Wikipedia Search Engine - Phase II

1. Directory Structure

1.1. Code Files -

1.1.1. **Search.py** - Main file containing all the code for **Query Processing**
1.1.2. **Driver.py** - Main file which runs the code for **Indexing**
1.1.3. **Preprocess.py** - File containing all functions related to XML parsing and text preprocessing.
1.1.4. **MultiwayMerge.py** - File with functions related to k-way mergesort algorithm.
1.1.5. **MultiLevelIndexing.py** - File containing functions related to making offset files and secondary index.
1.1.6. **Indexing_defaultdict.py** - File which performs the actual indexing
1.1.7. **TermHandling.py** - File with functions which split the term-term_id map into small files and sorts and performs external merge on these files. Also makes a secondary index to access this map.

1.2. Required Directories -

1.2.1. **Index** - Initial index gets created here
1.2.2. **IndexMerge** - They get merged here
1.2.3. **PrimaryIndex** - Merged file is split into smaller files here
1.2.4. **PrimaryOffset** - Offset files for these merged files are made here
1.2.5. **SecondaryIndex** - Secondary index file for these offset files are made here
1.2.6. **PageTitleMap** - Files containing page_id-title map are made here
1.2.7. **TermIdMap** - Small files of the term-term_id map are made here
1.2.8. **TermIdMerge** - These small files are merged here and split into many files
1.2.9. **TermIdMapSecondary** - Secondary index for these files are present here.

2. Execution of Code

2.1. Prerequisits -

2.1.1. All the directories mentioned above

2.1.2. A csv file containing all the stop words in the current directory of the code

2.2. For Query -

2.2.1. Run Search.py - An infinite loop runs expecting queries.

2.2.2. Types of Queries -

2.2.2.1. **Field query** - Assuming that fields are small letters(**b, i, c, t, r, e**) followed by colon and the fields are space separated.

“b:sachin i:2003 c:sports”

2.2.2.2. **Boolean query** - Assuming that the boolean operators are given in capitals (**AND, OR, NOT**) and remaining words are space separated.

“Sachin AND Dhoni NOT Kohli”

2.2.2.3. **Normal query** - Any sequence of words that doesn’t satisfy the above conditions is considered a normal query.

“Sachin Tendulkar”