COMP6714 Assignment

Question1:

Assume we could divide each document into set of paragraph, and each paragraph can be divided into set of sentences. Hence we could record the position of every term, its sentence id and paragraph id.

In other words, the /k extract the term position, /s extract the sentenceID and /p extract the paragraphID and process as usual. This process costs a lot of space.

Another way is including two tokens to record the end of the sentence position and the end of paragraph position.For example, the '.?'are end of sentence symbol and token them and'\n'is the symbol of end of paragraph.

For example: Positional inverted indexes of "end of sentence"(doc1): [5, 10, 30, ...],Positional inverted indexes of "end of paragraph"(doc1): [30, 60, ...] Then we can use the two Inverted index to find the result.

Question 2:

(1) First stage we performed a sequenctial search and the cost avoiding the big-o Notation is $\frac{n}{2}$

Second Stage we performed sequencial seach on one segment and the cost avoiding the big-O notation is $\frac{L}{2n}$

hence we have can calculate the total cost is

$$\frac{n}{2} + \frac{L}{2n}$$

where n is the number of pointers and L considered to be a constant variable in this function

$$f'(n) = \frac{1}{2} - \frac{L}{2n^2}$$

and we want to minimize it so set it equal to 0 then we get

$$n = \sqrt{L}$$

Hence choosing $\sqrt{L}$ skip pointers has the best performance

(2) We both perform binary search both in step1 and step2. The cost avoiding the big-O notation for step 1 is $log(n)$ and the second step is $log(\frac{L}{N})$ hence the tototal cost in a function of n is

$$f(n) = log(n) + log(\frac{L}{n}) = log(L)$$

Which is dependent to the n. Consider the space, we choose n=log(L).

(3)We perform binary search in step and avoiding the big-O notation the cost is $log(n)$ and second step we perform a sequential search in step 2 the cost is $\frac{L}{2n}$

Hence the total cost is

$$f(n) = log(n) + \frac{L}{2n}$$

where L is considered to be a constant variable here and n is the number of pointer. Calculate its derivatie

$$f'(n) = \frac{1}{n} - \frac{2L}{4n^2} = \frac{2n - L}{2n^2}$$

and set it equal to 0 we get

$$n = \frac{L}{2}$$

question3

We subtract the given numbers into the formula we get

$$maxsocre(d_{max}, \{t\}) = \sum_{t \in Q} idf_t \cdot \frac{(3tf_{t,d})}{2 + tf_{t,d}} \cdot \frac{3tf_{t,Q}}{2 + tf_{t,Q}} = idf_t \cdot \frac{3max_{tf}}{2 + max_{tf}}$$

considering the limits, we could regard the $\frac{3max_{tf}}{2+max_{tf}}$ as a function $f(x) = \frac{3x}{2+x}$ and it easy to see that

$$\lim_{x \to \infty} \frac{3x}{2 + x} = \lim_{x \to \infty} \frac{3}{\frac{2}{x} + 1} = 3$$

Hence we can easily conclude the score function will become

$$maxscore(d) = 3 \cdot idf_t$$

And maxscore for the terms are 18(A),6(B),3(C)

(2)

We can consruct a table

term	maxscore	idf						Postings					
			1	2	3	4	5	6	7	8	9	10	11
A	1 8	6	1	8			3			10			
B	6	2	1				4	1	4				
C	3	1	1	2		1	2	3		1	1	3	7

Recall the score formula given in the question. First we consider the D_1 which has the score

$$Score(d_1) = 6 \times \frac{3 \times 1 \times 3 \times 1}{3 \times 3} + 2 \times \frac{3 \times 1 \times 3 \times 1}{3 \times 3} + 1 = 9$$

Then we consider the D_2 which has the score

$$Score(d_2) = 6 \times \frac{3 \times 8 \times 3}{10 \times 3} + \frac{1 \times 3 \times 2 \times 3}{4 \times 3} = 15.9$$

Now the temp top-2 result become(15.9, 9) the treshhold $\tau^{'} = 9$ here.

We skip doc_4 because the $maxscore_c$ is 3 which less than the treshold.

Then we consider the score of d_5

$$Score(d5) = 6 \times \frac{3 \times 3 \times 3}{5 \times 3} + 2 \times \frac{3 \times 4 \times 3}{6 \times 3} + \times \frac{3 \times 2 \times 3}{4 \times 3} = 16.3$$

And now the Top-2 become(16.3,15.9) and new $\tau^{'} = 15.9$.

Similarly, we could skip the doc_6 and doc_7 and now we consider the doc_8

$$Score(d8) = 6 \times \frac{3 \times 10 \times 3}{12 \times 3} + 1 = 16$$

which greater than the treshold now the top 2 is d_5 and d_8 and the treshold $\tau^{'} = 16$

We can Skip doc_9,doc_{10} and doc_{11} due to its maxscore is smaller than the treshold.

