

TD03 - Information Retrieval and Algorithms

Third Year ISIL - University of M'Sila

Exercise 01: Positional Index and Phrase Queries

Consider the following document collection:

- **Doc1:** “The quick brown fox jumped over the lazy dog.”
- **Doc2:** “The brown dog chased the fox, but the fox was too quick.”

1. Construct the **positional index** for the collection. For each term, indicate the document IDs and positions.
2. Using the positional index, find the documents where the phrase:
 - a) “the fox”
 - b) “quick brown”
3. Find the documents where:
 - a) “fox” appears within 3 words of “dog”
 - b) “quick” appears within 2 words of “brown”

Exercise 02: Phrase and Proximity Queries

Consider the following positional index:

Term	Document IDs and Positions
apple	Doc1: 1, 3, 5 Doc2: 2, 4
banana	Doc1: 2, 4 Doc2: 1, 3, 5
cherry	Doc1: 3 Doc2: 2, 4

1. Find all documents that contain the phrase “apple banana”.
2. Find all documents that contain the pattern:
“banana * cherry”
where the maximum distance between the two terms is 2.

Exercise 03: Phrase Query Evaluation

Consider the following portion of a positional index:

Term	Document IDs and Positions
angels	2: (36,174,252,651); 4: (12,22,102,432); 7: (17);
fools	2: (1,17,74,222); 4: (8,78,108,458); 7: (3,13,23,193);
fear	2: (87,704,722,901); 4: (13,43,113,433); 7: (18,328,528);
in	2: (3,37,76,444,851); 4: (10,20,110,470,500); 7: (5,15,25,195);
rush	2: (2,66,194,321,702); 4: (9,69,149,429,569); 7: (4,14,404);
to	2: (47,86,234,999); 4: (14,24,774,944); 7: (199,319,599,709);
tread	2: (57,94,333); 4: (15,35,155); 7: (20,320);
where	2: (67,124,393,1001); 4: (11,41,101,421,431); 7: (16,36,736);

Determine the documents returned for the following queries:

- a) “fools rush in”
- b) “fools rush in” AND “angels fear to tread”

Exercise 04: Biword Index

Consider the following document collection:

- **Doc1:** “Information retrieval is important”
- **Doc2:** “Retrieval of information is useful”
- **Doc3:** “Search engines perform information retrieval”

1. Construct the **biword index** for the documents.
2. Identify the documents that contain the phrase:
“information retrieval”

Exercise 05: Positional Index and Proximity Search

Consider the following documents:

- **Doc1:** “Data science is an interdisciplinary field”
- **Doc2:** “The field of science is vast”
- **Doc3:** “Data is crucial in science”

1. Construct the **positional index**, storing the position of each term in each document.
2. Find the documents that match the phrase query:
“data science”
3. Find the documents where:
“science” appears within 2 words of “field”