Q2- Report

Problem Statement : To determine the most relevant sites in the provided dataset, by applying the PageRank and authority-hub algorithms.

Dataset: Wikivote data, which consists of 7115 nodes and 103689 edges, was utilised for this.

Graph : After importing the provided data using Python's standard file reading function, we divide it into informational and graphical components. The graph connection is in the Graph section, while the Information component offers details and statistics about the data in the variable ‘info’ .

We use Python networkx package, to create graph using dataframe. In order to generate the graph used nx.DiGraph(). We used the nx.add_edge() method to add edges.

PageRank :
We initially take equal probabilities of all nodes to be most relevant page. We then initialized a loop for all nodes in which we recalculate the probability of each node by dividing its current probability by the no. of outgoing edges.This value is added to all nodes having 0 score.

$P(a) = \Sigma p(i)/O(i)$
O(i) - number of outgoing edges
p(i) - current pagerank score.

For convergence criteria we use mean squared error and set threshold e=1e-15.

Authority and Hub :
Authority score is for incoming edges while hub score is for outgoing edges.
We take all hub and authority scores = 1 .
We then keep a list of authority scores and hub scores as previous pointers.

Formula :
authority scores : $auth(p) = \Sigma\ hub(q)$
hub scores : $hub(p) = \Sigma\ auth(q)$

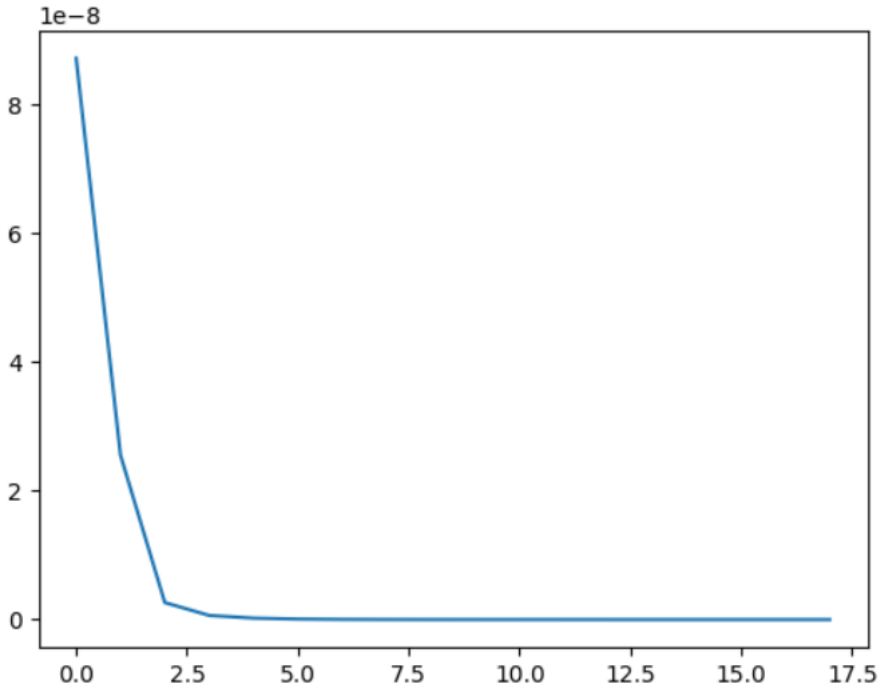
The buffer is updated using MSE errors. Threshold = 1e-20 and no of iter = 100 (max).

Results :

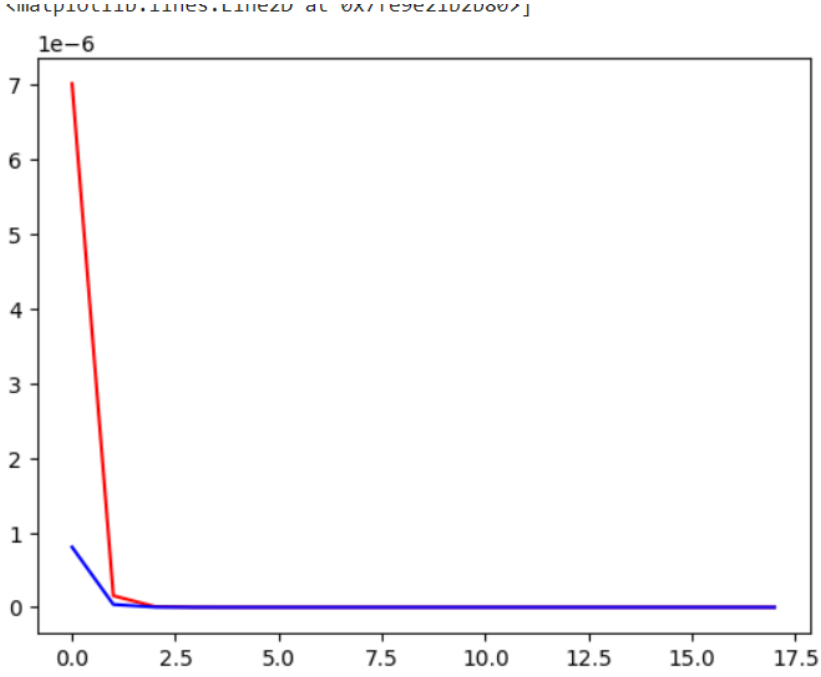
```
top 10 rank scores:
NodeId   Score
4037     0.0019241311389590733
15       0.0015369371893900383
6634     0.0014988609865573734
2625     0.001371535149427395
2398     0.0010896489086301506
2470     0.0010539097484338054
2237     0.001042634004914745
4191     0.0009472408397925709
7553     0.0009062507094909314
5254     0.0008980262120648992


top 10 authority scores:
NodeId   Score
2398     0.09211925178717026
4037     0.09187268402901715
3352     0.08313163612682965
1549     0.0822503546435512
762      0.08054172500769863
3089     0.08045360346596589
1297     0.08033713852810939
2565     0.07938813074309921
15       0.07860192662339409
2625     0.078471728898818


top 10 hub scores:
NodeId   Score
2565     0.2191839488482209
766      0.20907678925989998
2688     0.1777722438561388
457      0.17712691955651996
1166     0.1659116200118411
1549     0.15791179639870212
11       0.13584095614738909
1151     0.12620349562227773
1374     0.123328557131714
1133     0.10817414045591804
```

Error plots :

PageRank:



Authority-Hub :



Comparison :

By comparing the top nodes identified by each algorithm, we can see that they are not the same. The PageRank algorithm identifies nodes that have a high number of incoming links from other important nodes, while the Authority and Hub algorithm identifies nodes that have both high-quality incoming and outgoing links. However, we can see some overlap between the two algorithms in terms of the top nodes identified, which suggests that they are complementary measures of node importance in the network.