Assignment on Document Summarization

Data summarization is the process of creating a concise representation of a document. In extractive graph-based summarization tasks, you need to create a graph representation of the document and select some of the sentences of the document for forming the summary.

Task: Given 5 topics and 25 documents each within a topic, you have to synthesize a fluent, well-organized 250-word summary of the documents per topic using the following graph-based approaches:

- Text-Rank
- Degree centrality Based

Successful performance on the task will benefit from a combination of IR and NLP capabilities. The weighted graph has to be thresholded for three different thresholds: threshold = 0.1, 0.2 and 0.3.

Evaluation: Perform Rouge-N (with N values 1 and 2) and Rouge-L evaluation on the summaries.

Caution: You should implement your own versions of the two algorithms and NOT use ready-to-use online implementations. If we find that this is the case, your assignment will be outright cancelled. However, you can use the standard packages for computing the ROUGE scores.

Format of results (Table I):

Topic	Degree Centrality: Rouge-1 (threshold= 0.1, 0.2, 0.3)	Degree Centrality: Rouge-2 (threshold= 0.1, 0.2, 0.3)	Degree Centrality: Rouge-L (threshold= 0.1, 0.2, 0.3)	TextRank: Rouge-1 (threshold= 0.1, 0.2, 0.3)	TextRank: Rouge-2 (threshold= 0.1, 0.2, 0.3)	TextRank: Rouge-L (threshold= 0.1, 0.2, 0.3)
Topic 1						
Topic 2						
Topic 3						
Topic 4						
Topic 5						

Deadline: **23rd March 2018**

Things to submit:

A. Codes for (i) degree centrality based summary (ii) TextRank based summary. The codes should have separate modules for building the tf-idf vectors, constructing the graph, thresholding, generating summary etc.

B. A text file corresponding to Table I showing all the results.