Ankitkumar Kushwaha (MT21010)

Rutvikkumar Bandhaniya (MT21116)

Q-1).

- First of all, we have picked the dataset which has named “Wiki-Vote”.
- Wikipedia, is a free encyclopedia written collaboratively by volunteers around the world. A small part of Wikipedia contributors are administrators, who are users with access to additional technical features that aid in maintenance. In order for a user to become an administrator a Request for adminship (RfA) is issued and the Wikipedia community via a public discussion or a vote decides who to promote to adminship. Using the latest complete dump of Wikipedia page edit history (from January 3 2008) we extracted all administrator elections and vote history data. This gave us 2,794 elections with 103,663 total votes and 7,066 users participating in the elections (either casting a vote or being voted on). Out of these 1,235 elections resulted in a successful promotion, while 1,559 elections did not result in the promotion. About half of the votes in the dataset are by existing admins, while the other half comes from ordinary Wikipedia users.

- Dataset statistics

Nodes	7115
Edges	103689
Nodes in largest WCC	7066 (0.993)
Edges in largest WCC	103663 (1.000)
Nodes in largest SCC	1300 (0.183)
Edges in largest SCC	39456 (0.381)
Average clustering coefficient	0.1409
Number of triangles	608389
Fraction of closed triangles	0.04564
Diameter (longest shortest path)	7
90-percentile effective diameter	3.8

Analysis:

- The chosen dataset has data in form of text and the first four line contain the normal text related to the dataset. From fifth line onwards, the main information starts about an adjacency list for graph representation.

- Here, we have represented our network dataset in terms of its ‘adjacency matrix’ as you can see on the below snippet.

Matrix Representation of graph

+ Code + Text

```
[ ]  pd.DataFrame(wiki_matrix)
```

	0	1	2	3	4	5	6	7	8	9	...	8288	8289	8290	8291	8292	8293	8294	8295	8296	8297
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
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	0	0	0	1	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

8298 rows × 8298 columns

- Here, we have represented our network dataset in terms of its ‘edge list’ as you can see on the below snippet.

```
print(edgelist)
```

```
{30: [1412, 3352, 5254, 5543, 7478], 3: [28, 30, 39, 54, 108, 152, 178, 182, 214, 271, 286, 300, 348, 349, 371, 567, 581, 584, 586, 590, 604, 611, 8283], 25: [3,
```

- Also, we have listed down all the answer that has been asked in the question.

```
Number of Nodes  7115
```

```
Number of Edges  103689
```

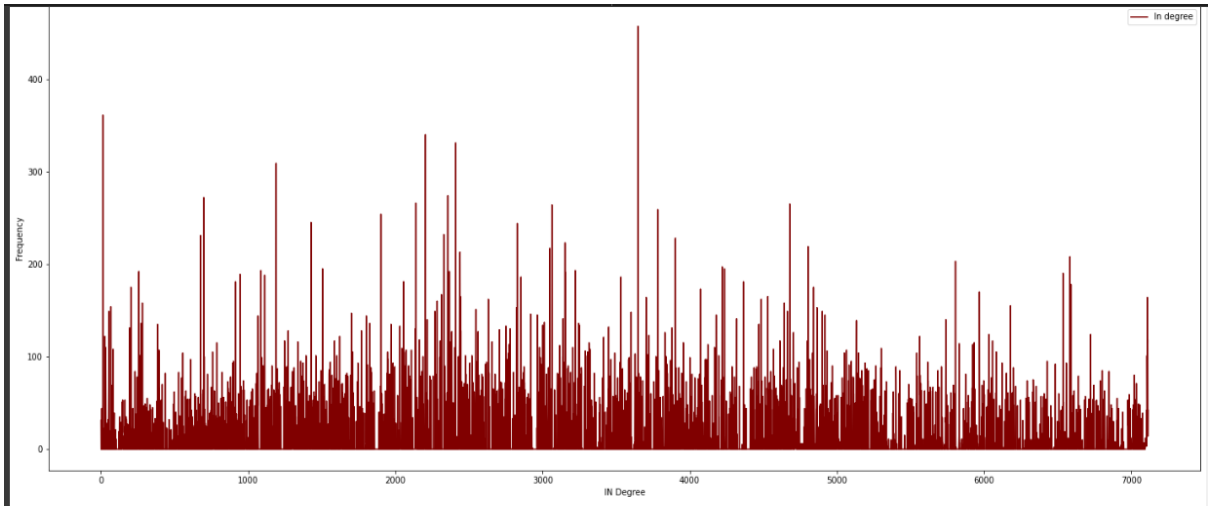
```
Average IN Degree  14.573295853829936
Average OUT Degree  14.573295853829936
```

