CS 242: Information Retrieval & Web Search Project.

Phase A Report prepared for Prof. Vagelis Hristidis.

By:

Joshua Potter	860159747
Ashwin Ramadevanahalli	861186399

April 29th, 2015

Index

Introduction

The design, implementation, data collection, and testing of Java-based WebCrawler and Indexer is the project that this paper will be discussing. The authors implemented the URL collection algorithm and mapping URL links from a given seed collection of URL's, along with additional efficiency and efficacy enhancements. Also, the authors implemented a HTML parsing and indexing schema that runs on a directory of HTML documents and outputs an index in an output directory. This paper will discuss these aspects of the WebCrawler and Indexer as constructed by the authors.

Collaboration Details

Joshua served as project head and designed and instrumented the framework for the Crawler and Indexer applications. Coding was divided between Joshua and Ashwin handling the algorithm implementation, function /class instantiations, and debugging, testing and error handling such that whomever identified errors, they would discuss the solution with their partner and code solutions and make corrections. The code was freely available to both parties throughout development for ease of versioning and code maintenance. Execution of applications were each handled on their respective machines with final code review and corrections handled by Ashwin. Project review and writing of the report was handled by Ashwin with minor input from Joshua.

Overview of the Systems

The system used for development is an Intel i7 and i5 based PCs with 8 GB RAM one running an Ubuntu v14.01 64-bit Linux distribution and the other 64-bit OS X 10.10. Software packages installed for Java development include Eclipse Luna v4.4.1, Java v7.0, JDK JavaSE v1.7, and the Apache Tomcat v7.0.59 Dynamic Web Content server. Additionally, Jsoup v1.8.1 package was downloaded and integrated into the WebCrawler project to assistance in the handling of URL's and HTML processing; Jericho 3.3 HTML parsing libraries and Lucene v3.4 packages were downloaded and integrated into the Indexer implementation to assistance in the handling and parsing HTML processing in the case of the Jericho package and Lucene for the indexing and searching.

Overview of the Crawling system

1. Architecture:

The WebCrawler application consists of several classes and packages described as follows:

Crawl.java
The Crawl.java file serves as the main entry point to the application and is what is invoked at the command line. All other threads, classes /objects and packages are instantiated from here.

Environment.java
Environment.java is a class file that defines /maintains the environment variables that the application uses to communicate between the main routine and each of the threads containing the 'Crawler' objects.

CrawlerThread.java
CrawlerThread.java is a class extending the standard Java Thread class. This class instantiates Thread objects to allow for concurrent processing of URL's and document parsing. Each thread instantiates an instance of the Crawler class during the life of the CrawlerThread.

Crawler.java
Crawler.java is the object-class that does the actual work of visiting a domain in a URL passed in, keeps track of hopCount, downloads the HTML doc, parses for other URL's, cleans and validates those URL's, and then sorts those URL's into links for the same domain which are then added to the Crawler's frontier queue are written to the global environment Frontier Queue along with an increase in it's hopCount.

CrawlURLObj.java
CrawlURLObj.java is an object-class used to abstract the relationship between the URL and the number of hops away from the seed URL's. This object is how URL's are placed in the global frontier Queue so that when a hopURL is pulled from the frontier, the hopCount can be checked to see if the number of maximum hops has been reached and terminate the routine.

2. Crawling Strategy and Algorithm:

The approach implemented by the Crawler is to take a set of seed URL's, and then process all valid URL's contained in the HTML docs downloaded at those links to discover other HTML docs while bounding for number of hops away from the seeds, number of threads instantiated, and the number of pages downloaded to enforce that the Crawler will, in fact, eventually terminate; which in turn will restrict the amount of data collected to be under the specified 10 Gigabytes.

The high-level algorithm is as follows:

```
Crawler (seedURLs) {

    Queue frontier := seedURLs
    HashMap visited
    Int hopCount, maxPages, maxThreads Int hops, pages, threads

    while ( !frontier.empty AND hops < hopCount AND pages < maxPages ) {
        URL := frontier.pop()
        new crawler(URL)
        download URL document
        list Links := parse and clean URL links
        foreach (link in Links) {
            if (link is in the same domain as URL) {
                crawler.frontier := link, hopCount
            } else {
                Crawler.frontier := link, hopCount++
            }
        }
    }
}
```

3. Data Structures:

The Crawler utilizes a global FIFO Queue named as frontier that stores URL's to be visited as hopURL objects where the URL is associated with a hopCount that indicates its distance from the seed URL's. The frontier resides in main memory and can be read and written by all Crawler Thread /Crawler objects.
A HashMap named visited is utilized as a means of tracking which URL's have been seen before. The HashMap is implemented as a <key, value> pair where the key is the URL and the value is an int representing the page number when it was accessed. visited is accessible to all threads for reading and writing and can be written to disk by the operating system as needed.

A custom data object called hopURL was implemented as a means to associate each URL with its respective hopCount. It simply consists of a String and an int that are set /read by the appropriate methods for the class. The intent was to make sure the maximum number of hops bound would be identified and enforced.

4. Limitations:

The implementation of a WebCrawler entails many components, dependencies, variations, data handling, and more. A discussion of the limitations of the WebCrawler follows.

The WebCrawler was developed as a multi-threaded application in order to increase throughput of URL's during the crawling process as there are inherent delays when accessing servers across a network. Both the Frontier Queue and the Visited HashMap are publicly accessible to all thread- resident Crawlers which can both read and write to the Frontier and Visited structures. Consequently, there is a potential for race conditions to result and the likelihood increasing as thread count increases. A token-based access system with an Crawler access Queue should be implemented as a way to control access to these structures on a Crawler-by-Crawler basis. Further, as each thread instantiates it's own Crawler, there is no direct accounting for the creation or tracking of the Crawlers instantiated to verify their resource usage or to verify their destruction leading to a possible memory leak as the application executes. There are also not timing watches or delays currently implemented in the Crawler.

Due to the wide variety of possible valid URL's, as well as the plethora of improper URL's, that the WebCrawler can encounter, only a cursory processing of URL's is performed. Several file types are checked for an excluded if found as a resource link, and some URL validation is performed, but error detection of bad URL's needs substantially more development.

5. Instructions for Running the Package:

The WebCrawler Java program has been packaged to run stand alone on a Linux machine. A shell script is not utilized as not all parameters are necessary to run the program, therefore making it difficult to enforce dynamic passing of parameters into the executable. We have implemented a flag system for allow parameter passing with maximum flexibility so that parameters can be specified in any order or not at all – save for the input file parameter which is required to run.

The resulting command line execution is of the format:

```
java jar Crawler.jar f <input file>
```

Optional parameters that can be passed in include:

```
h <max # of hops (default = 3)>
t <max # threads (default = 10)>
p <max # of pages (default = 10,000)>
o <output directory name (default = "output")>
```

For example, the following commands with execute the WebCrawler with various specified parameters:

```
1. java jar Crawler.jar f seeds.txt o downloads h 3 t 20
2. java jar Crawler.jar f seeds.txt
```

Command #1 will execute the WebCrawler on seeds.txt, create an output folder called “downloads” in the executable's directory, set a maximum of 3 hops from the URL's in the seeds.txt file, and instantiate no more than 20 simultaneous threads. The default value of 10,000 pages downloaded is set as the parameter was not set. Command #2 opts to just run the WebCrawler with just the required input file name, accepting all other parameters as the default values.

