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

Student ID: _____

THE UNIVERSITY OF NEW SOUTH WALES
Final Exam

MOCK EXAM
COMP6714
Information Retrieval and Web Search

TERM T3, 2020

- Time allowed: **10 minutes** reading time + **2 hours**
- Total number of questions: **3**.
- Total number of marks: **34**
- Total number of pages: **3 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 (14 marks)

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

Algorithm 1: Intersect(p_1, p_2)

```
1  answer ← ∅;
2  while p1 ≠ nil and p2 ≠ nil do
3      if docID(p1) = docID(p2) then
4          Add(answer, docID(p1));
5          p1 ← next(p1);
6          p2 ← next(p2);
7      else if docID(p1) < docID(p2) then
8          p1 ← next(p1);
9      else
10         p2 ← next(p2);
11 return answer;
```

(a) What is the time complexity of the algorithm?

(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?

Question 2 (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.

Question 3 (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
```

END OF EXAM PAPER