

Information Retrieval

Exercises session n°1: Inverted index and boolean retrieval

Exercise 1: Inverted Index Example

Here is a collection of 8 documents (one line, one document). Build its inverted index usable for a boolean search. Give a representation of this index both with postings lists (cf. lecture n°1, slide “Indexer steps: Dictionary & Postings”) and with an incidence matrix (cf. lecture n°1, slide “Term-document incidence matrices”).

```
Please Please Me
A Day in the Life
A Hard Day's Night
Long, Long, Long
The Long and Winding Road
Love Me Do
Love You To
Please Mr. Postman
```

Exercise 2: Complexity

Suppose that we have a collection of 1 million documents. For the two queries below, can we still run through the intersection in time $O(x + y)$, where x and y are the lengths of the postings lists for Brutus and Caesar? If not, what can we achieve?

```
Query1: Brutus AND NOT Caesar
Query2: Brutus OR NOT Caesar
```

Exercise 3: Boolean Processing Order Optimization

In an inverted index over 0.5 million documents, the following term-frequency statistics were observed:

Term	Document frequency
eyes	213312
kaleidoscope	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812

Recommend a query processing order for the following queries:

```
Query1:      (tangerine OR trees)
              AND (marmalade OR skies)
              AND (kaleidoscope OR eyes)

Query2:      tangerine
              AND (NOT marmalade)
              AND (NOT trees)
```