Web Retrieval and Mining 資工碩一 林佳彥 r08922170

Program HW1 report

1. Vector Space Model

1.1. TF-IDF

維度為 unigram+bigram 共 1193467 維。

$$\text{tf} = \frac{\text{tf\_raw}(\text{k}+1)}{tf\_raw + k(1 - b + b\ \frac{docLen}{AvgLen})}$$

$$idf_i = \log \frac{doc_{freq} - n(t_i) + 0.5}{n(t_i) + 0.5}$$

K = 1.5, b = 0.75

docVec 中 title 的 tf*9
queryVec 中 title, question, concepts 之 tf*5

1.2. Similarity

使用 cosine similarity。

2. Rocchio Feedback

$$Q_m = aQ_o + \left(b\frac{1}{|Dr|}\sum_{d \in Dr} d\right) + (\ c\frac{1}{|Dnr|}\sum_{d \in Dnr} d)$$

a = 1, b = 0.8, c = 0

原本沒 feedback map = 0.78363
有 feedback map = **0.8581**

3. Experiment

在 test data 上:
K = 1.5, b = 0.75, map = 0.676
K = 1.8, b = 0.75 map = 0.6722

將 docVec 之 title tf*9 後
K = 1.5, b= 0.75, map = 0.755

4. Discussion

原本只用 unigram map 很低，後改 bigram 有顯著提升，中間用 numpy array 存不下 Vecs, 考慮到其為 sparse matrix 而改用 csr_matrix 來當資料結構。
未來可以實驗 word2vec 之 deep learning 技巧來實作，將 vector 維度降低。