To create new index:

`python filterDocsDriver.py`

- Uses simhash to remove similar links
- Make sure the `DEV` folder path is correct before running
- Unique URLs will be added to `validDocs` file

`python partialIndexDriver.py`

- Creates sets of partial indices
- Make sure the `DEV` folder path is correct before running
- Make sure to create a `partialIndex` folder

`python mergeIndexDriver.py`

- Merges the partial indices together
- Make sure `partialIndex` folder path is correct before running
- Make sure to create `mergeIndex` folder

`python scoreIndexDriver.py`

- Scores each token for each document
- Make sure the `mergeIndex` folder path is correct before running
- Make sure to create a `scoredIndex` folder

`python splitIndexDriver.py`

- Creates an index that has a seek number for each token so only one line has to be read in (to reduce search time)
- Make sure to create `splitIndex` folder

To run search:

`python gui.py`

- Runs the search functionality assuming index is fully created
- Uses fully built `splitIndex`

File Tree

```
hw3
├──── OLD
│     └──── search.py
├──── docindex                  (map docNum to url)
├──── index                     (the merged index)
│     ├──── a
│     ├──── b
│     ├──── c
│     ├──── ...
│     ├──── 9
├──── indexer.py                (library for index functionality)
├──── mergeIndexDriver.py       (runs merge code)
├──── partialIndex              (the partial index, unmerged)
│     ├──── 1_a
│     ├──── 1_b
│     ├──── 1_c
│     ├──── ...
│     ├──── 5_9
├──── partialIndexDriver.py     (runs partial index code code)
├──── scoreIndexDriver.py       (runs score index code code)
├──── scoredIndex               (the scored index, merged)
│     ├──── a
│     ├──── b
│     ├──── c
│     ├──── ...
│     ├──── 
├──── search.py                 (runs for search functionality)
├──── seekdict                  (map first 2 char of tok to index)
├──── splitIndex                (the split index, merged)
│     ├──── a
│     ├──── b
│     ├──── c
│     ├──── ...
│     ├──── 9
├──── splitIndexDriver.py       (runs split index code code)
└──── tokenizer.py              (library for token functionality)
```

Project Statistics

	Statistics
Number of Documents Before Optimization	55,393 docs
Number of Documents After Simhash	38,809 docs
Number of Tokens After Simhash	754,446 tokens
Size of Index After Simhash	1,750,512 KB

Query Statistics

Query	Time (sec)
ICS	0.006855
Masters of Software Engineering (w/stopwords)	1.760504
Masters of Software Engineering (w/o stopwords)	0.213571
Cristina Lopes	0.007106
Machine Learning	0.214893
ACM	0.045252