IR HW1 Report

B05902010 張頌平

1. Describe your VSM (e.g., parameters….)

- 利用query中的concept所包含詞彙的bigram、和利用jieba斷詞後的narrative的bigram(有把標點符號去掉)當作document/query vector的vocabulary，字典(vector dimension)大小約為400~800 (根據query-train或query-test有所不同)
- 使用model/invert-file 去搜尋所有document且有存在於query的bigram，並且記錄Term Frequency和Document Frequency
- 計算TFIDF時，使用Okapi BM25 normalize我的Term Frequency (k1=1.8, b=0.75)

2. Describe your Rocchio Relevance Feedback (e.g., how do you define relevant documents, parameters…)

- 我有實作兩種方式來界定何為relevant documents:
  - Top k ranking: 取相似度前k名的documents為relevant documents(k=10)
  - High Similarity: 取相似度高於某threshold s時的documnet為relevant documents(s=0.6)
  - parameters: α=1, β=0.35, γ=0.01

3. Results of Experiments

以下實驗結果以query-train, public query-test, private query-test分數進行回報

- **3-1. MAP value under different parameters of VSM (BM25 parameters)**
  - k1=1.8, b=0.75: train: 0.76155, public test: 0.77572, private test: 0.68977
  - **k1=1.7**, b=0.75: train: 0.76160, public test: 0.77529, private test: 0.68961
  - **k1=1.9**, b=0.75: train: 0.76119, public test: 0.77843, private test: 0.68991
  - k1=1.8, **b=0.6**: train: 0.76169, public test: 0.77680, private test: 0.68968
  - k1=1.8, **b=0.4**: train: 0.75954, public test: 0.77657, private test: 0.68983

以下實驗結果以 k1=1.8, b=0.75 為 BM25 parameters

- ***3-2. Feedback vs. no Feedback***
  - no Feedback: train: 0.76155, public test: 0.77572, private test: 0.68977
  - Feedback(α=1, β=0.35, γ=0.01, k=10):
    train: 0.75528, public test: 0.77828, private test: 0.69211

- Feedback(α=1, β=0.35, **γ=0.05**, k=10):

  train: 0.75516, public test: 0.77749, private test: 0.69051
- Feedback(α=1, **β=0.5**, γ=0.05, k=10):

  train: 0.74968, public test: 0.77775, private test: 0.69361
- Feedback(α=1, β=0.35, γ=0.01, **s=0.6**):

  train: 0.75906, public test: 0.78104, private test: 0.69079
- Feedback(α=1, β=0.35, γ=0.01, **s=0.7**):

  train: 0.75357, public test: 0.77818, private test: 0.68594

以下實驗結果以 k1=1.8, b=0.75 為 BM25 parameters，以 α=1, β=0.35, γ=0.01, k=10 為 feedback parameters

- *3-3 Other experiments you tried*
  - Remove stop-words with feedback:

    train: 0.75508, public test: 0.77408, private test: 0.68882
  - Remove stop-words without feedback:

    train: 0.76093, public test: 0.77312, private test: 0.68846
  - Only use query concept(no narrative) with feedback:

    train: 0.76055, public test: 0.77341, private test: 0.71316
  - Only use query concept(no narrative) without feedback:

    train: 0.75376, public test: 0.77281, private test: 0.67842
  - Average Ranking with feedback
    (把僅用concept和用concept&narrative的分數進行平均後排名):
    train: 0.76339, public test: 0.78654, private test: 0.69817
  - Average Ranking without feedback:

    train: 0.75762, public test: 0.78498, private test: 0.69362

4. Discussion: what you learn in the homework.

- Rocchio Relevance Feedback在這次的作業中，效果不大，雖然在public query-test中有進步一點點，但在query-train裡面是退步的，原因可能是因為我設定的document vector的dimension不夠大(僅使用query的bigram)，導致在feedback的時候，沒有辦法透過相關的document來更接近query想搜尋的文本，或是narrative當中的存在過多無相關的詞彙，導致feedback時有雜訊混入。

- 移除停用詞也沒有使得結果更進步，代表停用詞或多或少仍然保有一些文本的含義，而並不是真的完全沒有效益。
- 在僅使用query的concept上，Relevance Feedback的結果會比沒有使用時好上一些(在train和private test上皆有進步)
- 將使用的兩種方法的相似度相加並進行平均，因為兩者在預測比較好分數的文本是不同的，可以理解為他們捕捉到的資訊也較不同，因此將其相似度取平均能夠更好的得到雙方的優點，而結果也確實有更為進步。