CSE508 Information Retrieval
Winter 2023
Assignment-3

Max. Marks: 80

Instructions:

1. The assignment is to be attempted in groups of max 3 members.
2. Each group member must do at least one task. All members should know the working of all the tasks. This will be evaluated during your code demo and viva.
3. Institute plagiarism policy will be strictly followed.
4. Programming language allowed: Python.
5. Your code should be well documented.
6. You are allowed to use libraries such as NLTK for data preprocessing, NumPy, Pandas and matplotlib.
7. You are required to use version control via GitHub:
    a. Make a GitHub repository with the name: **CSE508_Winter2023_A3_<Group_No.>**.
    b. Add your assignment TA as a contributor. The TA assigned (along with their GitHub handle) to your assignment group for this assignment can be found here. 8. You must make a detailed report with the name **Report.pdf** covering your methodologies, assumptions, and results.
9. Submission:
    a. A zipped folder **CSE508_Winter2023_A3_<Group_No.>** consisting of all your code files, dumped files and **Report.pdf**
    b. A text file **CSE508_Winter2023_A3_<Group_No.>.txt** consisting of the link to your GitHub repository.
10. Only one member from a group needs to submit.

Question 1 - [45 Points] Link Analysis

Pick a real-world directed 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 (lcc vs frequency of lcc) 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 pagerank and other APIs from networkx library 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.