SI 650/EECS 549 - Information Retrieval
Assignment 2 Part 2

Due Thursday, Oct 18th, 23:59 EDT.
Please submit attachments via Canvas.
General discussion encouraged, but everyone should come up with the solution independently.
If you received help from anyone, you should list the name(s) on the top of your submission.

The goal of this assignment is to understand how a real retrieval system works. Please download the ipython notebook in the attachment. You can load it locally or upload to google colab (`https://colab.research.google.com`), which is recommended.

1 Design Your Own Function

Now you know how to implement and evaluate a retrieval function. In the next step, you will be proposing new retrieval functions of your own. Remember that every reasonable retrieval function should be able to be decomposed into the form of Equation 1, the weight of every query term should contain some form of TF, IDF, Query TF, and document length normalization. You are encouraged to search the literature and implement the well performing functions reported by the researchers. Do remember, however, you can only access certain information from the index using metapy package (refer to the ipython notebook).

$$s(q,d) = g[f(w_1,q,d) + ... + f(w_N,q,d), q, d], \tag{1}$$

Task (40 points) Implement at least one retrieval function different from BM25, Dirichlet Prior, and Pivoted Normalization. You will be graded based on your best performing function. You'll get full credit if your retrieval function can beat the provided baseline in the dataset. By "beat", we mean that your implemented function and your choice of parameters should reach higher MAP than the baseline in both cranfield and CACM datasets. Report this information in your submission: the code to implement the retrieval function, the parameter you used that achieved the best performance, and the best performance. In addition, ***Explain*** **what you have explored and why you decide to try those**.
Note: Simply varying the value of parameters in Okapi/BM25, Dirichlet Prior or Pivoted Normalization does not count as a *new* retrieval function.
Note: Please submit your **code** instead of screenshots for the new retrieval function.

Table 1: Dataset Baseline

Collection	MAP
cranfield	0.283
cacm	0.271