Web Information Retrieval

Assignment 3

Team Name : Gamma

Members

Name	Matriculation Number
Mohammad Nizam Uddin	**216101140**
Md. Shohel Ahamad	**216203438**
MD Jakaria Nawaz	**216203442**
Shreya Chatterjee	**216100848**

1 Boolean Retrieval

Document Collection

Doc 1: “preliminary findings in cancer research”
Doc 2: “novel cancer research findings”
Doc 3: “new research in cancer healing”
Doc 4: “new optimism in cancer patients”

1.1 Index Construction

Term, Document-ID Pairs

Term	Document ID	Term	Document ID
preliminary	1	research	3
findings	1	in	3
in	1	cancer	3
cancer	1	healing	3
research	1	new	4
novel	2	optimism	4
cancer	2	in	4
research	2	cancer	4
findings	2	patients	4
new	3		

Term-Document Matrix

Document Collection

Doc 1: “preliminary findings in cancer research”
Doc 2: “novel cancer research findings”
Doc 3: “new research in cancer healing”
Doc 4: “new optimism in cancer patients”

Term		**D1**	**D2**	**D3**	**D4**
cancer	**T1**	1	1	1	1
findings	**T2**	1	1	0	0
healing	**T3**	0	0	1	0
in	**T4**	1	0	1	1
new	**T5**	0	0	1	1
novel	**T6**	0	1	0	0
optimism	**T7**	0	0	0	1
patients	**T8**	0	0	0	1
preliminary	**T9**	1	0	0	0
research	**T10**	1	1	1	0

Inverted Index

Document Collection

Doc 1: “preliminary findings in cancer research”
Doc 2: “novel cancer research findings”
Doc 3: “new research in cancer healing”
Doc 4: “new optimism in cancer patients”

“cancer” → d1, d2, d3, d4

“findings” → d1, d2

“healing” → d3

“in” → d1, d3, d4

“new” → d3, d4

“novel” → d2

“optimism” → d4

“patients” → d4

“preliminary” → d1

“research” → d1, d2, d3

1.2 Binary Search Tree

How Binary search tree works

Binary search starts from the root of the tree. Every internal node represents a binary test. Based on the result the search continues to one of the two sub-tree below that node. The numbers terms under two sub-trees of any node are either equal or differ by one. The principal issue here is that of rebalancing: as terms are inserted into or deleted from the binary search tree, it needs to be rebalanced so that the balance property is maintained. (D., C, 2008)

Document Collection

Doc 1: “preliminary findings in cancer research”
Doc 2: “novel cancer research findings”
Doc 3: “new research in cancer healing”
Doc 4: “new optimism in cancer patients”

