CS 600 Project: Search Engine | CWID 10430147 | Divyendra Patil | Username: dpatil3  
Date: 11/22/2017

I have used 6 links to read from the web using jsoup. Reading paragraph from the webpages I extracted words and stored them in a trie data structure,

The find the word entered by the user, trie data structure is used to search as follows:

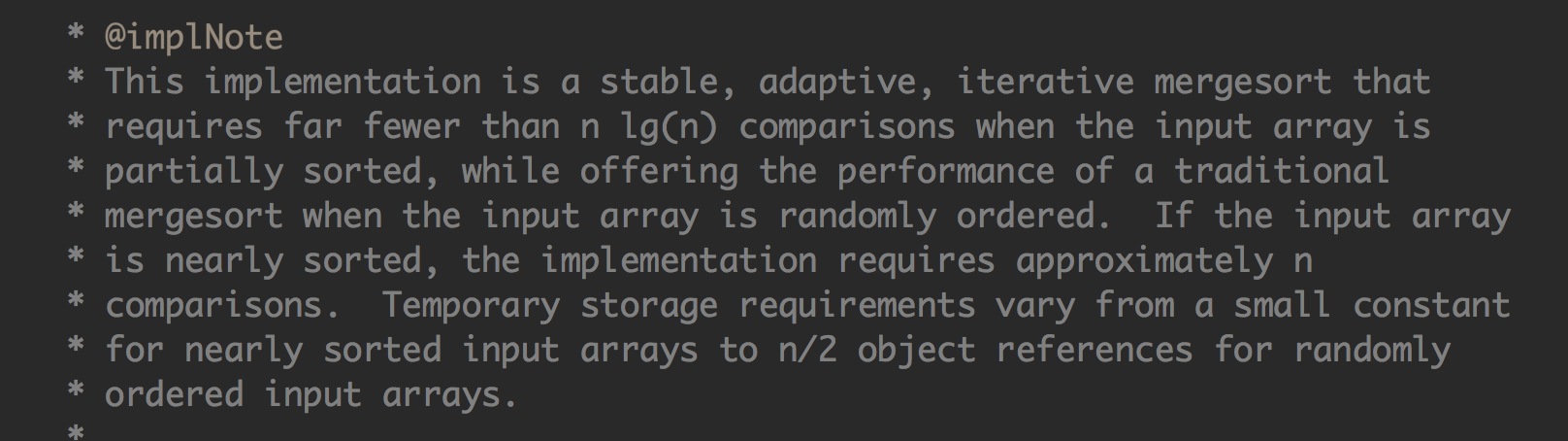


To find the count of occurrences (ranking) of word from each webpage I used a hashmap for each link which displays their word count respectively. If the word is not present in one of the document the count is shown as zero.

To remove the stop words, as we read each word from the webpage’s paragraph pass that word to the stopword function containing a list of all stop words (stopwords.txt). If a match is found that word is not inserted into a trie datastructure as well the hashmap.

Also, to display the total of all the words from all the documents I maintained a separate map which displays the words as well their total occurrences found from all the documents.

For ranking, we have ranked based on number of occurrences of the word on a page.   
For ranking, we have sorted using mergesort of it to be efficient and fast.



Used in here:

pages.**sort**(new PageComparator());

