IR Assignment -3

Group Number 22

Palak tiwari Rahul Gupta Deepti Gupta Mohit Ghai Ravi Rathee

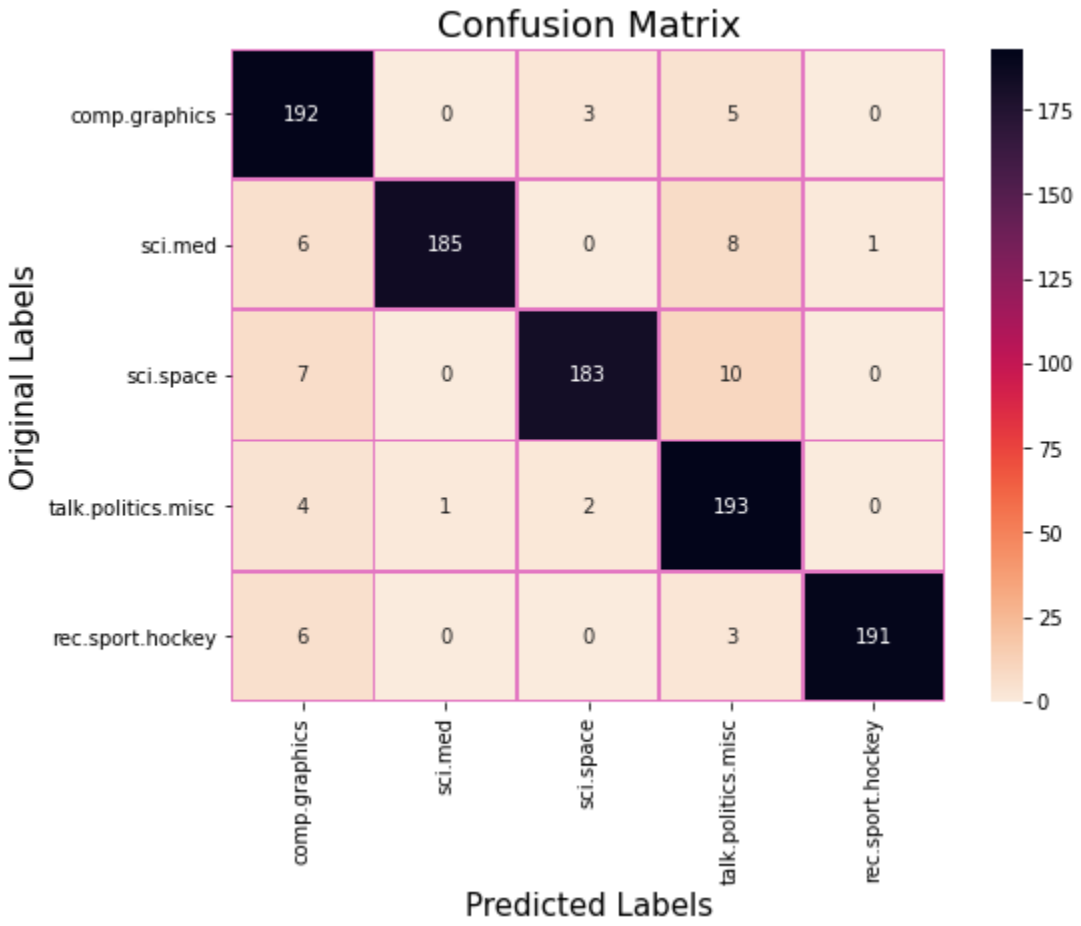
Analysis

Ans 1)

