Alex Ingrando & Michael Young

User Manual

Adding/Removing Files:

All StackOverflow tag and question files from 2008 to 2016 are required to be placed in the folder of the build directory. In the Document Parser .cpp, there is a section that lists each document that the program will run through. In order to add a new file to this, find the line above ‘this->argc = (a number)’ and write:

this->argv[x] = “<path of tag file>”;

this->argv[x+1] = “<path of questions file>”;

Where ‘x’ is the next number after the previous line. If the previous line says:

this->argv[11] = "2012-questions.psv";

Then x should equal ‘12’. It is important that two files are entered at the same time. The first file entered into the program will be the tag file. The file entered after that will be the questions file that corresponds to those tags. This file should have a number in the argv[x] that is one number greater than the x entered into the tag file. Once both files have been added, the line underneath it ‘this->argc = (a number)’ must be updated as well. The number argc should equal to should be one higher than the previous line, just like how x was calculated earlier.

In order to remove files, delete the lines with the questions and its corresponding tag file from the list. Any lines underneath the deleted lines should have their numbers updated according to how x was calculated above. Then argc should be updated once more. The first two lines:

this->argv[0] = "output.txt";

this->argv[1] = "stopwords.psv";

Should never be deleted and argc should always equal a number above 1.

The Start Menu:

The main menu prompts the user to choose between interactive or maintenance mode. In maintenance mode, the user is able to delete the index if one is stored, or to add a tag-file and questions-file to the existing index. The interactive will first provide a prompt that allows either an AVL Tree or a Hash Table to be used to store the information within the files. If the index has been stored in a file, the index type will be specified at the top of the menu. The Hash Table will run through the files faster, but the AVL Tree is still a valid option too. If '1' is pressed, then the AVL Tree will be selected to store the data. If '2' is pressed, then the Hash Table will be selected to store the data. If '0' is pressed, the user will be rerouted to the main menu. If any other button is pressed, it will prompt the same display over again until one of the options are chosen.

If the program has not been run before, it may take some time to parse through all the documents (this will go faster if the Hash Table was chosen).

If the program has been run before, it should have made a persistent index. This file is an output of the information gathered from the previous runtime. The program will read through this file first and add the data back into itself before running through any additional files that may have been added since it's last runtime. This should take a much shorter time than if it had not been run before.

Searching:

Once the program has finished collected the data from the files, the menu will prompt for a search query. When a word has been entered into here, the program will return a list of up to the top 15 different questions that the word has appeared in. The list will be ordered by relevance with the most relevant search at number 1 and the least relevant search at number 15.

The search query will also accept three commands alter the results. They are: AND, OR, and NOT. Each command should be added directly behind the words that the command should apply to, followed by a space. The AND command searches the files for questions where all of the following words appeared in at once. The AND command should be the first word, and it will apply to all words after it, unless the program encounters a NOT. The OR command searches the files for questions where all of the following words appeared in, whether they shared a question or not. Like the AND, OR should be the first word, and will apply to all words after it unless a NOT is found. The NOT command searches the files for questions where all the preceding word(s) appeared in, but removes all the questions that the following word also appeared in. No matter which command is given, the menu will display the top 15 results.

The menu will then prompt for a numerical value corresponding to the question that is desired to be opened. When a number has been pressed, the question that corresponds to that number will be displayed again, this time with the body of the document underneath it. Pressing 1 will take the user back to the interactive menu. Pressing any other key will return the user to the search results. Again, pressing 0 will take the user back to the interactive mode. The menu will then prompt for another search query to be entered.

Exiting

To exit, reroute to the main menu. Once at the main menu, press 0 to exit the search engine. If an index is currently stored, it will now be written to the output.txt file before the program ends.