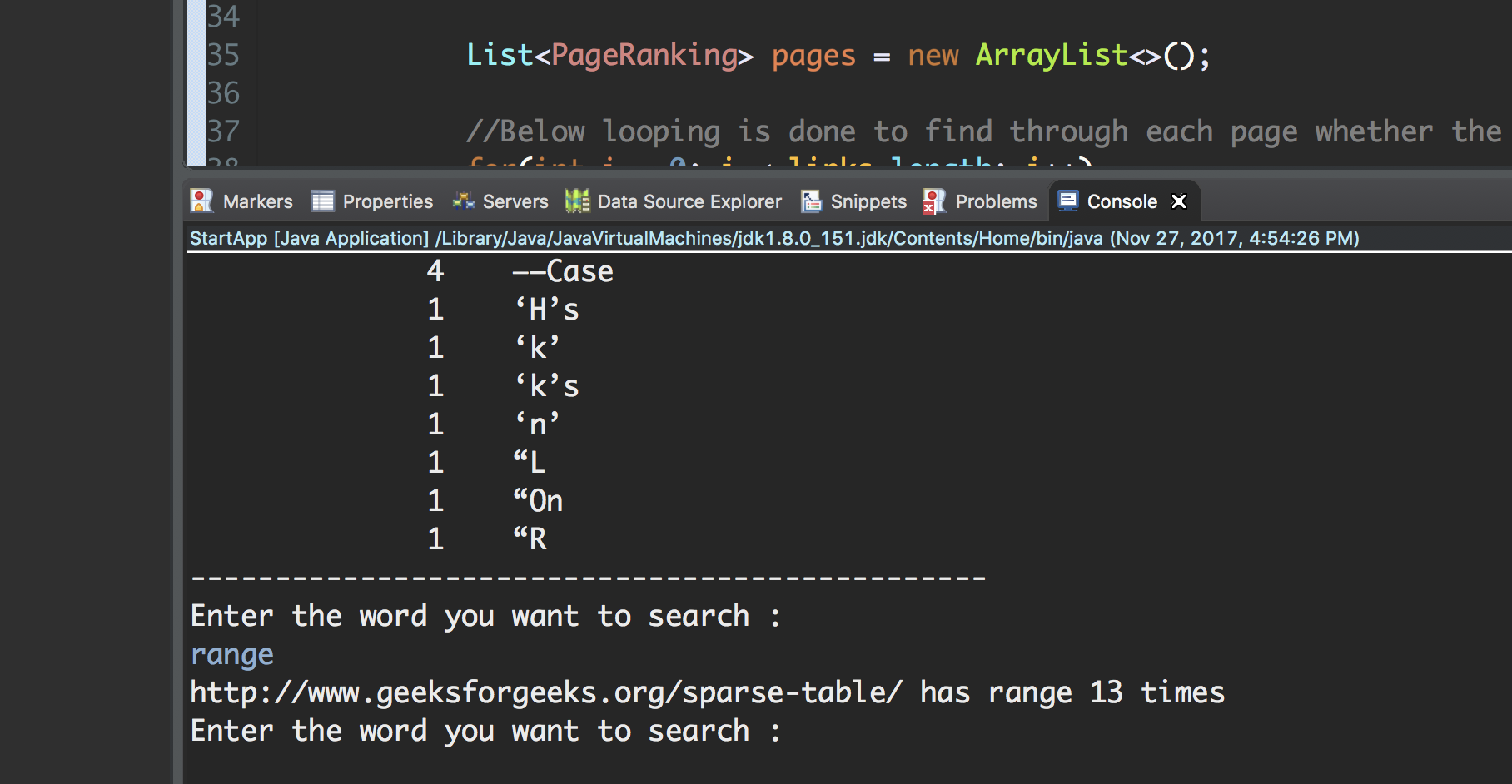
for (PageRanking page : pages) {

System.out.println(page.getName() + " has " + word + " " + page.getRank() + " times");