Outputs On different split -

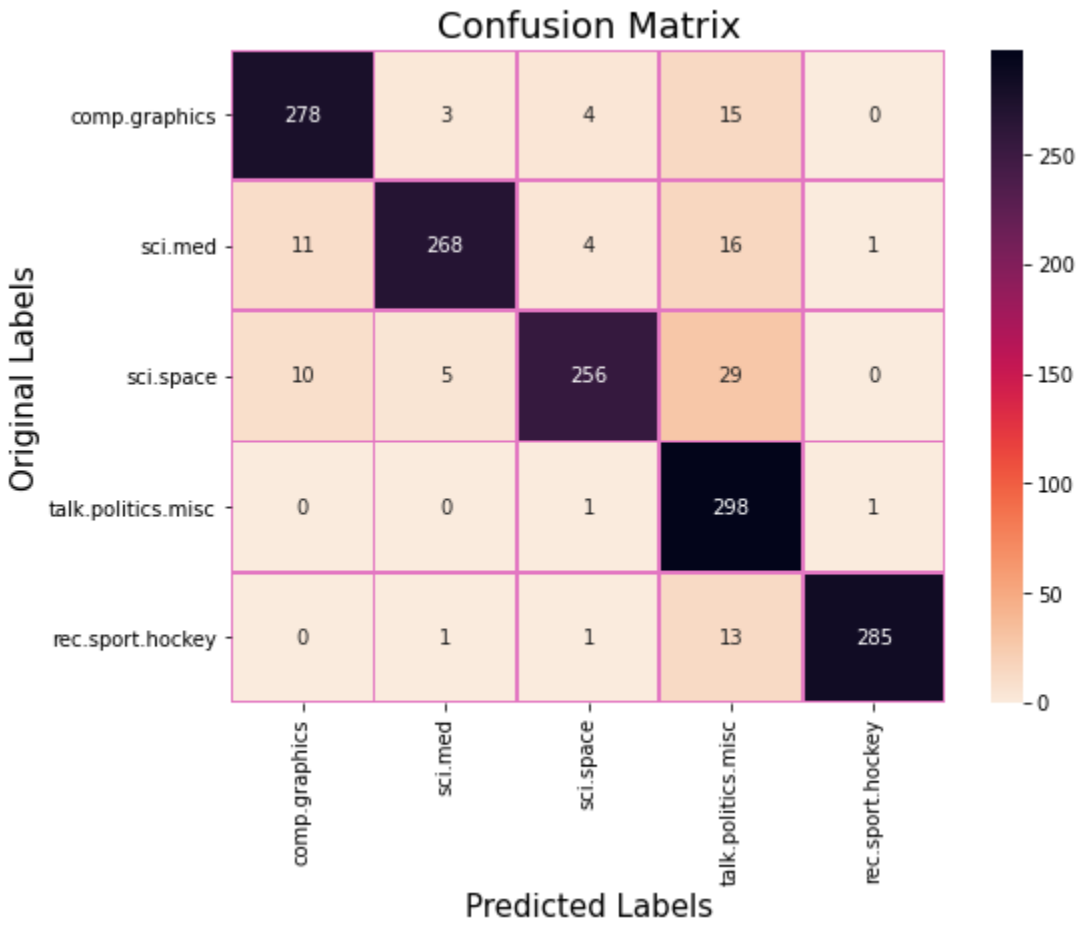
80:20 Split -

Accuracy achieved on Test Data was : **94.4%**



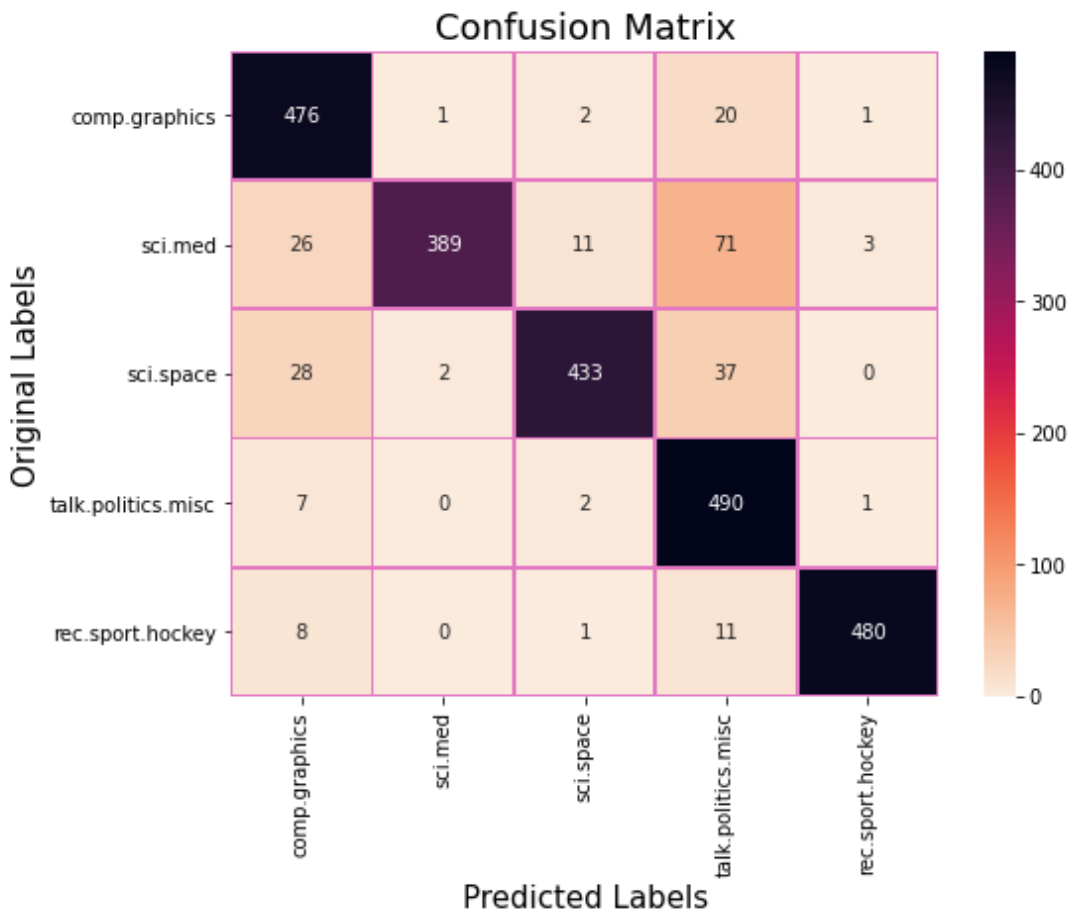
<u>70:30</u> Split -

Accuracy achieved on Test Data was **92.33%**



<u>50:50</u> Split -

Accuracy achieved on Test Data was **90.7%**



Analysis -

On different split ratios we observed the best accuracy on Test data was 