Name:	____________________ , (Family name)	____________________ (Given name)
Student ID:	____________________	

THE UNIVERSITY OF NEW SOUTH WALES
Final Exam

COMP6714
Information Retrieval and Web Search

SESSION 2, 2011

- Time allowed: **10 minutes** reading time + **3 hours**
- Total number of questions: **10+1**
- Total number of marks: **100+5**
- Only UNSW approved calculators are allowed in this exam.
- Answer **all** questions.
- You can answer the questions in any order.
- Start each question on a **new page**.
- Answers must be written in ink.
- Answer these questions in the script book provided.
- Do **not** write your answer in this exam paper.
- If you use more than one script book, fill in your details on the front of **each** book.
- You may **not** take this question paper out of the exam.

Question 1 (20 marks)

Briefly answer the following questions (1-4 sentences) in your script book. Lengthy but irrelevant answers will be *penalized.*

(a) How does stemming typically affect recall? Why?

(b) Given at least two reasons why *language identification* is important when indexing documents.

(c) Why specialized algorithms are needed to construct inverted index for large document collections?

(d) What are the largest gap that can be encoded in 2 bytes using variable byte code? You also need to show the encoded two bytes.

(e) Why is cosine a better similarity metric than the inverse of Euclidean distance in vector space model?

(f) Why is vector space model generally considered a better retrieval model than the boolean model?

(g) List the advantage(s) of using NDCG to evaluate Web search results over measures such as MAP.

(h) List one problem with the probabilistic ranking principle.

(i) In the early age of Web search engines (definitely pre-Google era), some system uses the following term frequency weighting $\frac{2 \cdot tf}{2+tf}$ to fight spam Web pages. Explain why this worked.

(j) What is a "shingle", and describe briefly the shingling method to detect near duplicate documents.

(k) List at least three requirements that complicate the design and implementation of an industrial strength crawler.

(l) Define the terms "hub" and "authority" in the context of the HITS algorithm. Can a page be both a hub and authority page at the same time?

Question 2 (5 marks)

Consider the algorithm (from the textbook) to intersect two postings lists p_1 and p_2.

Algorithm 1: Intersect(p_1, p_2)

```
answer ← ∅;
while $p_1 \neq$ nil and $p_2 \neq$ nil do
    if $docID(p_1) = docID(p_2)$ then
        Add(answer, $docID(p_1)$);
        $p_1 \leftarrow next(p_1)$;
        $p_2 \leftarrow next(p_2)$;
    else if $docID(p_1) < docID(p_2)$ then
        $p_1 \leftarrow next(p_1)$;
    else
        $p_2 \leftarrow next(p_2)$;
return answer;
```

(a) What is the time complexity of the algorithm? $O(|p_1| + |p_2|)$

(b) Modify the algorithm so that it can answer queries like A AND NOT B in time $O(|p_1| + |p_2|)$, where A and B are two terms.

(c) Is it possible to modify the algorithm so that it can answer queries like A OR NOT B in time $O(|p_1| + |p_2|)$? If not, what complexity can you achieve?

NO $O(N)$

```
IntersectAndNot (p1, p2)
  answers = < >
  while p1 != NULL and p2 != NULL do
    if (docID(p1) = docID(p2)) then
        p1 ← next(p1)
        p2 ← next(p2)
    else if ( docID(p1) < docID(p2) ) then
        Add( answers, docID(p1) ) ;
        p1 ← next(p1)
    else
        p2 ← next(p2)
  end if
end while
return answers
```

if p1 != NULL do

剩余加进去

Question 3

(10 marks)

Consider a casual user who input the boolean query "A OR B AND C". Our system deems the query as ambiguous, as either the OR or the AND operator can be executed first. To be on the safe side, the system decides to retrieve those results that belong to both interpretations only (i.e., no matter which interpretation the user intended, it will include our system's result). Describe how to support such query efficiently by accessing the inverted lists of tokens A, B, and C at most once.

