DCR-B Project Part 1

Summary

‘list.txt’ contains the information about files and directories in my main directory, called DCR-B. To create the table on MySql, I extracted all the information I needed from the file ‘list.txt’ with a hand-written function which, for every relevant row of the file, extracts a tuple that will be a row of the final table. Every tuple is characterized by these attributes:

* <FILE> or <DIR>
* Last\_modified
* Size (in bytes)
* File name
* File path
* Text corpus

The corpus of the file has been cleaned using the library ‘BeautifulSoup’, in order to eliminate the non-relevant tags of the HTML file for the search (<href>,<span>…).

The tuples are then inserted into the MySql database, 5 tuples per transaction.

Once the database has been built, I created two indexes to speed up the search:

* B-TREE on the Filename
* Fulltext index on the corpus

The code ‘query.py’ is the actual search engine, which takes in input a string from the user and provides a response if:

* The string is found in a name of a <FILE> or <DIR>
* The string is found in the corpus

Responses have this shape:

TYPE Last\_modified Size\_bytes Filename Path #of occurrences

The # of occurrences is computed by replacing the string in the corpus with an empty string, subtracting the length of the new corpus to the length of the original one, then dividing by the number of characters of the searched string.