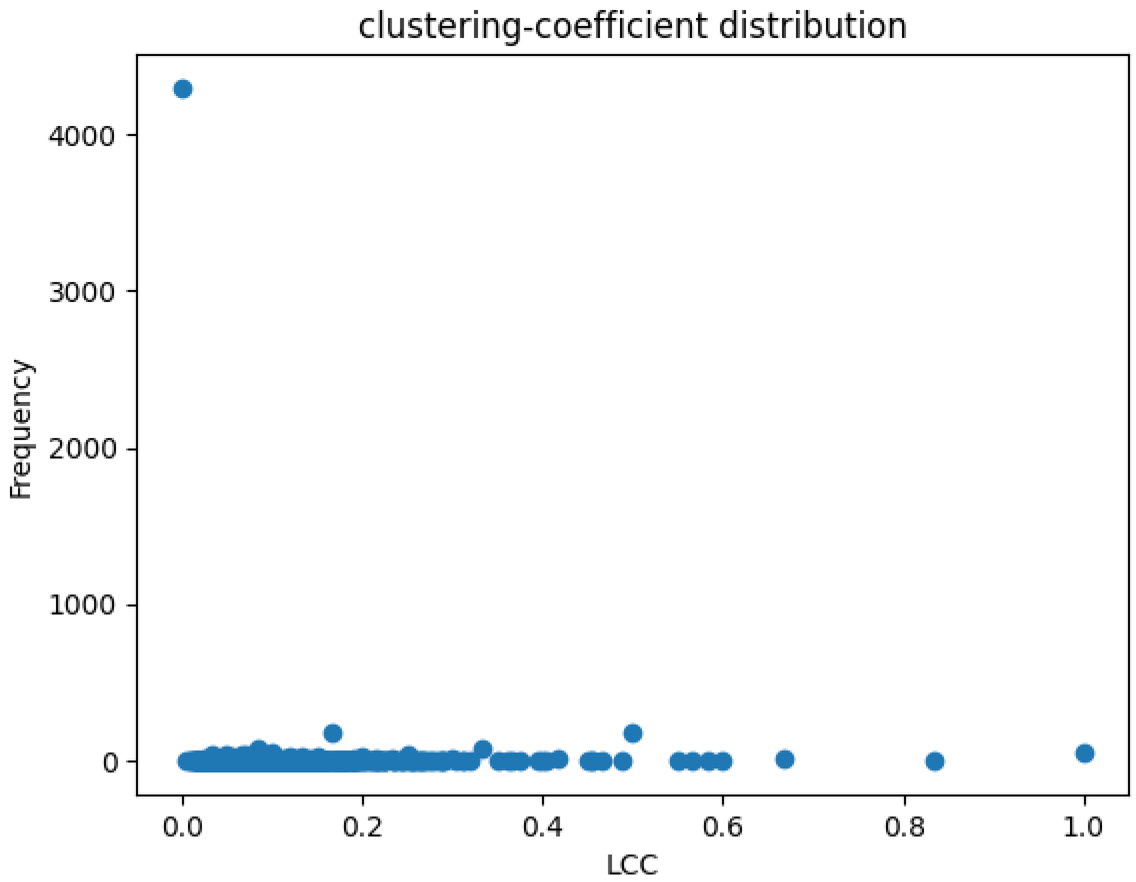
Out-degree distribution of the network:



Local clustering coefficient of each node:

