Foundations of Information Retrieval

Exercise 4

Exercise session: 20.11.2014
Questions to: Helge Holzmann, holzmann@L3S.de

1. Index Statistics

Given are the statistics of the Reuters RCV1 collection from the lecture:
(averaged and rounded values)

Documents	800,000
Tokens per document	200
Terms (= word types)	400,000
Bytes per token (incl. spaces/punct.)	6
Bytes per token (without spaces/punct.)	4.5
Bytes per term (= word type)	7.5
Non-positional postings	100,000,000

a) 1. How many tokens belong to one term?
   2. What is the average frequency of a term?
   3. What is the average number of terms per document?

b) Why is the average size of a term larger than the average size of a token?
   (Intuitively, one would expect the other way around)

c) How many positional postings would a positional index for this collection contain?
   We consider a positional posting every (term, docID, pos) tuple.

2. Distributed Indexing

Sketch a *MapReduce* schema for counting document frequencies of terms in a text collection.
Explain the input and output for the *map* as well as the *reduce* step.

How can this be optimized?