
Midterm (30 pts)

1 Boolean Queries Complexity (2+4+4 pts)

For the queries below, can we run through the intersection in time $O(b + c)$, where b and c are the lengths of the postings lists for **Brutus** and **Caesar** respectively? If not, what best bound can we achieve?

1. Brutus OR Caesar
2. Brutus AND NOT Caesar
3. Brutus OR NOT Caesar

2 True or False with Justification (2*4 pts)

1. In a Boolean retrieval system, stemming never lowers precision.
2. In a Boolean retrieval system, stemming never lowers recall.
3. Stemming increases the size of the vocabulary.
4. Stemming should be invoked at indexing time, but not while processing a query.

3 Estimating Time for Sorting (4+2 pts)

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)? What is the closest time unit that captures the order of magnitude of the time required: microseconds, seconds, minutes, hours, days, or months?

How does this compare with in-memory sort?

Use the system parameters and dataset parameters reproduced below.
($\log_2 10 = 3.3$ and $\log_{10} 2 = 0.3$)

Symbol	Statistic	Value
m	memory transfer time per byte	5 ns
s	average seek time	5 ms
b	disk transfer time per byte	20 ns
p	low-level op.	10 ns
M	Number of tokens	1,000,000
N	Number of terms	400,000

4 Simple Map Reduce Task (6 pts)

Given a directed graph as an adjacency list: $N \rightarrow \text{list}(N)$

```
src1 : dest11, dest12, ...
src2 : dest21, dest22, ...
...
```

determine the list of pairs $(\text{src}, \text{count})$, such that count is the number of (in-)links incident on src , using MapReduce paradigm. E.g., for the following graph, the expected pairs are $(a,1)$, $(b, 1)$, $(c,2)$, $(d,1)$.

```
a : b, c.
b : a, c, d.
```

Explain clearly the *map* task and the *reduce* task, defining these functions in a convenient notation of your choice.