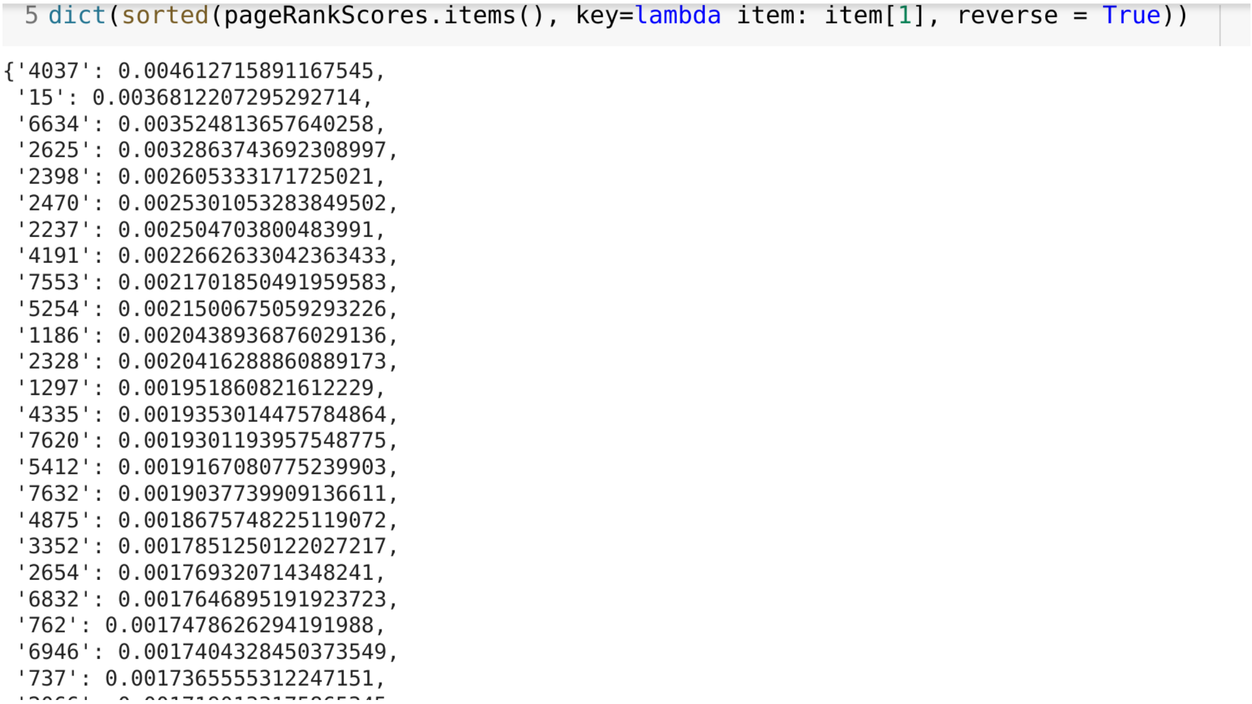
```
{3: 0.07509881422924901, 4: 0.13669950738916256, 5:
0.191699604743083, 6: 0.030659391432531737, 7:
0.09239130434782608, 8: 0.04319713435735535, 9:
0.05416666666666667, 10: 0.06757865937072503, 11:
0.02612160941473519, 12: 0.051201923076923075, 13:
0.11818181818181818, 14: 0.030740568234746156, 15:
0.06693877551020408, 16: 0.11904761904761904……}
```

Plot of local clustering coefficient vs frequency of LCC:



Question 2

Page Rank score of each node assigns a value defining the relative importance of that node with respect to all other nodes in that network.



```
dict(sorted(pageRankScores.items(), key=lambda item: item[1], reverse = True))
```

```
{'4037': 0.004612715891167545,
 '15': 0.0036812207295292714,
 '6634': 0.003524813657640258,
 '2625': 0.0032863743692308997,
 '2398': 0.002605333171725021,
 '2470': 0.0025301053283849502,
 '2237': 0.002504703800483991,
 '4191': 0.0022662633042363433,
 '7553': 0.0021701850491959583,
 '5254': 0.0021500675059293226,
 '1186': 0.0020438936876029136,
 '2328': 0.0020416288860889173,
 '1297': 0.001951860821612229,
 '4335': 0.0019353014475784864,
 '7620': 0.0019301193957548775,
 '5412': 0.0019167080775239903,
 '7632': 0.0019037739909136611,
 '4875': 0.0018675748225119072,
 '3352': 0.0017851250122027217,
 '2654': 0.001769320714348241,
 '6832': 0.0017646895191923723,
 '762': 0.0017478626294191988,
 '6946': 0.0017404328450373549,
 '737': 0.0017365555312247151,
```

