Group members: Duaa Omer : 331320 Kamlesh Kumar : 348395 Mastora Alizada : 359696 Abdullah Fazeel : 337188

Description : This is a simple search engine that can be used to index a dataset containg json files to obtain relevant search results when a query is entered by a user. It sotres a lexicon, inverted index in form of small barrels, comp file and some reference files and uses them to obtain search results response to a query.

\*\* comp\_file contains the record of files which are already been read by the program, last doc\_id and last lexicon index. (These things are being used when a new file is inserted in between already read file)

Lexicon , Doc\_ref, comp file and Inverted\_index: All of these are saved in json files. Lexicon contain all the unique words from all files being read, with each word assigned a specific word id. Doc\_ref stores the URLs of the files against their doc\_ids. Forward index is created as a dictionary but not stored in a file. Forward index contains the word id of each word present in an article, along with the lists of word positions against every word id. And in Inverted\_index against every word it stores the article id in which it occurs. Inverted\_index will also contain the lists of positions of words in a particular article as well. Inverted index is stored in the form of barrels, each named as the highest word\_id stored in it. (So that when searching for a word literal only that particular barrel is opened)

**NOTE :** Before that we were storing the complete Inverted Index in one file. And it was working very well for 100k. But when we tried for 160k articles there were an memory error. There-fore to make it scalable we did a trade-off between speed and scalability. After making the barrels we suffered a little slow down in speed but achieved scalability.

Program hierarchy: To beautify the project directory we have created hierarchy using folders and partitions. There is folder with name “sample\_data” to store the json files, within the same directory as of main code. After the execution of main file a new folder with name “Generated\_files” will be created in directory containg main code. if not exist already, this folder will store the files for storing Lexicon , Doc\_ref, comp\_file and the barrels of inverted index. The Generated\_files folder also contains a reference document that stores the names of already indexed files and some other relevant information. For inserting a new file there is button in GUI which opens a dialogue box and allow user to selected the json file from anywhere in the disk. A single new barrel is created for storing the data of all next incoming file.