Due: 2023/3/10 12:00

# Homework 3

**Student Number:**
**Name:**

**Problem 1.** (40 points) For n = 15 splits, r = 10 segments, and j = 3 term partitions, how long would distributed index creation take for Reuters-RCV1 in a MapReduce architecture? Base your assumptions about cluster machines on Table below.

| Symbol | Statistic | Value |
|---|---|---|
| $s$ | average seek time | $5ms = 5 \times 10^{-3}s$ |
| $b$ | transfer time per byte | $0.02\mu s = 2 \times 10^{-8}s$ |
| | processor's clock rate | $10^9 s^{-1}$ |
| $p$ | low-level operation(e.g., compare & swap a word) | $0.01\mu s = 10^{-8}s$ |
| | size of main memory | several GB |
| | size of disk space | 1TB or more |

**Problem 2.** (30 points) Here are a few simple examples of interesting programs that can be easily expressed as MapReduce computations. Please fill in the '?' in the following programs.

(a) **Inverted Index:** The map function parses each document, and emits a sequence of (word, ?) pairs. The reduce function accepts all pairs for a given word, sorts the corresponding document IDs and emits a (?, ?) pair. The set of all output pairs forms a simple inverted index. It is easy to augment this computation to keep track of word positions.

(b) **Count of URL Access Frequency:** The map function processes logs of web page requests and outputs (URL, ?). The reduce function adds together all values for the same URL and emits a (?, ?) pair.

(c) **Reverse Web-Link Graph:** The map function outputs (?, ?) pairs for each link to a target URL found in a page named source. The reduce function concatenates the list of all source URLs associated with a given target URL and emits the pair: (?, ?).

You may read original publication on MapReduce: Dean and Ghemawat (2004), to answer this question.

**Problem 3.** (30 points) Estimate the space usage of the Reuters dictionary with blocks of size $k = 8$ and $k = 16$ in blocked dictionary storage.