COMP6714 ASSIGNMENT 1

DUE ON 20:59 4 NOV, 2020 (WED)

Some Boolean retrieval systems://eduassistpro.gith.io/restricts the occurrences matches to be within the same sentence.

Assume that we have created an additional positional list for \$, which records the positions of the end Atle sent grant the Property Exam Help A B C. D E.

the position list for \$ is [4, 7].

You are required to engine an algorithm to support the query use finds solved in the property of the query use finds solved in the property of the query use finds solved in the property of the query use finds solved in the property of the query use finds solved in the property of the query use finds solved in the property of the query use finds solved in the property of the query use finds solved in the query use finds at the query use f

- the occurrences of A and B must be within the same sentence.
- the occurrence o

For example, the abounding the state of the

- make simple modifications to the pseudocode shown in Alg actly the algorithm in Figure 2.12 in the textbook. Note that we mod algorithm slightly death analyging exclash to the start of the start of
 - you need to insert some code between Lines 6 and 7, and p ifications to some lines afterwards.
 - In your submitted algorithm pseudocode (named $Q1(p_1, p_2, p_{\$})$), clearly mark the modifications using color or boxes.
- You can assume that there is a function $\mathsf{skipTo}(p, docID, pos)$, which move the cursor of list p to the first position such that (1) the position belongs to a document docID, and (2) the position is no smaller than pos.

Q2. (25 marks)

Consider the scenario of dynamic inverted index construction. Assume that t sub-indexes (each of M pages) will be created if one chooses the no-merge strategy.

- (1) Show that if the logarithmic merge strategy is used, it will result in at most $\lceil \log_2 t \rceil$ sub-indexes.
- (2) Prove that the total I/O cost of the logarithmic merge is $O(t \cdot M \cdot \log_2 t)$.

Algorithm 1: PositionalIntersect(p_1, p_2, k)

```
1 answer \leftarrow \emptyset;
2 while p_1 \neq \mathbf{nil} \land p_2 \neq \mathbf{nil} \mathbf{do}
      if doclD(p_1) = doclD(p_2) then
3
          l \leftarrow [];
4
          pp_1 \leftarrow \mathsf{positions}(p_1); pp_2 \leftarrow \mathsf{positions}(p_2);
          while pp_1 \neq \text{nil do}
6
             while pp_2 \neq \text{nil do}
 7
                 if pos(pp)
                              pos(pp)
                                         k then
 8
 9
                            s://eduassistpro.github.io/
10
11
                        break;
12
13
                                           oject Exam Help
14
                delete(l[1]);
15
16
             for each ps \in l do
                                          het edu_assist_pro
17
18
19
         p_1
      else
20
              https://eduassistpro.github.io/
21
22
          else
23
                           WeChat edu_assist_pro
25 return answer;
```

Q3. (25 marks)

- Decode the sequence of numbers the compressed list represents.
- List the document IDs in this list.

Q4. (25 marks)

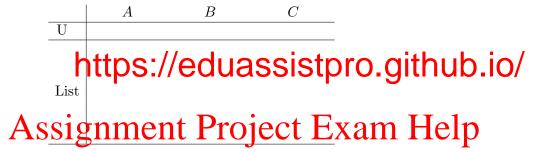
Consider the WAND algorithm described in Section 2.4 in the original paper.¹. There is a typo in the algorithm in Line 21: it should use "terms[0 .. (pTerm-1)]".

¹Efficient Query Evaluation using a Two-Level Retrieval Process.

However, even with this fixed, there is a bug in the algorithm (Figure 2) in which the algorithm will end up in an infinite loop.

You need to

- Identify the **single** lines in Figure 2 that causes the bug and describe concisely why this will lead to a bug.
- Give a simple example illustrating this bug. You should use three terms (named A, B, C) and k = 1. Do not include unnecessary entries in the lists.



You need Sist guranteent the country and sist yoro must

- include your
- the file can be open on the

Note: Collaboration is allowed. However, each person must independently write up his/her own solution.

You can then submitted by the contacts eau_assist_pro to 5MB.

Late Penalty: -10% per day for the first two days, and -20% per day for the following days.