REPORT - ASSIGNMENT 3

Ankit Mishra & Rajat Prakash

DATASET DESCRIPTION

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.

The network contains all the Wikipedia voting data from the inception of Wikipedia till January 2008. Nodes in the network represent wikipedia users and a directed edge from node *i* to node *j* represents that user *i* voted on user *j*.

PREPROCESSING

1. Data is given in .txt file format.
2. Splitted the text data to create nodes.
3. Creation of set of nodes.
4. Created an adjacency matrix graph.

METHODOLOGY

Q1.

1. Number of Nodes -
Count of unique separate vertices.
7115

2. Number of Edges -
All connections between nodes.
103689

3. Avg In-degree -
Sum of Number of incoming edges for all nodes / Number of nodes
14.573295853829936

4. Avg. Out-Degree -
14.573295853829936

5. Node with Max In-degree -
Node number with Max in-degree
4037

REPORT - ASSIGNMENT 3

Ankit Mishra & Rajat Prakash

6. Node with Max out-degree -

Node number with Max out-degree
2565

7. The density of the network -

Formula used - len(edge_list)/(len(s)*(len(s)-1))
0.0020485375110809584

Local Clustering Formula

- The local clustering coefficient of a node in a graph is given by dividing the number of links between nodes within its neighborhood by the number of possible links that could exist between them.
- For a directed graph, for each neighborhood, there are K * (K - 1) where K is the number of nodes in its neighborhood.

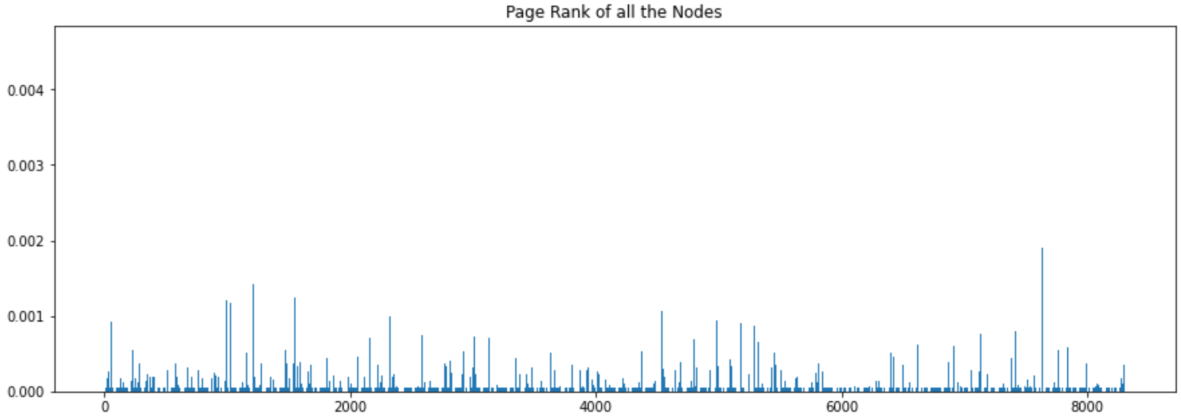
LCF = |Links between vertices in neighborhood| / (K * (K - 1))

Q2.

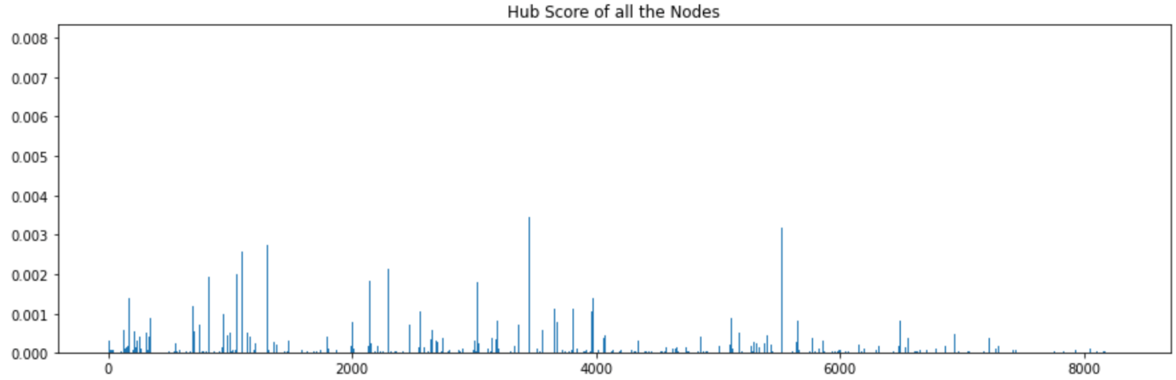
1. Created a digraph using networkx library.
2. Plotting of page rank for all the nodes. (Using python matplotlib library)

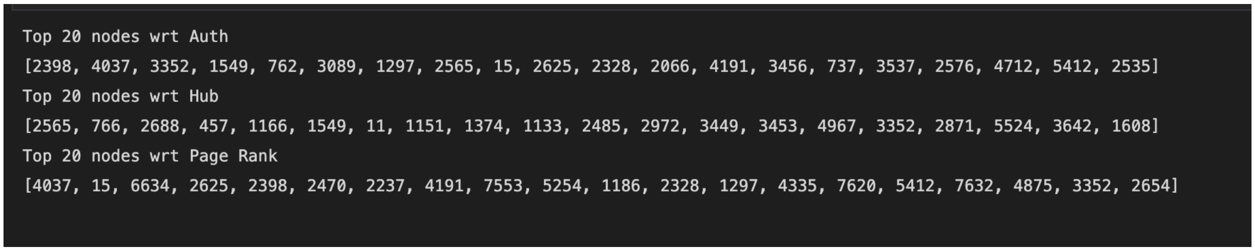
ANALYSIS

PageRank (PR) is a calculation, created by Larry Page and Sergey Brin for the evaluation of any website on the scale 1 to 10.
Page rank scores for most of the nodes are around or under 0.001. There are some nodes whose page rank scores lie between 0.001 and 0.002.



