Web Retrieval and Mining spring 2019

Programming HW 1

b04902004 王佑安

1. VSM

我用的ranking function是Okapi BM25，對每個document的ranking function如下

$$\sum_{t \in Q} (\ln \frac{df+1}{N+1} + 1) \cdot \frac{(k_1+1)tf}{(k_1(1-b+b\frac{dl}{avdl}))+tf} \cdot \frac{(k_3+1)qtf}{k_a+qtf}$$

其中
df是term總共出現的次數
Tf是term在該document中出現的次數
qtf是term在query中出現的次數

而k1、b、k3為可調的參數，我最後使用的參數為k1=2.0,b=0.75,k3=500

而Query處理的部分，我將question跟concepts全部切成bigram，並刪去不在inverted-file字典裡的字。

2. Rocchio Relevance Feedback

Rocchio algorithm的公式如下

$$\overrightarrow{q_m} = \alpha \overrightarrow{q} + \frac{\beta}{|D_r|} \sum_{\overrightarrow{d_j} \in D_r} \overrightarrow{d_j} - \frac{\gamma}{|D_n|} \sum_{\overrightarrow{d_j} \in D_n} \overrightarrow{d_j}$$

其中Dr、Dn分別是相關的Documents跟不相關的Documents。
而alpha、beta、gamma是可調參數。我最後使用的參數為alpha=beta=1、gamma=0。

相關跟不相關的documents的選擇方式，我是先用前面VSM的BM25將所有documents作排序，選擇前k個跟後k個分別作為Dr跟Dn。而在進行feedback時，只挑選所query中出現頻率最高的n個term進行更新。n跟k都是可調參數，我最後的選擇是n=k=10。

3. Experiments

Parameter Tuning:

我先分別對Okapi BM25的k1、k3的數值做了實驗

k1/k3	100	250	500
1.2	0.80957	0.81031	0.81060
1.5	0.81752	0.81764	0.81798
2.0	0.81754	0.81810	**0.81832**
*b=0.75, no feedback, score=train MAP			

發現k=2.0,k3=500時train MAP最高

接著在固定k1=2.0, k3=500下實驗b的數值

b	0.25	0.5	0.75
Train MAP	0.80028	0.81199	**0.81832**

發現還是b=0.75的train MAP最高

With Feedback & Without Feedback:

我在k1=2.0,b=0.75,k3=500，feedback中的n=10,k=10的設置下實驗有無feedback差異

	With Feedback	Without Feedback
Train MAP	0.81832	**0.82105**

發現使用feedback會使train MAP稍微上升

Feedback Parameters Tuning:

Feedback有n跟k兩個參數可以調，我在k1=2.0,b=0.75,k3=500的設置下實驗feedback中n和k的數值

n/k	5	10	50
5	0.82029	0.82055	0.81733
10	**0.82106**	**0.82105**	0.81747
50	0.81743	0.81980	0.81059

可以發現feedback的參數調整影響很大，參數沒調好甚至使MAP下降。而最後在n=10,k=5 or 10的MAP幾乎是同樣最高的。

Selection of Private Submissions

在kaggle上傳submmsion時，我發現train MAP很高的model在public scoreboard上有時候反而不高，因此我推測這train、public、private這三個dataset的distribution可能差異很大。由於private scoreboard只能選2筆submission，我在挑選這個submssion時並沒有選擇train或public分數最高的，而是train跟public"平均"起來最高的submssion，降低我的model overfit到某個dataset上的可能性。

Score	Train	Public	Private
Submmsion1	0.78277	**0.81125**	0.72660
Submmsion2	0.82105	0.79040	**0.75280**

最後果然我本來在public上最高的0.81125在private上只有0.72660，而最後選擇的在public上只有0.79040，到了private卻有0.75280，在private scoreboard上最後取得了第7名的成績。

4. Discussion

以前有用過其它套件的tfidf跟BM25，卻不知道他們的實作細節。這次作業自己實作一遍後才知道，原來VSM有那麼多種，以及哪些參數的意義，在這次作業的實驗中嘗試了各種參數的組合發現原來參數對結果影響那麼大。

另外由於我之前有過其他kaggle競賽的經驗，在這次的競賽中特別注意到了data distribution的問題。有些model只在train/public其中一個dataset上分數特別高或特別低，表示他有可能overfit了，在挑選最後的submmsion時要盡量避免。在private公布時原本前幾名的幾乎都掉到後面去，更是證實了這一點。