Overview of the Indexing system.

1. Architecture:

The Indexer application consists of several classes and packages described as follows:

Indexer.java
The Indexer.java file serves as the main entry point to the application and is what is invoked at the command line. All other classes /objects and packages are instantiated from here. It handles the passing of the command-line parameters specifying the input and output directories of the application. It then reads the file listing in the input directory and instantiates a Parser object, passing in the file location. The Parser then instantiates an Indexer_IndexedHTML object and passes it to the Indexer for indexing.

Indexer_Parser.java
Indexer_Parser.java is a class file that instantiates the Indexer_Parser object and performs the parsing of the HTML document passed in from Indexer.java using the Jericho HTML Parsing library. During parsing, the Parser does some cleanup of the HTML markup, instantiates a Indexer_IndexedHTML object, assigns the parser values and then returns it.

Indexer_Obj.java
The Indexer_Obj.java indexer app takes in a Indexer_IndexedHTML object for indexing. The Indexer uses the Lucene library for performing the indexing to the specified output directory for the index.

Indexer_IndexedHTML.java
Indexer_IndexedHTML.java is the class file that describes the object that contains all the HTML document information output by the Parser.

2. Data Structures , Indexing Strategy and Algorithm:

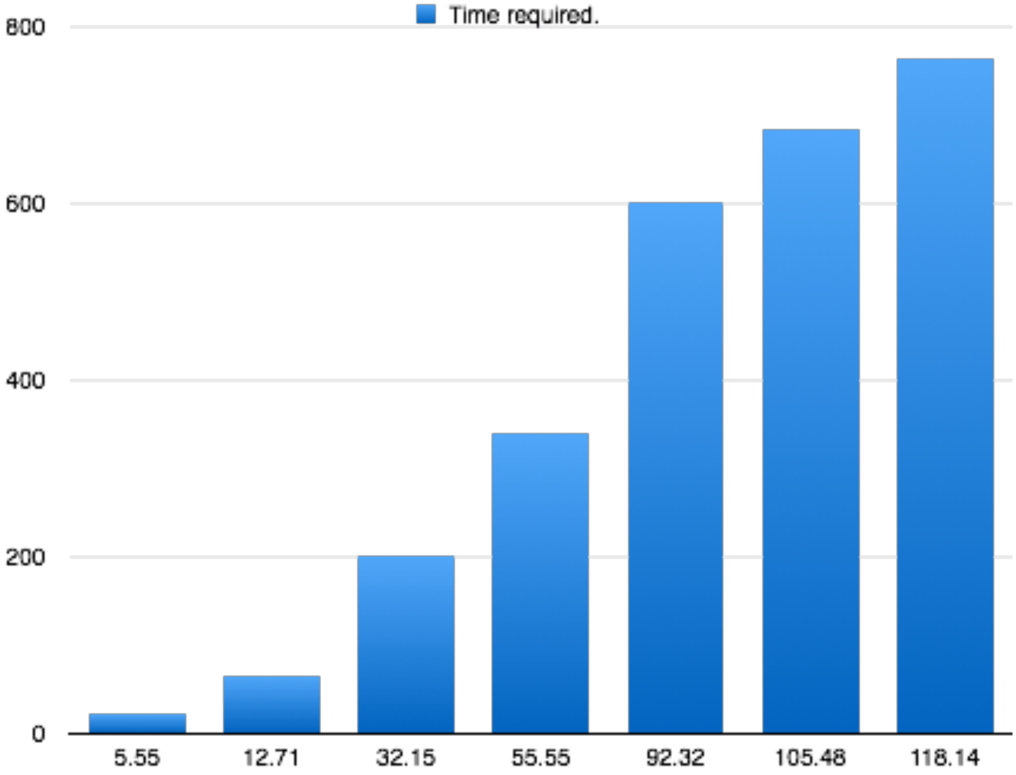
The Indexer uses a custom data object called Indexer_IndexedHTML was implemented as a means to encapsulate the data the Lucene-based Indexer was going to process. As the Jericho-based Parser processed the HTML documents, it would instantiate a new Indexer_IndexedHTML object and assign the available values to it and set default values as well to avoid null value errors during parsing. The object has fields and methods for setting, storing, and retrieving String values passed from the Parser for the title, META description, META keywords, URL, and body of the HTML file parsed. During parsing, the Indexed_Parser does some cleanup of the HTML markup, instantiates a Indexer_IndexedHTML object, assigns the parser values and then returns it.

Fields in the Lucene index:
The following fields are incorporated:

1.Text
2.Keywords
3.Descriptions
4.Url
5.Title

3. Run Time Analysis:

x-axis: Time in seconds.
y-axis: Number of Files.



4. Limitations:

The implementation of an HTML Indexer entails many components, dependencies, variations, data handling, and more. A discussion of the limitations of both the Indexer application is as follows:
Despite to the wide variety of possible valid HTML, as well as the plethora of improper HTML, that the Parser can encounter, the Jericho-based Parser managed to handle quite a lot of the HTML files. However, the Parser did not have the URL from the Crawler phase as that was not handled during collection, consequently this field is missing in the indexed HTML. Future development will re-visit the Crawler project to include this functionality.

5. Instructions for Running the Package:

The Indexer Java program has been packaged to run stand alone on a Linux machine. A shell script is not utilized as not all parameters are necessary to run the program, therefore making it difficult to enforce dynamic passing of parameters into the executable. We have implemented a flag system for allow parameter passing with maximum flexibility so that parameters can be specified in any order or not at all – save for the input file parameter which is required to run.

The resulting command line execution is of the format:

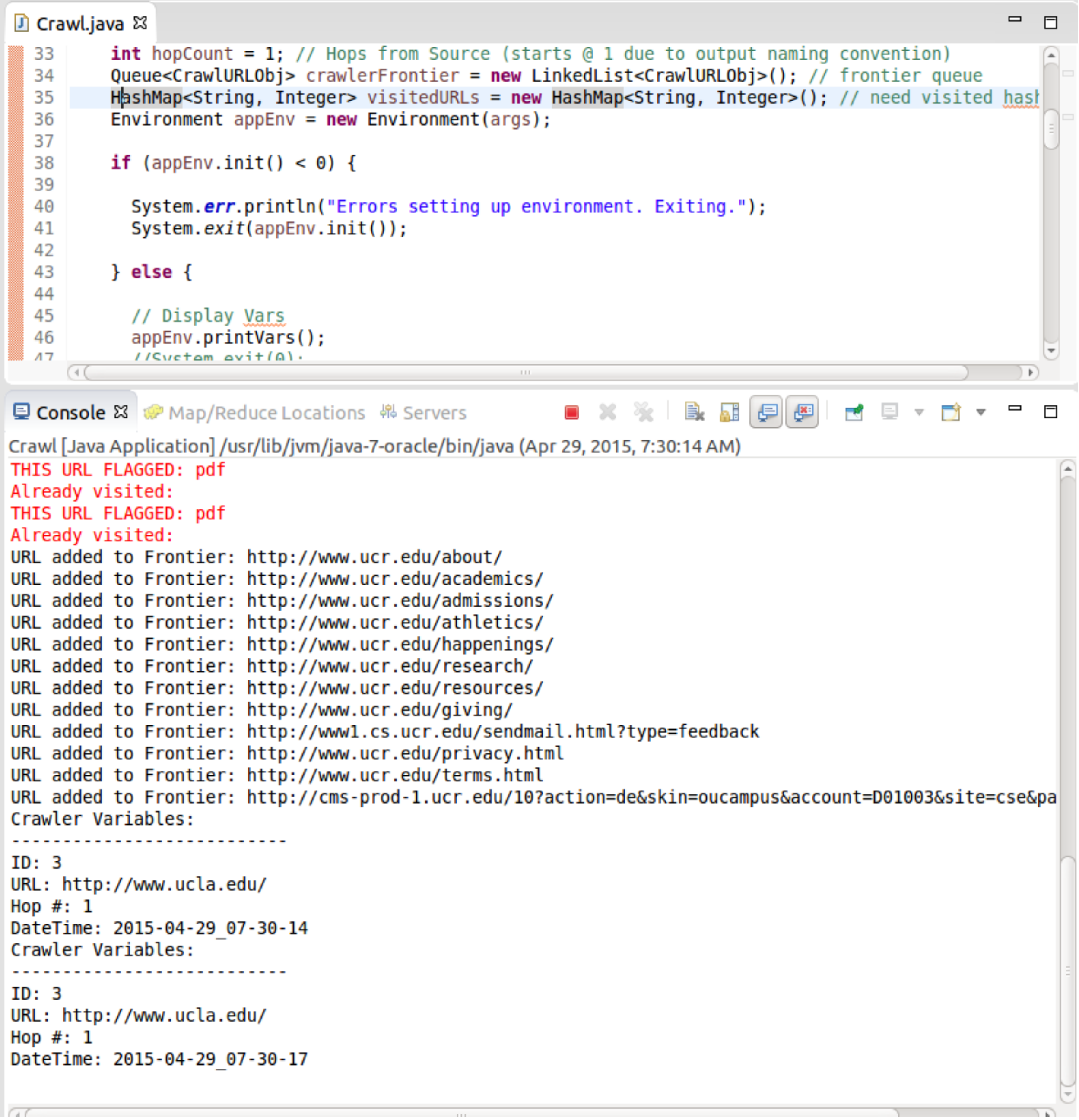
```
java -jar Indexer.jar -i <input directory> -o <input directory>
```

For example, the following commands with execute the Indexer with various specified parameters:

```
1. java -jar Indexer.jar -i html -o index1

2. java -jar Indexer.jar -i downloads -o index2
```

Command #1 will execute the Indexer on the directory html, and create an output folder called index1 in the executable's directory,. Command #1 will execute the Indexer on the directory downloads, and create an output folder called index2 in the executable's directory.

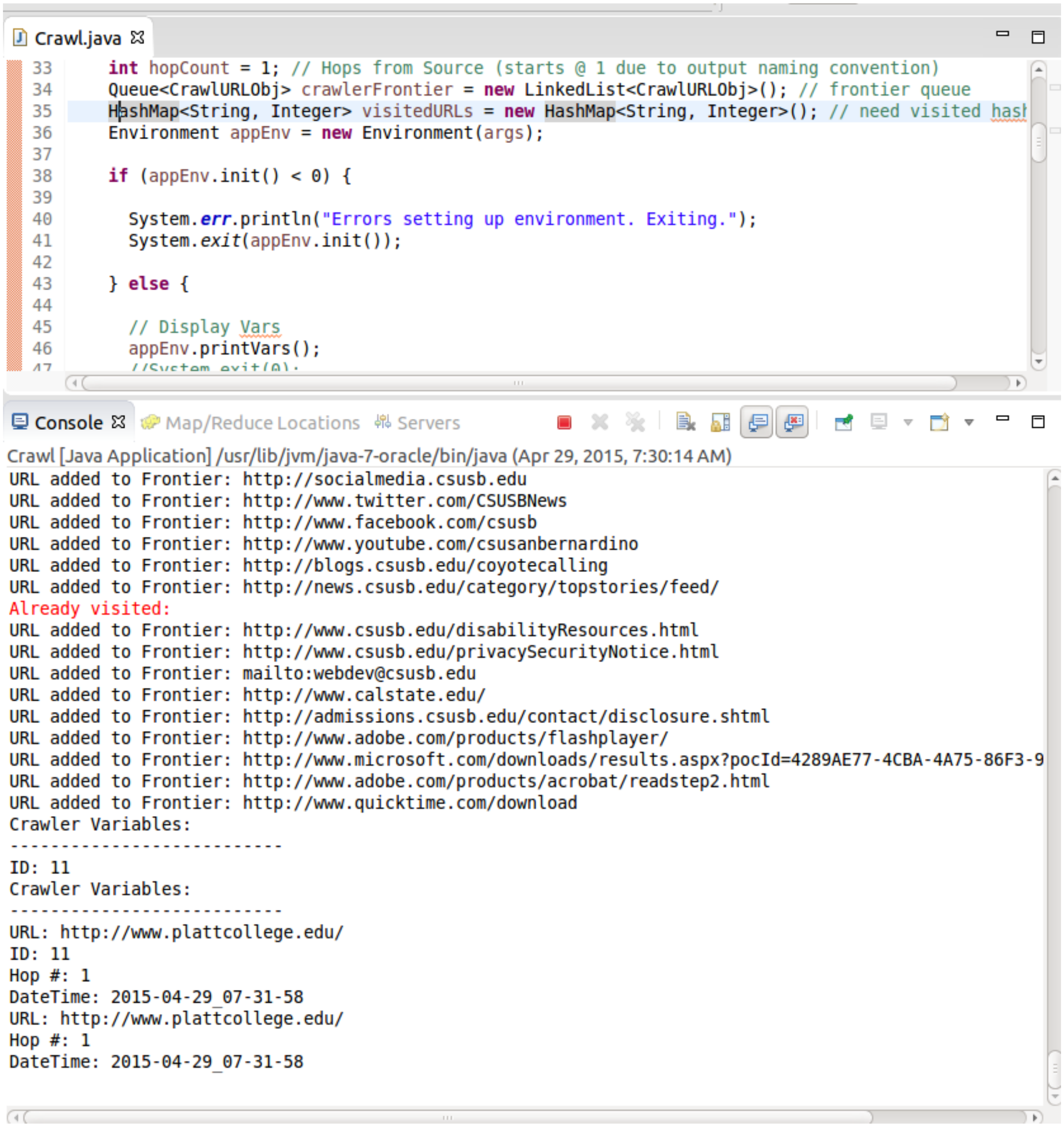
Screenshots



```
    int hopCount = 1; // Hops from Source (starts @ 1 due to output naming convention)
    Queue<CrawlURLObj> crawlerFrontier = new LinkedList<CrawlURLObj>(); // frontier queue
    HashMap<String, Integer> visitedURLs = new HashMap<String, Integer>(); // need visited hasl
    Environment appEnv = new Environment(args);

    if (appEnv.init() < 0) {

      System.err.println("Errors setting up environment. Exiting.");
      System.exit(appEnv.init());

    } else {

      // Display Vars
      appEnv.printVars();
      //System.exit(0);
```