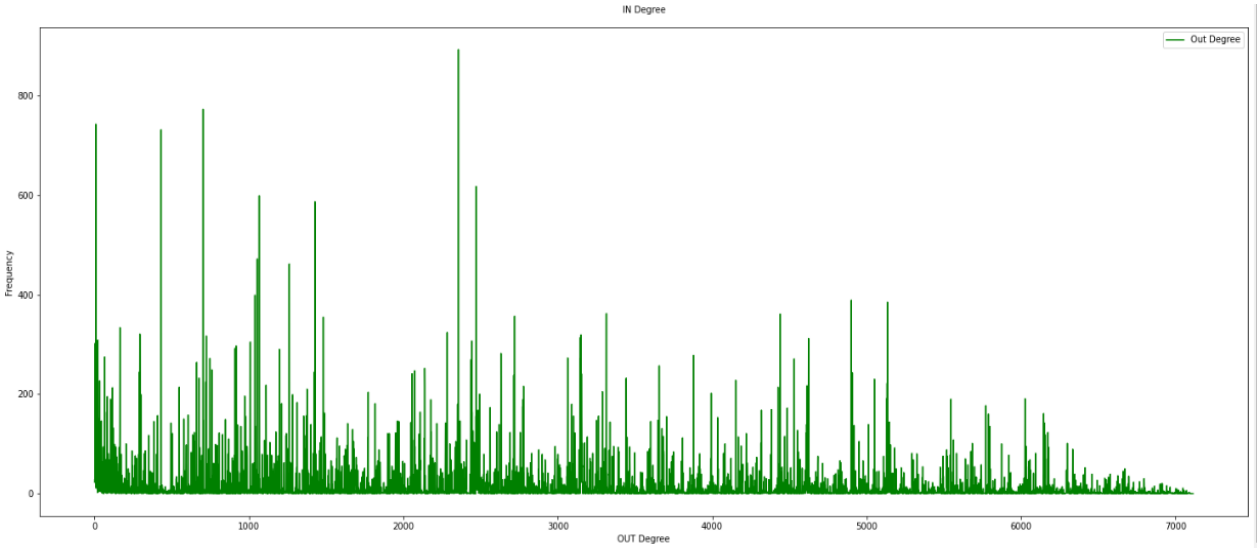
```
 Maximum IN degree node:   4037  Maximum IN degree:   457
 Maximum OUT degree node:  2565  Maximum OUT degree:  893
```

```
Density of Network:  0.002048537511080958 4
```

1. Plot degree distribution of the network (in case of a directed graph, plot in-degree and out-degree separately).

_____





2. Calculate the local clustering coefficient of each node and plot the clustering-coefficient distribution of the network.

