CSE508: Information Retrieval
Assignment 3

Max Marks: 80

Instructions-

- The assignment is to be attempted in groups (max 2 members).
- Language allowed: Python
- For plagiarism, institute policy will be followed.
- You need to submit README.pdf and code files. The code should be well commented.
- You are allowed to use libraries such as NLTK for data preprocessing, NumPy, Pandas and matplotlib.
- Mention methodology, preprocessing steps, and assumptions you may have in README.pdf.
- You will be required to use Github for code management.
  - Each group will create a GitHub repository with the name IR2022_A3_GroupNo (Eg - IR2022_A3_1 for Group No-1).
  - Each group would add the assigned TA as a collaborator to the GitHub repository. TAs' GitHub handles would be shared shortly.
  - While uploading on Classroom, each group would need to upload a link of the GitHub repository. Only one member needs to submit.
- You will have 10 days to complete the assignment.

Question 1 - [45 Points] Link Analysis

Pick a real-world network dataset (with number of nodes > 100) from here.

[2 points] Represent the network in terms of its 'adjacency matrix' as well as 'edge list'.

[28 points] Briefly describe the dataset chosen and report the following:

1. Number of Nodes
2. Number of Edges
3. Avg In-degree
4. Avg. Out-Degree
5. Node with Max In-degree
6. Node with Max out-degree
7. The density of the network

Further, perform the following tasks:

1. [5 points] Plot degree distribution of the network (in case of a directed graph, plot in-degree and out-degree separately).
2. [10 points] Calculate the local clustering coefficient of each node and plot the clustering-coefficient distribution of the network.

NOTE:

1. You are NOT allowed to use any library to perform the tasks for this question.
2. Mention the formula for calculating the metrics in your report.

Question 2 - [35 points] PageRank, Hubs and Authority

For the dataset chosen in the above question, calculate the following:

1. [15 points] PageRank score for each node
2. [15 points] Authority and Hub score for each node

[5 points] Compare the results obtained from both the algorithms in parts 1 and 2 based on the node scores.

HINT: Note that PageRank computes a ranking of nodes in the graph based on the structure of the incoming links. On the other hand, the HITS algorithm computes the authority score for a node based on the incoming links and computes the hub score based on outgoing links.

NOTE: You CAN use libraries like networkx to solve this question.

You are allowed to subsample the dataset in case it is not processable on your machine. Ensure that you use an approach like random walk to subsample the nodes so that you get a connected network.