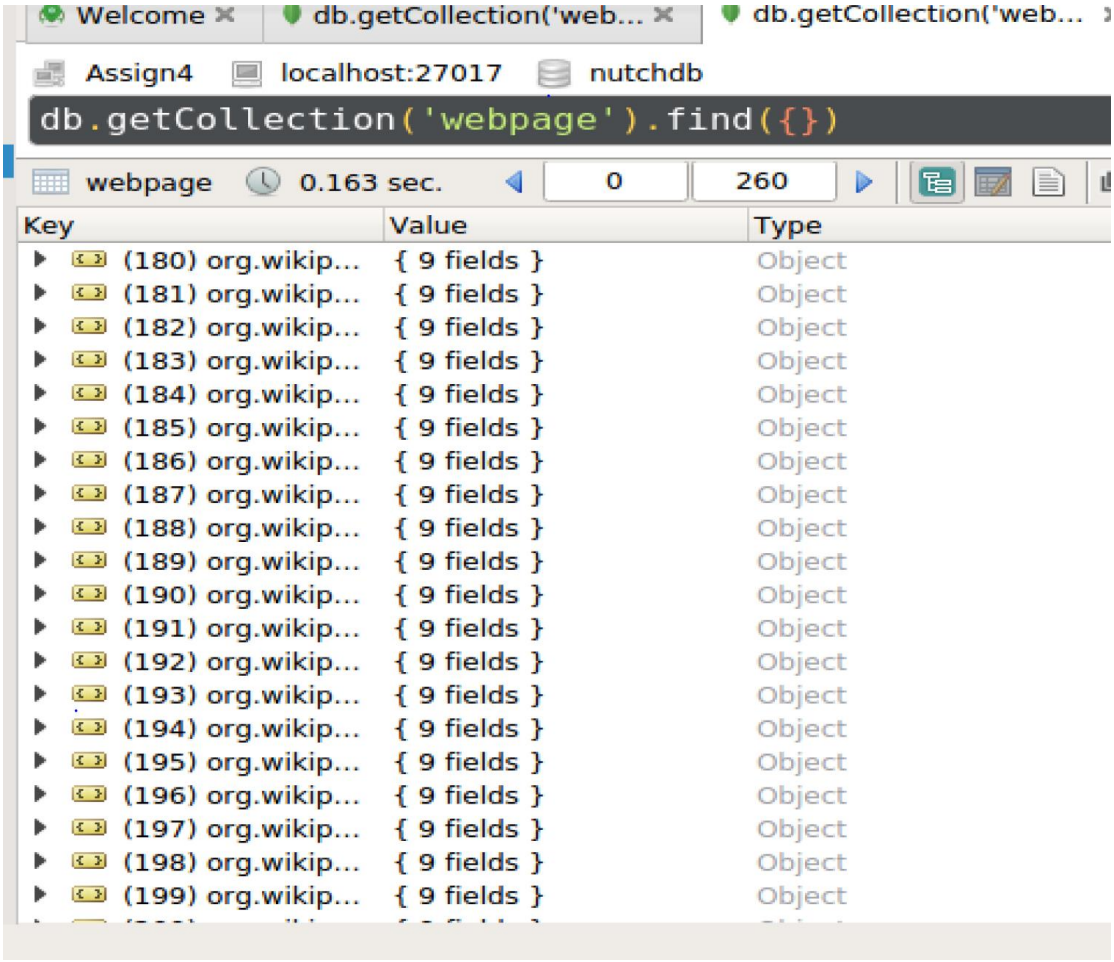
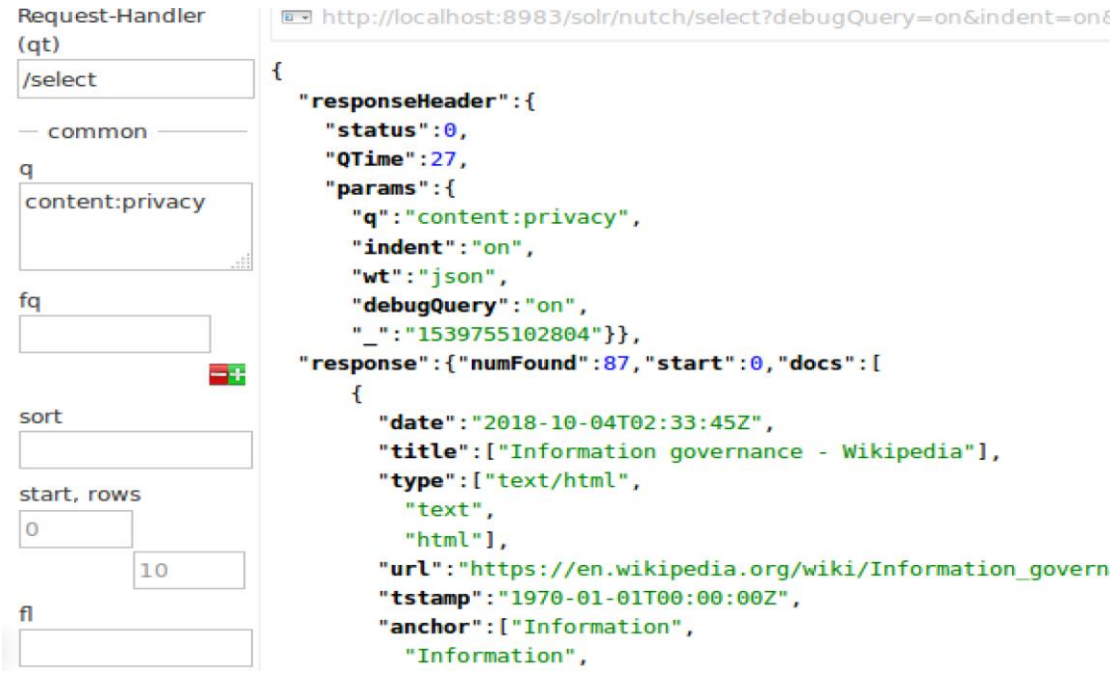
Project Report

