ASSIGNMENT 2

Due date: 4th Oct 2020, 11.59PM (Hard deadline)

Marks: 10

Consider the corpus of 100 Wikipedia articles that you used for the Assignment 1. You may also reuse the preprocessing performed on the corpus earlier for this assignment.

a. For your corpus, demonstrate the process of Term Weight modelling for each index term as per the standard TF-IDF weighting scheme.

b. Experiment with the different variants of TF (log normalization and double normalization K) and IDF (inverse frequency smooth and probabilistic inverse frequency), and note the changes in the term weights when the different combinations are used.

c. Represent the document corpus using the standard TF-IDF weights as per the formalism used by the Vector Space IR Model. Define a set of queries (at least three different queries with different lengths, e.g. 2 words, 3 words, 5 words etc). Demonstrate the process of generating the ranked list for the given query using an appropriate closeness measure.

d. Use the different term weights obtained from at least two TF-IDF variant schemes (computed for Question *b*), and report if any changes are observed in the ranked list, for the same queries used in Question *c*.

Note:

1. Submit a detailed report on your observations and analysis, supported by the necessary code snippets w.r.t your program and results.
2. Upload your report and code (well documented) on Moodle to the folder provided before the deadline of **11.59PM on 4th October 2020 (No extension will be given).**