CSE508 : Information Retrieval Assignment 6

Max Marks: 80

Instructions

- The assignment is to be attempted individually
- Language allowed: Python
- For Plagiarism, institute policy will be followed
- You need to submit README.pdf, Code files (it should include both .py files and .ipynb files), and Analysis.pdf
- You are allowed to use libraries such as NLTK for data preprocessing.
- Mention methodology, preprocessing steps and assumptions you may have in README.pdf.
- Mention your outputs, analysis done (if any) in Analysis.pdf
- Submit code, readme and analysis files in ZIP format with the following name: **A6_<roll_no>.zip**
- Save all your precomputed indexes and tables which may take time to compute.
- ***Note: Due to the Covid-19 outbreak and lockdown, it may so happen that assignment demos cannot be taken, hence you are advised to prepare a well documented IPYNB file and report/analysis with all the justifications that may be necessary.***

Q1: Pick a real-world network dataset (number of nodes > 100) from
https://snap.stanford.edu/data/index.html
Represent the network in terms of its ‘adjacency matrix’ as well as ‘edge list’.
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. Density of the network

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.