S44872452 Jiayong Kuang

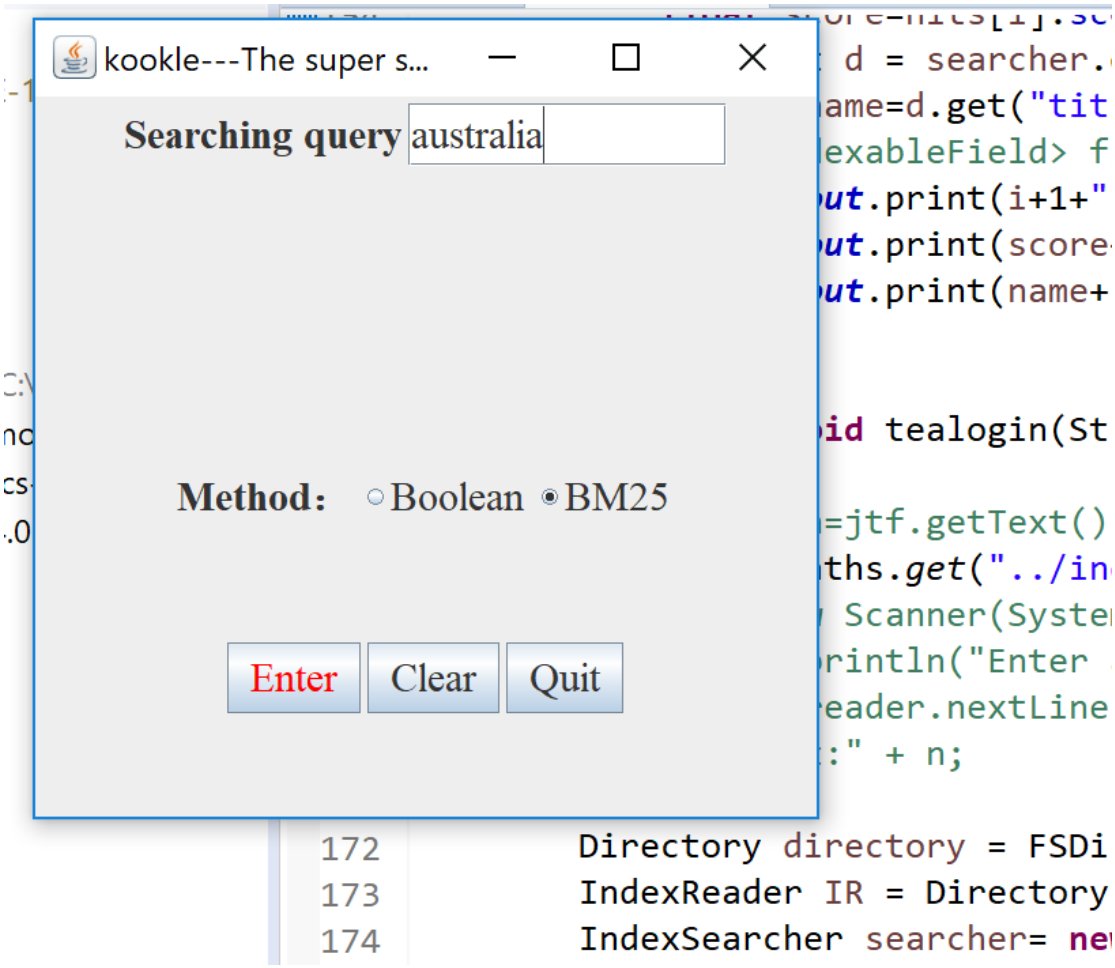
This project is based on JDK8.
After creating a linux vitual environment and installing all software necessary. Some webpages in Wikipedia have been crawled using nutch to build an individual searching engineer. The original seed webpage was set as
https://en.wikipedia.org/wiki/States_and_territories_of_Australia
Among the process, the 4 steps (generate, fetch, parse and update) were repeated 3 times. 50 webpages were generated in the first round and 150 in the second and third round.
After that, totally 10479 webpages were updated in mongodb, while only 259 pages can really be used in searching, which have more than 9 fields.
Then, Solr was used to generate index and perform some basic query searching.
In order to create a search engine based on the index generated, lucene 7.4.0 and some build- in packages was used.
There are 2 model were implemented in the app, Boolean model to search and BM25 model to search.
As you can see in the outcome when using Boolean model, it can only tell whether the keyword is in the file or not (the score is either 1 or 0). It is why the rank does not actually matter. But BM25 is basically based on Bayes' theorem, it would calculate how relevant the documents are based on probabilities (frequencies of the query in documents) and return a specific score for each document. The rank of this model is technically important
Thus, the conclusion is, when doing precise search or searching rare words, Boolean model performs better or BM25 is preferred.



