CSCI 517 (Natural Language Processing)

Text Representation Learning A (TF-IDF) – Homework 1

Tasks/Questions with (***) Prefix are only for Graduate Students

Task 1: Building TF-IDF embedding matrix from scratch. In this task, you are only allowed to use external that are already included in the Jupyter Notebook Template. You will use the same corpus as Homework 1. Each text in the corpus will be a document. *To make your life easier, I have prepared the flow of the program which already includes basic text processing. Your job is to read the template code, understand it, and finish all the #TODO tasks to calculate the TF-IDF embedding matrix.*

Evaluation: Your TF-IDF embeddings will be evaluated using an information retrieval task. The code of this task is already prepared for you in the template file. There is a test set of several queries and their respective ground truth of the top-10 most relevant documents from the corpus. You are evaluated on how many you can collect or recall from the provided ground-truth documents for each query. If your TF-IDF embeddings are correct, you should achieve a reasonable average recall score.

(**) **Performance Optimization:** You are asked to make several changes in the current notebook template to further improve the recall score. For example, you can consider removing stop-words v.s. not removing stop-words, using lemmatization v.s. not using lemmatization, using stemming v.s. not using stemming, using a different tokenizer, etc.

(BONUS) Runtime Optimization: You are asked to optimize the template codes to reduce the speed of calculating TF-IDF. You might want to find all the possible bottlenecks in the workflow and optimize them one by one. The current template code is not optimized and can be very slow if we have an extremely large corpus (millions of rows). Please note down your optimization changes in the notebook in the submission.

(BONUS) The students who achieve the best average recall will get bonus points and a small prize.