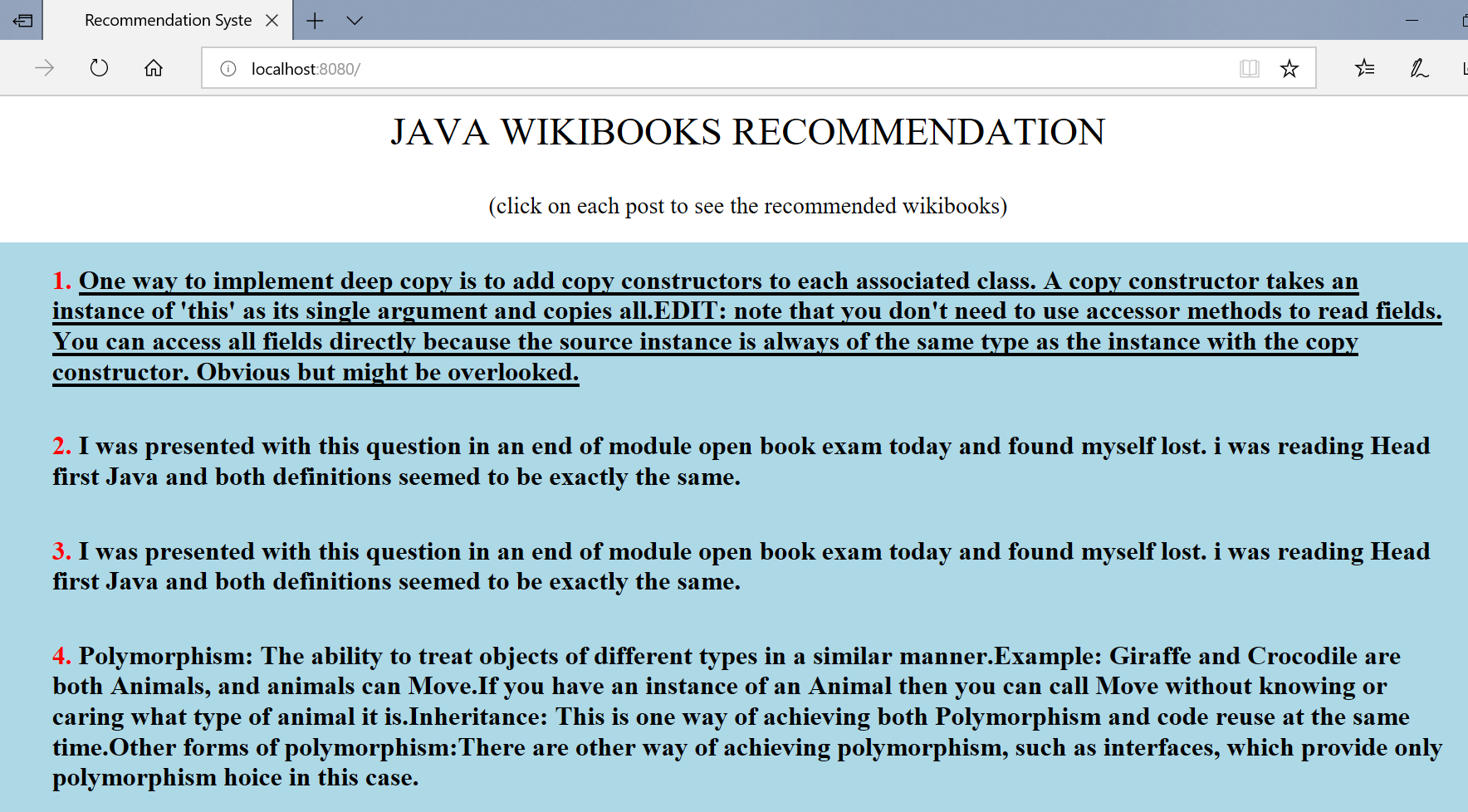
JAVA WIKIBOOKS CRAWLING AND INDEXING

Readme:

1. Run the starter.java file using a local server. Here I am using apache tomcat. The starter.java file calls the index.jsp file in the tomcat Catalina and opens index.jsp on the registered browser.
2. Import the required libraries for Jsoup and apache lucene in case of errors.
3. The webpage has the different posts displayed on it. Click on each post to view the java wiki book recommendations for it.





The project contains 3 main files:

1. crawling: it helps crawl the wikibooks on java from <https://en.wikibooks.org/wiki/Java_Programming> and saves the file as .txt files which is readable by apache lucene
2. SimpleLuceneIndexing: it is the file provided by professor for apache lucene indexing with two other functions for converting to tokens and stems to be read by keyword file. The keyword file is taken from <http://stackoverflow.com/questions/17447045/java-library-for-keywords-extraction-from-input-text> that is java library to extract keywords from input text.

It indexes the files and searches the .txt files for the given question.

1. starter: the starter class calls the HTTP request to start the servlet class and initiates apache lucene for indexing. It sends the text data to the webpage.