(As you can see in the pictures, lots of crawled webpages have only 9 fields and can not be used)



(The keyword “privacy” was used to perform a simple search)



(Enter a query, chose a method and press enter, the Rank, Score and the Document title would be show in the console)

```
Rank    Score   Docname
1       1.0     States and territories of Australia - Wikipedia
2       1.0     Northern Territory - Wikipedia
3       1.0     Queensland - Wikipedia
4       1.0     South Australia - Wikipedia
5       1.0     Tasman Sea - Wikipedia
6       1.0     Template:Administrative divisions of Australia - Wikipedia
7       1.0     Territory - Wikipedia
8       1.0     Timor Sea - Wikipedia
9       1.0     129th meridian east - Wikipedia
10      1.0     138th meridian east - Wikipedia
```

(The outcome using Boolean method)

```
gui (1) [Java Application] C:\Program Files\Java\jre1.8.0_191\bin\javaw.exe (Oct 19, 2018, 9:55:56 PM)
Rank    Score    Docname
1       0.85296685      States and territories of Australia - Wikipedia
2       2.3704113E-15   Australia Act 1986 - Wikipedia
3       2.0365507E-15   Time in Australia - Wikipedia
4       2.0365507E-15   Time in Australia - Wikipedia
5       2.0365507E-15   Time in Australia - Wikipedia
6       2.0365507E-15   Time in Australia - Wikipedia
7       1.7026899E-15   South Australia - Wikipedia
8       1.4689873E-15   Western Australia - Wikipedia
9       1.3020569E-15   File:Coat of arms of South Australia.svg - Wikipedia
10      1.2018987E-15   File:Coat of arms of Western Australia.svg - Wikipedia
```

(The outcome using BM25 method)