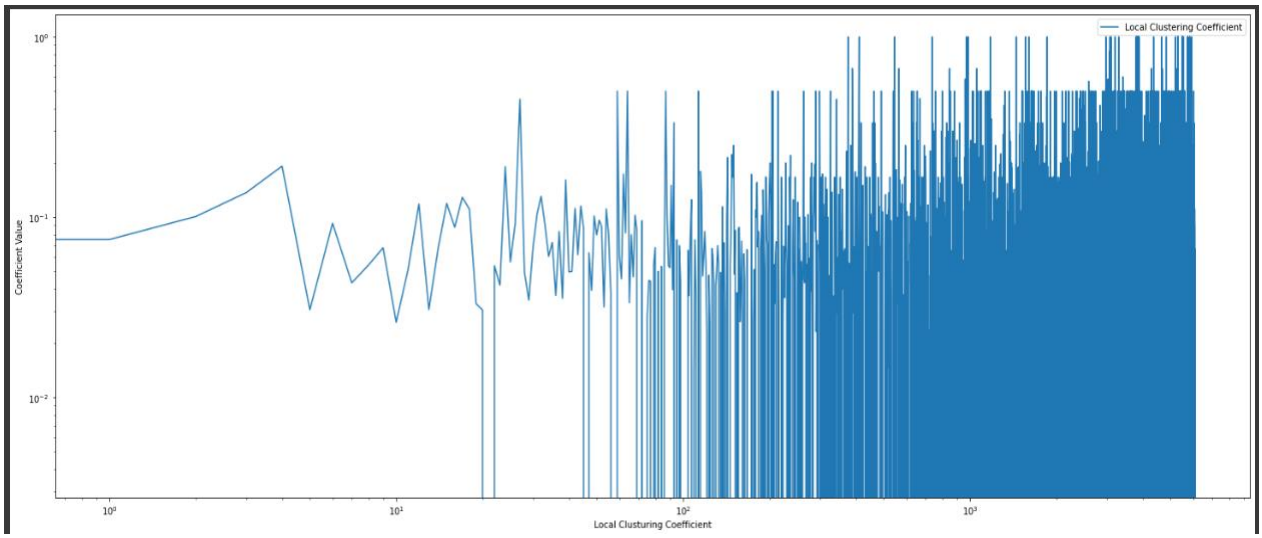
- Here is the calculation of the local clustering coefficient for each and below snippet shows the same.

Local Clustering Coefficient Calculation

clustering coefficient for directed graph = Li/(di(di-1))

"di" is the number of neighbours of a vertex

"Li" is the number of edges between the "ki" neighbors of node i**



Q-2).

- Here we have chosen the same dataset which we have in question 1 and calculate the below part.
- We have used NetworkX library which helps us for creation, manipulation, and study for the various network.
- Also, we have used HITS algorithm for calculating the hub and authority score of nodes.

1.

- We have taken default value and maximum number of nodes of iteration is 100.
- Here we have calculated the PageRank score for each node in the reverse sorted order and you can see in the below snippet.

```
[('2565', 0.009438010519173635),
 ('1166', 0.0054741331852336626),
 ('1549', 0.005257993008557954),
 ('1374', 0.0044513297508182335),
 ('1151', 0.004064182817336744),
 ('5524', 0.003981891848595786),
 ('5802', 0.0038489062118703997),
 ('2972', 0.0036585632725494163),
 ('1608', 0.0033514143824402405),
 ('2658', 0.003183308039213882),
 ('6', 0.0031053690065793176),
 ('5189', 0.0030765166009585993),
 ('2485', 0.0030485066459949514),
 ('3453', 0.0030448100594621295),
 ('722', 0.00297549913518371 3),
 ('1305', 0.0029303930116693754),
 ('789', 0.0028883136248367435),
 ('2871', 0.00282418359518539 2),
 ('4310', 0.0027483425155386183),
 ('3352', 0.0027478998710408573),
 ('3447', 0.00272295423345969),
 ('5079', 0.0026065279264447907),
 ('2651', 0.0025635647599803264),
 ('737', 0.002558753867938459),
 ('813', 0.0024919059607141905),
 ('5800', 0.0024321508597427116),
 ('826', 0.0024264417276790035)
```

2.

