SI 650: Homework 2 Part 2

Name: Chenyun Tao Uniqname: cyuntao Kaggle Username: chenyuntao

My own scoring function is

$$\sum_{w \in Q \cap D} \frac{(k+1) \cdot c(w,Q)}{k + c(w,Q)} \cdot \frac{1 + ln\left(\sqrt{c(w,D)}\right)}{1 - b + b\frac{|D|}{avgdl}} ln\left(\frac{N + ln(df(w) + 1)}{df(w) + 1}\right)$$

and the parameters I used that achieved the best performance were k = 7, b = 0.4, and the best performance was 0.21586 on Kaggle.

My own function reaches a higher performance than the untuned BM25, and is also composed of 3 parts: QTF, TF and IDF. I kept the QTF part in BM25, as I think this normalized QTF is quite reasonable, and modified the TF and IDF parts based on parts from BM25 and Pivoted Length Normalization. For the TF part, I was inspired by the TF part in Pivoted Length Normalization, and decided to use a common function, square root function instead of ln. As I find `pyserini` has removed a lot of stop words, I think it makes sense to increase the effect of term frequency in documents. Then for the IDF part, I would like to normalize it in a different way, and I decided to use ln as it seems to be quite useful when normalizing TF.

For the hyperparameters, I just tuned them based on trials and performance on Kaggle. I chose k = 7 from {5, 7, 9}, and b = 0.4 from {0.3, 0.4, 0.5}.