COMP6714 ASSIGNMENT 1

DUE ON 23:59 12 AUG, 2016 (MON)

Q1. (*30 marks*)

Answer the following questions. Note that when we refer to Google, we mean searches via `www.google.com.au`, and when we refer to Bing, we mean the `www.bing.com` search service, but with "Region" choice set to "Only from Australia".

(1) Search "DFA" using Google and Bing. Compare their top-10 results (*not* including the ads). List the web sites returned by both search engines in their top-10 results (you need to include visible screenshots).

(2) Search "ioauen" using Google and Bing. Compare their top-10 results. Describe the possible differences the two search engines have in terms of token normalization, query expansion, and query suggestion.

(3) Translate and write down the following Boolean searches to queries using the (advanced) query syntax provided by Google[1]. Make sure that you disable google's automatic query expansion (e.g., from `otta` to `otto`). You should also record the result numbers estimated by Google.
   (a) Neuro-linguistic
   (b) otta
   (c) Neuro-linguistic AND otta (note: AND means conjunction)
   (d) Neuro-linguistic OR otta (note: OR means disjunction)
   (e) Neuro-linguistic /1 otta (note: /1 means the occurrence of the two terms must be within distance of 1)
   (f) bugle
   (g) bugle bugle
   (h) bugle bugle bugle

   Do the estimated numbers make sense in terms of Boolean logic? What is the upper and lower bounds for the number of query results of the third query based on the number returned from the first and the second queries?

Q2. (*20 marks*)

Shown below is a portion of a positional index in the format:

```
term: doc1: <position1, position2, ...>;
doc2: <position1, position2, ...>;
```

[1] `http://www.google.com.au/support/websearch/bin/answer.py?answer=136861`

```
angels: 2: <36,174,252,651>;   4: <12,22,102,432>;     7: <17>;
fear  : 2: <87,704,722,901>;   4: <13,43,113,433>;     7: <18,328,528>;
fools : 2: <1,17,74,222>;      4: <8,78,108,458>;      7: <3,13,23,193>;
in    : 2: <3,37,76,444,851>;  4: <10,20,110,470,500>; 7: <5,15,25,195>;
rush  : 2: <2,66,194,321,702>; 4: <9,69,149,429,569>;  7: <4,14,404>;
to    : 2: <47,86,234,999>;    4: <14,24,774,944>;     7: <199,319,599,709>;
tread : 2: <57,94,333>;        4: <15,35,155>;         7: <20,320>;
where : 2: <67,124,393,1001>;  4: <11,41,101,421,431>; 7: <15,35,735>;
```

(1) Which document(s) (if any) match each of the following queries where each expression within quotes is a phrase query?
   (a) “fools rush in”
   (b) “fools rush in” AND “angels fear to tread”.
   At which positions do the queries match?

(2) There is something wrong with this positional index. What is the problem?

Q3. (*25 marks*)

(1) What is the worst-case time complexity of the algorithm depicted in Figure 2.12 in the MRS08 textbook? Describe a simple modification that improve the time complexity of the algorithm with respect to k.

(2) 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.

Q4. (*25 marks*)

How many sub-indexes will the three dynamic indexing methods, namely *immediate merge*, *no merge*, and *logarithmic merge*, create, respectively? Assume that

- We start from scratch.
- We use $|C|$ to denote the total size of the document collection, and M to denote the memory size.
- You can simply assume that by you can create a sub-index of size M after consuming documents of size M.

You need to show your steps.

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`.

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