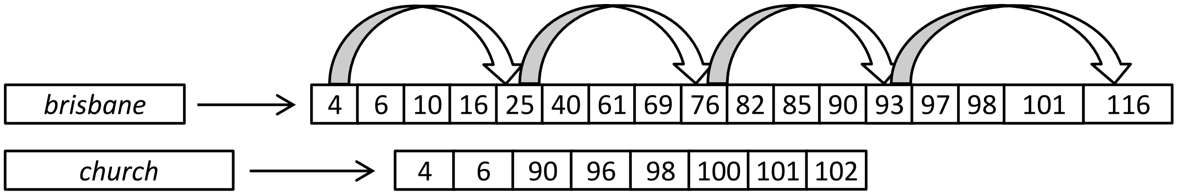
Information Retrieval (CS60092)
Mid-semester examination, Autumn 2012 – 2013

Time: 2 hours, **Full Marks:** 50

Attempt all questions.
Use of calculator is allowed.
State any assumptions made clearly.

Q. 1> Consider the Boolean query *brisbane* AND *church*. The postings lists of the two words are



brisbane ⟶ 4 | 6 | 10 | 16 | 25 | 40 | 61 | 69 | 76 | 82 | 85 | 90 | 93 | 97 | 98 | 101 | 116

church ⟶ 4 | 6 | 90 | 96 | 98 | 100 | 101 | 102

shown. The list for *brisbane* is augmented with skip pointers.

(i) How many comparisons are required if neither of the lists had skip pointers? List the comparisons sequentially.

(ii) How many comparisons are required in the situation shown, when *brisbane* has skips? List the comparisons sequentially.

(iii) How often is a skip pointer used in the above situation?

(iv) What is the trade-off issue faced involving time and space complexities when choosing the number of skip pointers for a given postings list?

(v) What is a common heuristic for deciding the number of skip pointers, given that the size of the postings list is *P*?

(vi) Are skip pointers useful for processing Boolean OR queries? Justify. **(3 + 3 + 1 + 1 + 1 + 1 = 10)**

Q. 2> List two cases of IR where achieving perfect recall is more important than high precision. **(2)**

Q. 3> (i) What is the Levenshtein distance between: (a) *house* and *musket* (b) *basket* and *stables*?

(ii) What is the time complexity of the dynamic programming approach that is commonly used to calculate the Levenshtein distance between two strings of length $|s_1|$ and $|s_2|$?

(iii) What is the Jaccard coefficient between: (i) *idea* and *sidearm*? (ii) *boat* and *aboard*? Assume alphabet bigrams as the unit and take into account terminal dollar ($) for the computations (a word *cat* is assumed to be denoted as *cat$*). **(2 + 1 + 2 = 5)**

Q. 4> Consider the query *catholic church brisbane*. Assume the vector space model and the TF-IDF weighting scheme. The corpus consists only of the following documents:

d_1: *roman catholic church brisbane roman*

d_2: *church brisbane church church*

d_3: *catholic catholic protestant protestant all all brisbane*

d_4: *catholic church welcome catholic church brisbane*

Assign vector indices 1 – 7 to the vocabulary words in the following order:

roman, catholic, church, brisbane, protestant, all, welcome

(i) Now, assuming this order, clearly write the complete weight vectors for the query and each document. State the formula used for IDF.

(ii) What are the lower and upper bounds for IDF of a term in a corpus according to the stated formula?

(iii) Rank the documents w.r.t. the query assuming the overlap score measure [Hint: The score for a query-document pair is the sum of the weights of the query terms in the document vector].

(iv) Rank the documents assuming the cosine similarity as a scoring function. Show all steps of the computation. **(3 + 1 + 2 + 4 = 10)**

Q. 5> (i) Define a champion list for a term.

(ii) How is a champion list useful in IR?

(iii) What is an alternate term for a champion list? **(1 + 1 + 1 = 3)**

Q. 6> Consider two queries for which there are 4 and 6 relevant documents in the collection respectively. Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result):

System 1, Query 1: R N R R R N N N N N

System 1, Query 2: R R R R N N N R N R

System 2, Query 1: N N N N R R R N N R

System 2, Query 2: N N N N R R R R R R

(i) What is the MAP of each system? Which system has a higher MAP?

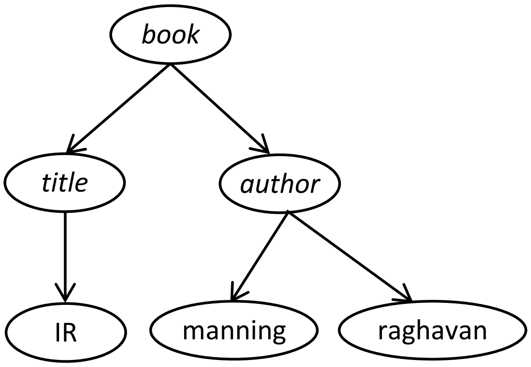
(ii) What does the result say about what is important in getting a good MAP score?

(iii) How is R-precision of a system defined? What is the R-precision of each system here? Does it rank the systems in the same order as MAP? **(6 + 1 + 3 = 10)**

Q. 7> (i) Draw the XML DOM for the document below.

```
<play>
<author>Shakespeare</author>
<title>Macbeth</title>
<act number="i">
<scene number="vii">
<title>Macbeth’s hostel</title>
<verse>Will I with wine</verse>
</scene>
</act>
</play>
```

(ii) What are the structural terms (represented as lexicalized subtrees) of the XML document below?



(iii) Name the following (expand abbreviations wherever applicable):

(a) Most popular format for XML queries

(b) Evaluation forum for XML

(c) Two types of information needs (or topics) at (b)

(d) Two orthogonal dimensions of relevance assessment in XML **(3 + 3 + 4 = 10)**