Programming Assignment 2- Information Retrieval (CS F469)
Deadline: Nov 3, 2019 12:00 PM, Max Marks: 40

This assignment covers the Term weighting scheme and clustering models. The assignment is comprised of three tasks. First task is mandatory since it will be required to complete the other tasks. Students are allowed to use **nltk**, **spacy**, **numpy**, **sklearn**, **scipy** libraries. No other libraries will be allowed. All the tasks in the assignment are basic IR tasks and do not require any advanced libraries. Use the same datasets (queries and documents) sampled in Assignment 1.

Task 1 **[4 Points]** Design a raw term frequency based model to retrieve the search results. Report the precision and recall of your model and compare them with the values acquired in Assignment 1 submission.

Task 2 **[4 Points]** Design a weighted TF-IDF based model and report the precision and recall.

$$W_{t,d} = \begin{cases} 1 + \log_{10} TF_{t,d}, & \text{if } TF_{t,d} \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

$$IDF_t = \begin{cases} \log_{10}(\frac{N}{df_t}), & \text{if } df_t \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

Compare the values with the results acquired in Task 1.

Task 3 **[6 Points]** Apply Clustering methods on the dataset (weighted TF-IDF) with the following combinations:

(a) K-means clustering with euclidean distance. Select K on your own.
(b) K-means clustering with cosine similarity. Select K on your own.
(c) Agglomerative hierarchical clustering with single linkage
(d) Agglomerative hierarchical clustering with complete linkage
(e) Agglomerative hierarchical clustering with average linkage
(f) Agglomerative hierarchical clustering with ward's method

for k-means clustering, display the clusters of documents in form of table or list.
for hierarchical clustering, display the clusters of documents in form of dendrogram. Dendrogram library is available under scipy. Write your program in modules and functions. Give required comments in the code.

Task 4 Empirical Analysis

(a) **[15 Points]** Find the best value of K for all clustering algorithms in Task 3 (six). Use intra-similarity and inter-similarity measures as parameters for deciding best K. Hint: try different values of K (start from minimum and increase the value in each trial).
(b) **[6 Points]** Plot values of K, cluster intra-similarity and inter-similarity combination values to justify your answer.

Task 5 **[5 Points]** Search for the same queries on dataset (now available in form of clusters) and commpute the precision and recall for each query. Compare the results with Task 2.

Do not submit assignments over email.