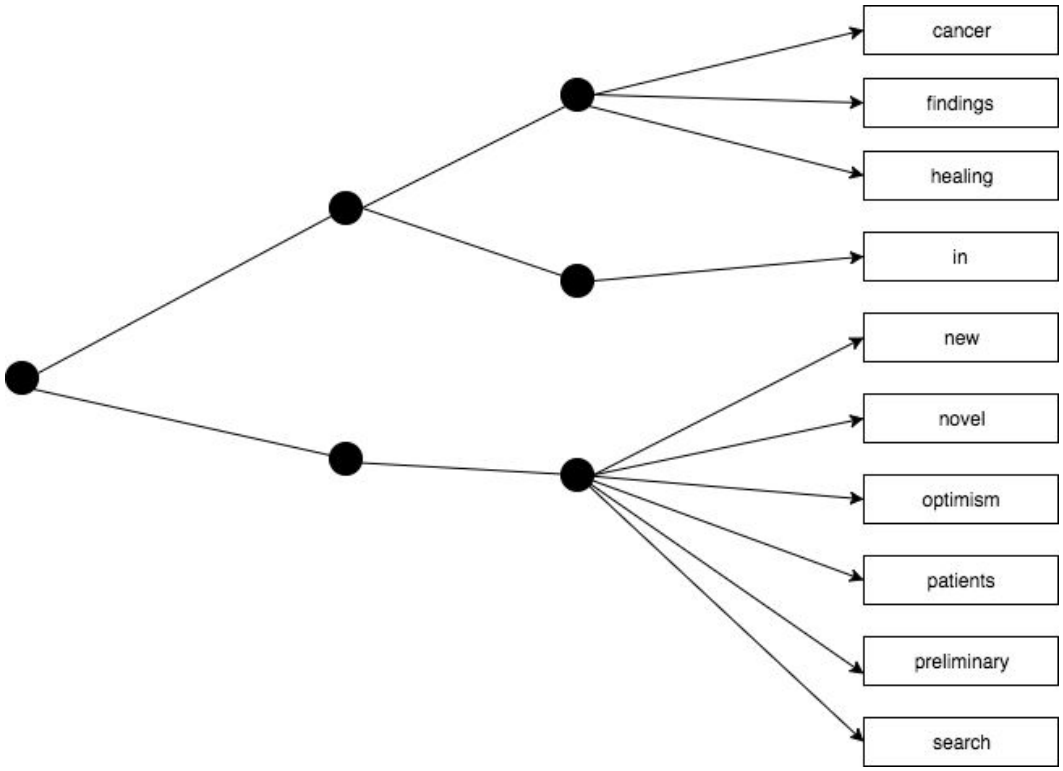

Figure-01: Binary search tree for given corpus.

1.3 Retrieval

Document Collection

Doc 1: “preliminary findings in cancer research”
Doc 2: “novel cancer research findings”
Doc 3: “new research in cancer healing”
Doc 4: “new optimism in cancer patients”

Documents that matches the query **“cancer” AND “research”**

“cancer” → { d1, d2, d3, d4 }
“research” → { d1, d2, d3 }

{ d1, d2, d3, d4 } ∩ { d1, d2, d3 } = { d1, d2, d3 }

- Doc1
- Doc2
- Doc3

Documents that matches the query **“in” AND NOT (“research” OR “healing”)**

“in” → { d1, d3, d4 }
“research” → { d1, d2, d3 }
“healing” → { d3 }

“research” OR “healing”
{ d1, d2, d3 } U { d3 } = { d1, d2, d3 }

“in” AND NOT (“research” OR “healing”)
{ d1, d3, d4 } **AND NOT** { d1, d2, d3 } = { d4 }

- Doc4

2 Preprocessing

Document Collection

Doc 1: “My name is Bond, James Bond.”
Doc 2: “That was probably one of the most well known
'James Bond' quotes, right?”

Tokenization :

Tokenization is the process of chopping up character sequence into pieces, called tokens and at the same time throwing away characters, such as punctuation. (*D., C, 2008)*

Doc1 : [“My” ,“name”, “is” ,“Bond”, “James”, “Bond”]
Doc2 : [“That”, “ was” , “ probably”, “one”, “of”, “the”, “most”, “well”, “known”, “James”, “Bond”, “quotes”, “right”]

Lemmatization :

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. (*D., C, 2008)*

Doc1 : [“My”, “name”, “is”, “Bond”, “James”, “Bond”]
Doc2 : [“That”, “was”, “probable”, “one”, “of”, “the”, “most”, “well”, “known”, “James”, “Bond”, “quotation”, “right”]

Case Folding

The simplest heuristic is to convert to lowercase words at the beginning of a sentence and all words occurring in a title that is all uppercase or in which most or all words are capitalized. These words are usually ordinary words that have been capitalized. Mid-sentence capitalized words are left as capitalized (which is usually correct). (*D., C, 2008)*

Doc1 : [“my”, “name”, “is”, “Bond”, “James”, “Bond”]
Doc2 : [“that”, “was”, “probably”, “one”, “of”, “the”, “most”, “well”, “known”, “James”, “Bond”, “quotes”, “right”]

Stop Word Removal

The general strategy for determining a stop list is to sort the terms by collection frequency (the total number of times each term appears in the document collection), and then to take the most frequent terms, often hand-filtered for their semantic content relative to the domain of the documents being indexed, as a stop list , the members of which are then discarded during indexing. (*D., C, 2008)*

Doc1 : [“My”, “name”, “Bond”, “James”]
Doc2 : [“probably”, “one”, “most”, “well”, “known”, “James”, “Bond”, “quotes”, “right”]

References and Bibliography

D., C, 2008. *Introduction to Information Retrieval*. 1. Cambridge University Press.