Hub score is the sum of all the authority scores of pages it points to.
We can observe that most of the nodes' scores lie under 0.001 and others are under 0.003. Further, scores for 2 of them are between 0.003 and 0.004.

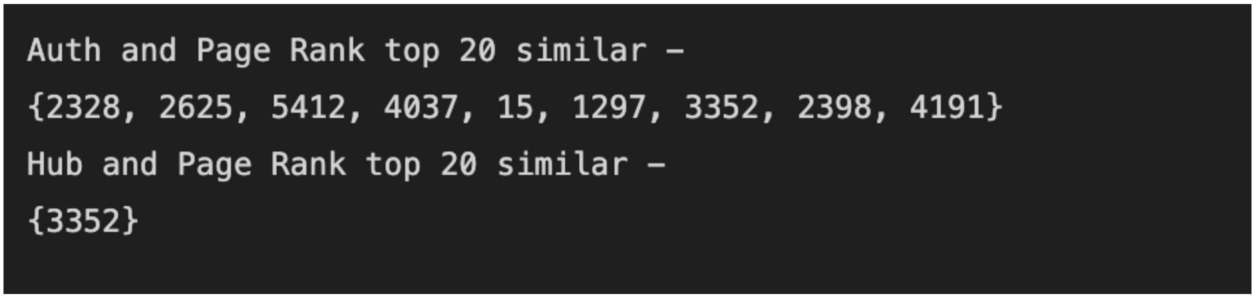




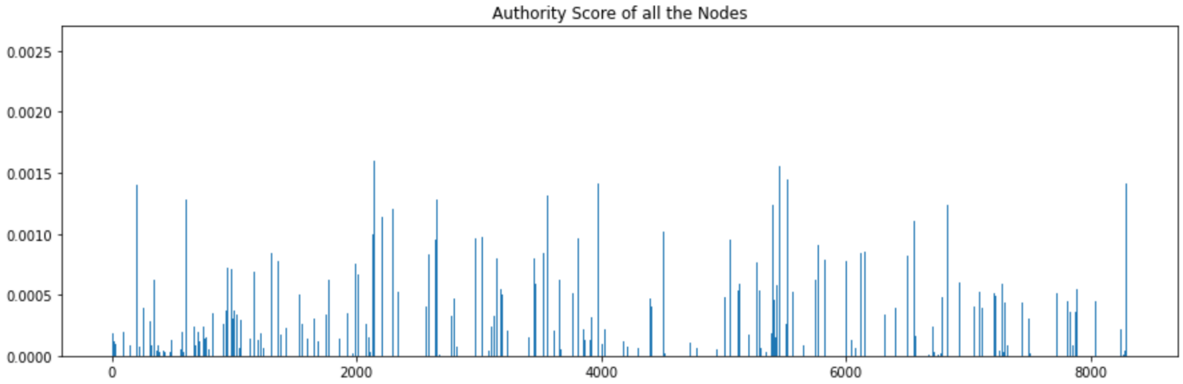
```
Top 20 nodes wrt Auth
[2398, 4037, 3352, 1549, 762, 3089, 1297, 2565, 15, 2625, 2328, 2066, 4191, 3456, 737, 3537, 2576, 4712, 5412, 2535]
Top 20 nodes wrt Hub
[2565, 766, 2688, 457, 1166, 1549, 11, 1151, 1374, 1133, 2485, 2972, 3449, 3453, 4967, 3352, 2871, 5524, 3642, 1608]
Top 20 nodes wrt Page Rank
[4037, 15, 6634, 2625, 2398, 2470, 2237, 4191, 7553, 5254, 1186, 2328, 1297, 4335, 7620, 5412, 7632, 4875, 3352, 2654]
```

REPORT - ASSIGNMENT 3

Ankit Mishra & Rajat Prakash



```
Auth and Page Rank top 20 similar -
{2328, 2625, 5412, 4037, 15, 1297, 3352, 2398, 4191}
Hub and Page Rank top 20 similar -
{3352}
```



- Pagerank votes vertices 15 and 2565 higher signifying that a lot of important pages have their links in them.
- These are voted slightly lower by authority score showing that they point to a lot of pages but there are more pointed at.
- Pagerank and hit scores differ a bit but have more than 5 nodes in common in the top 20.
- Nodes that may be less scored by page rank are being more ranked by authority and hub scores
- Authority and page rank scores have a similar graph and distribution.

This shows the basic similarities between the two algorithms, and how Auth are the real root nodes, which are being pointed by many hubs(related nodes) which is the same as what page rank algorithms try to find.