Finally, we could conclude the top-2 documents are D_5 and D_8. This algorithm scored 4 documents, and accessed 10 postings.

question4

(1) According to the given in the questions, we could construct the chart

	Relevant	Nonerelevant
Retrieved	6	14
Not retrieved	2	9986

Hence the Precision $= \dfrac{tp}{tp + fp} = \dfrac{6}{20} = 30\%$

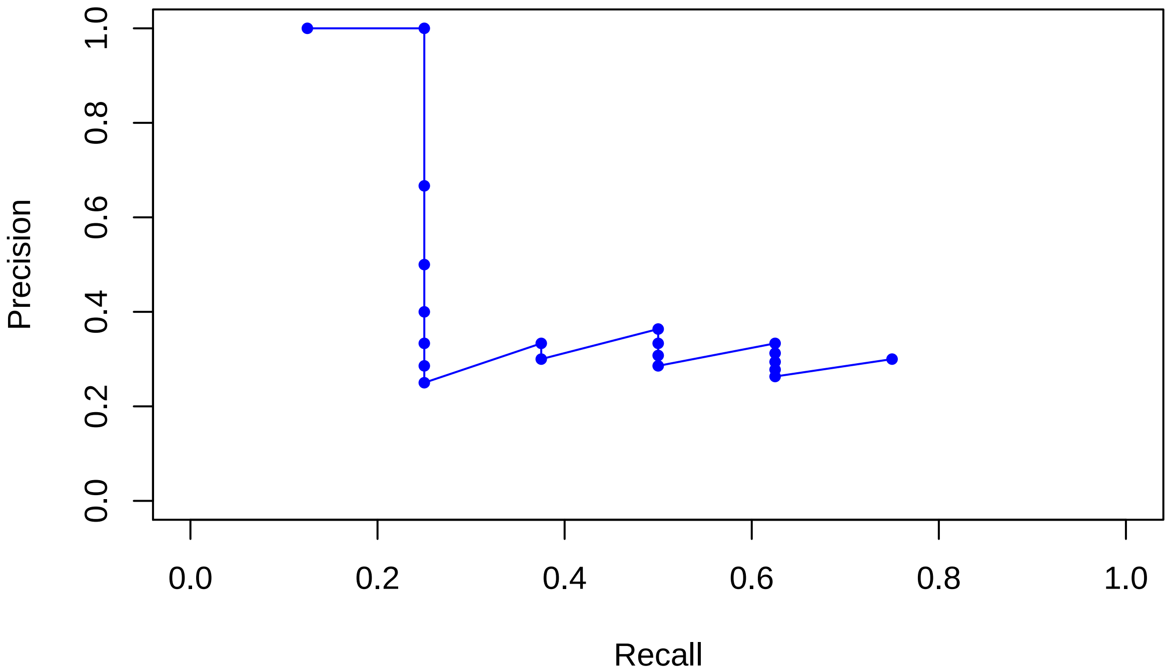
(2) The Recall $= \frac{tp}{tp+fn} = \dfrac{6}{8} = \dfrac{3}{4}$

Hence $F_1 = \dfrac{2RP}{P+R} = 0.4286$

(3)

First we calculate the precision and recall for each output and summarize in the table

k	Judgement	Precesion	Recall
1	R	1	1/8
2	R	1	1/4
3	N	2/3	1/4
4	N	1/2	1/4
5	N	2/5	1/4
6	N	1/3	1/4
7	N	2/7	1/4
8	N	1/4	1/4
9	R	1/3	3/8
10	N	3/10	3/8
11	R	4/11	1/2
12	N	4/12	1/2
13	N	4/13	1/2
14	N	4/14	1/2
15	R	1/3	5/8
16	N	5/16	5/8
17	N	5/17	5/8
18	N	5/18	5/8
19	N	5/19	5/8
20	R	6/20	3/4



Hence we can see from above. The un-interpolated precision are

$$1, \frac{2}{3}, \frac{1}{2}, \frac{2}{5}, \frac{1}{3}, \frac{2}{7}, \frac{1}{4}$$

(4) The interpolated precision for 33% recall is the biggest pecision could achieve when $k > 9$. Seeing the graph obtained above, the maximum is $\frac{4}{11}$

(5)

$$MAP = \frac{1 + 1 + \frac{1}{3} + \frac{4}{11} + \frac{1}{3} + \frac{3}{10}}{8} = 0.4163$$

(6)The larggest possible MAP will be got if the rest two relevant documents appear at k=21,22 respectively.

$$MAP_{max} = \frac{1 + 1 + \frac{1}{3} + \frac{4}{11} + \frac{1}{3} + \frac{3}{10} + \frac{1}{3} + \frac{4}{11}}{8} = 0.5034$$

(7) In contrast, we will et the smallest MAP if the 2 relevant document are in 9999th and 1000th

$$MAP_{min} = \frac{1 + 1 + \frac{1}{3} + \frac{4}{11} + \frac{1}{3} + \frac{3}{10} + \frac{7}{9999} + \frac{8}{10000}}{8} = 0.4165$$

(8) $0.5034 - 0.4163 = 0.0871$