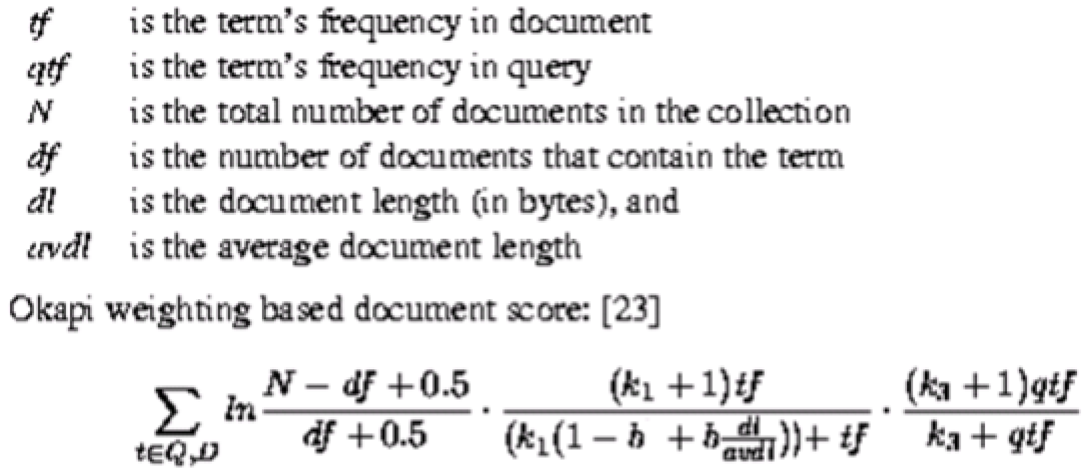
IR Pogramming Report #1

Name: 劉家維 **ID: b05902052**

VSM

The score between a document and a query is calculated as below:



tf is the term's frequency in document
qtf is the term's frequency in query
N is the total number of documents in the collection
df is the number of documents that contain the term
dl is the document length (in bytes), and
avdl is the average document length

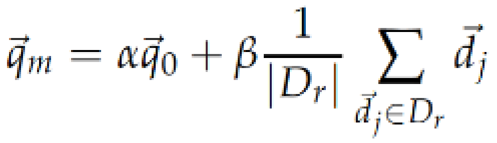
Okapi weighting based document score: [23]

$$\sum_{t\in Q,D} ln\frac{N-df+0.5}{df+0.5}\cdot\frac{(k_1+1)tf}{(k_1(1-b\ +b\frac{dl}{avdl}))+tf}\cdot\frac{(k_3+1)qtf}{k_3+qtf}$$

where $k_1 = 1.6, k_a = 5.0, b = 0.75$

Rocchio Relevance Feedback

The modified query is calculated as below:



$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j\in D_r}\vec{d}_j$$

where $\alpha = 1.0, \beta = 0.75, D_r = \{\text{the top 10 documents retrieved by original query}\}$

Results of Experiment

- MAP value under different parameters of VSM (1-gram only, the MAP of `query_train.xml`)

ka \ k_1	**1.0**	**1.2**	**1.4**	**1.6**	**1.8**	**2.0**
0.00	0.7467	0.7489	0.7480	0.7451	0.7464	0.7444
5.00	0.7668	0.7735	0.7755	**0.7781**	0.7741	0.7652
10.00	0.7682	0.7684	0.7728	0.7672	0.7635	0.7616
50.00	0.7612	0.7499	0.7525	0.7491	0.7435	0.7374
100.00	0.7613	0.7467	0.7504	0.7437	0.7390	0.7357
300.00	0.7480	0.7480	0.7482	0.7415	0.7380	0.7352
500.00	0.7481	0.7467	0.7439	0.7412	0.7367	0.7355
1000.00	0.7483	0.7468	0.7419	0.7399	0.7352	0.7355

- With Feedback vs. without Feedback

  (doc_len=byte length, 1-gram+2-gram)

Without Feedback	**With Feedback**
0.81810 / 0.74097	0.81925 / 0.66859

- Other experiments

  (doc_len=byte length, without feedback)

1-gram only	**2-gram only**	**1-gram + 2-gram**
0.74721 / 0.64046	0.80068 / 0.74258	0.81810 / 0.74097

  (1-gram+2-gram, without feedback)

doc_len=token numbers	**doc_len=byte length**
0.81709 / 0.74174	0.81810 / 0.74097

Discussion

- It seems that there is a score gap between public and private test data, since the performances of them are 0.818 v.s. \$0.741\$, and thus effect the result of the experiment testing Roccchio Relevance Feedback.
- The parameter $k_1 = 1.6, k_a = 5.0$ is found by the method below, since it takes long time to test the parameters under the settings of 2-gram, I just find the parameters under 1-gram settings:
  - Use different parameter values to generate the weights of documents.
  - Use the query in `query_train.xml` to retrieve documents.
  - Calculate MAP of the result with the grown truth (`ans_train.csv`).
  - Choose the paraemeters which make the MAP highest.

- The difference between performances using token numbers or byte length as doc length is not obvious.
- The difference between performances using 1-gram or 2-gram is obvious, and using 2-gram is much better, though it will take more time to calculate.
- The difference between performances using feedback or not is obvious, and using the feedback worsen the performance. An appropriate guess is that if the retrieval result is not so good, then Rocchio Relevance Feedback can make the modified query worse.