Crawl [Java Application] /usr/lib/jvm/java-7-oracle/bin/java (Apr 29, 2015, 7:30:14 AM)

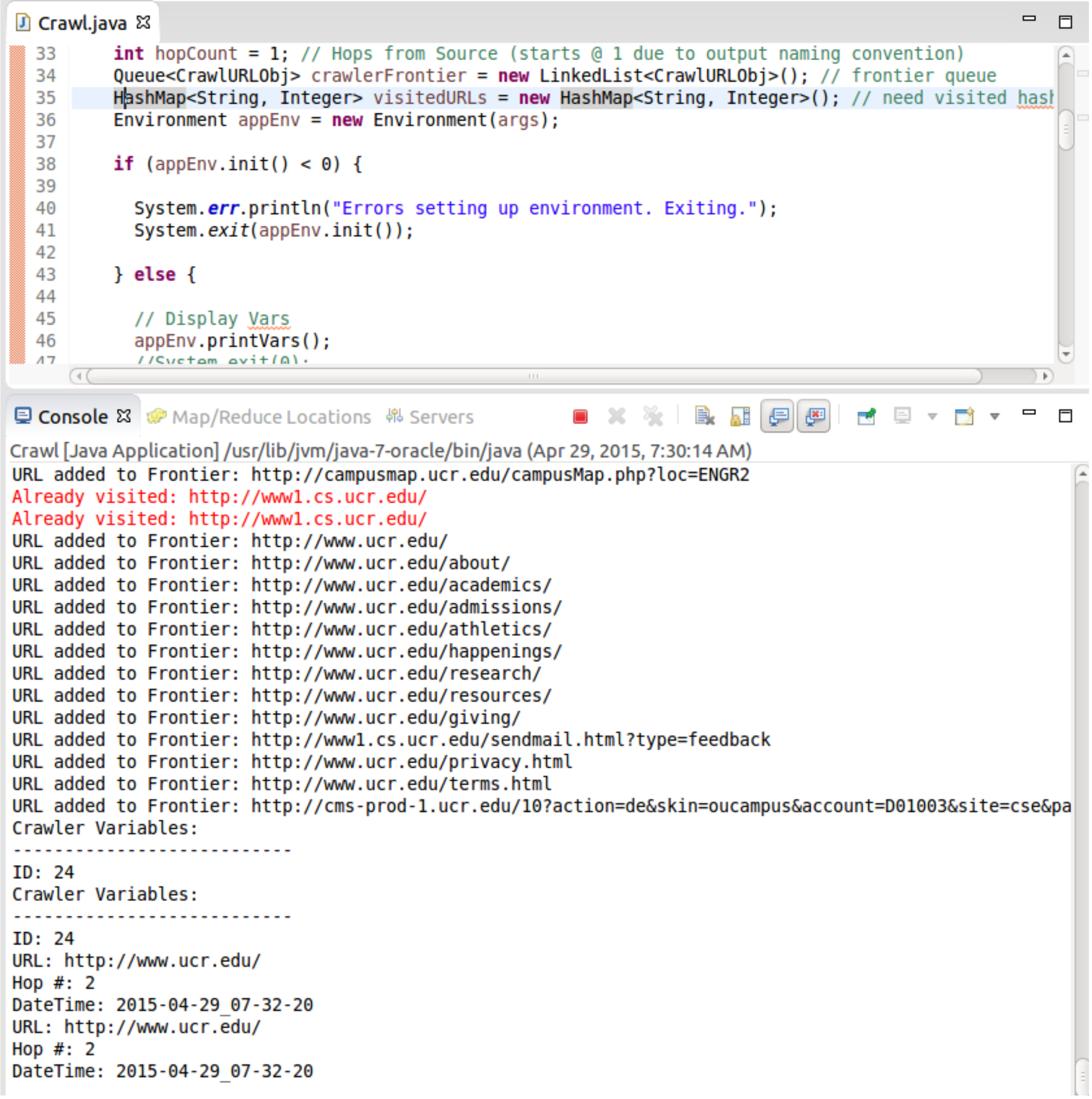
```
THIS URL FLAGGED: pdf
Already visited:
THIS URL FLAGGED: pdf
Already visited:
URL added to Frontier: http://www.ucr.edu/about/
URL added to Frontier: http://www.ucr.edu/academics/
URL added to Frontier: http://www.ucr.edu/admissions/
URL added to Frontier: http://www.ucr.edu/athletics/
URL added to Frontier: http://www.ucr.edu/happenings/
URL added to Frontier: http://www.ucr.edu/research/
URL added to Frontier: http://www.ucr.edu/resources/
URL added to Frontier: http://www.ucr.edu/giving/
URL added to Frontier: http://www1.cs.ucr.edu/sendmail.html?type=feedback
URL added to Frontier: http://www.ucr.edu/privacy.html
URL added to Frontier: http://www.ucr.edu/terms.html
URL added to Frontier: http://cms-prod-1.ucr.edu/10?action=de&skin=oucampus&account=D01003&site=cse&pa
Crawler Variables:
---------------------------
ID: 3
URL: http://www.ucla.edu/
Hop #: 1
DateTime: 2015-04-29_07-30-14
Crawler Variables:
---------------------------
ID: 3
URL: http://www.ucla.edu/
Hop #: 1
DateTime: 2015-04-29_07-30-17
```



```java
    int hopCount = 1; // Hops from Source (starts @ 1 due to output naming convention)
    Queue<CrawlURLObj> crawlerFrontier = new LinkedList<CrawlURLObj>(); // frontier queue
    HashMap<String, Integer> visitedURLs = new HashMap<String, Integer>(); // need visited has
    Environment appEnv = new Environment(args);

    if (appEnv.init() < 0) {

      System.err.println("Errors setting up environment. Exiting.");
      System.exit(appEnv.init());

    } else {

      // Display Vars
      appEnv.printVars();
      //System.exit(0);
```

Crawl [Java Application] /usr/lib/jvm/java-7-oracle/bin/java (Apr 29, 2015, 7:30:14 AM)