[(A OR B) and C] and [A OR (B and C)]

= [(A OR B) and C] and [(A OR B) and (A OR C))

= (A OR B) and C and (A OR B) and (A OR C)

= (A OR B) and C and (A OR C)

= (A OR B) and C

```
Intersect($p_1$, $p_2$, $p_3$)        ($p_1$ OR $p_2$) and $p_3$
answers = < >
while ($p_3$ != NULL) do.
    while (docID($p_1$) < docID($p_3$)) do
        $p_1 \leftarrow$ next($p_1$)
    end while
    while (docID($p_2$) < docID($p_3$)) do
        $p_2 \leftarrow$ next($p_2$)
    end while
    if docID($p_1$) = docID($p_3$) or docID($p_2$) = docID($p_3$) then
        Add(answers, docID($p_3$))
    end if
    $p_3 \leftarrow$ next($p_3$)
end while
return answers
```

Question 4 (10 marks)

From the following sequence of γ-coded gaps, reconstruct first the gap sequence and then the postings sequence (assume that *docid* starts from 1). Note that spaces were deliberately added for clarity purpose only. You need to illustrate your steps.

```
1110 1101 1111 1001 0111 1111 1110 1000 1111 1001
```

$1+2+4+16$

$2^3 + 6 = 14$

$2^6 + 23 = 87$

$2^7 + 71 = 199$

$2^2 + 1 = 5$

$2^7 +$

64
23
87

$1+2+4+64$

64
128
71
199

15
87

102
199
201

1, 15, 102, 201, 206.

Question 5 (10 marks)

The figure below shows the output of two information retrieval systems on the same two queries in a competitive evaluation. The top 15 ranks are shown. Crosses correspond to a document which has been judged relevant by a human judge; dashes correspond to irrelevant documents. There are no relevant documents in lower ranks.

System 1:

Rank	Q1	Q2
1	-	X
2	X	-
3	X	-
4	X	-
5	-	-
6	-	-
7	-	-
8	X	-
9	X	-
10	X	-
11	X	-
12	-	-
13	-	X
14	-	X
15	X	-

System 2:

Rank	Q1	Q2
1	X	X
2	X	-
3	X	-
4	-	X
5	X	X
6	X	-
7	-	-
8	-	-
9	-	-
10	-	-
11	X	-
12	X	-
13	-	-
14	-	-
15	X	-

(a) Explain the following evaluation metrics and give results for query Q1 for both systems.

system 1 : $\frac{6}{10}$ system 2 : $\frac{5}{10}$

1. Precision at rank 10.
2. Recall at precision 0.5. System 1 = $\frac{1}{8}$, $\frac{3}{8}$, $\frac{4}{8}$ $\frac{7}{8}$ system 2 : $\frac{5}{8}$, $\frac{7}{8}$

(b) The metrics in part (a) above are not adequate measures of system performance for arbitrary queries. Why not? What other disadvantages do these metrics have?

(c) Give the formula for mean average precision (MAP), and calculating MAP for both systems.

(d) For each system, draw a precision-recall curve. Explain how you arrived at your result.

(b) E-measure. ⇒ unrank

rankresult

(c) $MAP = \sum_{Q}^{|Q|} \frac{1}{|R|} \cdot \sum_{Rel} P_r$

Sys1 $\begin{cases} Q_1 : MAP_1 = \frac{1}{8}\left(\frac{1}{2}+\frac{2}{3}+\frac{3}{4}+\frac{4}{8}+\frac{5}{9}+\frac{6}{10}+\frac{7}{11}+\frac{8}{15}\right) \\ Q_2 = MAP_2 = \frac{1}{3}\left(1+\frac{2}{13}+\frac{3}{14}\right) \end{cases}$

$MAP = \frac{1}{2}(MAP_1 + MAP_2)$

Sys2 $\begin{cases} Q_1 = MAP_1 = \frac{1}{8}\left(1+1+1+\frac{4}{5}+\frac{5}{6}+\frac{6}{11}+\frac{7}{12}+\frac{8}{15}\right) \\ Q_2 : MAP_2 = \frac{1}{3}\left(1+\frac{2}{4}+\frac{3}{5}\right) \\ MAP = \frac{1}{2}(MAP_1 + MAP_2) \end{cases}$

Question 6 (10 marks)

Determine the new query vector determined by the Rocchio relevant feedback algorithm ($\alpha = \beta = \gamma = 1.0$), given that the initial query is "$t_1\, t_3$" and we have the following documents and user feedback.

docid	t_1	t_2	t_3	t_4	feedback
1	2	1	0	0	R
2	3	2	1	0	NR
3	0	3	0	3	R
4	2	1	2	2	NR
5	0	1	2	3	NR

Note: "R" standards for relevant and "NR" stands for non-relevant.