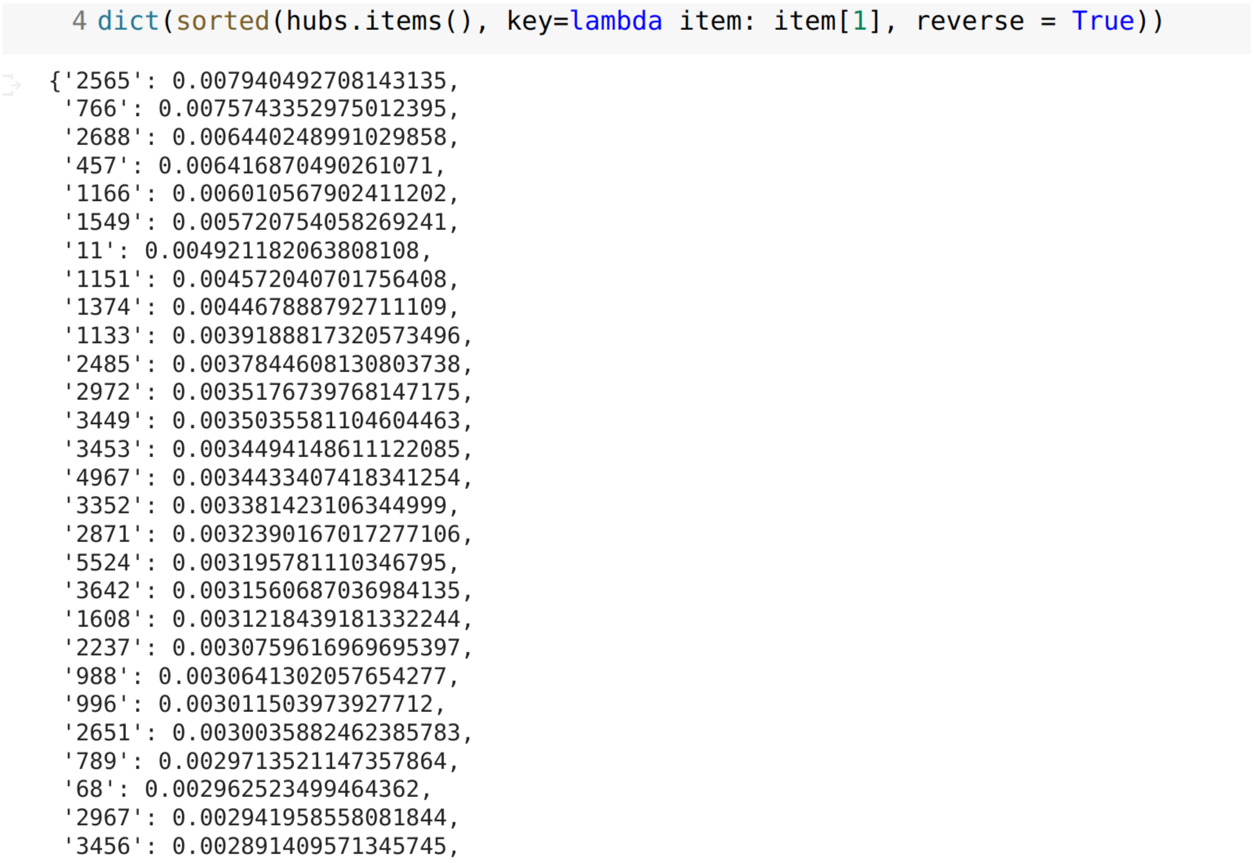
The Hits algorithm assigns two scores hub, authority to every node in the network.

A good authority score means, lots of pages point to you for your information.(in-degree).

A good hub score means, you act as a good compilation of a lot of pages on a specific category you claim upon.(out-degree)

The Authority score of a node is the InDegree(incoming nodes) from the hub whereas the Hubs score of a node is the Out-Degree(outgoing nodes) from the authorities. Hubs and Authorities score is calculated using the HITS algorithm built-in in the Networkx library.

