IT458: Information Retrieval
Lab Assignment 1
Inverted Index

Name: Chinmayi C. Ramakrishna
Roll No.: 181IT113

A corpus of 100 news on various categories is taken like health, finance, politics etc.

Part 1: It is observed that Vocabulary Sizes after various preprocessing steps:

Initial vocabulary size: 251633
After tokenization: 263647
After removing stop words: 127227
After Stemming: 127227

Part 2: Different data structures that can be used for Inverted Index:

Data Structure Type	Description	Insertion	Updation
Sorted Array	Contains a list of keywords classified by language, as well as a link to each keyword's corresponding document.	Sorting the array after insertion can take O(nlogn)	Sorting the array after updation can take O(nlogn)
B Trees	B-trees use more space. Updates are much easier and the search time is generally faster especially if secondary storage is used.	Can be done in O(logn).	Can be done in O(logn).
Tries	For this structure, what is represented are the keywords after digital decomposition.	Happens in O(n) where n is index length	Happens in O(n) where n is index length
Linked Lists	It used pointers to point to the next index word.	In an unsorted fashion insertion can be done in O(1).	Happens in O(n) for unsorted linked lists.

Optimal Solution

A document vector file can be taken containing the concept vectors for each document in the collection to be indexed.

A dictionary stores all the unique index terms after three preprocessing steps.

The three preprocessing steps are:

1. Replacing punctuation marks with white spaces.
2. Removal of stopwords.
3. Stemming the index terms.

For every word in the corpus after preprocessing, the dictionary stores two data: the first index contains the frequency of the word and then a list of documents containing the index term.

This gives a time complexity of O(d*n) where d represents the number of documents in the corpus and n is the number of tokens after preprocessing steps.

This approach would take O(d*n).

Part 3: The time taken for boolean querying:

```
Vocabulary size before preprocessing:  251633
263647
127227
Vocabulary Sizes after various preprocessing steps
Initial vocabulary size:  251633
Vocabulary after tokenization:  263647
Vocabulary after removing stop words:  127227
Vocabulary after Stemming 127227
Number of final index terms:  127227
```

```
Query: bright OR health AND care
bright [41, {66, 69, 37, 167, 138, 170, 12, 113, 84, 185}]
health [12097, {1, 129, 4, 134, 8, 140, 141, 14, 142, 143, 17, 145, 20, 148, 22, 152, 25, 26, 153, 28, 158, 159, 160, 33, 161, 39, 40, 41, 42, 44, 174, 47, 177, 5
1, 52, 57, 58, 59, 60, 193, 195, 200, 73, 76, 92, 94, 99, 102, 105, 109, 115, 118, 121, 123, 126, 127}]
care [19553, {1, 2, 3, 4, 8, 9, 10, 15, 21, 22, 24, 25, 26, 27, 28, 32, 33, 34, 38, 39, 40, 41, 42, 44, 49, 50, 51, 52, 57, 58, 59, 60, 62, 65, 68, 72, 73, 84, 87
, 88, 89, 90, 92, 93, 94, 95, 96, 102, 103, 104, 105, 109, 110, 111, 116, 122, 123, 125, 126, 127, 128, 129, 133, 134, 135, 139, 140, 141, 142, 143, 145, 150, 151
, 152, 153, 158, 159, 160, 161, 163, 166, 169, 173, 174, 185, 188, 189, 190, 191, 193, 194, 195, 196, 197}]
Documents satisfying the boolean query: {1, 2, 3, 4, 8, 9, 10, 12, 14, 15, 17, 20, 21, 22, 24, 25, 26, 27, 28, 32, 33, 34, 37, 38, 39, 40, 41, 42, 44, 47, 49, 50,
 51, 52, 57, 58, 59, 60, 62, 65, 66, 68, 69, 72, 73, 76, 84, 87, 88, 89, 90, 92, 93, 94, 95, 96, 99, 102, 103, 104, 105, 109, 110, 111, 113, 115, 116, 118, 121, 1
22, 123, 125, 126, 127, 128, 129, 133, 134, 135, 138, 139, 140, 141, 142, 143, 145, 148, 150, 151, 152, 153, 158, 159, 160, 161, 163, 166, 167, 169, 170, 173, 174
, 177, 185, 188, 189, 190, 191, 193, 194, 195, 196, 197, 200}
```

```
Query: music OR travel
music [2029, {3, 4, 9, 138, 139, 22, 24, 157, 37, 38, 169, 171, 177, 56, 184, 68, 196, 70, 76, 83, 95, 104, 105, 110, 123, 125}]
travel [1985, {3, 133, 15, 19, 156, 32, 163, 165, 168, 170, 172, 173, 180, 55, 183, 62, 190, 64, 191, 67, 69, 71, 72, 199, 79, 82, 89, 90, 98, 104, 116, 120}]
Documents satisfying the boolean query: {3, 4, 133, 9, 138, 139, 15, 19, 22, 24, 156, 157, 32, 163, 37, 38, 165, 168, 169, 170, 171, 172, 173, 177, 180, 55, 184,
56, 183, 62, 190, 64, 191, 67, 68, 196, 70, 69, 71, 72, 199, 76, 79, 82, 83, 89, 90, 95, 98, 104, 105, 110, 116, 120, 123, 125}


Query: music AND travel
music [2029, {3, 4, 9, 138, 139, 22, 24, 157, 37, 38, 169, 171, 177, 56, 184, 68, 196, 70, 76, 83, 95, 104, 105, 110, 123, 125}]
travel [1985, {3, 133, 15, 19, 156, 32, 163, 165, 168, 170, 172, 173, 180, 55, 183, 62, 190, 64, 191, 67, 69, 71, 72, 199, 79, 82, 89, 90, 98, 104, 116, 120}]
Documents satisfying the boolean query: {3, 4, 133, 9, 138, 139, 15, 19, 22, 24, 156, 157, 32, 163, 37, 38, 165, 168, 169, 170, 171, 172, 173, 177, 180, 55, 184,
56, 183, 62, 190, 64, 191, 67, 68, 196, 70, 69, 71, 72, 199, 76, 79, 82, 83, 89, 90, 95, 98, 104, 105, 110, 116, 120, 123, 125}


Total Time: 198.85122156143188 seconds
```

Total time: 198.85 seconds

Time complexity of boolean querying: O(n) where n is the number of words in the query.