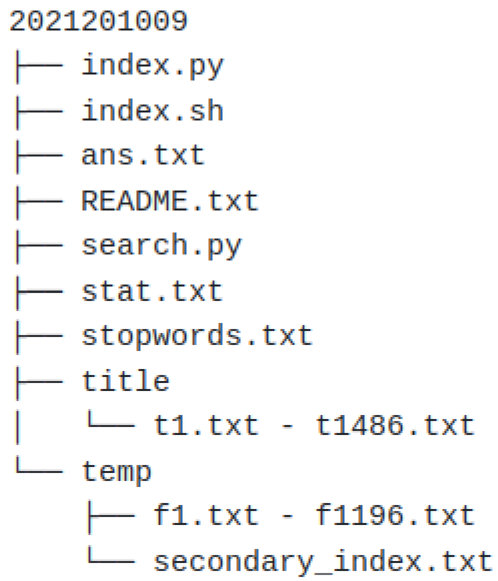
IRE – MINI PROJECT REPORT

Directory Structure



```
2021201009
├── index.py
├── index.sh
├── ans.txt
├── README.txt
├── search.py
├── stat.txt
├── stopwords.txt
├── title
│   └── t1.txt - t1486.txt
└── temp
    ├── f1.txt - f1196.txt
    └── secondary_index.txt
```

Optimizations Used

1. I have used Secondary indexing to ensure that searching becomes fast as the linear searching was taking too much time
2. All the index files are in sorted order in such a way that if all the files are merged then they will also be in sorted order. It was done to ensure that we could use binary search for searching words in the file

Improvements in terms of time and space from those optimizations

The time complexity of searching has been reduced to O (log n) from O(n) resulting in faster searching

Total Index Size – 17.8 GB

Index Creation Time – 7 hr.

Index Format

Possible category name = {‘t’,’i’,’b’,’c’,’r’,’l’}

Key Value = The word present in the dump

<Key_Value> d<Document_Number>< category _name><count of the category> <category _name><count of the category>………..#d<Document_Number><category _name><count of the category> <category_name><count of the category>#…..............................#d<Document_Number><category _name><count of the category> <category_name><count of the category>#

For example,

deleni d270864t1#d298488t1b1#d386304b1#d394237i1#d402290t1b2i2#d402294i1#d425506b2#