- Hyperlink Induced Topic Search algorithm is basically the link search algorithm which helps us to rates the webpages by calculating the Hub and Authority Score.
- You can see the Authority Score in below snippet.

```
[('6950', 0.12283410394734937),
 ('6661', 0.10653852166080613),
 ('5880', 0.09243736483739717),
 ('6721', 0.041648455079773036),
 ('7765', 0.03669042375810286),
 ('5756', 0.029872942575773703),
 ('5514', 0.0291715896085008),
 ('5388', 0.027747062827709345),
 ('5778', 0.0277400218117607 2),
 ('8090', 0.02666576854702607),
 ('7952', 0.024849915364120328),
 ('5368', 0.024613239388955283),
 ('6088', 0.023962386903405093),
 ('4779', 0.021143293988495942),
 ('8169', 0.017644181968514114),
 ('4625', 0.015374994667565015),
 ('8024', 0.015111571374520627),
 ('7257', 0.015072272495179896),
 ('5260', 0.014292341550572438),
 ('4778', 0.012706782527714146),
 ('8296', 0.011371677235715744),
 ('8025', 0.011300819459283854),
 ('5406', 0.010243322284270488),
 ('7088', 0.009651261644437486),
 ('7520', 0.009601689847088698),
 ('4809', 0.009282499863184817),
 ('4216', 0.008653977178675904),
 ('4815', 0.008586889948399058),
```

- You can see the Hub Score for each node in below snippet.

```
[('6483', 0.06875071619575998),
 ('6660', 0.05973482460250689),
 ('6132', 0.04952365829624222),
 ('7818', 0.04647976616379795),
 ('7682', 0.035965613515672325),
 ('5692', 0.03492567567071872),
 ('5203', 0.03013335775093676),
 ('6916', 0.02898995439540561),
 ('4967', 0.02496098362783597),
 ('6460', 0.02021832376744933),
 ('7614', 0.01992553447038253),
 ('5085', 0.019217575385328263),
 ('2346', 0.018901559661122773),
 ('7900', 0.01707325487573716 3),
 ('7132', 0.016084623094404626),
 ('7086', 0.0160285104782263),
 ('5531', 0.015855920904449993),
 ('7743', 0.015425659904582133),
 ('7964', 0.014944338443040824),
 ('7057', 0.014302661665654927),
 ('7727', 0.0138013414935524 3),
 ('6620', 0.0133827594561573 8),
 ('5190', 0.011146723932348957),
 ('5755', 0.010987855062174793),
 ('6615', 0.010034651901237077),
 ('5961', 0.009322797804678803),
 ('7055', 0.00837819284725244),
 ('5468', 0.0083645514213361 4),
 ('5320', 0.008048433043828957),
```

Comparison:

- Based upon the result we get and the analysis we have done here, the nodes with are different marked considered as a highly recommendable and important for the score metrics. However, both the score of PageRank and authority are helpful to understand the relevance and important of web page according to their link which are incoming.
- So, PageRank score refer to the node which has maximum in-degree (4037 node).
- Other side, Authority score is high for those who has link which is incoming from the maximum number of hub nodes. This may not be included all in-degree from the network.
- For Hub score, we have referred node with outgoing nodes and as a result the node with maximum outgoing links is 2565.

Results for 80:20

```
[ ] solution(0.8,10)

    Accuracy  0.91991991991992

    Matrix
     [[188  25   5   3  15]
     [  0 157   5   1   6]
     [  0   1 195   0   1]
     [  0   2   6 200   1]
     [  3   2   4   0 179]]
```

Results for 70:30

```
[ ] solution(0.7,10)

    Accuracy  0.9166110740493663

    Matrix
     [[299  38   9   5  23]
     [  1 233   5   1   8]
     [  0   1 266   0   3]
     [  0   3  14 308   2]
     [  4   1   7   0 268]]
```

Results for 50:50

```
[ ]  solution(0.5,10)

     Accuracy  0.8947157726180944

     Matrix
      [[480  62  18   8  47]
      [  1 403  21   1  14]
      [  0   3 401   1   4]
      [  3   8  34 513   5]
      [  9   4  20   0 438]]
```

Learnings: -

- From this assignment, we have learnt a lot of things like calculating the number of nodes, edges, Average In-degree, Average out-degree and density as well.
- Also, we have learnt how to calculate the PageRank, authority and hub score for each node.
- We have taken a look on HITS algorithm and seen that how it works with the nodes and many such things we have learnt from this assignment.
- Overall, learning experience id quite good from it and it has gained our knowledge in a way or other.