

Information Retrieval – Exercise 3 IST 441

This exercise is worth 6 points.

In this exercise, you will create a simplified Lucene index. To get partial credit in case of miscalculations, please give detail to your solutions.

Given the following documents:

D1: You say goodbye, I say hello

D2: You say stop, I say go

D3: Hello, hello, you say goodbye

D4: I say yes, you say no

1. (4 points) Build the inverted index for the documents.

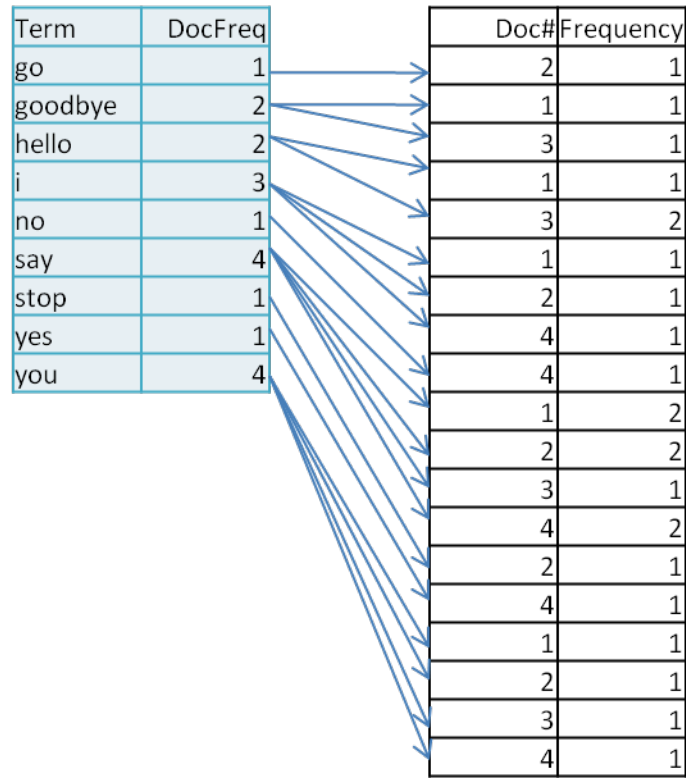
a. Dictionary file

list all the terms:

| Term | Doc# | Term | Doc# |
|-------------|-------------|-------------|-------------|
| you | 1 | go | 2 |
| say | 1 | goodbye | 1 |
| goodbye | 1 | goodbye | 3 |
| i | 1 | hello | 1 |
| say | 1 | hello | 3 |
| hello | 1 | hello | 3 |
| you | 2 | i | 1 |
| say | 2 | I | 2 |
| stop | 2 | i | 4 |
| I | 2 | no | 4 |
| say | 2 | say | 1 |
| go | 2 | say | 1 |
| hello | 3 | say | 2 |
| hello | 3 | say | 2 |
| you | 3 | say | 3 |
| say | 3 | say | 4 |
| goodbye | 3 | say | 4 |
| i | 4 | stop | 2 |
| say | 4 | yes | 4 |
| yes | 4 | you | 1 |
| you | 4 | you | 2 |
| say | 4 | you | 3 |
| no | 4 | you | 4 |

| Term | DocFreq |
|-------------|----------------|
| go | 1 |
| goodbye | 2 |
| hello | 2 |
| i | 3 |
| no | 1 |
| say | 4 |
| stop | 1 |
| yes | 1 |
| you | 4 |

b. Posting file (terms are implicit)



c. Position file (terms are implicit)

| Term | DocFreq | D1 | D2 | D3 | D4 |
|---------|---------|-----|-----|-----|-----|
| go | 1 | 0 | 6 | 0 | 0 |
| goodbye | 2 | 3 | 0 | 5 | 0 |
| hello | 2 | 6 | 0 | 1,2 | 0 |
| i | 3 | 4 | 4 | 0 | 1 |
| no | 1 | 0 | 0 | 0 | 6 |
| say | 4 | 2,5 | 2,5 | 4 | 2,5 |
| stop | 1 | 0 | 3 | 0 | 0 |
| yes | 1 | 0 | 0 | 0 | 3 |
| you | 4 | 1 | 1 | 3 | 4 |

d. For a given query Q: say goodbye, present the process to search the inverted index.

1. Tokenize query Q: term1=say, term2=goodbye
2. Lookup terms in dictionary:
 - a. DocFreq(say) = 4,
 - b. DocFreq(goodbye) = 2.
3. Getting the record from the posting file for each term:

| <i>tf</i> | D1 | D2 | D3 | D4 | |
|-----------|-----------|-----------|-----------|-----------|---|
| say | | 2 | 2 | 1 | 2 |
| goodbye | | 1 | 0 | 1 | 0 |

4. Calculate relevance scores for each document
5. Sort the documents based on the scores.
6. Present the documents.

2. (2 points)

- a. Estimate the total size of the inverted index files in bytes. Numbers are counted for 4 bytes. Strings are counted for the number of characters multiplying 4 bytes. For example, the size of string "hello" is $5 \times 4 = 20$ bytes.

Dictionary file:

9 terms (30 characters) + 9 numbers = 39

Total size: $39 \times 4 = 156$ (bytes)

Posting file:

19 rows, each row has 2 numbers

Total size: $19 \times 2 \times 4 = 152$ (bytes)

Position file:

9 rows and 4 columns = 36 cells

4 cells have 2 number, so

Total size = $40 \times 4 = 160$ (bytes)

If position files are stored with document delta, 0 cells do not occupy any space. So the total size for not counting the 0 cells are: $23 \times 4 = 92$ (bytes)

- b. Size of the documents: $94 \times 4 = 376$ (bytes)
Index size: $156 + 152 + 92 = 400$ (bytes)
The total size is larger than the original documents.

If we ignore the position file, index size = $152 + 156 = 308$ (bytes), the total size is smaller than the original documents.