ASSIGNMENT 2

Due date: Sunday, 2nd October 2022, 11.59PM (Hard deadline)

Consider the corpus of 100 articles from the Incredible India website (articles that you used for the Assignment 1). You may also reuse the preprocessing performed on the corpus earlier, for this assignment. Define a set of queries (at least three different queries with different lengths, e.g. 2 words, 3 words, 5 words etc), for your experimentation.

a. Demonstrate the process of generating term weights for the vocabulary as per the standard TF-IDF weighting scheme.

b. Represent your document corpus using the standard TF-IDF weights as per the formalism used by the Vector Space IR Model. Demonstrate the process of generating the ranked list for sample queries using any one distance measure and a similarity measure. Compare and contrast the ranking generated by each.

c. 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.

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

Note:

1. Submit a detailed report on your observations and analysis, supported by the necessary code snippets w.r.t your program and results. **Plagiarised assignments will not be graded.**
2. Upload your report and code (well documented) on Moodle to the folder provided before the deadline of **11.59PM on 2nd October 2022 (No extension will be given).**

EVALUATION RUBRICS:

Part a: Corpus used, code and demo	- 10 marks
Part b: VSM representation, ranking analysis	- 10 marks
Part c: Tf-idf computation, analysis (2 variants)	- 10 marks
Part d: Corpus used, code, observations/analysis	- 10 marks
Detailed report	-10 marks