WM Programming Hw1

b07902124 資工三 鄭世朋

1. VSM model

Document

- Let $c(t, d)$ be the frequency count of term t in doc d, n denote total number of documents and $k(t)$ equals the number of document has term t
  (t can be unigram or bigram)

$$TF(t, d) = c(t, d), \;\; IDF(t) = \log \frac{n+1}{k(t)+1} + 1$$

Query

If t does not appear in a query q

$$TF(t, q) = 0$$

If t is in q and t is unigram

$$TF(t, q) = 5 \times c\left(t,\ title\right) + 3 \times c\left(t,\ question\right) + 3 \times c\left(t,\ concepts\right)$$

If t is in q and t is bigram

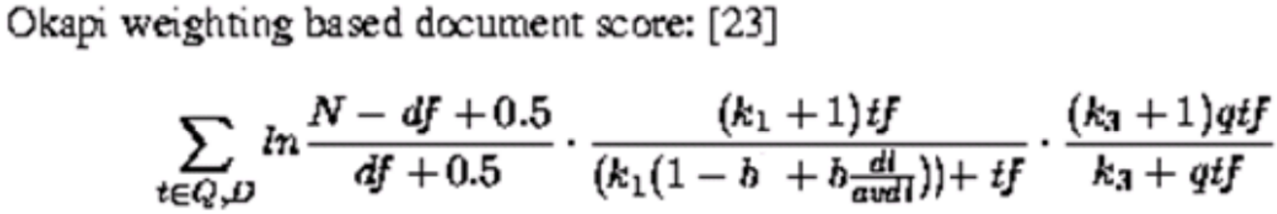
$$TF(t, q) = 9 \times c\left(t,\ title\right) + 6 \times c\left(t,\ question\right) + 6 \times c\left(t,\ concepts\right)$$

$$IDF(t) = \log \frac{n+1}{k(t)+1} + 1$$

Where n is total number of queries and $k(t)$ equals the number of queries has term t.

Okapi:

(In Lecture slides vsmodel2020 page 38)



Okapi weighting based document score: [23]

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

$k_1 = 1.5, b = 0.6, k_3 = 2$

2. Rocchio Relevance Feedback

- First 12 documents are thought to be relevent and last 50 documents are thought to be not relevent.
- $\alpha = 1$, $\beta = 0.8$, $\gamma = 0.2$
- 3 iterations

3. Experiment

- Default setting: $k_1 = 1.5,\ b = 0.6, k_3 = 2, \alpha = 1, \beta = 0.8, \gamma = 0.2$

$$Score = 0.78914$$

Different k_1 and b

$k_1 = 1.5, b = 0.6$	$k_1 = 2.5, b = 0.5$	$k_1 = 2, b = 0.5$
0.78914	0.78339	0.78398

- As we can see from table above, as long as k_1 and b is in reasonable scope, there's only little difference with different k_1 and b

Normalization

Normalized	Unnormalized
0.78914	0.77893

- According to the result, we can see length of document have impact on the score, so we can penalize more on long documents when normalizing

Feedback

With Feedback	Without Feedback
0.78914	0.75950

- This table shows if we don’t apply feedback, score may drop a lot.

4. What I’ve learned

- 在這次作業中，因為有五分鐘的限制，所以很多矩陣乘法的地方不能直接相乘，如果這樣做跑一天都跑不完，要先將sparse matrix轉成壓縮過後的格式才能在時間內完成，可見儲存資料的格式影響之大。