Task 1

Note: In this assignment I have defined stopwords to be the words provided as stopwords in assignment 3 (common-english-words-txt).

A)

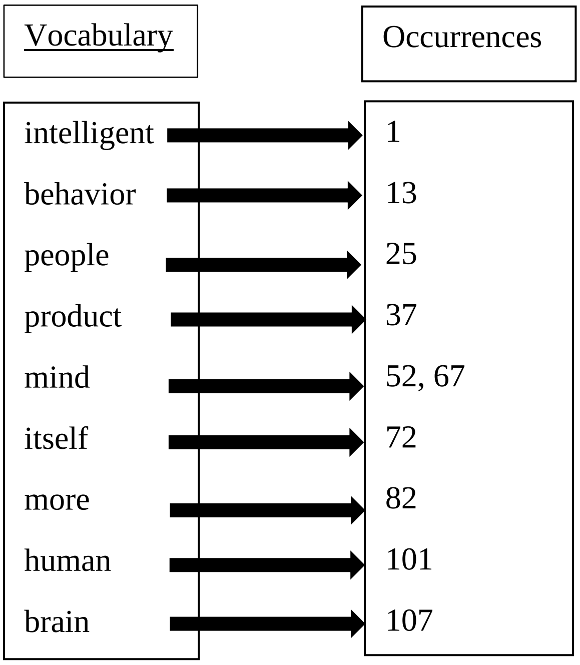
Text:

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

Non-stopwords of the text:

Intelligent, behavior, people, product, mind, itself, more, human, brain.

Inverted file:



Vocabulary	Occurrences
intelligent	1
behavior	13
people	25
product	37
mind	52, 67
itself	72
more	82
human	101
brain	107

B)

Intelligent behavior in people is	a product of the mind.	But the mind itself

is more like what	the human brain does.

Vocabulary	Occurrences
intelligent	1
behavior	1
people	1
product	2
mind	2, 3
itself	3
more	4
human	5
brain	5

C)

Start by sorting the words (not stop-words but including punctuation), alphabetically:

behavior

brain

human

intelligent

itself

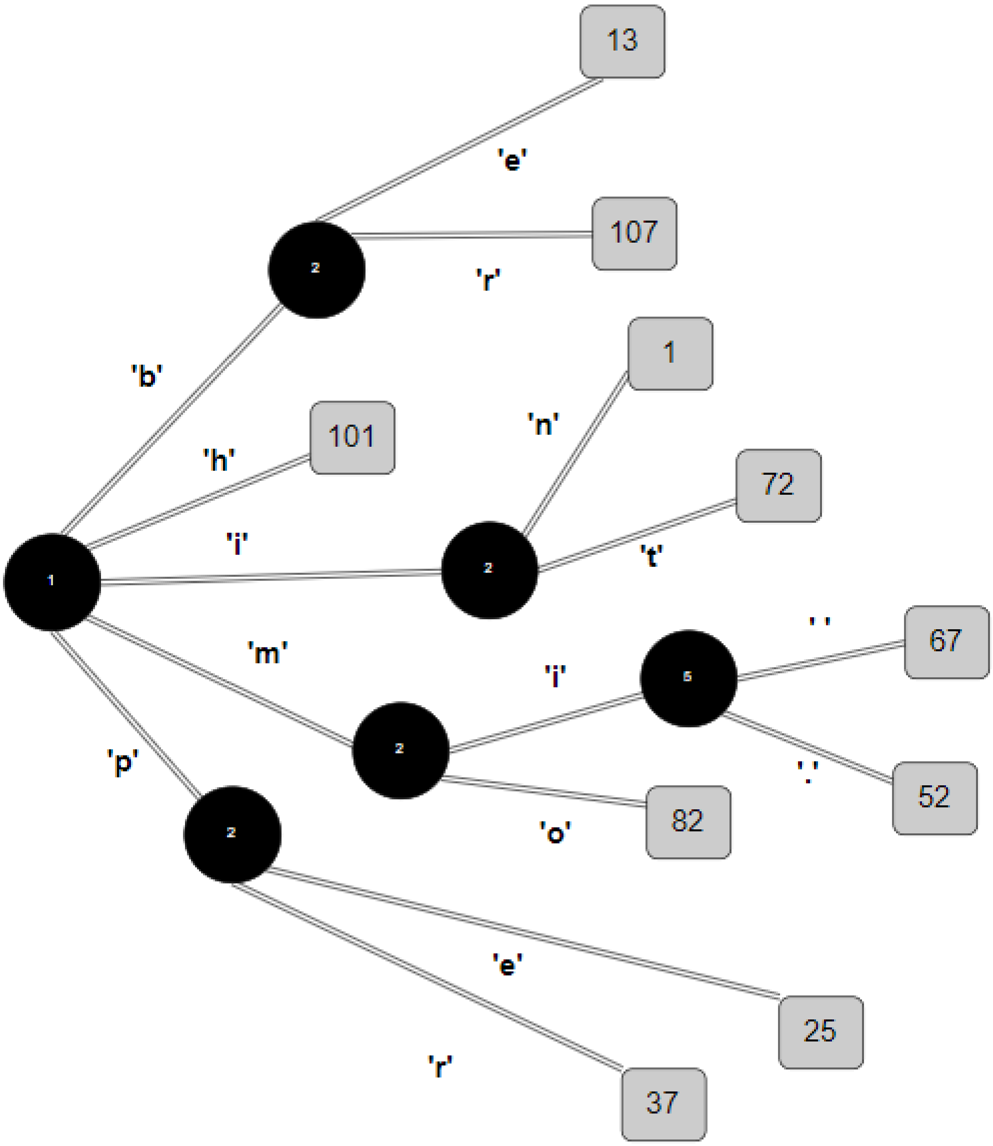
mind

mind.

more

people

product



d)

Wordlists without stopwords:

D1:

although, know, much, more, human, brain, even

D2:

ten, years, ago, thinking, engages, remains, pretty, much, total

D3:

mystery, big, jigsaw, puzzle, see, many

D4:

pieces, put, together, much

D5:

understand

List of all unique words across documents (no stopwords), in alphabetical order:

ago, although, big, brain, engages, even, human, jigsaw, know, many, more, much, mystery, pieces, pretty, put, puzzle, remains, see, ten, thinking, together, total, understand, years

ago – 2:1

although – 1:1

big – 3:1

brain – 1:1

engages – 2:1

even – 1:1

human – 1:1

jigsaw – 3:1

know – 1:1

many – 3:1

more – 1:1

much – 1:1, 2:1, 4:1

mystery – 3:1

pieces – 4:1

pretty – 2:1

put – 4:1

puzzle – 3:1

remains – 2:1

see – 3:1

ten – 2:1

thinking – 2:1

together – 4:1

total – 2:1

understand – 5:1

years – 2:1

Task 2

A)

The ELK stack consists of Elasticsearch, Logstash and Kibana. Elasticsearch is a search and analytics engine, Logstash is a server-side data processing pipeline, and Kibana is a tool for visualizing data with charts and graphs.

Lucene is a full-text search engine. It is capable of many things, including providing ranked search, fielded searching, simultaneous update and searching. Elasticsearch is built on Lucene.

B)

For a given list of document, I want to create a vocabulary of all words in the documents, except for stopwords. This vocabulary should map to a list of which documents contain the word.

E)

In my simple implementation, “claims of duty” only return document 6. However, the ELK stack returns all documents except for document 3.

Yes, for the search claim*, both my implementation and the ELK stack return documents 2 and 6.

I would argue that returning results for claim* is more straightforward than for the other searches, as it is a precise regex expression. Therefore I would expect this.

F)

I was not able to download the gzip file with the emails, as there was a timeout error.