Wynnter Ong
Phase 3 - Search Engine & GUI

To run this java program, the user must have the Java SDK installed on the computer.
To make sure the command prompt can run the java compiler (javac) and program

- First, find where your java SDK is located, an example of where it can be found is **C:\Program Files\Java\jdk1.8.0_2417\bin**
  Keep this in mind for later
- Then go to Control Panel > System and Security > System
  On the left side click Advanced system settings, then Environment Variables.
- If there is a Path variable already there, click on it then edit. Once the new window opens click New and copy/paste the location of the Java SDK location there.
    - If there is no Path variable, you can create it by clicking New, and entering Path for the variable name, and the SDK location for the value.

Once in the command prompt, use 'cd' to change directories to reach the location where the java files are (SearchEngine.java, ItemIndex.java, and Porter.java should be in the same location).
The user should first enter 'javac SearchEngine.java' to compile with the java compiler, then the user can enter the command as shown below. The flags can be used in any order.

The command line used: java **SearchEngine** -CorpusDir **PathOfDir** -PremadeInvertedIndex **NameOfPremadeInvertedIndex** -Queries **QueryFile** -Results **ResultsFile** -Output **TEXTFILE / GUI / BOTH** -InvertedIndex **NameOfIIndexFile** -StopList **NameOfStopListFile** -Stemming **YesOrNo**

- CorpusDir & **PathOfDir** (REQUIRED) - Enter the absolute path to the chosen PathOfDir where the corpus is stored, where all the HTML files and the StopList should be located. The same PathOfDir will hold the InvertedIndex/Results text files after the program has finished running.
- PremadeInvertedIndex & **NameOfPremadeInvertedIndex** (OPTIONAL) - If included, the program will read the text file of the premade inverted index and use it for the queries.
    - This means the program will ignore the InvertedIndex, StopList, and Stemming flags because it will not create an inverted index.
- Queries & **QueryFile** (REQUIRED) - Enter the name of the text file within the same directory. The text file should be formatted like this per line:
    - Query <term>
    - Frequency <term>
- Results & **ResultsFile** (OPTIONAL) - If included, the results of the queries on the InvertedIndex will be saved to the name given, otherwise it will be by default Results.txt
    - Results are in the format:
      <filename> <frequency> (If Frequency is the query) <Locations> (If Locations is YES)

    - The locations are an array of the lines where the <term> appears, if the html file was opened in a text editor.
- Output & **TEXTFILE / GUI / BOTH** (OPTIONAL) - If not included, it will default to output to the results text file only.
    - If "**TEXTFILE**" is entered, it will output only to the results text file
    - If "**GUI**" is entered, it will display the results in a GUI, but not save the results in a text file
    - If "**BOTH**" is entered, it will both display the results in a GUI and save the results to a text file.
- InvertedIndex & **NameOfIndexFile** (OPTIONAL) - If included, the InvertedIndex will be saved to the name given, otherwise it will be by default InvertexIndex.txt.
    - Will only work if PremadeInvertedIndex is not provided, because it does not need to create an inverted index.
- StopList & **NameOfStopListFile** (OPTIONAL) - If included, the named text file will be the stop list, otherwise there will be no stop list to filter out words during parsing.
    - Will only work if PremadeInvertedIndex is not provided, because it does not need to parse the stoplist .
- Stemming & **YesOrNo** (OPTIONAL) - If "YES" is entered, then the inverted index will use Porter's Stemming Algorithm. If "NO" is entered or the Stemming flag is not included, the stemming algorithm will not be used.
    - Will only work if PremadeInvertedIndex is not provided, because it does not need to create an inverted index.
    - Porter's Stemming Algorithm will turn every <term> into LowerCase, so "Get" will become "get" in the InvertedIndex, along with applying the stemming algorithm to shorten some words.
- **NOTE**: Each flag can only be used once, if multiple of the same flags are inputted then the program will not run.
- **NOTE**: If NameOfIndexFile & ResultsFile already exists in the CorpusDir, it will be overwritten by the new one.
- **NOTE**: The term(s) in the Queries.txt will be case-sensitive when searching the InvertedIndex. This means that "Get" is different from "get".
- Example: java SearchEngine -CorpusDir C:\Users\23617034\eclipse-workspace\IRPhase1\Corpus -StopList StopList.txt -Queries Queries.txt
    - InvertedIndex and Results will be saved to InvertedIndex.txt and Results.txt respectively by default because not spoken

## InvertedIndex

The inverted index is saved as one file with 2 parts. The first part at the top are the filenames of the HTML files indexed, and the second part is the index list.