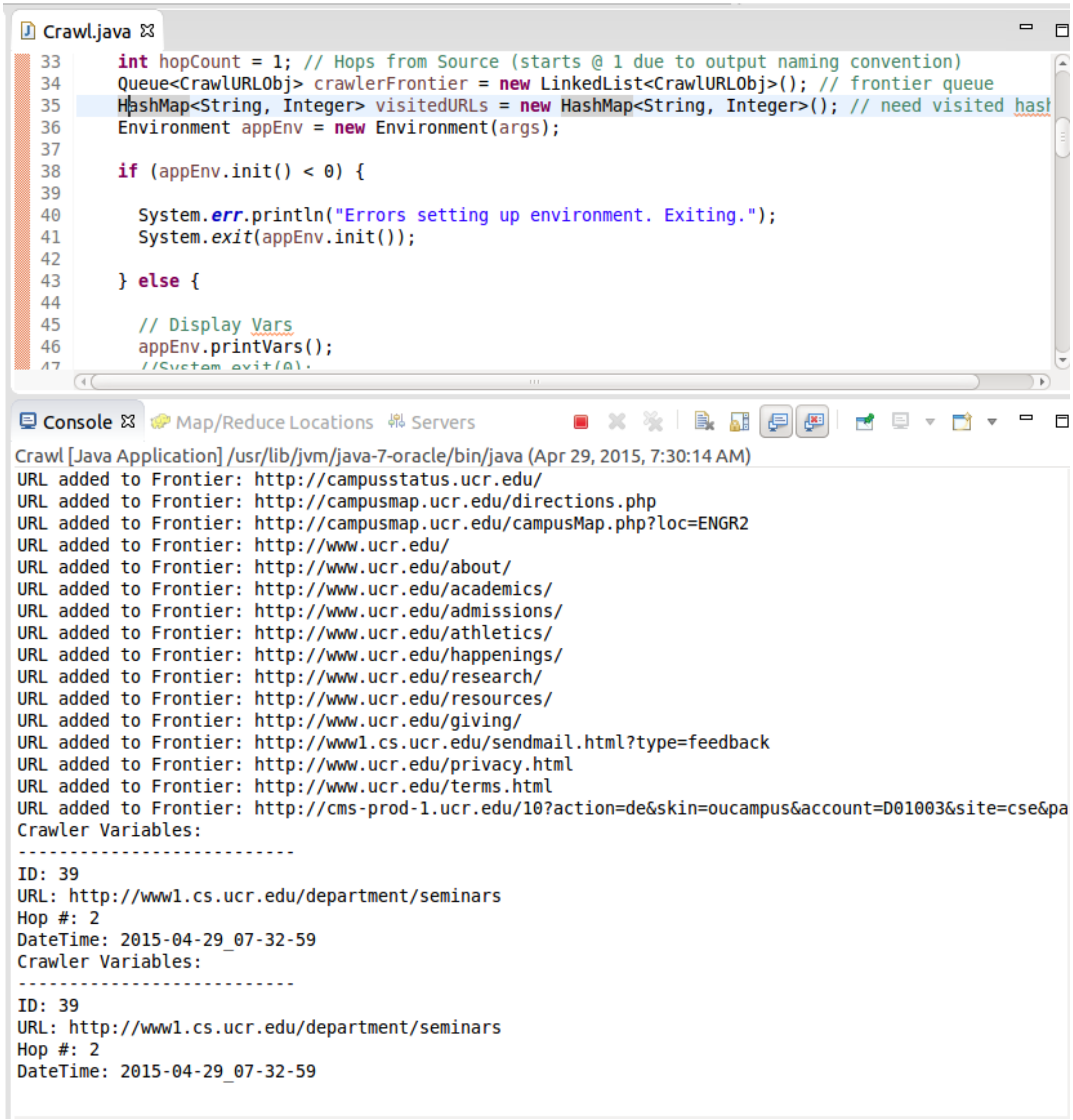
```
URL added to Frontier: http://socialmedia.csusb.edu
URL added to Frontier: http://www.twitter.com/CSUSBNews
URL added to Frontier: http://www.facebook.com/csusb
URL added to Frontier: http://www.youtube.com/csusanbernardino
URL added to Frontier: http://blogs.csusb.edu/coyotecalling
URL added to Frontier: http://news.csusb.edu/category/topstories/feed/
Already visited:
URL added to Frontier: http://www.csusb.edu/disabilityResources.html
URL added to Frontier: http://www.csusb.edu/privacySecurityNotice.html
URL added to Frontier: mailto:webdev@csusb.edu
URL added to Frontier: http://www.calstate.edu/
URL added to Frontier: http://admissions.csusb.edu/contact/disclosure.shtml
URL added to Frontier: http://www.adobe.com/products/flashplayer/
URL added to Frontier: http://www.microsoft.com/downloads/results.aspx?pocId=4289AE77-4CBA-4A75-86F3-9
URL added to Frontier: http://www.adobe.com/products/acrobat/readstep2.html
URL added to Frontier: http://www.quicktime.com/download
Crawler Variables:
---------------------------
ID: 11
Crawler Variables:
---------------------------
URL: http://www.plattcollege.edu/
ID: 11
Hop #: 1
DateTime: 2015-04-29_07-31-58
URL: http://www.plattcollege.edu/
Hop #: 1
DateTime: 2015-04-29_07-31-58
```



```
    int hopCount = 1; // Hops from Source (starts @ 1 due to output naming convention)
    Queue<CrawlURLObj> crawlerFrontier = new LinkedList<CrawlURLObj>(); // frontier queue
    HashMap<String, Integer> visitedURLs = new HashMap<String, Integer>(); // need visited has
    Environment appEnv = new Environment(args);

    if (appEnv.init() < 0) {

      System.err.println("Errors setting up environment. Exiting.");
      System.exit(appEnv.init());

    } else {

      // Display Vars
      appEnv.printVars();
      //System.exit(0);
```

Crawl [Java Application] /usr/lib/jvm/java-7-oracle/bin/java (Apr 29, 2015, 7:30:14 AM)

