CSE508 : Information Retrieval
Assignment 5
Deadline : 21st April'18, 2359 hrs

Total: 100 marks

Instructions

- Assignment is to be attempted individually. Please keep the discussions on an abstract level
- Language allowed : Python
- For Plagiarism, institute policy will be followed
- You need to submit ReadMe, code files and analysis.pdf
- You folder should be renamed in the NameRollNo_HW5 format before zipping

Question 1

Download 20_newsgroup dataset from
https://drive.google.com/file/d/1VA4a-wveTVXEy0J_NNv8oZ_YG2smxvPL/view

You need to pick documents of comp.graphics, sci.med, talk.politics.misc, rec.sport.hockey, sci.space [5 classes] for text classification.

You need to use the below as feature vectors

1) Bag of Words Model
2) Word2Vec representation from Google News Pretrained Word2Vec model [you can refer to: http://mccormickml.com/2016/04/12/googles-pretrained-word2vec-model-in-python/]

For both of these features set, implement K-means clustering algorithm *[you cannot use any library for k-means] [don't use groundtruth information]* and report the error.
Draw your inferences

Question 2

Choose any real world network [eg from https://snap.stanford.edu/data/index.html]
Describe your network briefly in terms of nodes, edges etc
Make sure to choose a network of less than 1000 nodes or randomly subsample nodes from available data. You need to *[don't use any library for any of these tasks]*

- Plot degree distribution of the network
- Calculate clustering coefficient for each node
- Calculate betweenness and closeness centrality for each node

What can you infer about the network. State your observations.