Question 7 (10 marks)

(a) State and *justify briefly* the assumptions made to derive Equations (3) from (2) and Equation (6) from (5) in the Binary Independence Model.

(b) State which values need to be estimated for a document collection in the final Equation (8) (i.e., other parts can be discarded safely without affecting the ranking).

Let $\vec{x}$ be the binary term incidence vector representing document D, $O(p)$ be the odd ratio of probability p, Q be the query, R and NR stand for "relevant" and "non-relevant", respectively, V is the vocabulary.
In addition, we use the shorthand notations: $p_i = p(x_i = 1|R, Q)$ and $r_i = p(x_i = 1|NR, q)$.

$$O(R|Q,\vec{x}) = \frac{p(R|Q,\vec{x})}{p(NR|Q,\vec{x})} \quad (1)$$

$$= \frac{p(R|Q)}{p(NR|Q)} \cdot \frac{p(\vec{x}|R,Q)}{p(\vec{x}|NR,Q)} \quad (2)$$

independence assumption

$$= O(p(R|Q)) \cdot \prod_{i=1}^{|V|} \frac{p(x_i|R,Q)}{p(x_i|NR,Q)} \quad (3)$$

$$= O(p(R|Q)) \cdot \prod_{x_i=1} \frac{p(x_i=1|R,Q)}{p(x_i=1|NR,Q)} \cdot \prod_{x_i=0} \frac{p(x_i=0|R,Q)}{p(x_i=0|NR,Q)} \quad (4)$$

$$= O(p(R|Q)) \cdot \prod_{x_i=1} \frac{p_i}{r_i} \cdot \prod_{x_i=0} \frac{1-p_i}{1-r_i} \quad (5)$$

when $q_i=0$, $p_i=r_i$

$$= O(p(R|Q)) \cdot \prod_{x_i=1, x_i\in Q} \frac{p_i}{r_i} \cdot \prod_{x_i=0, x_i\in Q} \frac{1-p_i}{1-r_i} \quad (6)$$

$$= O(p(R|Q)) \cdot \prod_{x_i=1, x_i\in Q} \frac{p_i}{r_i} \cdot \left(\frac{\prod_{x_i\in Q} \frac{1-p_i}{1-r_i}}{\prod_{x_i=1, x_i\in Q} \frac{1-p_i}{1-r_i}} \right) \quad (7)$$

$$= O(p(R|Q)) \cdot \boxed{\prod_{x_i=1, x_i\in Q} \frac{p_i(1-r_i)}{r_i(1-p_i)}} \prod_{\underset{q_i=1}{\boxed{x_i\in Q}}} \frac{1-p_i}{1-r_i} \quad (8)$$

$p_i = p(x_i=1|R,q)$

$r_i = p(x_i=0|NR,q)$

(a) independence assumption / when $q_i=0$, $p_i=r_i$.

(b)

Question 8 (5 marks)

Suppose we have a document collection with an extremely small vocabulary with only 6 words $w_1, w_2, \ldots, w_6$. The following table shows the estimated background language model $p(w|C)$ using the whole collection of documents (2nd column) and the word counts for document d_1 (3rd column) and d_2 (4th column), where $c(w, d_i)$ is the count of word w in document d_i. Let $Q = \{w_1, w_2, w_3, w_4\}$ be a query.

Word	$p(w\|C)$	$c(w, d_1)$	$c(w, d_2)$
w_1	0.800	2	7
w_2	0.100	3	1
w_3	0.025	1	1
w_4	0.025	2	1
w_5	0.025	2	0
w_6	0.025	0	0

(a) Suppose we do not smooth the language model for d_1 and d_2. Compute the likelihood of the query for both d_1 and d_2, i.e., $p(Q|d_1)$ and $p(Q|d_2)$ (Do *not* compute the log-likelihood. You should use the scientific notation (e.g., 0.0061 should be 6.1×10^{-3}) Which document would be ranked higher?