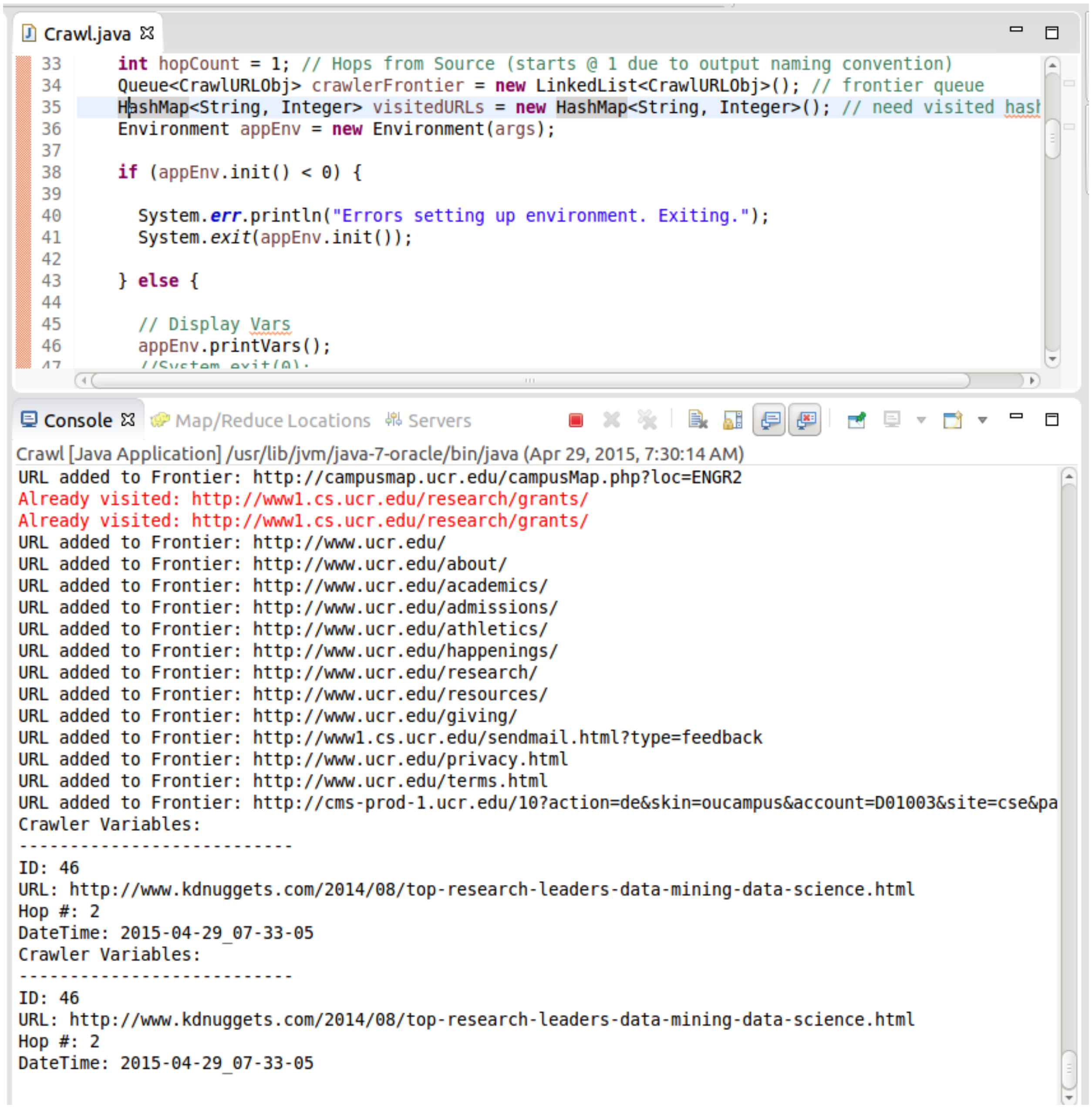
```
URL added to Frontier: http://campusmap.ucr.edu/campusMap.php?loc=ENGR2
Already visited: http://www1.cs.ucr.edu/
Already visited: http://www1.cs.ucr.edu/
URL added to Frontier: http://www.ucr.edu/
URL added to Frontier: http://www.ucr.edu/about/
URL added to Frontier: http://www.ucr.edu/academics/
URL added to Frontier: http://www.ucr.edu/admissions/
URL added to Frontier: http://www.ucr.edu/athletics/
URL added to Frontier: http://www.ucr.edu/happenings/
URL added to Frontier: http://www.ucr.edu/research/
URL added to Frontier: http://www.ucr.edu/resources/
URL added to Frontier: http://www.ucr.edu/giving/
URL added to Frontier: http://www1.cs.ucr.edu/sendmail.html?type=feedback
URL added to Frontier: http://www.ucr.edu/privacy.html
URL added to Frontier: http://www.ucr.edu/terms.html
URL added to Frontier: http://cms-prod-1.ucr.edu/10?action=de&skin=oucampus&account=D01003&site=cse&pa
Crawler Variables:
---------------------------
ID: 24
Crawler Variables:
---------------------------
ID: 24
URL: http://www.ucr.edu/
Hop #: 2
DateTime: 2015-04-29_07-32-20
URL: http://www.ucr.edu/
Hop #: 2
DateTime: 2015-04-29_07-32-20
```



```
    int hopCount = 1; // Hops from Source (starts @ 1 due to output naming convention)
    Queue<CrawlURLObj> crawlerFrontier = new LinkedList<CrawlURLObj>(); // frontier queue
    HashMap<String, Integer> visitedURLs = new HashMap<String, Integer>(); // need visited hash
    Environment appEnv = new Environment(args);

    if (appEnv.init() < 0) {

      System.err.println("Errors setting up environment. Exiting.");
      System.exit(appEnv.init());

    } else {

      // Display Vars
      appEnv.printVars();
      //System.exit(0);
```

Crawl [Java Application] /usr/lib/jvm/java-7-oracle/bin/java (Apr 29, 2015, 7:30:14 AM)

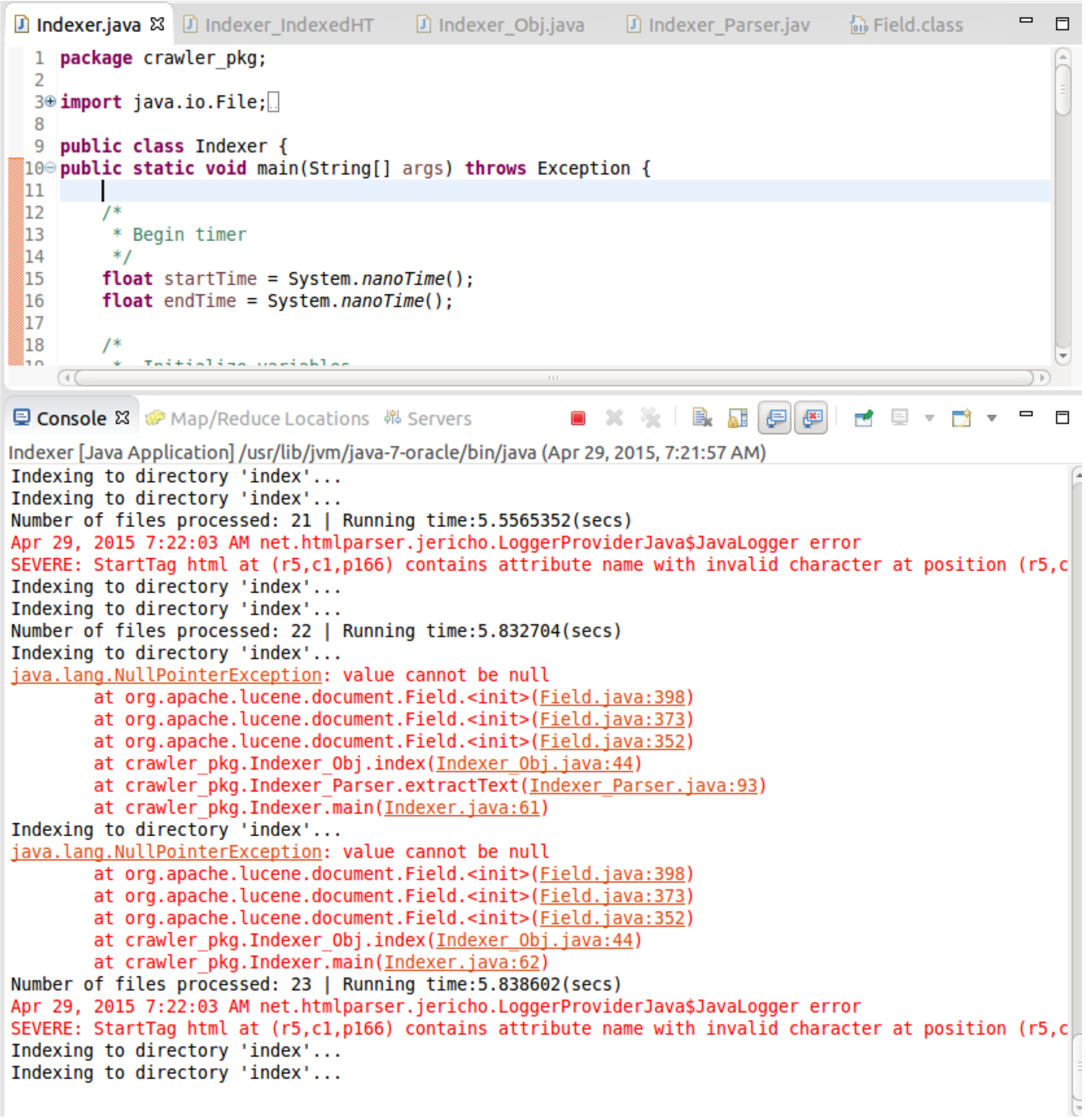
```
URL added to Frontier: http://campusstatus.ucr.edu/
URL added to Frontier: http://campusmap.ucr.edu/directions.php
URL added to Frontier: http://campusmap.ucr.edu/campusMap.php?loc=ENGR2
URL added to Frontier: http://www.ucr.edu/
URL added to Frontier: http://www.ucr.edu/about/
URL added to Frontier: http://www.ucr.edu/academics/
URL added to Frontier: http://www.ucr.edu/admissions/
URL added to Frontier: http://www.ucr.edu/athletics/
URL added to Frontier: http://www.ucr.edu/happenings/
URL added to Frontier: http://www.ucr.edu/research/
URL added to Frontier: http://www.ucr.edu/resources/
URL added to Frontier: http://www.ucr.edu/giving/
URL added to Frontier: http://www1.cs.ucr.edu/sendmail.html?type=feedback
URL added to Frontier: http://www.ucr.edu/privacy.html
URL added to Frontier: http://www.ucr.edu/terms.html
URL added to Frontier: http://cms-prod-1.ucr.edu/10?action=de&skin=oucampus&account=D01003&site=cse&pa
Crawler Variables:
---------------------------
ID: 39
URL: http://www1.cs.ucr.edu/department/seminars
Hop #: 2
DateTime: 2015-04-29_07-32-59
Crawler Variables:
---------------------------
ID: 39
URL: http://www1.cs.ucr.edu/department/seminars
Hop #: 2
DateTime: 2015-04-29_07-32-59
```



```
    int hopCount = 1; // Hops from Source (starts @ 1 due to output naming convention)
    Queue<CrawlURLObj> crawlerFrontier = new LinkedList<CrawlURLObj>(); // frontier queue
    HashMap<String, Integer> visitedURLs = new HashMap<String, Integer>(); // need visited has
    Environment appEnv = new Environment(args);

    if (appEnv.init() < 0) {

      System.err.println("Errors setting up environment. Exiting.");
      System.exit(appEnv.init());

    } else {

      // Display Vars
      appEnv.printVars();
      //System.exit(0);
```