80:20 split, as the more data in the training set better is the learning and hence better is the performance on Test Data. However the performance of the classifier also depends on the value of 'K' as well as splitted documents, as the splitting was random.

Observation : On increasing the Train data the accuracy is better, on a smaller test set.

Ans 2)

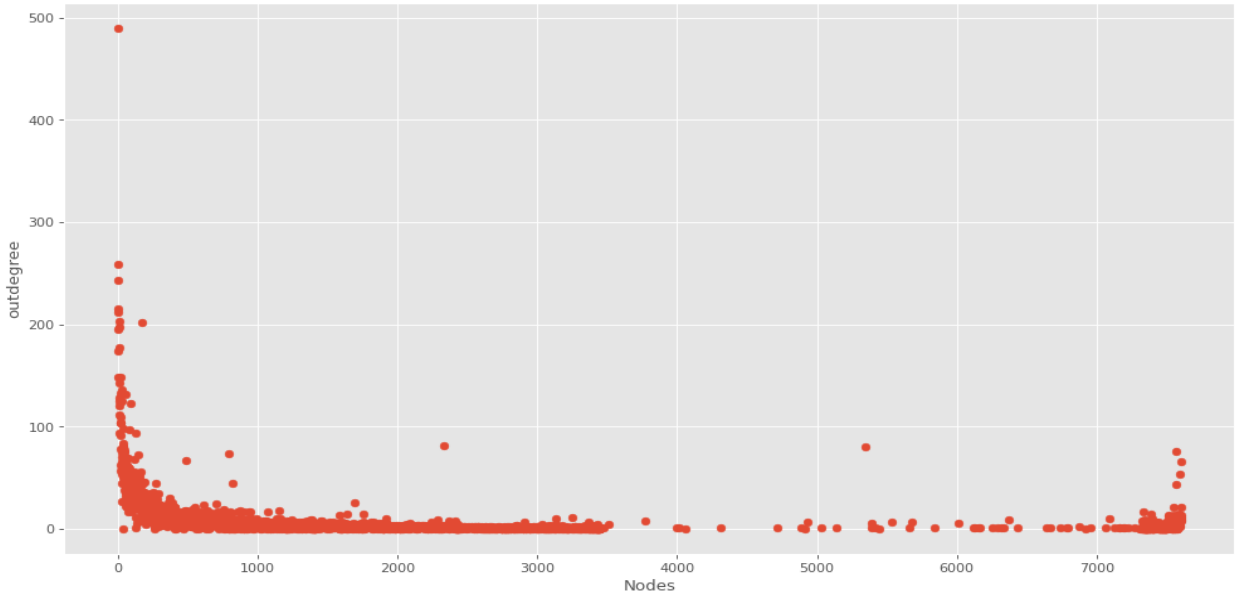
(A)