(b) Suppose we now smooth the language model for d_1 and d_2 using the Jelinek-Mercer smoothing method with $\lambda = 0.8$ (i.e., $p(w|d) = \lambda \cdot p_{\text{mle}}(w|M_d) + (1-\lambda) \cdot p_{\text{mle}}(w|M_c)$). Recompute the likelihood of the query for both d_1 and d_2, i.e., $p(Q|d_1)$ and $p(Q|d_2)$ (Do *not* compute the log-likelihood. You should use the scientific notation) Which document would be ranked higher?

(a). $P(Q|d_1) = \prod_{x \in Q} p(x|d_1) = \frac{2}{10} \cdot \frac{3}{10} \cdot \frac{1}{10} \cdot \frac{2}{10} = \frac{12}{10^4} = 1.2 \times 10^{-3}$

$p(Q|d_2) = \prod_{x \in Q} p(x|d_2) = \frac{7}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} = \frac{7}{10^4} = 7 \times 10^{-4}$

Therefore, document 1 should be ranked higher.

b) $P(Q|d_1) = (0.8 \cdot \frac{2}{10} + 0.2 \cdot 0.8) \cdot (0.8 \cdot \frac{3}{10} + 0.2 \cdot 0.1) \cdot (0.8 \cdot \frac{1}{10} + 0.2 \cdot 0.025) \cdot (0.8 \cdot \frac{2}{10} + 0.2 \cdot 0.025)$

$p(Q|d_2) = (0.8 \cdot \frac{7}{10} + 0.2 \cdot 0.8) \cdot (0.8 \times \frac{1}{10} + 0.2 \cdot 0.1) \cdot (0.8 \cdot \frac{1}{10} + 0.2 \cdot 0.025) \cdot (0.8 \cdot \frac{1}{10} + 0.2 \cdot 0.025)$

Question 9 (10 marks)

Consider the following web graph:

```
Page A points to page B, C, and D.
Page B points to C and D.
Page C points to A and E.
Page D points to E and F.
Page E points to G.
Page F points to G and H.
```

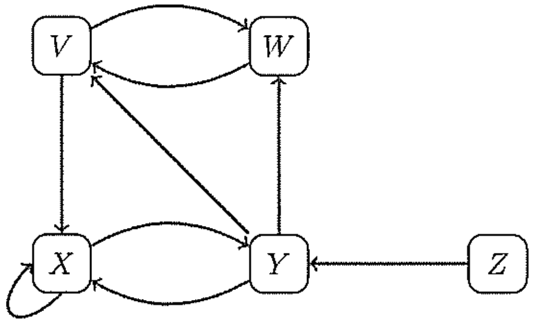
Consider a crawler that starts from page A

(a) Give the order of the indexing, assuming the crawler uses a URL frontier with duplicate detection, and all the pages are at different web sites.

(b) Assume pages B, C, F, H are on web site α, pages D, E, G are on web site β, and page A is on web site γ. The politeness policies on these three web sites all specify at least 3 seconds between each visit (i.e., if the crawler visit a web site at the i second, the earliest time it can revisit the web site is the $i+3$ second). We assume that (1) the crawler can only fetch a page every one second, and all the processing (including physically getting the page, extracting and processing the links, etc.) can be completed before the next fetch. (2) the crawler process links in the order mentioned above.

The crawler still uses a ULR frontier with duplicate detection, and also uses back queues to adhere to the politeness policies. Give the order of the indexing. (If two pages can be visited at the same time, we always choose the smaller one according to the alphabetical order)

Question 10 (10 marks)



(a) Explain the concept of PageRank, and how it is calculated in practice.

(b) Why is it relevant for Web search?

(c) Give, and briefly explain, the corresponding matrix notation of the PageRank computation.

(d) Show the final matrix that will be used for the PageRank calculation for the above graph, if the random teleporting probability is 0.2.

(e) Perform two iterations starting from the initial probability distribution vector of $(0.2, 0.2, 0.2, 0.2, 0.2)$.

BONUS

Question 11 (5 marks)

Explain analytically why galloping search (aka. double binary search) is preferred to the normal binary search when implementing the `skipTo(docid)` method on a sorted list of `docids`. Make sure you state clearly the meaning of variables and any assumption you use.

END OF EXAM PAPER