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。

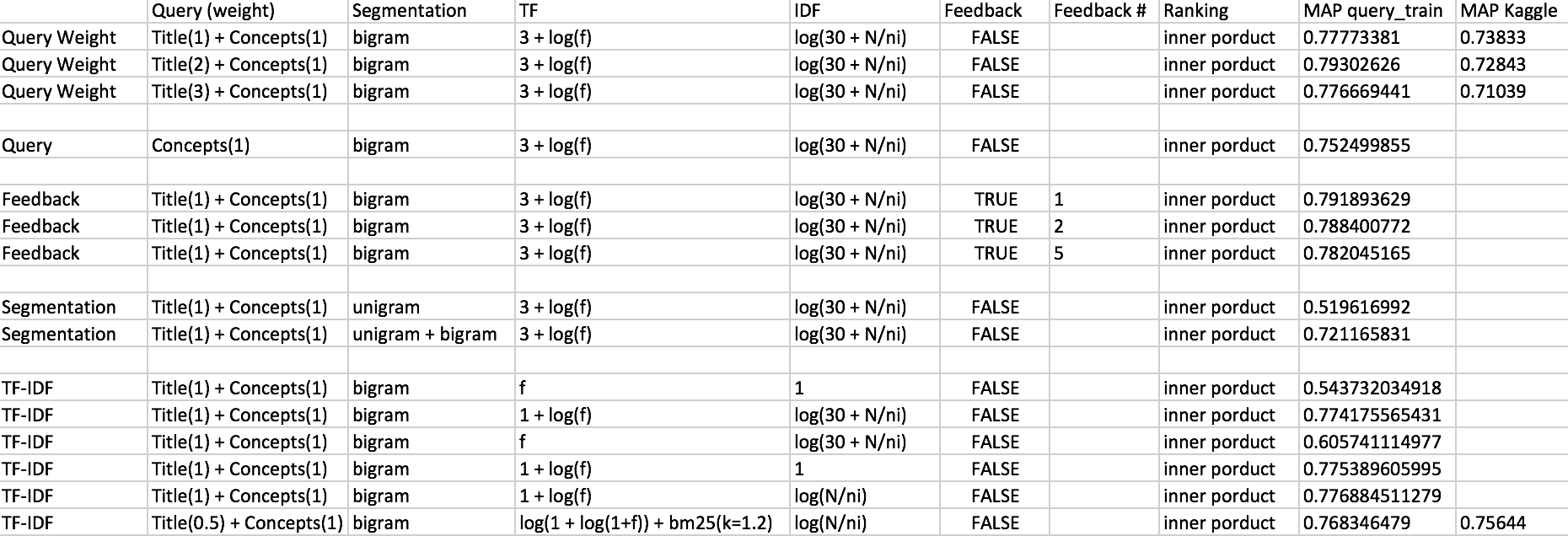
* Results of Experiments
  + Feedback vs. no Feedback

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。



* + - 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。