Due: 2019/3/22

Homework 3

Student Number:
Name:

Problem 1. (20 points) Estimate the space usage of the Reuters dictionary with blocks of size $k = 8$ and $k = 16$ in blocked dictionary storage.

Problem 2. (20 points)

a. Write down the entries in the permuterm index dictionary that are generated by the term *conflict*.

b. Consider the query *conf*ct*, what Boolean query on a bigram index would be generated for this query?

c. Can you think of a term that satisfies the Boolean query in question b. but does not match the permuterm query *ct$conf**? What about the reverse case?

Problem 3. (30 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	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

Problem 4. (30 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?