COMP6714 ASSIGNMENT 1

DUE ON 23:59 25 SEPT, 2018 (TUE)

Q1. (*25 marks*)

Some Boolean retrieval systems (e.g., Westlaw) support the following proximity operators: `/k`, `/S`, and `/P`. Describe a simple modification to the positional inverted index to support all these three proximity operators.

Q2. (*25 marks*)

The textbook recommendation for placing evenly-spaced skip pointers for a posting list of length L is $\sqrt{L}$. This is based on the assumption that:

- **Step 1**: We perform sequential search in the skip pointers array, and then
- **Step 2**: We perform sequential search in the target segment.

You can also assume that

- every posting in the posting list has the same probability of being searched.
- the entire list is in the memory and we only consider the CPU cost.

Answer the following questions:

(1) Prove that choosing $\sqrt{L}$ skip pointers has the best worst case performance.
(2) Find the best number of skip pointers if we perform binary searches both Step I and II.
(3) Find the best number of skip pointers if we perform binary search in Step I and sequential search in Step II.

Q3. (*25 marks*)

Consider using the maxscore algorithm to find top-2 results for a query with three different terms $\{A, B, C\}$. The scoring function is the following function with $k_1 = k_3 = 2.0$ and $b = 0$ (commonly known as the *Okapi BM25*).

$$score(d, Q) = \sum_{t \in Q} idf_t \cdot \frac{(k_1 + 1)tf_{t,d}}{k_1((1 - b) + b\frac{L_d}{L_{ave}}) + tf_{t,d}} \cdot \frac{(k_3 + 1)tf_{t,Q}}{k_3 + tf_{t,Q}}$$

Answer the following questions. You need to show major steps.
The posting lists are shown below. Each posting consists of document ID and tf.

(1) Show that the maxscore for each keyword can be computed *without* examining the postings list.

term	idf	postings
A	6	$(D_1:1),(D_2:8),(D_5:3),(D_8:10)$
B	2	$(D_1:1),(D_5:4),(D_6:1),(D_7:4)$
C	1	$(D_1:1),(D_2:2),(D_4:1),(D_5:2),(D_6:3),(D_8:1),(D_9:1),(D_{10}:3),(D_{11}:7)$

TABLE 1. Posting Lists

(2) Using the maxscore obtained above, determine the postings that are accessed *for scoring* by the algorithm. You need to assume that each `skipTo(x)` call "magically" moves the cursor **directly** to the first posting with document ID at least x (i.e., it does *not* access any other postings).

Hint 1. *Calculate the maxscore if you know that the maximum tf is 1, 10, 100, and 1000, respectively.*

Q4. (*25 marks*)

The following list of Rs and Ns represents relevant (R) and nonrelevant (N) returned documents in a ranked list of 20 documents retrieved in response to a query from a collection of 10,000 documents. The top of the ranked list is on the left of the list. This list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection.

```
R R N N N     N N N R N     R N N N R     N N N N R
```

(Note that spaces above are just added to make the list easier to read)

(1) What is the precision of the system on the top-20?
(2) What is the F_1 on the top-20?
(3) What is/are the uninterpolated precision(s) of the system at 25% recall?
(4) What is the interpolated precision at 33% recall?
(5) Assume that these 20 documents are the complete result set of the system. What is the MAP for the query?

Assume, now, instead, that the system returned the entire 10,000 documents in a ranked list, and these are the first 20 results returned.

(6) What is the largest possible MAP that this system could have?
(7) What is the smallest possible MAP that this system could have?
(8) In a set of experiments, only the top-20 results are evaluated by hand. The result in (5) is used to approximate the range (6) to (7). For this example, how large (in absolute terms) can the error for the MAP be by calculating (5) instead of (6) and (7) for this query?

SUBMISSION INSTRUCTIONS

You need to write your solutions to the questions in a pdf file named `ass1.pdf`. You **must**

- include your **name** and **student ID** in the file, and
- the file can be opened correctly on CSE machines.

You need to show the key steps to get the full mark.

Note: Collaboration is allowed. However, each person must independently write up his/her own solution.

You can then submit the file by `give cs6714 ass1 ass1.pdf`. The file size is limited to 5MB.

Late Penalty: **-10%** per day for the first two days, and **-20%** per day for the following days.