Assignment3 Report

Name: Ke Wang NetID: kw427

1. Implementation

MiniSearchEngine.java: the main class of the core code. *Test.java:* GUI and main function.

Part1: Extracting hyperlinks from the input data

readFile(): read the test file and use *HashMap<String, Integer> pages* to store the url and the corresponding page id. *setMatrix():* use *jsoup.jar* to access each URL, analyze, get all the reference links, and store them to the link matrix. After that check for dead end and solve it by function *solveDeadEnd()*.

Part2: PageRank

getPageRank(): use the matrix B from Part1 and *getG()* to get the Google matrix. Iterate W=G*W until W converges. It may take some 56 times.

Part3: Indexing

When reading URL and parsing href in part1, I have already stored the URL, title and anchors in the structure *HashMap<String,ArrayList<String>> anchors* (the first element in anchors value is the title and the rest are the anchor text). For those image links, we get the value of tag "alt" as the anchor. Call function *writeFile()* to write short index order in file *metadata.txt.*

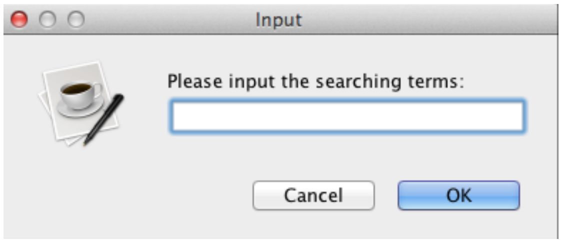
Part4: Searching

I have processed all the anchor text into tokens and stored in the structure *ArrayList<String> allAnchorText* in part1.

For each page, set up a term-document incidence matrix and query vector for the input terms(similar to Assignment1). Call *sim()* to calculate the similarity between the query vector and page vector. Select the top 10 matching pages, sort by their pageRank and then display.

2. How to run

Method1: Run the runnable file *MiniSearchEngine.jar* by double clicking or in command line. (It read the webpage repeatedly, so sometimes time-out happens.) The execution may take some time for large computation. **Method2:** Import the project "MiniSearchEngine" into Eclipse, add external jars *jsoup.jar and jama.jar* in "lib" directory, and finally run Test.java. You would get the searching layout as :



3. Test Case

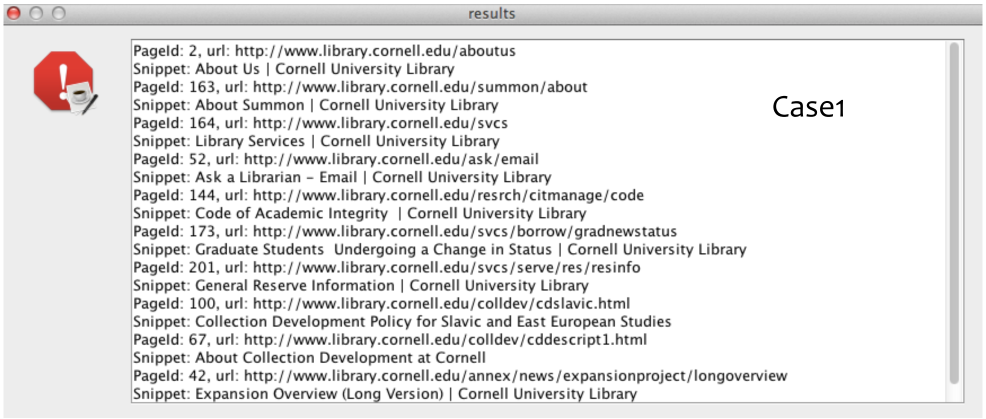
Case1: a single term: about, we get the result sorted by page rank.

Case2: repeated terms: about about, I treat repeated terms as the same term, so I get the same result as in **Case1**.

Case3: multiple terms: cornell university, it will use the query vector to search the pages.

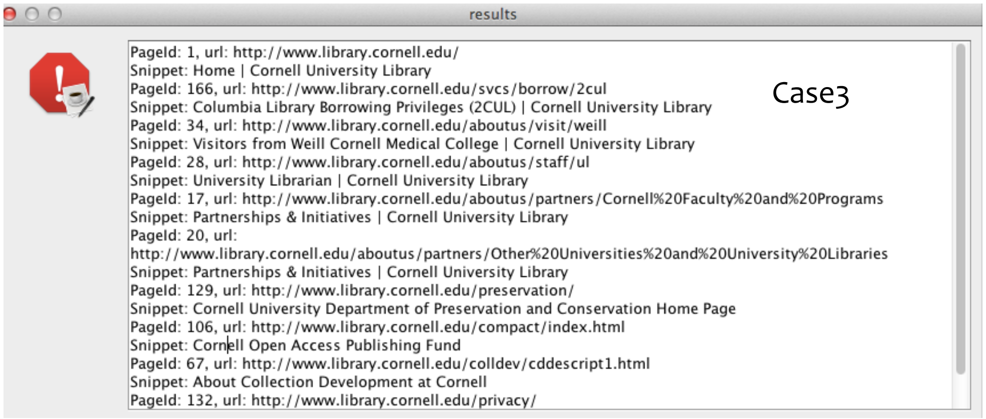
Case4: cornell, it is similar to **Case3**, just need to find the influence of the second term, it is slightly different from **Case3**.

Case5: input nothing, the program would exit automatically.