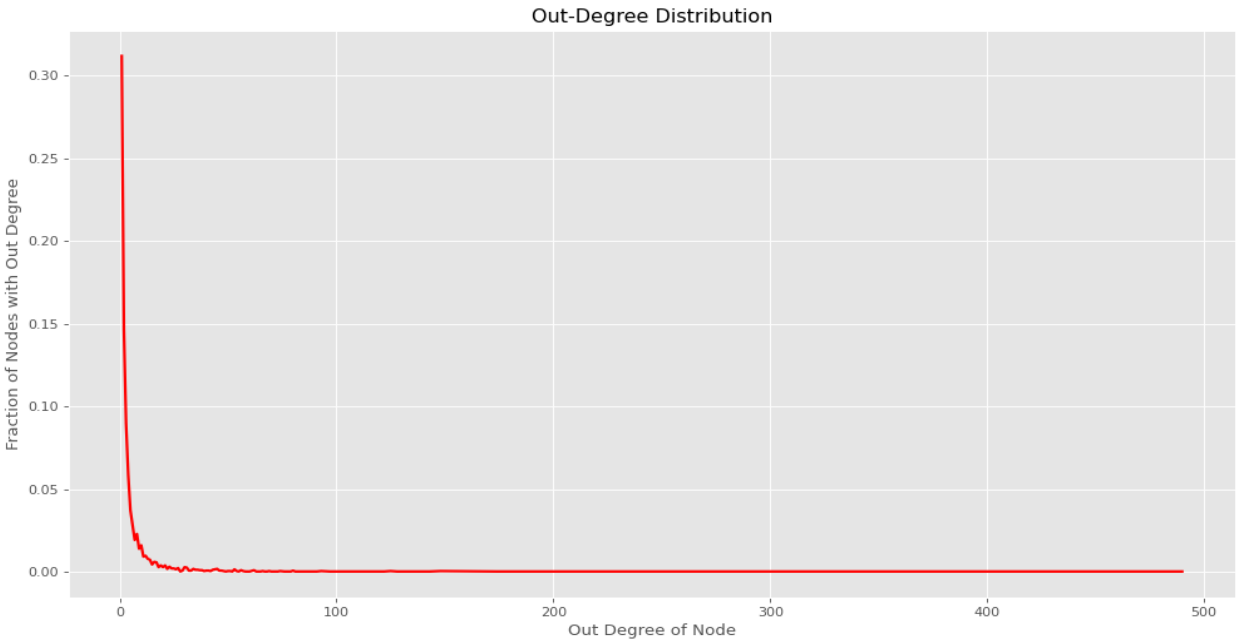
Outputs -

1. Number of Nodes: **3783**
2. Number of Edges: **24186**

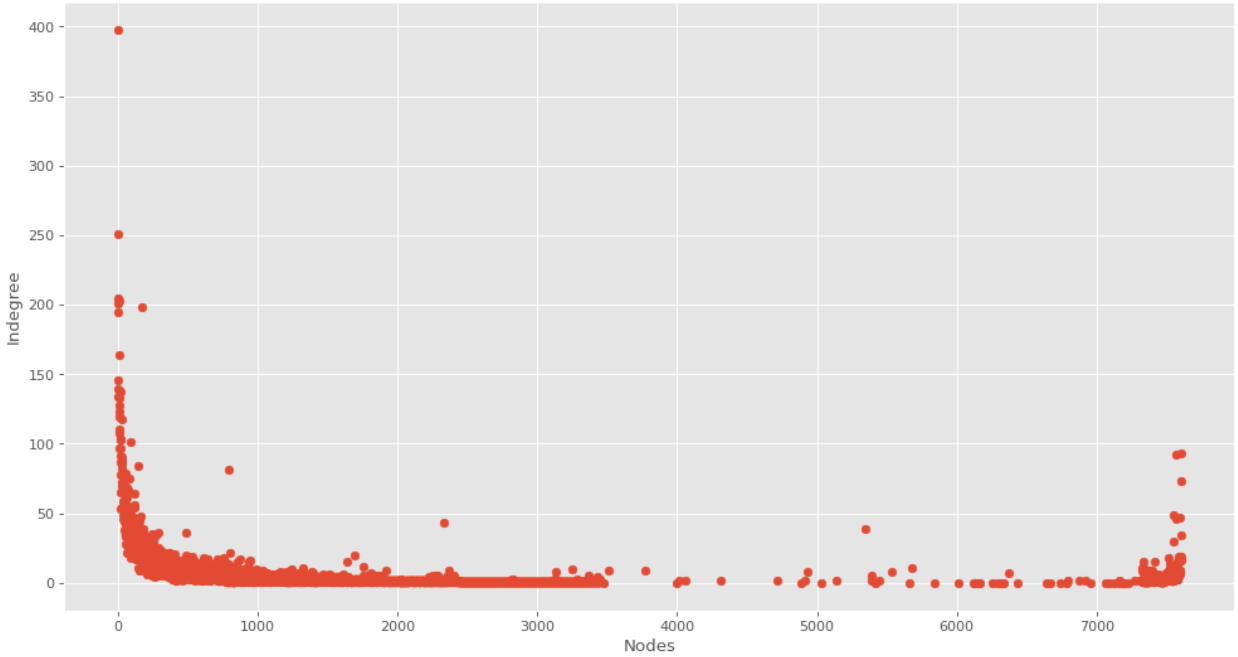
3. Average In Degree: **6.393338620142744**
4. Average Out Degree: **6.393338620142744**
5. Node with Maximum In-Degree (398) : **1**
6. Node with Maximum Out-Degree (490): **1**
7. Density of the Network: **0.0016904649973936393**

