R05922096 資工碩一 李哲安

Web Retrieval
Programming HW1 Report

- Describe your VSM (e.g., parameters....)
  - VSM + Rocchio Relevance Feedback (pseudo version)
  - Ranking: vector inner product
  - Parameters:
    Feedback or not、TF、IDF、Segmentation、Query(concept, title)

- Describe your Rocchio Relevance Feedback (e.g., how do you define relevant documents, parameters...)
  - Relevant Document:
    將第一次 Retrieve 回來的 Article 中 Rank1 的 Article 視為 Relevance，並將此 Article 的 Title 加入 Query 中做第二次 search。
  - Parameters:
    no。

- Results of Experiments
  - Feedback vs. no Feedback
    以上述方法做 feedback 後，在 query_train 中的 MAP 提高了 0.01，但是在 query_test 中 kaggle 上的分數卻降低了 0.01。
    Relevance Feedback 是利用 relevant set 中的資訊來 improve search 的準確度，但在這次的資料中卻沒有得到好的效果，推測可能因為這次的資料在做 search 時原本的準確率就已經相當高(0.7)，所以在 feedback 後的效果有限，反而可能將 noise include 進來導致準確率下降。

- MAP value under different parameters of VSM
  以下表格主要比較在 query_train 的 performance。

	Query (weight)	Segmentation	TF	IDF	Feedback	Feedback #	Ranking	MAP query_train	MAP Kaggle
Query Weight	Title(1) + Concepts(1)	bigram	3 + log(f)	log(30 + N/ni)	FALSE		inner porduct	0.77773381	0.73833
Query Weight	Title(2) + Concepts(1)	bigram	3 + log(f)	log(30 + N/ni)	FALSE		inner porduct	0.79302626	0.72843
Query Weight	Title(3) + Concepts(1)	bigram	3 + log(f)	log(30 + N/ni)	FALSE		inner porduct	0.776669441	0.71039
Query	Concepts(1)	bigram	3 + log(f)	log(30 + N/ni)	FALSE		inner porduct	0.752499855	
Feedback	Title(1) + Concepts(1)	bigram	3 + log(f)	log(30 + N/ni)	TRUE	1	inner porduct	0.791893629	
Feedback	Title(1) + Concepts(1)	bigram	3 + log(f)	log(30 + N/ni)	TRUE	2	inner porduct	0.788400772	
Feedback	Title(1) + Concepts(1)	bigram	3 + log(f)	log(30 + N/ni)	TRUE	5	inner porduct	0.782045165	
Segmentation	Title(1) + Concepts(1)	unigram	3 + log(f)	log(30 + N/ni)	FALSE		inner porduct	0.519616992	
Segmentation	Title(1) + Concepts(1)	unigram + bigram	3 + log(f)	log(30 + N/ni)	FALSE		inner porduct	0.721165831	
TF-IDF	Title(1) + Concepts(1)	bigram	f	1	FALSE		inner porduct	0.543732034918	
TF-IDF	Title(1) + Concepts(1)	bigram	1 + log(f)	log(30 + N/ni)	FALSE		inner porduct	0.774175565431	
TF-IDF	Title(1) + Concepts(1)	bigram	f	log(30 + N/ni)	FALSE		inner porduct	0.605741114977	
TF-IDF	Title(1) + Concepts(1)	bigram	1 + log(f)	1	FALSE		inner porduct	0.775389605995	
TF-IDF	Title(1) + Concepts(1)	bigram	1 + log(f)	log(N/ni)	FALSE		inner porduct	0.776884511279	
TF-IDF	Title(0.5) + Concepts(1)	bigram	log(1 + log(1+f)) + bm25(k=1.2)	log(N/ni)	FALSE		inner porduct	0.768346479	0.75644

  - 不同的斷詞對於準確率有很大的影響，由於這次的文章大多數為中文，以 Bigram 斷詞效果看起來較好。
  - TF-IDF 對於準確率的影響也很大，對 TF 做 log normalize + bm25 之後準確率有提升。IDF 則影響不明顯。
  - Query 的選擇也很重要，除了 Query 中的 concepts 外再加入 title 資訊並且給予不同權重後，有助於整體 MAP 的提升。
  - Feedback 同上題。

- Other experiments you tried
  同上題表格。

- Discussion: what you learn in the homework
  一個好的 IR system 必須要根據 Corpus 的不同來使用不同的的 parameter (segmentation、TF-IDF、feedback、ranking 等等)。例如中文的斷詞和英文的斷詞差異，當文章長短不一的時候 TF-IDF 的選擇，還有何時該利用 search 的結果來做 feedback，不同的 ranking 方法之間的差異，都非常的重要。