Due: 2018/3/23

Homework 3

Problem 1. (30 points) Estimate the space usage of the Reuters dictionary with blocks of size $k = 8$ and $k = 16$ in blocked dictionary storage.

Solution:
As K=8 , We will have: $(8 - 1) * 3 = 21$ bytes for term pointer.
We need additional K=8 for term length so space reduced by 13 bytes per 8 term block.
Total Space reduced $= 4\,00\,000 * \frac{13}{8} = 0.65$ MB
Total Space is: $7.6 - 0.65 = 6.95$ MB
As K=16 then,
We will have : $(16 - 1) * 3 = 45$ bytes for term pointer.
Need additional K=16 for term length so space reduced by 29 bytes per 16 term block.

Problem 2. (35 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?? clock rate	$10^9 s^{-1}$
p	lowlevel operation(e.g., compare & swap a word)	$0.01\mu s = 10^{-8}s$
	size of main memory	several GB
	size of disk space	1TBormore

Solution:
For Map-Reduce distributed index creation, Number of splits = 15
Number of machines = 10
Number of partitions = 3
Size of a split Reuters RCV1 to be parsed $= \frac{800}{15}$ MB
MAP Phase: 10 machines process simultaneously
Time spent by a machine $= \frac{800}{15} * 10^6$ bytes $* (10^{-7}(reading) + 10^{-7}(comparisonop.))\ \frac{s}{byte} \approx 10s$
Time to parse entire data $= 10 * 2$ (2 stages of MAP phase are required) = 20 s
<u>REDUCE Phase:</u>

For Reuters-RCV1, Number of postings per inverter = $\frac{100}{3}$ million
For an inverter, Time spent in reading = $\frac{800}{3} * 10^6$ bytes $* 10^{-7} s/bytes \approx 26s$
Time spent in sorting = $(\frac{100}{3} * 10^6) * \mathit{log}\ (\frac{100}{3} * 10^6) * 10^{-7} = 83s$
Size of the index to be written = $(\frac{4*10^5}{3} * 4) + (\frac{100*10^6}{3} * 4) = \frac{4}{3} * 10^8$
Time spent in Writing = $\frac{4}{3} * 10^8$ bytes $* 10^{-7} s/bytes = 13s$
Total Time in Distributed Index Creation = 20+26+83+13 = 162s $\approx$ 3 min

Problem 3. (35 points) Assume that machines in MapReduce have 100 GB of disk space each. Assume further that the postings list of the term the has a size of 200 GB. Then the MapReduce algorithm as described cannot be run to construct the index. How would you modify MapReduce so that it can handle this case?

Solution:
We can partition it by DOC_ID as well as term for very frequent terms