COMP6714 ASSIGNMENT 1

DUE ON 20:59 29 NOV, 2019 (FRI)

Q1. (*25 marks*)

Consider the following pseudo code which performs list intersection based on the divide-and-conquer paradigm. **Note** that the input lists are not necessarily sorted. 参照Lab练习题

Algorithm 1: Intersect(A, B)

```
1 if ... then
      /* Deal with the boundary case                                  */
2     ...;
3     return ...;
4 else
      /* Recursively break down each list into two parts and recurse  */
5     ...;
6     return ...;
```

(1) Complete the above pseudo code. You can assume that you can invoke the following member methods on a `List` object L: 用递归写出合并两个列表的算法
   - L.len returns the length of the list L.

   You can also use the usually indexing and slicing operation on the list (as in `python`).

(2) Think of a method to divide each input list into k sub-lists ($k \geq 2$) without changing the main logic of the algorithm you implemented in the first part. You should be able to describe the only change succinctly. 在上个算法基础上，增加一部 把输入列表拆分为k个子列表，然后再合并的过程

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 $O(t \cdot M \cdot \log_2 t)$.

Q3. (*25 marks*)

The following list of Rs and Ns represents relevant (R) and nonrelevant (N) returned documents in a ranked list of 20 documents retrieved in response to a query from a collection of 10,000 documents. The top of the ranked list is on the left of the list. This list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection.

```
R R N N N     N N N R N     R N N N R     N N N N R
```

(Note that spaces above are just added to make the list easier to read)

(1) What is the precision of the system on the top-20?
(2) What is the F_1 on the top-20?
(3) What is/are the uninterpolated precision(s) of the system at 25% recall?
(4) What is the interpolated precision at 33% recall?
(5) Assume that these 20 documents are the complete result set of the system. What is the MAP for the query?

Assume, now, instead, that the system returned the entire 10,000 documents in a ranked list, and these are the first 20 results returned.

(6) What is the largest possible MAP that this system could have?
(7) What is the smallest possible MAP that this system could have?
(8) In a set of experiments, only the top-20 results are evaluated by hand. The result in (5) is used to approximate the range (6) to (7). For this example, how large (in absolute terms) can the error for the MAP be by calculating (5) instead of (6) and (7) for this query?

Q4. (*25 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

(1) 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?

(2) 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?

Submission Instructions

You need to write your solutions to the questions in a pdf file named `ass1.pdf`. You **must**

- include your **name** and **student ID** in the file, and
- the file can be opened correctly on CSE machines.

You need to show the key steps to get the full mark.

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

You can then submit the file by `give cs6714 ass1 ass1.pdf`. The file size is limited to 5MB.

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