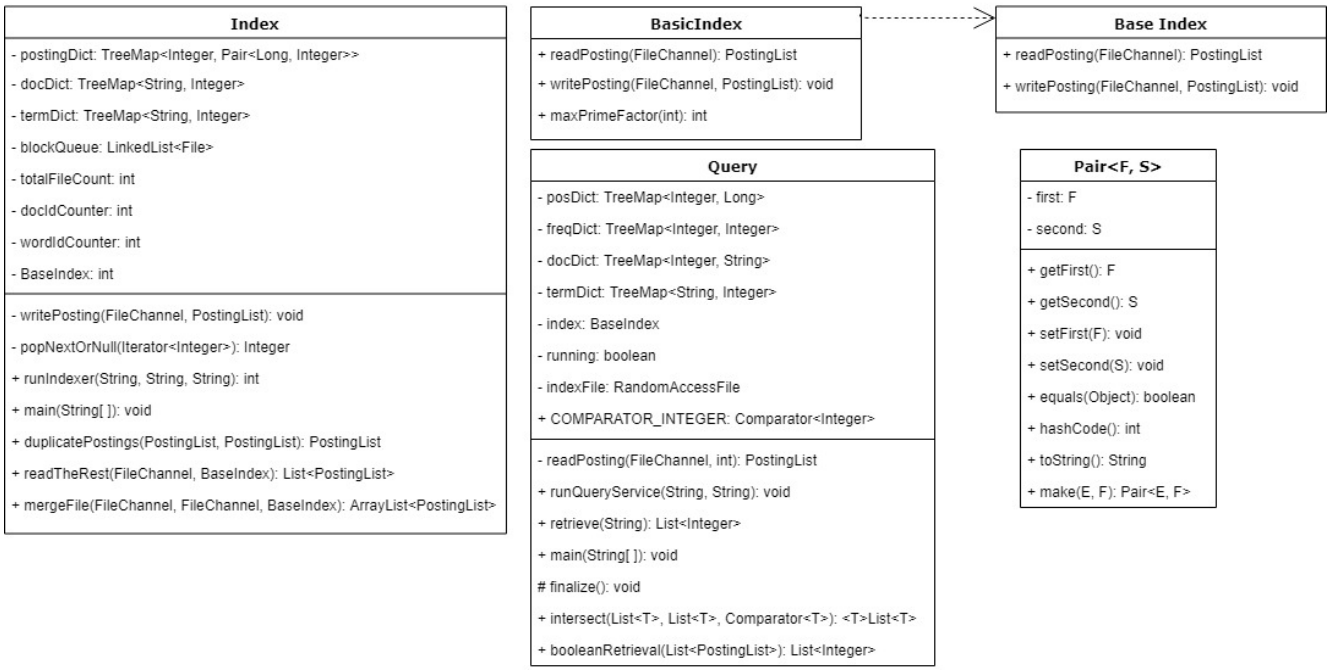
Thanakorn Pasangthien 6088109 Sec 1
Tanaporn Rojanaridpiched 6088146 Sec 3
Nontapat Pintira 6088118 Sec 1

Project 1 Report
YouGle: Your First Search Engine

1. A brief description of how your program is structured, and how the key steps in your indexing and retrieval algorithms work. Make sure you report statistics on the size of the index and statistics of retrieval time for the development queries.




```
"C:\Program Files\JetBrains\IntelliJ IDEA
Indexing Test Result: ./index/small:
    Total Files Indexed: 6
    Memory Used: 2.625008 MBs
    Time Used: 0.25 secs
    Index Size: 2.17437744140625E-4 MBs
    Alright. Good Bye.
```

```
Query Test Result: [hello, bye, you, how are you, how are you ?]:
        Memory Used: 2.692672 MBs
        Time Used: 0.065 secs
        No problem. Have a good day.
```

```
"C:\Program Files\Java\jdk-12.0.1\bin\java.e
Indexing Test Result: ./index/large:
    Total Files Indexed: 98998
    Memory Used: 176.819744 MBs
    Time Used: 158.878 secs
    Index Size: 55.37303924560547 MBs
    Alright. Good Bye.
```

```
Query Test Result: [we are, stanford class, stanford students, very cool, the, a, the the,
stanford computer science]:
        Memory Used: -170.12112 MBs
        Time Used: 0.652 secs
        No problem. Have a good day.
```

```
"C:\Program Files\Java\jdk-12.0.1\bin\java.exe"
Indexing Test Result: ./index/citeseer:
    Total Files Indexed: 18824
    Memory Used: 393.135648 MBs
    Time Used: 215.389 secs
    Index Size: 64.60990905761719 MBs
    Alright. Good Bye.
```

```
Query Test Result: [shortest path algorithm, support vector machine, random forest,
convolutional neural networks, jesus, mahidol, chulalongkorn, thailand, polar bears
penguins tigers, algorithm search engine, innovative product design social media,
suppawong, tuarob, suppawong tuarob, suppawong tuarob conrad tucker]:
        Memory Used: 78.321504 MBs
        Time Used: 0.182 secs
        No problem. Have a good day.
```

Thanakorn Pasangthien 6088109 Sec 1
Tanaporn Rojanaridpiched 6088146 Sec 3
Nontapat Pintira 6088118 Sec 1

2. How the key steps in your indexing and retrieval algorithms work.

Question A

1. **The trade-off of different sizes of blocks -** The larger block size is required a large memory to store, but it comes with the perk of lower write overhead. However, the smaller block size generally creates more blocks to be merged which creates more overhead in the disk. Therefore, we should balance time and block size to manage file reading and writing.
2. **Minimize indexing time by balancing memory usage -** Writing and reading into the disk required such an expensive task in terms of time. In order to minimize time and memory usage, the choice to find the right buffer size is crucial. Initially, we tried to allocate 4 bytes for buffer. However, in the case of a larger index file. we have discovered that this process is not efficient enough for a very large data set. After several attempts, we found the solution in which we could find the maximum common prime number from the size of the posting as the buffer size. This lower indexing time and reduce disk IO overhead.

Question B

1. **Large index files** - The large index file can get very large when working with a large data set which can cause the fragmentation on the disks and reduce read and write time.
2. **Slow merging algorithm -** Merging process is the most time consuming due to the fact that it iterates through a file and comparing posting list one by one. The algorithm also allows redundant because of the way the pointer moves to where it was before reading. Both of these are the overhead that we need to consider as our limitation of scalability.
3. **Sequential block indexing -** Iterating through in a sequential manner all of the blocks which could lead to linear time complexity. As a result, the large dataset will cause a limitation of the system.

Question C

1. **Introduce parallelism into the project -** We can introduce parallel computation into the project which will help us greatly because we don't have to process the block sequentially anymore. We can distribute each block can processing them concurrently which is the perk of parallel computing.
2. **Improve on merging algorithm -** Merging algorithm is the one process that takes the longest time. Therefore, if we can improve the merging algorithm, it would result in a greatly improved in performance.

Answer to question 2