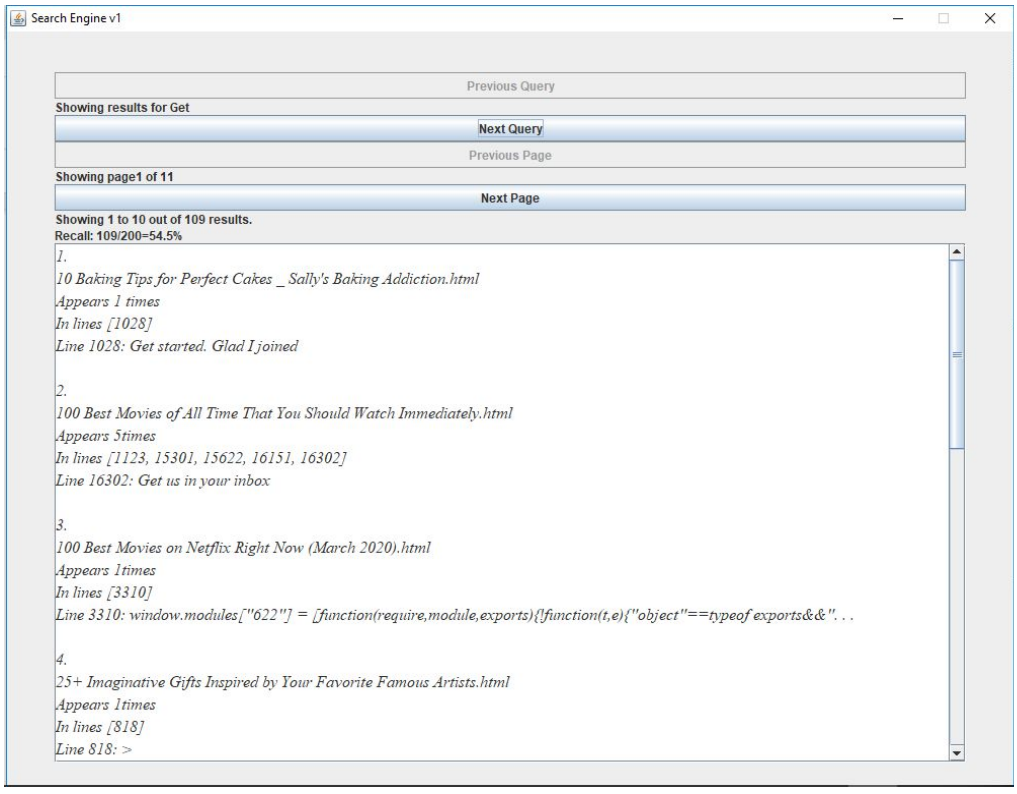
The first part, the filenames, are sorted in alphabetical order.
The second part, the index list, are stored in this format:

- <FileIndex> <Word> <WordCount> <Locations>
- FileIndex - the # corresponds to the placement in the filenames list, ex. 1 corresponds to the first filename at the top.
- <Word> - the word being indexed for that website
- <WordCount> - the # of times the word appears in that file
- <Locations> - the locations are the line(s) where the word appears, if opened in a text editor.

## Images

### Creating an Inverted Index from the Corpus

GUI



InvertedIndex - Notepad

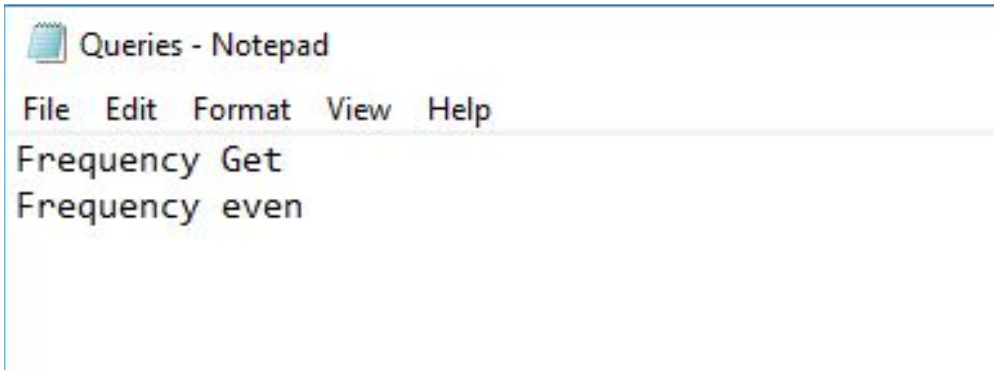
File Edit Format View Help

```
10 Baking Tips for Perfect Cakes _ Sally's Baking Addiction.html
100 Best Games of the Decade _ Den of Geek.html
100 Best Movies of All Time That You Should Watch Immediately.html
100 Best Movies on Netflix Right Now (March 2020).html
20 Secret Cake Baking Tips We Learned From Grandma _ Taste of Home.html
25+ Imaginative Gifts Inspired by Your Favorite Famous Artists.html
30 Modern & Contemporary Artists - Artland Magazine.html
312 Famous Artists And Their Studios _ Bored Panda.html
4 Safe Ways to Treat a Fever.html
4 Ways to Bake a Cake - wikiHow.html
4 Ways to Cure a Fever at Home - wikiHow.html
42 maps that explain World War II - Vox.html
5 Diets That Are Supported by Science.html
5 Tips for How to Deal with a Fever _ MOTRIN®.html
50 Best Movies of 2019 - Top 2019 Movies List of the Year.html
54 Famous Paintings Made by Famous Artists.html
7 Famous Artists Who Are Known for Mixed Media Paintings – Smart Art.html
7 Rules For Baking The Perfect Cake (And How To Fix Any Mistakes).html
8 Famous Artists Who Were Self-Taught - Artsy.html
8 Home Remedies for Fever - How to Get Rid of a Fever.html
A Strange New World_ Has Virtual Reality Gaming Lived Up To Its Promise_ _ Here & Now.html
AFI's 100 YEARS…100 MOVIES _ American Film Institute.html
Artcyclopedia_ Most Popular Artists.html
BBC - History_ World War Two.html
Backpacking France Travel Guide for 2019_ Sights, Costs, & Ways to Save.html
Basic 1-2-3-4 Cake - Bake from Scratch.html
Basic Vanilla Cake Recipe _ Baked by an Introvert.html
Basics to know before starting to bake a cake _ The Splendid Table.html
Best Computer Science Programs - Top Science Schools - US News Rankings.html
Best Diet Plans That Work - Weight Loss Plans to Help You Lose Weight Fast.html
Best Movies and Film News - Classic Movies You'll Love.html
Best Movies of 2020_ Good Movies to Watch From This Year So Far - Thrillist.html
Best Movies of All Time - Metacritic.html
Best Netflix movies_ March 2020 _ TechRadar.html
```