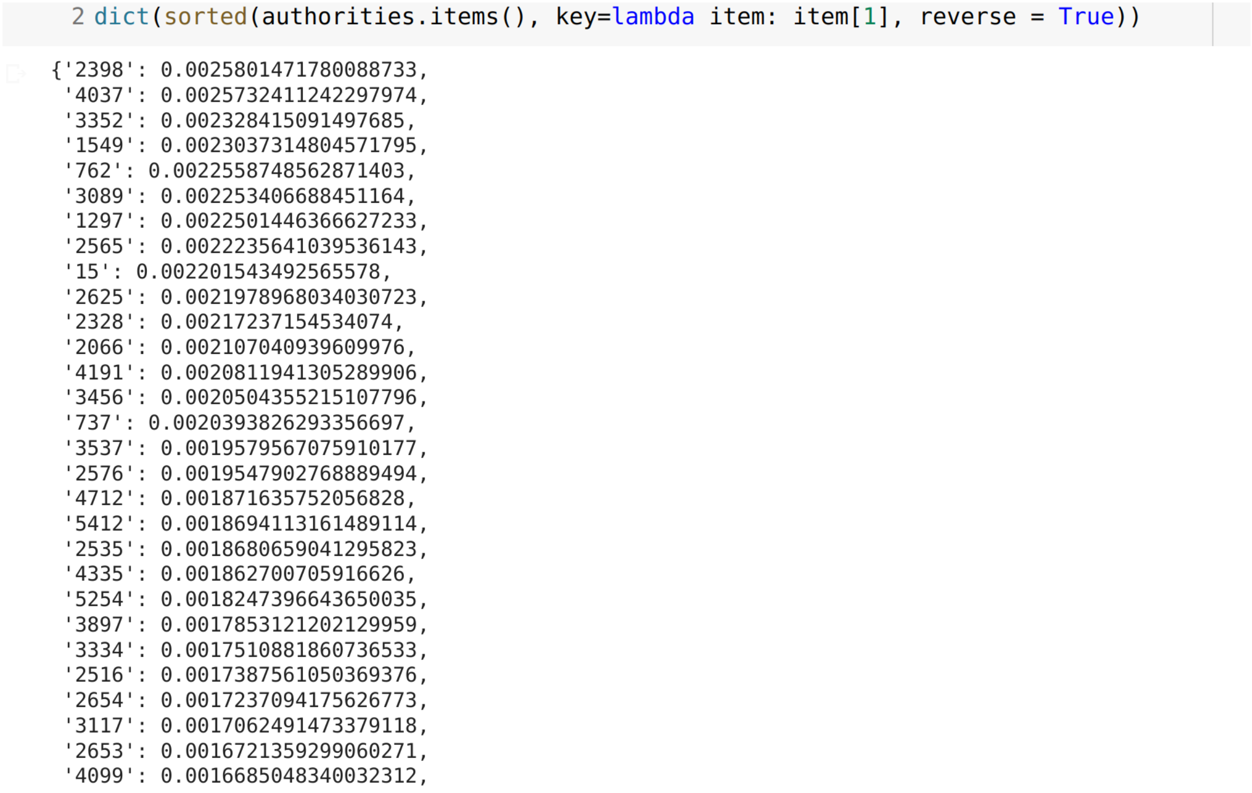
Hub score for every node



```
dict(sorted(hubs.items(), key=lambda item: item[1], reverse = True))
```

```
{'2565': 0.007940492708143135,
 '766': 0.0075743352975012395,
 '2688': 0.006440248991029858,
 '457': 0.006416870490261071,
 '1166': 0.006010567902411202,
 '1549': 0.005720754058269241,
 '11': 0.004921182063808108,
 '1151': 0.004572040701756408,
 '1374': 0.004467888792711109,
 '1133': 0.0039188817320573496,
 '2485': 0.0037844608130803738,
 '2972': 0.0035176739768147175,
 '3449': 0.0035035581104604463,
 '3453': 0.0034494148611122085,
 '4967': 0.0034433407418341254,
 '3352': 0.003381423106344999,
 '2871': 0.0032390167017277106,
 '5524': 0.003195781110346795,
 '3642': 0.0031560687036984135,
 '1608': 0.0031218439181332244,
 '2237': 0.0030759616969695397,
 '988': 0.0030641302057654277,
 '996': 0.003011503973927712,
 '2651': 0.0030035882462385783,
 '789': 0.0029713521147357864,
 '68': 0.002962523499464362,
 '2967': 0.002941958558081844,
 '3456': 0.002891409571345745,
```

Authority score of every node



```
dict(sorted(authorities.items(), key=lambda item: item[1], reverse = True))
```

```
{'2398': 0.002580147178008873З,
 '4037': 0.0025732411242297974,
 '3352': 0.002328415091497685,
 '1549': 0.0023037314804571795,
 '762': 0.002255874856287140З,
 '3089': 0.002253406688451164,
 '1297': 0.0022501446366627233,
 '2565': 0.0022235641039536143,
 '15': 0.002201543492565578,
 '2625': 0.0021978968034030723,
 '2328': 0.00217237154534074,
 '2066': 0.002107040939609976,
 '4191': 0.0020811941305289906,
 '3456': 0.0020504355215107796,
 '737': 0.0020393826293356697,
 '3537': 0.0019579567075910177,
 '2576': 0.0019547902768889494,
 '4712': 0.001871635752056828,
 '5412': 0.0018694113161489114,
 '2535': 0.0018680659041295823,
 '4335': 0.001862700705916626,
 '5254': 0.0018247396643650035,
 '3897': 0.0017853121202129959,
 '3334': 0.0017510881860736533,
 '2516': 0.0017387561050369376,
 '2654': 0.0017237094175626773,
 '3117': 0.0017062491473379118,
 '2653': 0.0016721359299060271,
 '4099': 0.0016685048340032312,
```

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

From the results above, both the pagerank and authorities work upon incoming nodes but still shows different nodes as important because pagerank takes nodes with maximum in-degree i.e. for node 4037 whereas, authority score is calculated using maximum incoming nodes from the hubs, so that doesn’t includes total incoming nodes.Here. node with maximum authority score is 2398.

In Hubs, the Outgoing nodes are referred and thus, 2565 has maximum OutDegree nodes.