Information Retrieval (TDT4117)

Assignment 4

Submitted by

Name	**:**	Md Anwarul Hasan
Student ID	**:**	583233

Task 1 – Text Indexing

Text: Intelligent behavior in people is a product of the mind. But the mind itself is more like what the human brain does.

a) Answer:

Text

1 13 22 25 32 35 37 45 48 52 58 62 66 71 78 81 86 91
Intelligent behavior in people is a product of the mind. But the mind itself is more like what
96 100 106 112
the human brain does.

Vocabulary	Occurrences
behavior	13
brain	106
does	112
human	100
Intelligent	1
itself	71
mind	52, 66
people	25
product	37

b) Answer:

Text

Block 1	Block 2	Block 3	Block 4
Intelligent behavior in people is a	product of the mind. But the	mind itself is more like	What the human brain does.

Vocabulary	Occurrences
behavior	1
brain	4
does	4
human	4
Intelligent	1
itself	3

mind	2
people	1
product	2

I divided the total text in 4 blocks as the word count is 22 which is approximately closely divisible by 6 and approximately the answer is 4. That is why I divided the text in 4 blocks.

c) Answer:

Text

1 13 22 25 32 35 37 45 48 52 58 62 66 71 78 81 86
Intelligent behavior in **people** is a **product** of the **mind**. But the **mind itself** is more like
91 96 100 106 112
what the **human brain does**.

String 1: Intelligent behavior in people is a product of the mind. But the mind itself is more like what the human brain does.

String 13: behavior in people is a product of the mind. But the mind itself is more like what the human brain does.

String 25: people is a product of the mind. But the mind itself is more like what the human brain does.

String 37: product of the mind. But the mind itself is more like what the human brain does.

String 52: mind. But the mind itself is more like what the human brain does.

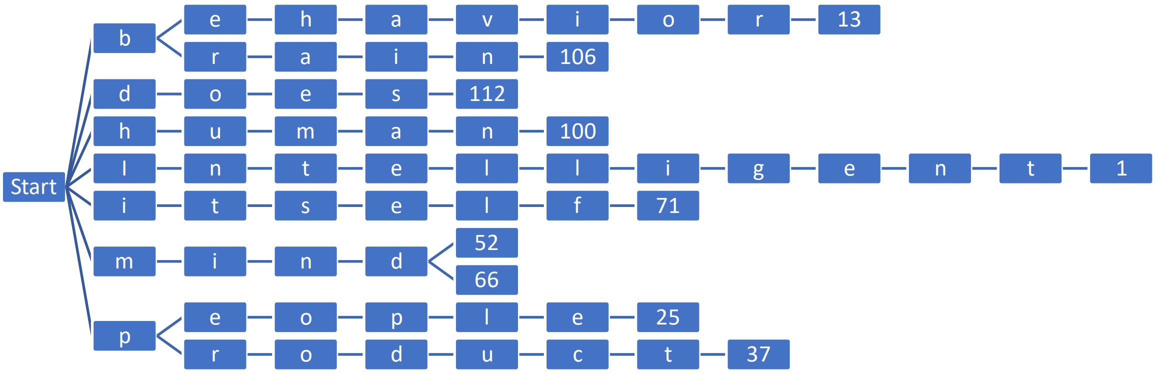
String 66: mind itself is more like what the human brain does.

String 71: itself is more like what the human brain does.

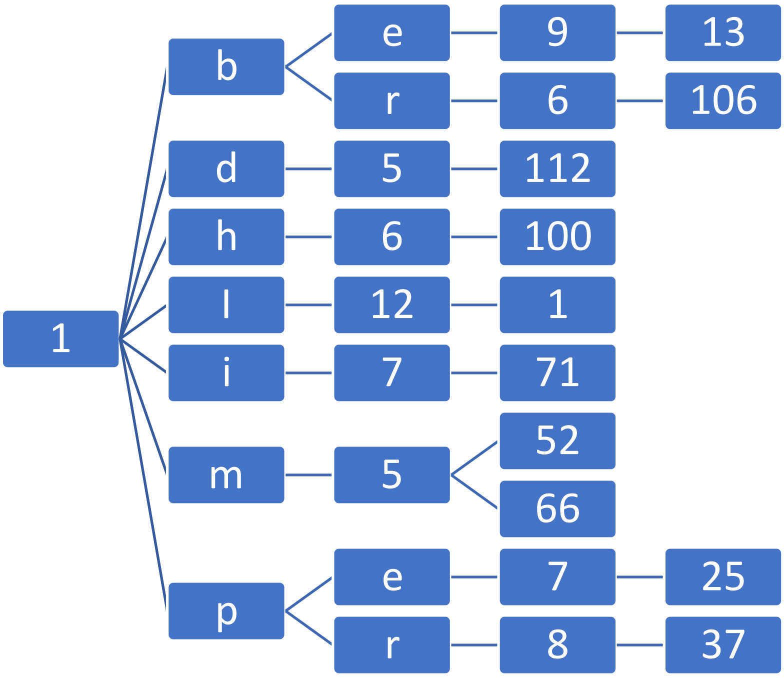
String 100: human brain does.

String 106: brain does.

String 112: does.



Suffix trie



Suffix tree

d) Answer:

Documents are:

D1: Although we know much more about the human brain than we did even

D2: ten years ago, the thinking it engages in remains pretty much a total

D3: mystery. It is like a big jigsaw puzzle where we can see many of the

D4: pieces, but cannot yet put them together. There is so much about us

D5: that we do not understand at all.

A simple inverted index by using a posting list is given below,

Vocabulary	Occurrences as inverted lists
ago	[2:1]
brain	[1:1]
engages	[2:1]
human	[1:1]
jigsaw	[3:1]
know	[1:1]
like	[3:1]
many	[3:1]
mystery	[3:1]
pieces	[4:1]
pretty	[2:1]
put	[4:1]
puzzle	[3:1]
remains	[2:1]
see	[3:1]
ten	[2:1]
them	[4:1]
thinking	[2:1]
together	[4:1]
total	[2:1]
understand	[5:1]
us	[4:1]
we	[1:2][3:1][5:1]
years	[2:1]

We eliminated all the stop words, because it reduced size of our indexes. Besides indexes we also eliminated the punctuations.

Task 2 – Index Analysis Using Lucene

a) Answer:

Elastic Stack (ELK): Is an open-source tool to analyse data. It has the capacity to aggregate data from different data source. It can also store data to a common location and provide the power of analyse to the user.

Lucene: Is a search engine library in java. It is suitable for almost application that require search feature in it. Lucene provides easy implementation of search capability on the existing software and make the work of the implementor easy who require a search capacity in the application.