InvertedIndex - Notepad

File Edit Format View Help

```
World War II - Wikipedia.html
World War II Fast Facts - CNN.html
World War II Photos _ National Archives.html
World War II _ Facts, Summary, Combatants, & Causes _ Britannica.html
World War II _ Time.html
World War II for Kids.html
World War II.html
World War II_ After the War - The Atlantic.html
World War II_ Causes and Timeline _ HISTORY.com - HISTORY.html
World War II_ In Depth _ The Holocaust Encyclopedia.html
World War II_ Summary, Combatants & Facts - HISTORY.html
World War Two - American Memory Timeline- Classroom Presentation _ Teacher Resources - Library of Congress.html
computer science _ Definition, Fields, & Facts _ Britannica.html
1 AFFILIATE 1 [835]
1 Ads 1 [2444]
1 Akismet 1 [1965]
1 Also 1 [1366]
1 Another 1 [1655]
1 Any 1 [1856]
1 April 8 [2014, 2026, 2038, 2050, 2062, 2074, 2086, 2098]
1 Arthur 1 [1410]
1 August 8 [2010, 2022, 2034, 2046, 2058, 2070, 2082, 2094]
1 BEST 1 [1788]
1 Baking 11 [15, 731, 832, 1611, 2166, 2167, 2209, 2314, 2383, 2402]
1 Basically 1 [1655]
1 Besides 1 [1655]
1 Blueberry 2 [1983, 2292]
1 Brazil 1 [1655]
1 Bring 1 [1319]
1 CONTAIN 1 [835]
1 Cake 1 [1944]
1 Cakes 2 [15, 1611]
1 Candy 1 [2403]
1 Canvas 2 [2389, 2438]
```




Results - Notepad

File Edit Format View Help

```
Frequency Get
10 Baking Tips for Perfect Cakes _ Sally's Baking Addiction.html 1 [1028]
100 Best Movies of All Time That You Should Watch Immediately.html 5 [1123, 15301, 15622, 16151, 16302]
100 Best Movies on Netflix Right Now (March 2020).html 1 [3310]
25+ Imaginative Gifts Inspired by Your Favorite Famous Artists.html 1 [818]
30 Modern & Contemporary Artists - Artland Magazine.html 1 [234]
312 Famous Artists And Their Studios _ Bored Panda.html 1 [454]
4 Ways to Bake a Cake - wikiHow.html 5 [1503, 1770, 3285, 3453, 3676]
4 Ways to Cure a Fever at Home - wikiHow.html 3 [2913, 2931, 3099]
42 maps that explain World War II - Vox.html 2 [1856, 1949]
50 Best Movies of 2019 - Top 2019 Movies List of the Year.html 2 [1962, 3494]
7 Famous Artists Who Are Known for Mixed Media Paintings – Smart Art.html 2 [966, 1120]
7 Rules For Baking The Perfect Cake (And How To Fix Any Mistakes).html 1 [1366]
8 Famous Artists Who Were Self-Taught - Artsy.html 1 [278]
8 Home Remedies for Fever - How to Get Rid of a Fever.html 19 [7, 67, 68, 662, 1941, 1947, 1950, 1953, 1956]
A Strange New World_ Has Virtual Reality Gaming Lived Up To Its Promise_ _ Here & Now.html 1 [17]
```