Crawl [Java Application] /usr/lib/jvm/java-7-oracle/bin/java (Apr 29, 2015, 7:30:14 AM)

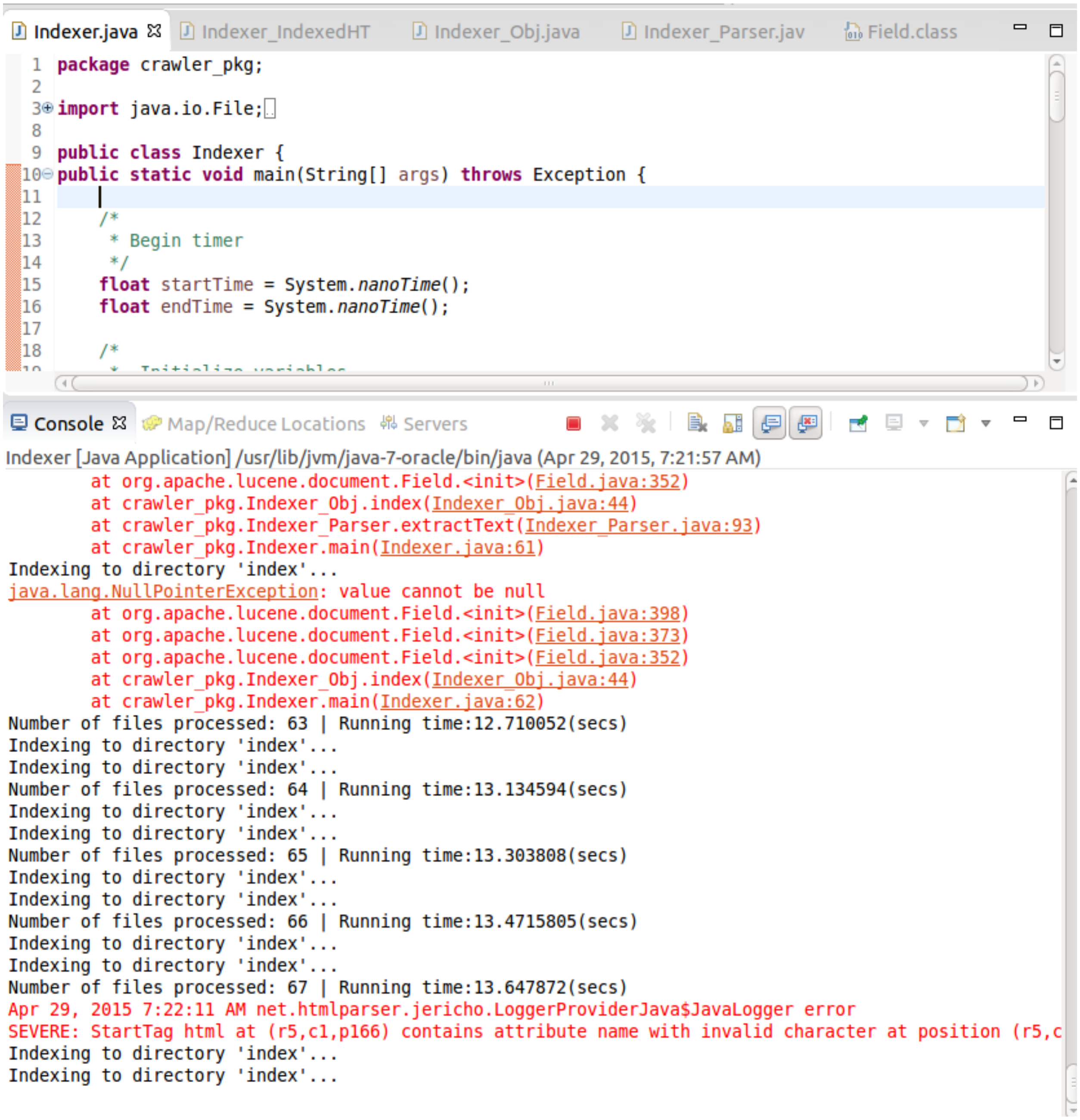
```
URL added to Frontier: http://campusmap.ucr.edu/campusMap.php?loc=ENGR2
Already visited: http://www1.cs.ucr.edu/research/grants/
Already visited: http://www1.cs.ucr.edu/research/grants/
URL added to Frontier: http://www.ucr.edu/
URL added to Frontier: http://www.ucr.edu/about/
URL added to Frontier: http://www.ucr.edu/academics/
URL added to Frontier: http://www.ucr.edu/admissions/
URL added to Frontier: http://www.ucr.edu/athletics/
URL added to Frontier: http://www.ucr.edu/happenings/
URL added to Frontier: http://www.ucr.edu/research/
URL added to Frontier: http://www.ucr.edu/resources/
URL added to Frontier: http://www.ucr.edu/giving/
URL added to Frontier: http://www1.cs.ucr.edu/sendmail.html?type=feedback
URL added to Frontier: http://www.ucr.edu/privacy.html
URL added to Frontier: http://www.ucr.edu/terms.html
URL added to Frontier: http://cms-prod-1.ucr.edu/10?action=de&skin=oucampus&account=D01003&site=cse&pa
Crawler Variables:
---------------------------
ID: 46
URL: http://www.kdnuggets.com/2014/08/top-research-leaders-data-mining-data-science.html
Hop #: 2
DateTime: 2015-04-29_07-33-05
Crawler Variables:
---------------------------
ID: 46
URL: http://www.kdnuggets.com/2014/08/top-research-leaders-data-mining-data-science.html
Hop #: 2
DateTime: 2015-04-29_07-33-05
```



```java
package crawler_pkg;

import java.io.File;

public class Indexer {
public static void main(String[] args) throws Exception {

    /*
     * Begin timer
     */
    float startTime = System.nanoTime();
    float endTime = System.nanoTime();

    /*
     * Initialize variables
```

