CSE508: Information Retrieval
Assignment 6

Instructions

- The assignment is to be attempted individually
- Language allowed: Python
- For Plagiarism institute policy will be followed
- You need to submit ReadMe, code files and analysis.pdf

Q1: Pick a real-world network dataset from https://snap.stanford.edu/data/index.html
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

Further, perform the following tasks

1. Plot degree distribution of the network
2. Calculate the clustering coefficient of each node
3. Find any 1 centrality measure for each node

NOTE: You are not allowed to use any library for this question.

Q2: For the dataset chosen in the above question, calculate the following:

1. PageRank score for each node
2. Authority and Hub score for each node

Compare the results obtained from both these parts.

NOTE: You CAN use libraries like networkx (https://networkx.github.io/) to solve this question.

For both the questions, you are allowed to subsample the dataset so that it is processable on your machine. Ensure that you use an approach like random walk to subsample the nodes so that you get a connected network.