IR Programming homework 1 report

B04902025 資工四 施博瀚

VSM: (b=0.75)

$$TF(doc, term) = \frac{(k+1) * term\ frequency}{term\ frequency + k * \frac{(1 - b + b * doc\ length)}{average\ doc\ length}}$$

$$IDF(term) = \ln \frac{\#\ of\ docs - doc\ frequency + 0.5}{doc\ frequency + 0.5}$$

score = TF*IDF

Best similarity function : np.sum(doc_score, axis=1)

Relevance Feedback:

Using Rocchio Feedback, and set 0.8 to alpha, 0.1 to beta, 0.1 to gamma

Consider top 100 performance docs as relevant docs, while 100~200 as irrevelant docs.

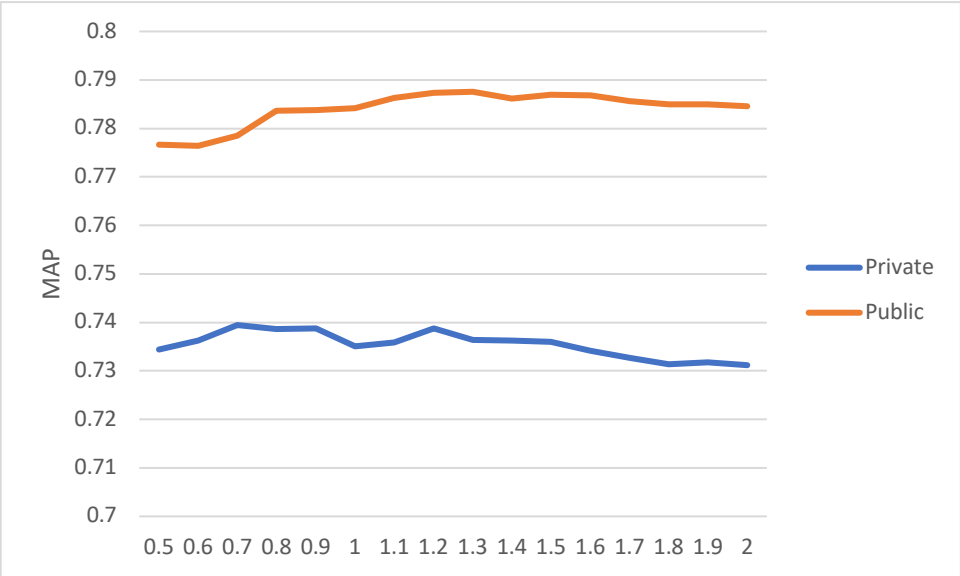
Results:

1. Using Cosine Similarity as similarity function:

   Private: 0.71689, Public: 0.76905

2. Without Feedback (np.sum())

   k = np.arange(0.5, 2.1, 0.1)



When k = 1.2, it performs best at both private and public score.

3. With Feedback (Update queries vectors 3 times, cosine similarity):
   Private: 0.72371, Public: 0.76544
   No improvement comparing to result 1.

Discussion:

It might behave worth when I update my queries vectors with relevance feedback too many times, and there's no huge improvement when I update my queries as well.