INFORMATION RETREIVAL

ASSIGNEMENT - 3

GROUP NUMBER - 44

GROUP MEMBERS-

KUSHAGRA MITTAL - MT22108
VAIBHAV NAURIYAL - MT22129
SHUBHAM SAHANI - 2020132

QUESTION-1

LIBRARIES USED:

1. os: for importing dataset folder from directory.
2. networkx: python library for studying graph networks.
3. pandas: for implementing dataFrames.
4. tqdm: used for creating progress bars
5. pretty table: creates relational tables in python

Dataset chose: Wikipedia Vote Network

Results-

1. The average indegree and outdegree will be the same because the nodes with in-degree will get balanced by the nodes with one outdegree.
2. The network density tells us that if the density is 0, then the network has no edges, and if the density is 1, then the network is a complete graph.
3. Network density is calculated as the total count of edges / (n)*(n-1) ; n = the total count of nodes in the network (for directed graph).

- The clustering coefficient of each node. The clustering coefficient lies between 0 and 1.
- The clustering coefficient more skewed towards 1 gives higher certainty. Count of nodes with clustering coefficient 0: Count of nodes with clustering coefficient 1: Overall clustering coefficient of the network:
- The formula for calculating the clustering coefficient of the network for directed graph: N/n*(n-1) here n = total node neighbors, N = number of edges among n neighbors of that node in the network.

QUESTION-2

Page rank scores: Page rank ranks the web pages and returns them in the order of relevance. For the nodes with higher incoming edges, a high page rank is assigned.

Hubs and Authority scores: Used to measure the importance of web pages. Root nodes are the highly related web pages for the query provided.
Non-relevant pages pointing to the root nodes are called hubs. A good authority has many hubs pointing to it. A page that many hubs link joined to. A set of highly relevant web pages are called Roots. They are also known as potential.

Comparison between Algorithm 1 and Algorithm 2: The time taken for evaluate the scores in the HITS algorithm is greater than the time taken to evaluate the scores in the Pagerank algorithm. As the HITS creates mutual reinforcement between authority and hub scores and page rank just does it based on authority, the HITS results are less relevant than the page rank scores. This popularity is due to the features like efficiency, feasibility, less query time cost, etc., which are absent in the HITS algorithm.