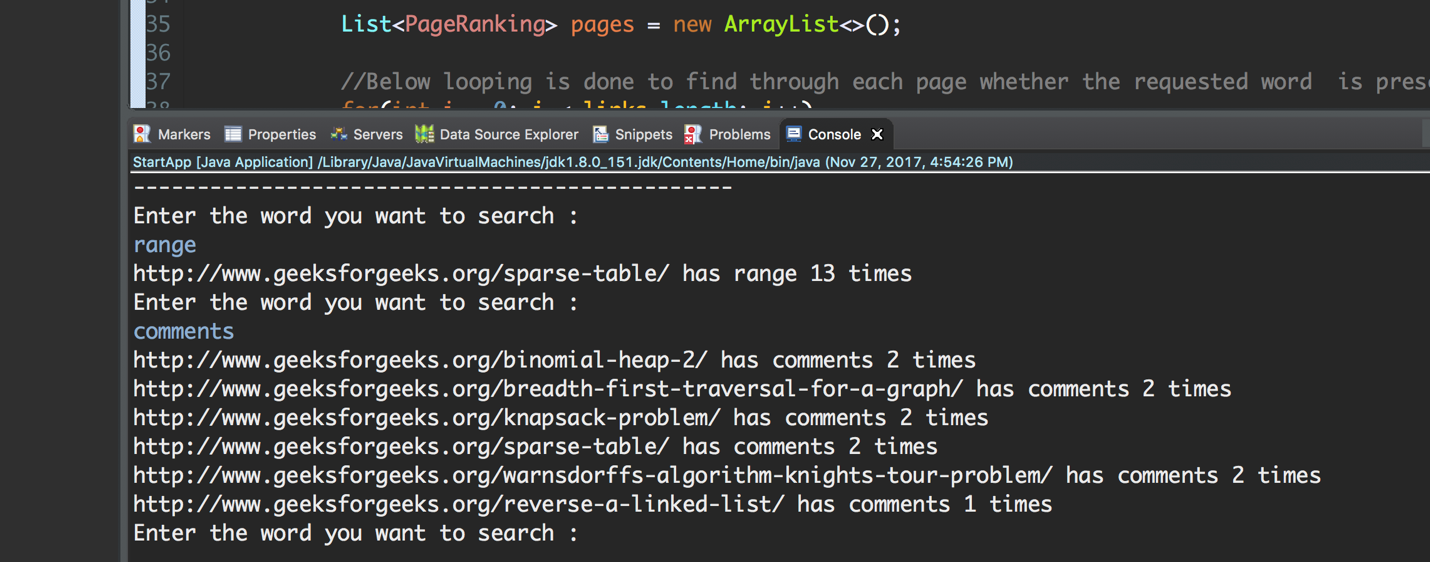
}

Sample Output:

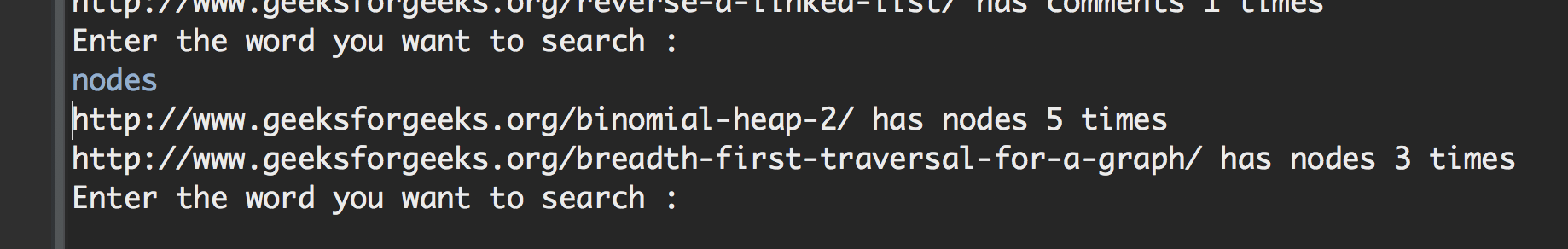
1] Since the word “range” is only available in one of the links of sparse table it only shows that link.