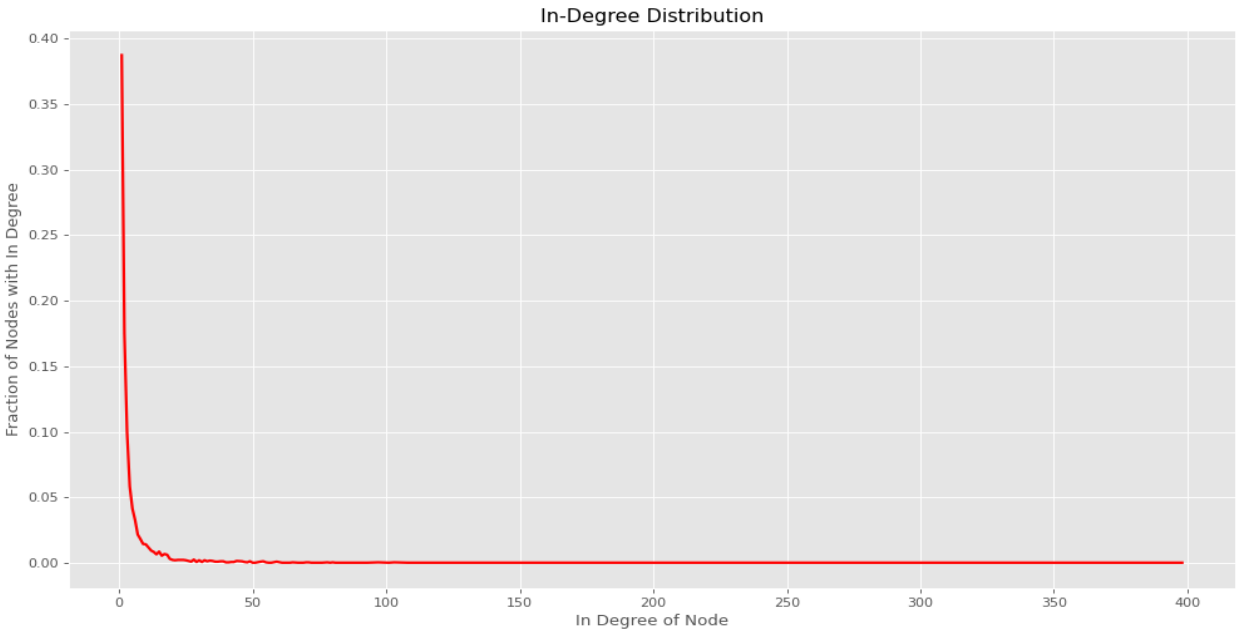
Out-Degree Distribution Curve





In-Degree Distribution Curve



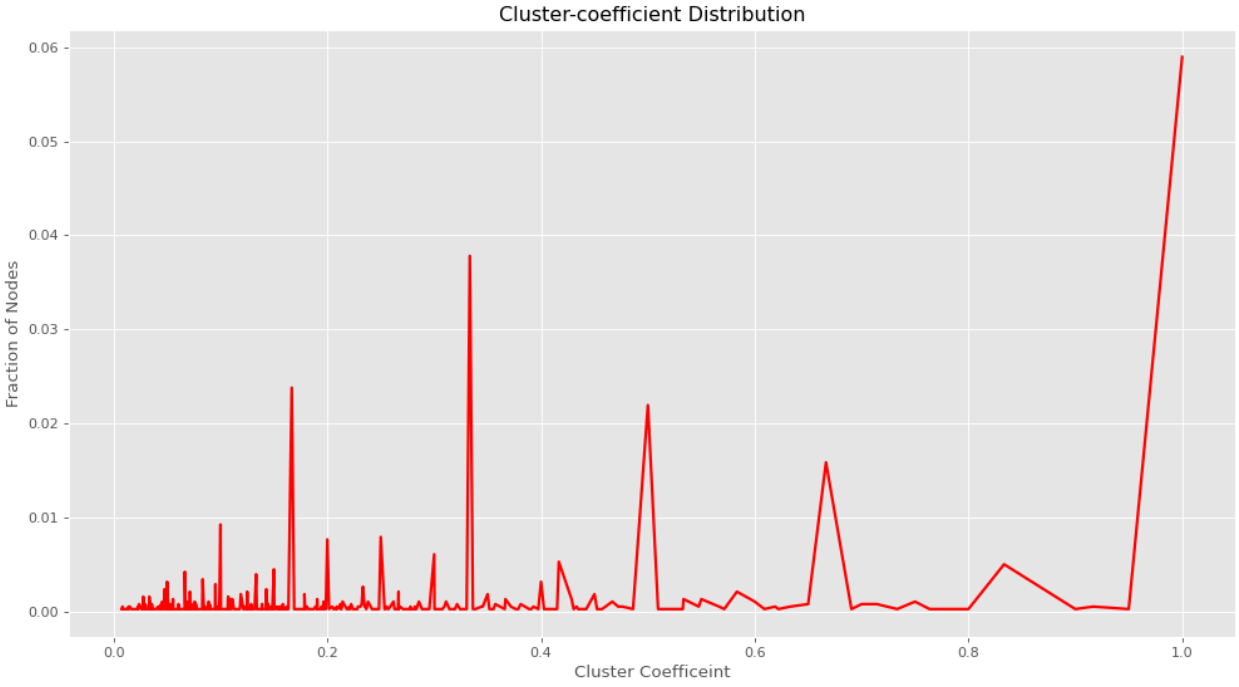


As we can see lower in-degree and out-degree hold by more number of nodes than higher degree.