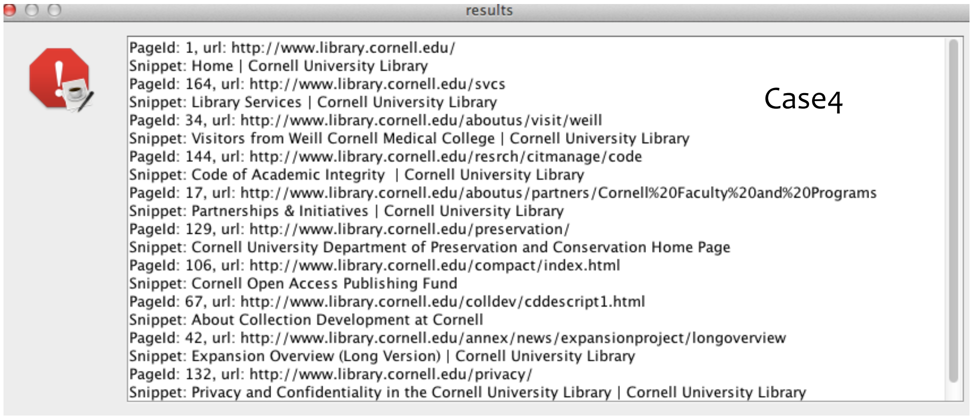
Case1

```
PageId: 2, url: http://www.library.cornell.edu/aboutus
Snippet: About Us | Cornell University Library
PageId: 163, url: http://www.library.cornell.edu/summon/about
Snippet: About Summon | Cornell University Library
PageId: 164, url: http://www.library.cornell.edu/svcs
Snippet: Library Services | Cornell University Library
PageId: 52, url: http://www.library.cornell.edu/ask/email
Snippet: Ask a Librarian - Email | Cornell University Library
PageId: 144, url: http://www.library.cornell.edu/resrch/citmanage/code
Snippet: Code of Academic Integrity  | Cornell University Library
PageId: 173, url: http://www.library.cornell.edu/svcs/borrow/gradnewstatus
Snippet: Graduate Students  Undergoing a Change in Status | Cornell University Library
PageId: 201, url: http://www.library.cornell.edu/svcs/serve/res/resinfo
Snippet: General Reserve Information | Cornell University Library
PageId: 100, url: http://www.library.cornell.edu/colldev/cdslavic.html
Snippet: Collection Development Policy for Slavic and East European Studies
PageId: 67, url: http://www.library.cornell.edu/colldev/cddescript1.html
Snippet: About Collection Development at Cornell
PageId: 42, url: http://www.library.cornell.edu/annex/news/expansionproject/longoverview
Snippet: Expansion Overview (Long Version) | Cornell University Library
```



Case3

```
PageId: 1, url: http://www.library.cornell.edu/
Snippet: Home | Cornell University Library
PageId: 166, url: http://www.library.cornell.edu/svcs/borrow/2cul
Snippet: Columbia Library Borrowing Privileges (2CUL) | Cornell University Library
PageId: 34, url: http://www.library.cornell.edu/aboutus/visit/weill
Snippet: Visitors from Weill Cornell Medical College | Cornell University Library
PageId: 28, url: http://www.library.cornell.edu/aboutus/staff/ul
Snippet: University Librarian | Cornell University Library
PageId: 17, url: http://www.library.cornell.edu/aboutus/partners/Cornell%20Faculty%20and%20Programs
Snippet: Partnerships & Initiatives | Cornell University Library
PageId: 20, url:
http://www.library.cornell.edu/aboutus/partners/Other%20Universities%20and%20University%20Libraries
Snippet: Partnerships & Initiatives | Cornell University Library
PageId: 129, url: http://www.library.cornell.edu/preservation/
Snippet: Cornell University Department of Preservation and Conservation Home Page
PageId: 106, url: http://www.library.cornell.edu/compact/index.html
Snippet: Cornell Open Access Publishing Fund
PageId: 67, url: http://www.library.cornell.edu/colldev/cddescript1.html
Snippet: About Collection Development at Cornell
PageId: 132, url: http://www.library.cornell.edu/privacy/
```



Case4

```
PageId: 1, url: http://www.library.cornell.edu/
Snippet: Home | Cornell University Library
PageId: 164, url: http://www.library.cornell.edu/svcs
Snippet: Library Services | Cornell University Library
PageId: 34, url: http://www.library.cornell.edu/aboutus/visit/weill
Snippet: Visitors from Weill Cornell Medical College | Cornell University Library
PageId: 144, url: http://www.library.cornell.edu/resrch/citmanage/code
Snippet: Code of Academic Integrity  | Cornell University Library
PageId: 17, url: http://www.library.cornell.edu/aboutus/partners/Cornell%20Faculty%20and%20Programs
Snippet: Partnerships & Initiatives | Cornell University Library
PageId: 129, url: http://www.library.cornell.edu/preservation/
Snippet: Cornell University Department of Preservation and Conservation Home Page
PageId: 106, url: http://www.library.cornell.edu/compact/index.html
Snippet: Cornell Open Access Publishing Fund
PageId: 67, url: http://www.library.cornell.edu/colldev/cddescript1.html
Snippet: About Collection Development at Cornell
PageId: 42, url: http://www.library.cornell.edu/annex/news/expansionproject/longoverview
Snippet: Expansion Overview (Long Version) | Cornell University Library
PageId: 132, url: http://www.library.cornell.edu/privacy/
Snippet: Privacy and Confidentiality in the Cornell University Library | Cornell University Library
```

4. Others

a). I deal with various kinds of URL, including relative URL, anchors within a file and some redirect URLs.

b). To calculate the similarity between the query and page, I use cosine of the two vectors in this case. We can also use tf.idf as the weight as we do in assignment1.

c). I have already written the short index file in the *metadata.txt.* If you want to rerun the writeFile(), you can add "ws.writeFile()" in the Test.java.