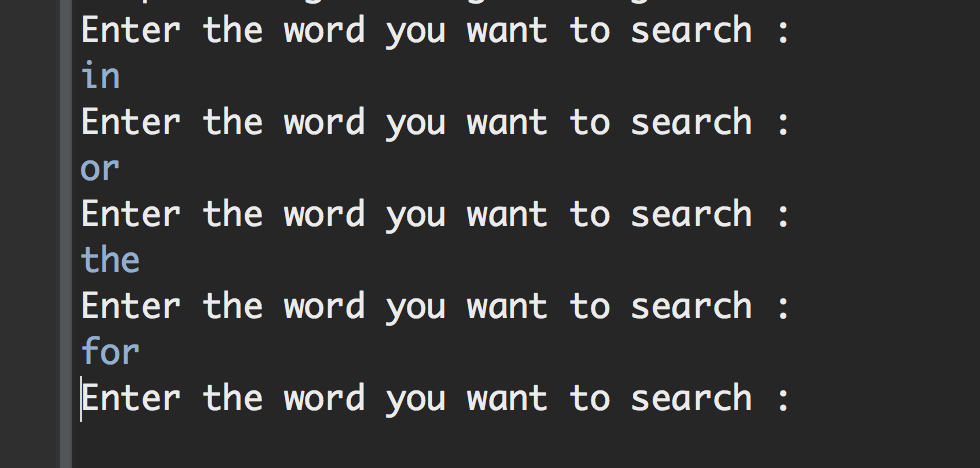
2] Word: Comments



3] Word: nodes



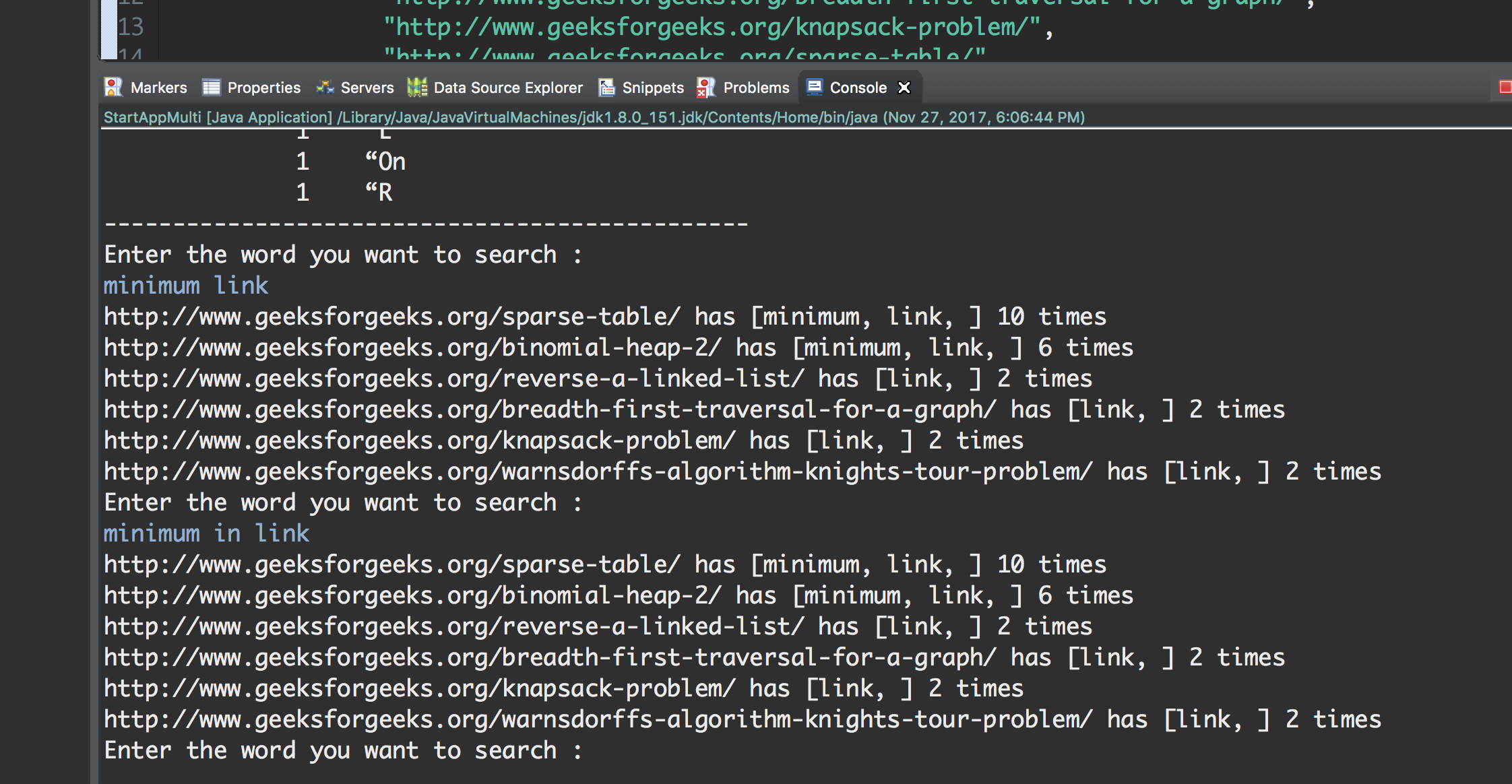
4] Stop Words being eradicated in output results (Stored in stopwords.txt):



FOR MULTIPLE KEYWORDS

For more than one keywords, we search the pages containing the word by trie and sort them first based on number of keywords out of search query and then in buckets by number of their individual occurrences.

Start with: **StartAppMulti.java**

****

When we put “Minimum in Link”… “in” gets disbarred from the search results since it’s a stop word.