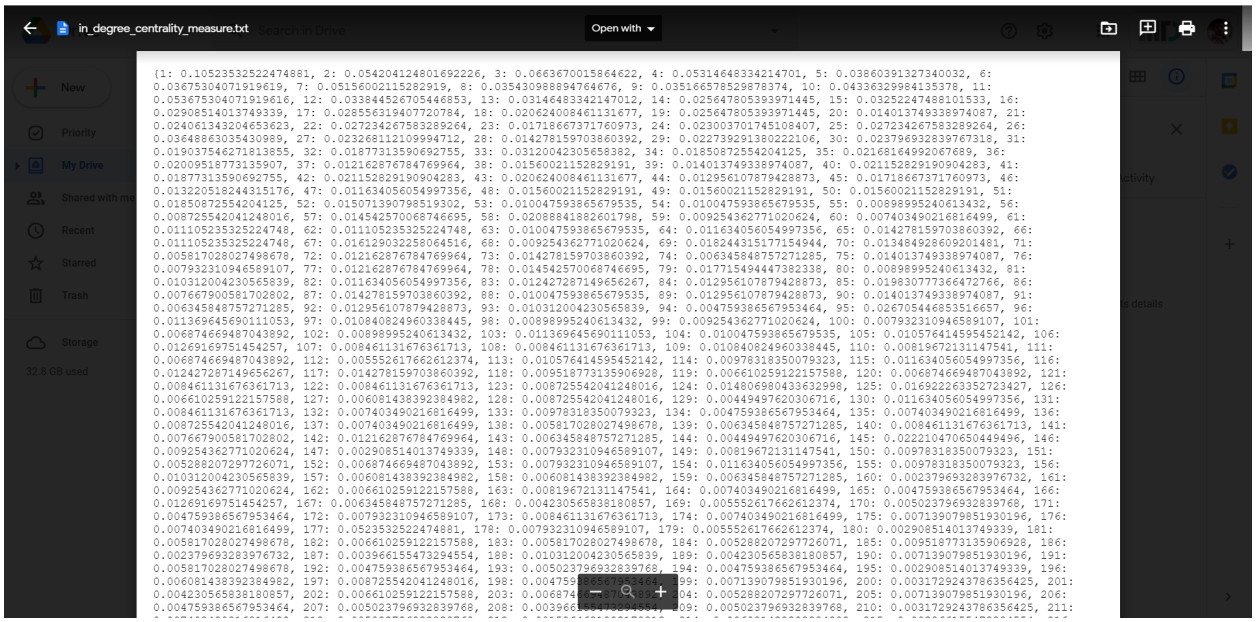
Cluster Coefficient of Some Node -

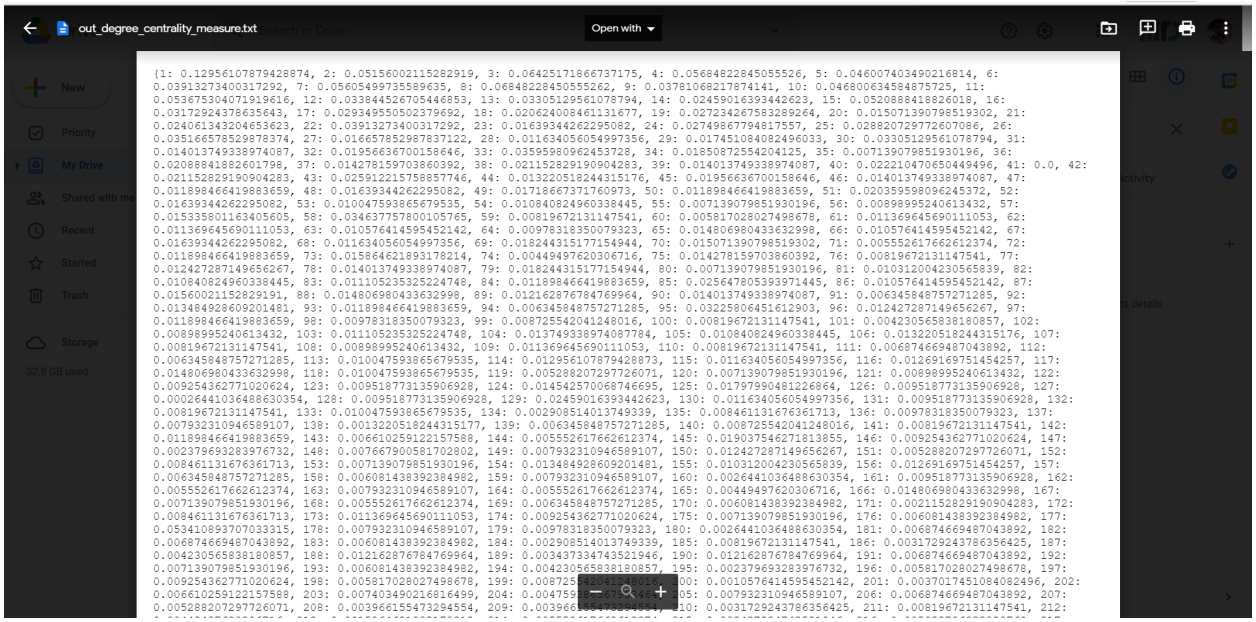
Nodes	Scores
1	0.008088300444286926
2	0.05282161645145863
3	0.034438247011952194
4	0.027860696517412936
5	0.067312234293812
6	0.05948284850380565
7	0.04171292624900872

