COMP6714 ASSIGNMENT 1 SAMPLE SOLUTION

Q1.

We assume that each document has been divided into a set of paragraph, and each paragraph divided into a set of sentences.

There are at least two possible solutions.

M1. We record the position of a term in a document in the following scheme:

`paragraphid.sentenceid.position`

where the `paragraphid` is the paragraph id within the document, and the `sentenceid` is the sentence id within the paragraph, and the `position` is the position (of the term) in the *document* (not in the sentence).

Then

- For `/k` queries, we just extract the `position` and process as usual.
- For `/S` queries, we just extract the `sentenceid` and process as usual.
- For `/P` queries, we just extract the `paragraphid` and process as usual.

M2. We include two special tokens, one corresponding to each sentence end position and the other to each paragraph end position. In order to answer, e.g., query Q with `/S` operator, we just perform list intersection of keywords in Q and the inverted list for sentence-end. Some changes are needed so that whenever we encounter another sentence end, we output the current "valid" occurrences of keywords in Q (if any), and then force their cursors to move to the first position beyond the current sentence end. This method is also known as the *extent list approach* (c.f., Chap 5.3.4 in [CMS09]).

Note: it is instrumental to think about the pros and cons of both methods.

Q2.

Note: In the following derivation, we conveniently assume that (1) L can be divided by x, and the gaps are all the same and are $\frac{L}{x}$[1]. This assumption makes the derivation simpler and revealing its **essence**. In addition, the results are the same under the $O()$ notation in most cases.

[1] not $\frac{L}{x+1}$ or that number -1 or -2.

1. Consider putting x evenly spaced pointers. Hence the gap between two consecutive pointers is $\frac{L}{x}$. Consider the worst case scenario to find an item, the cost is:

$$f(x) = x + \frac{L}{x}$$

To find the minimum value of $f(x)$, we take the derative of $f(x)$ and make it equal to 0, i.e.,

$$f'(x) = 1 - \frac{L}{x^2} = 0$$
$$\Longrightarrow x = \sqrt{L}$$

Easy to verify that this does achieve the minimum value.

2. Consider the worst case scenario to find an item in this double-binary search setting, the cost is:

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

While we can still use the above method, it is easy to see that $f(x) = \log(x \cdot \frac{L}{x}) = \log L$. Hence, the cost is the same no matter which x to choose.

Note: if one consider $f(x) = \log\lceil x \rceil + \log\lceil \frac{L}{x} \rceil$, there will be several x that really achieve the minium cost of $\lceil \log(L) \rceil$, while many other xs will achieve slightly higher cost.

3. Consider the worst case scenario to find an item in this double-binary search setting, the cost is:

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

To find the minimum value of $f(x)$, we take the derative of $f(x)$ and make it equal to 0, i.e.,

$$f'(x) = \frac{x - L}{x^2} = 0$$
$$\Longrightarrow x = L$$

This essentially says that since binary search is always no worse than sequential search (under our cost model), we should always use binary search (hence $x = L$).

Q3.

(1) The BM25 formula essentially limits the impact of tfs (the value converges when $tf \to \infty$). In our case, the scoring function is

$$score(d) \leq 6f(tf_1) + 2f(tf_2) + f(tf_3)$$

where $f(x) = \frac{3x}{2+x}$. Since $\lim_{x\to\infty} f(x) = 3$, we can find the maxscores for the terms are 18, 6, and 3.

(2) We first consider D_1, with score

$$score(D_1) = 6f(1) + 2f(1) + f(1) = 9$$

Then we consider D_2

$$score(D_2) = 6f(8) + 2f(0) + f(2) = 15.90$$

At this stage, both of them become the current top-2 results, and $\tau' = 9$. Since $3 + 6 \leq \tau'$, we only need to consider A. (hence no need to score D_4)

Driven by A, the next document to score is D_5. We need to probe the lists of B and C for D_5, and compute its score as

$$score(D_5) = 6f(3) + 2f(4) + f(2) = 16.30$$

Similarly, since now $\tau' = 15.90$.

The next document to consider is D_8

$$score(D_8) = 6f(10) + 2f(0) + f(1) = 16.00$$

Since A's postings list is now exhausted, we conclude that the final top-2 documents are D_5 and D_8. The algorithm scored 4 documents, and accessed 10 postings.

Q4.

k	1	2	3	4	5	6	7	8	9	10
precision (%)	100.00	100.00	66.67	50.00	40.00	33.33	28.57	25.00	33.33	30.00
recall (%)	12.50	25.00	25.00	25.00	25.00	25.00	25.00	25.00	37.50	37.50
k	11	12	13	14	15	16	17	18	19	20
precision (%)	36.36	33.33	30.77	28.57	33.33	31.25	29.41	27.78	26.32	30.00
recall (%)	50.00	50.00	50.00	50.00	62.50	62.50	62.50	62.50	62.50	75.00

(1) precision@20 is $\frac{6}{20}$.

(2) recall@20 is $\frac{6}{8}$. $F_1 = \frac{2 \cdot \frac{3}{10} \cdot \frac{3}{4}}{(\frac{3}{10} + \frac{3}{4})} = 0.4286$

(3) 25% recall corresponds to uninterpolated precisions of 100%, 66.67%, 50.00%, 40.00%, 33.33%, 28.57%, 25.00%.

(4) the interpolated precision for 33% recall is the maximum precision achieved for $k \geq 9$. Obviously, the maximum value is $\frac{4}{11} = 0.3636$.

(5) MAP is $\frac{1}{8} \cdot (\frac{1}{1} + \frac{2}{2} + \frac{3}{9} + \frac{4}{11} + \frac{5}{15} + \frac{6}{20}) = 0.4163$.

(6) The largest possible MAP is $\frac{1}{8} \cdot (\frac{1}{1} + \frac{2}{2} + \frac{3}{9} + \frac{4}{11} + \frac{5}{15} + \frac{6}{20} + \frac{7}{21} + \frac{8}{22}) = 0.5034$.

(7) The smallest possible MAP is $\frac{1}{8} \cdot (\frac{1}{1} + \frac{2}{2} + \frac{3}{9} + \frac{4}{11} + \frac{5}{15} + \frac{6}{20} + \frac{7}{9999} + \frac{8}{10000}) = 0.4165$.

(8) $0.5034 - 0.4163 = 0.0871$