

Exercise 4.1

- If we need $T \log_2 T$ comparisons (where T is the number of termID–docID pairs) and two disk seeks for each comparison,
- how much time would index construction for Reuters-RCV1 take if we used disk instead of memory for storage and an unoptimized sorting algorithm (i.e., not an external sorting algorithm)?
- Use the system parameters in Table 4.1. (page 68 IIR).

Exercise 4.4

- For $n = 2$ and $1 \leq T \leq 30$, perform a step-by-step simulation of the algorithm in Figure 4.7. Create a table that shows, for each point in time at which $T = 2 * k$ tokens have been processed ($1 \leq k \leq 15$), which of the three indexes I_0, \dots, I_3 are in use. The first three lines of the table are given below

	I_0	I_1	I_2	I_3
2 tokens	1	0	0	0
4 tokens	0	1	0	0
6 tokens	1	1	0	0

Exercise 4.11

- Apply MapReduce to the problem of counting how often each term occurs in a set of files.
- Specify map and reduce operations for this task. Write down an example along the lines of Figure 4.6.
- Use the same example of files' collection as in 4.6:
 - $d2$: C died.
 - $d1$: C came, C c'ed.