Clustering-Coefficient Distribution of the Network -



Centrality Measure -





Formulas -

Average In-Degree = Sum of number of incoming edges for every node / total number of nodes

Average Out-Degree = Sum of number of outgoing edges from every node / total number of nodes

Network Density (For Directed Graph) =

$$e / n * (n-1)$$

Where e -number of edges

n - number of nodes

Local Cluster Coefficient (For Directed Graph) =

CC_i = (No of pairs of neighbors of i that are connected) /(No of pairs of neighbors of i)

Normalised Degree Centrality (For Directed Graph) =

Cin_di = (in-degree of Node i)/(max degree of a node can have i.e. n-1)

Cout_di = (out-degree of Node i)/(max degree of a node can have i.e. n-1)

Where n is number of nodes in network

(B)

PageRank Scores for Each Node -

The PageRank scores for some nodes are as -

Nodes	Scores
7188	4.9739967085536536e-05
1	0.016993099228405292
430	0.00028751507169557397
3134	0.00011917626603597539
3026	7.937879011849084e-05
3010	7.937879011849084e-05
804	0.00034797608896262324

Authority and Hub Scores for Each Node -

The authority and hub scores for some nodes are -

Nodes	Authority Scores	Hub Scores
7188	0.0	0.00015064556213535568
1	0.005881193763593962	0.0064088516365644315
430	0.00012057995517272793	0.00035917665852851346
3134	0.0003912912178082634	0.00035052262848803324
3026	0.00013944393576446274	0.00015064556213535568
3010	0.00013944393576446274	0.00015064556213535568
804	0.0005047716629344154	0.0005473505087191831

Results Comparison for PageRank and HITS -

	Page Rank	HITS
By	Larry Page and Sergey Brin	Jon Kleinberg in 1998
Year	1998	1998
Working	PageRank computes a ranking of nodes in the graph based on the structure of the incoming links.	HITS algorithm computes the authority score for a node based on the incoming links and computes the hub score based on outgoing links.
Calculating Scores	As PageRank calculates scores based on incoming links, higher the incoming links from the good nodes.	As hub score is based on outgoing links, the higher the scores means it points to many good nodes.
Usage Plan	PageRank is used for ranking all the nodes of the complete graph and then applying a search.	HITS is applied on a subgraph after a search is done on the complete graph.
Structure Idea	PageRank is based on the 'random surfer' idea and the web is seen as a Markov Chain.	HITS defines hubs and authorities recursively.
Known Issues	PR is distributed from a node to their outlinks, therefore there is an issue when the "surfer" reaches the dangling nodes in a graph.	Links to a large number of separate topics may receive a high hub rank which is not relevant to the given query.
Historical Day Uses	Rank web pages in google's search engine results.	HITS has been used to analyse co-citation and co-reference in the field of citation analysis and bibliometrics.
Modern Day Uses	VisualRank - Google's application of PageRank to image-search	In recent years, researchers have been using the HITS algorithm to solve problems such as spam detection.