Name: _____ , _____
(Family name) (Given name)

Student ID: _____

THE UNIVERSITY OF NEW SOUTH WALES
Final Exam

COMP6714
Information Retrieval and Web Search

TERM T3, 2020

- Time allowed: **10 minutes** reading time + **2 hours**
  - Exception: students with extra exam time approved by **Equitable Learning Services (ELS)** can make submissions after 14:30, 2 December 2020 within their **approved extra time**.
- Total number of questions: **8**.
- Total number of marks: **100**
- Total number of pages: **8 excluding this cover page**
- This is an open-book exam. You are allowed to use textbook(s), lecture notes and other study materials. However, you are **not** allowed to (1) communicate with anyone else or (2) use the Internet during the exam.
- Items allowed: UNSW approved calculators.
- You can answer the questions in any order.
- Start each question on a **new page**.
- Answers must be written in ink on A4 papers and scanned into a PDF file. Alternatively, you can use any software to directly generate the answers in a PDF file.

Question 1 (10 marks)

Write down the pseudo-code of answering the Boolean query `not` A `and` B. You need to use the function skipTo(id) wherever it is possible.

You can refer to the following skeleton code. **Note** that it is **by no means** the complete pseudo-code, and you should add multiple lines or modify existing line(s) to complete this task.

Algorithm 1: Q1(p_1, p_2)

```
1  answer ← ∅;
2  while ... do
3      if docID(p1) > docID(p2) then
4          ;
5      else
6          if docID(p1) < docID(p2) then
7              ;
8          else
9              ;
10 return answer;
```

Question 2 (14 marks)

Consider applying γ-encoding to a posting list of n document IDs (within the range of $[1, N]$).

Prove that:

- For a value x, its γ-encoded value takes at most $2\log_2(x) + 1$ bits.
- The compressed posting list (using γ codes on the gaps) takes at most $n \cdot \log_2\left(\frac{2N^2}{n^2}\right)$ bits.

Question 3 (14 marks)

(a) Consider a dictionary that contains v terms, and each term has exactly l characters. Assume we build both a permuterm index and a bi-gram index for the dictionary. What are the sizes of these two indexes (note, do not include the size of the dictionary and the inverted lists), respectively? You may assume that a pointer (to a term in the dictionary) is 4-bytes and all terms only contain characters from `a` to `z`.

(b) Consider a query of `P*Q*R`, where `P`, `Q`, and `R` are a sequence of characters. Briefly describe how a permuterm index can be used to efficiently answer this query.

(c) The above query can also be answer **without** using list interaction. Briefly describe how this can be done, and give a reason why this might be more efficient than the previous query processing method.

Question 4 (14 marks)

Consider the logarithmic merge algorithm for dynamic index construction. The sub-indexes created on the disk are: I_3, I_2, I_1, and I_0 (note: I_0 is the smallest sub-index on the disk). Assume the current in-memory index is full and needs to be dumped to the disk.

(a) What are the sub-indexes after dumping the current in-memory index to the disk?

(b) How many sub-indexes will the algorithm create to index a document collection of size $|C|$ when the memory size is M.

(c) Let $|C| = 14 \cdot M$. What is the total number of times sub-indexes are merged? You need to include merges of a sub-index on the disk and the index in the memory.

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 10 ranks are shown. Crosses correspond to a document which has been judged relevant by a human judge; dashes correspond to irrelevant documents. Assume that System 1 retrieved all the relevant documents for both queries.

System 1:

Rank	Q1			Q2		
1	X			X		
2	X			-		
3	X			-		
4	X			-		
5	-			-		
6	-			X		
7	-			-		
8	-			-		
9	X			X		
10	X			X		

System 2:

Rank	Q1			Q2		
1	X			X		
2	X			-		
3	X			-		
4	-			X		
5	X			X		
6	X			-		
7	-			-		
8	-			-		
9	X			-		
10	-			-		

(a) Explain the following evaluation metrics and give results for query Q2 for both systems.

1. Precision at rank 8.
2. Recall at precision $\frac{1}{3}$.

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

(c) Consider Q1 for System 1. Compute the interpolated precisions at recall levels 0.5 and 0.8, respectively.

Question 6 (14 marks)

Consider the following documents and the ground-truth of their relevance to the query $\{x_1, x_3\}$. Give the order that these documents will be ordered under the BIM model with add $\frac{1}{2}$ smoothing? You need to justify your answer.

Document	**Relevant?**	x_1	x_2	x_3
D_1	T	1	1	1
D_2	T	0	1	0
D_3	F	1	0	0
D_4	T	0	0	1
D_5	F	0	1	0

Question 7 (10 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 $Q = \{w_1, w_2, w_3, w_4, w_5, w_6\}$ be a query.

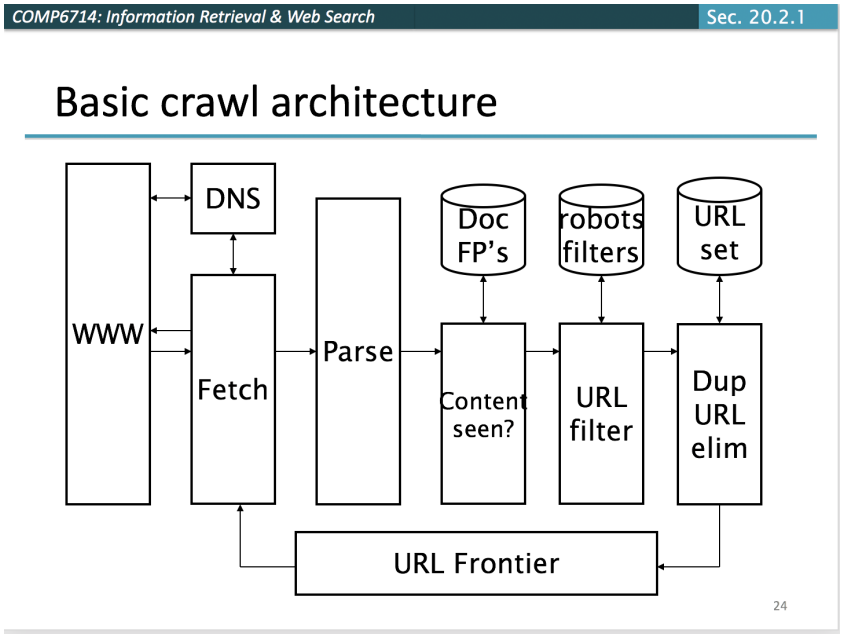
Word	$p(w\|C)$	$c(w, d_1)$	$c(w, d_2)$
w_1	0.800	2	7
w_2	0.100	3	1
w_3	0.025	1	1
w_4	0.025	2	1
w_5	0.025	2	0
w_6	0.025	0	0

(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?

Question 8 (14 marks)

Consider the basic architecture of a crawler in the figure below.



(a) Explain why the "content seen?" module is needed before the "Dup URL elim" module?

(b) Assume that the search engine uses MinHash algorithm to detect near-duplicate documents. Given a document with following hashed shingles: $\{1, 7, 15, 81\}$, and two universal hashing functions:

- $h_1(x) = (7x + 1 \bmod 31) \bmod 13$
- $h_2(x) = (18x + 26 \bmod 31) \bmod 13$

What are the resulting MinHash signatures of the document?

END OF EXAM PAPER