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

Student ID: _____

THE UNIVERSITY OF NEW SOUTH WALES
Final Exam

COMP6714
Information Retrieval and Web Search

TERM T3, 2021

- Time allowed: **10 minutes** reading time + **2 hours** + **20 minutes** submitting time
  - Exception: students with extra exam time approved by **Equitable Learning Services (ELS)** can make submissions after 15:30, 8 December 2021 within their **approved extra time**.
- Total number of questions: **7**.
- Total number of marks: **100**
- Total number of pages: **7 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 (13 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 either interpretations (i.e., no matter which interpretation the user intended, it will be included in 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.

Question 2 (16 marks)

Consider a collection of 3 documents as below:

D1: hello world

D2: hello word

D3: hi lord

(a) Consider a dictionary for this collection. Assume we build a permuterm index and a bi-gram index for the dictionary. Which one has a larger size? You may assume that a pointer (to a term in the dictionary) is 4-bytes. You need to show the steps.

(b) Illustrate the *complete* steps to use the permuterm index to answer the query `*e*lo`.

(c) Illustrate the *complete* steps to use a bi-gram index to answer the query `*e*lo`.

Question 3 (13 marks)

Consider using the noisy channel model to correct the non-word "wirdy". The table below gives all the tokens in the corpus that has no more than 3 Damerau-Levenshtein edit distance to "wirdy", and their occurrence probabilities.

Word	$p(w)$
ware	0.065
weird	0.045
windy	0.045
wired	0.100
wiry	0.029
word	0.025
wordy	0.037

(a) What are the set of candidates the model will consider using Damerau-Levenshtein edit distance with a threshold of 2?

(b) Using the following error model to compute the error probability

$$P(err) = \frac{\exp(-\beta \cdot err)}{\sum_{i=0}^{2} \exp(-\beta \cdot i)}$$

where err is the edit distance between the observed word and a candidate word, and $\beta = 1.0$, which word will be chosen as the correct word? You need to show your steps.

Question 4 (16 marks)

(a) What is the largest number that can be stored in no more than 3 bytes using Elias-γ encoding?

(b) What is the largest number that can be stored in no more than 3 bytes using Elias-δ encoding?

(c) Assume Elias-γ codes are used to encode the gap sequences of the postings lists. Let the encoded sequence for term A be: 1111 1111 1011 1100 1101 0011 1110 0000 0 and the encoded sequence for term B be: 1111 1111 1100 0000 0011 1011 1110 0000 0. What should be the result for the query "`A AND NOT B`".

Question 5 (13 marks)

Consider the following collection of documents:

D1: do you like green eggs and ham

D2: i do not like them

D3: i do not like eggs or ham

D4: i do not like eggs

D5: why are they green

Suppose we run the vector space query `green eggs` and we want to show the user 4 documents. The results will be ranked by the cosine similarity based on the td-idf weight.

What are the final 4 documents and their ranks? You need to show the complete steps of the computation.

Question 6 (16 marks)

(a) Is it possible that there is a horizontal line segment in the recall-precision graph? If yes, give a simple example; otherwise, state the reason concisely.

For subquestions (b) to (d), we consider the figure below which 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	-			-		
5	X			X		
6	-			-		
7	-			-		
8	-			-		
9	X			X		
10	X			X		

System 2:

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

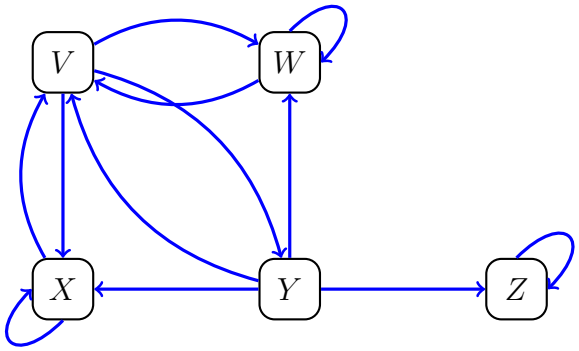
(b) Give the results for the following evaluation metric for query Q2 for both systems.

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

(c) Calculate MAP for both systems.

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

Question 7 (13 marks)



(a) Consider the markov chain of the above graph, is it ergodic? Why?

(b) Show the transition probability matrix of the markov chain.

(c) Show the final matrix that will be used for the PageRank calculation for the above graph, given the random teleporting probability is 0.5.

(d) Perform one iteration starting from the initial probability distribution vector of $(0.2, 0.2, 0.2, 0.2, 0.2)$.

END OF EXAM PAPER