This file includes information on the submitted files as well as instructions how how to use our bot and CSVs.

Included Files-

Note: All files are based out of the PersonalizedSearchBot directory

- ./build- A build directory that the java compiler places the class files
- ./CS 7270 Final Results.pdf- A file containing our project's final writeup
- ./CS 7270 README.pdf- This file with instructions and tips
- ./compile- A script that will compile the bot on Linux. Windows users will have to change the paths within the script to use the Windows style to make it run correctly.
- ./data- The raw CSV files where our data was stored.
- ./data/Analysis.xlsx- A CSV file with data summaries and graphs for our final paper
- ./compileAndRun- A script that both compiles the bot and runs it
- ./libs- A directory with the libraries needed for our project. This is where the Selenium code is located
- ./run- A script that will run the bot on Linux. Like the compile script, this will need to be altered to work on Windows.
- ./src/cs7270/personalizedSearch/FixOldCsvs.java- A short program used to update old CSV files into the ones currently in use by the project. This isn't needed anymore, but we included it just to help see how we progressed.
- ./src/cs7270/personalizedSearch/PersonalizedSearchBot.java- The source code for our Selenium Bot.
- ./whoIsWhich.txt- A file showing which files go with which user.

Prerequisites-

There are two things that may need to happen before running our bot. First, Firefox must be installed on your computer since the bot uses the FirefoxDriver. The normal installation should be enough for this, but if its not you may need to make sure that Selenium can find it somewhere in your bin folder. Second, our compile and run scripts assume the user is running on Linux. If running on Windows, those files may need to be updated to use the Windows path style so the libraries and sources can be found correctly.

Running-

Steps 1, 2, 3 and 5 require user input. The rest of the steps are automated

1. Run the "run" script in the PersonalizedSearchBot directory.
2. When prompted, enter the search you plan on running and click OK.
3. When prompted, click yes for personalized search or no for standard search.
4. Wait for Firefox to launch and access Google
5. If personalized, enter your login information
    a. Wait for the login page to load
    b. Enter user name and password
    c. Click Sign In

    d. If using two factor authentication, enter that information and click Ok
    e. Wait for Google's home page to load
6. Google will enter the search and begin the search.
7. Google will go through the first five pages of results and record them.
8. The bot will write the results to the CSV.
9. The bot will close Firefox and end the program.

Reading the CSV Files

Our bot creates custom CSV files as the output of its web crawling. These CSVs use semicolon as a separator (since commas are often in web page titles) and new lines to indicate new rows. To read the data more easily, you can open this file using spreadsheet software which can automatically format the rows and columns (although you will likely need to manually tell it to use ; instead of , as the separators). How do do this depends on the software. Microsoft Excel will require you to add sep=; as the new first line of the CSV, whereas OpenOffice will just require selecting semicolon separators when the file preview page comes up.

Once opened in spreadsheet software, you will see a row for each web page that showed up in a search results and a column for each search. The first column is the URL relevance score that was manually entered by the person who ran the search. The second is the url of the result which is used to identify a page in a search and determine relevance. The third is the title which is used to determine relevance. Each column after that is the result of a single search, with the date of that search at the top. The numbers of the search indicate how high the search result is on the first five pages- with the largest number in a column being the first result and the result with rank 1 being the last result of page 5. A 0 in this column indicates that the page did not show up in the first five pages on this result.