Indexer [Java Application] /usr/lib/jvm/java-7-oracle/bin/java (Apr 29, 2015, 7:21:57 AM)

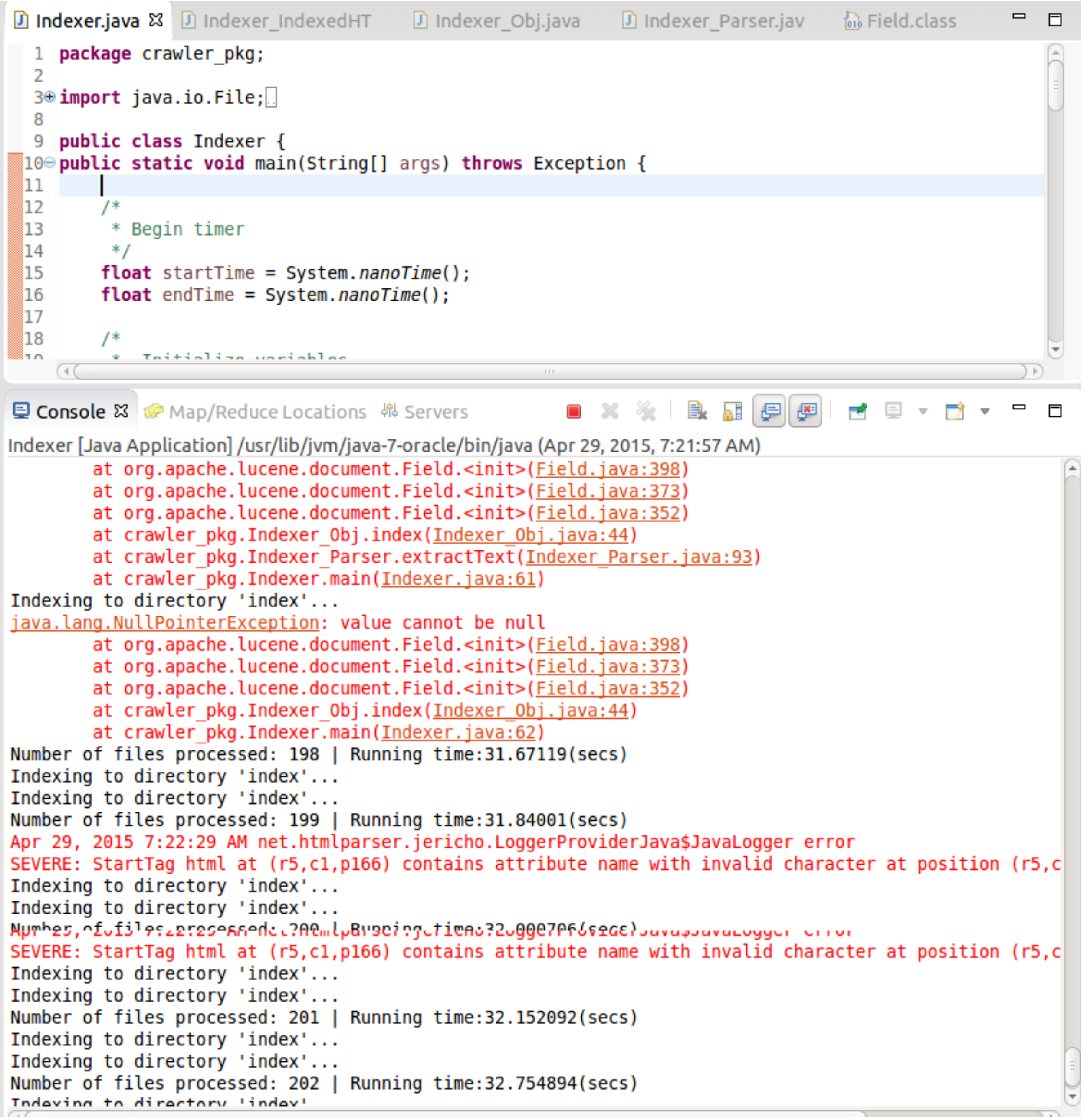
```
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 21 | Running time:5.5565352(secs)
Apr 29, 2015 7:22:03 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 22 | Running time:5.832704(secs)
Indexing to directory 'index'...
java.lang.NullPointerException: value cannot be null
        at org.apache.lucene.document.Field.<init>(Field.java:398)
        at org.apache.lucene.document.Field.<init>(Field.java:373)
        at org.apache.lucene.document.Field.<init>(Field.java:352)
        at crawler_pkg.Indexer_Obj.index(Indexer_Obj.java:44)
        at crawler_pkg.Indexer_Parser.extractText(Indexer_Parser.java:93)
        at crawler_pkg.Indexer.main(Indexer.java:61)
Indexing to directory 'index'...
java.lang.NullPointerException: value cannot be null
        at org.apache.lucene.document.Field.<init>(Field.java:398)
        at org.apache.lucene.document.Field.<init>(Field.java:373)
        at org.apache.lucene.document.Field.<init>(Field.java:352)
        at crawler_pkg.Indexer_Obj.index(Indexer_Obj.java:44)
        at crawler_pkg.Indexer.main(Indexer.java:62)
Number of files processed: 23 | Running time:5.838602(secs)
Apr 29, 2015 7:22:03 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
Indexing to directory 'index'...
```



```
package crawler_pkg;

import java.io.File;

public class Indexer {
public static void main(String[] args) throws Exception {

    /*
     * Begin timer
     */
    float startTime = System.nanoTime();
    float endTime = System.nanoTime();

    /*
     * Initialize variables
```

Indexer [Java Application] /usr/lib/jvm/java-7-oracle/bin/java (Apr 29, 2015, 7:21:57 AM)

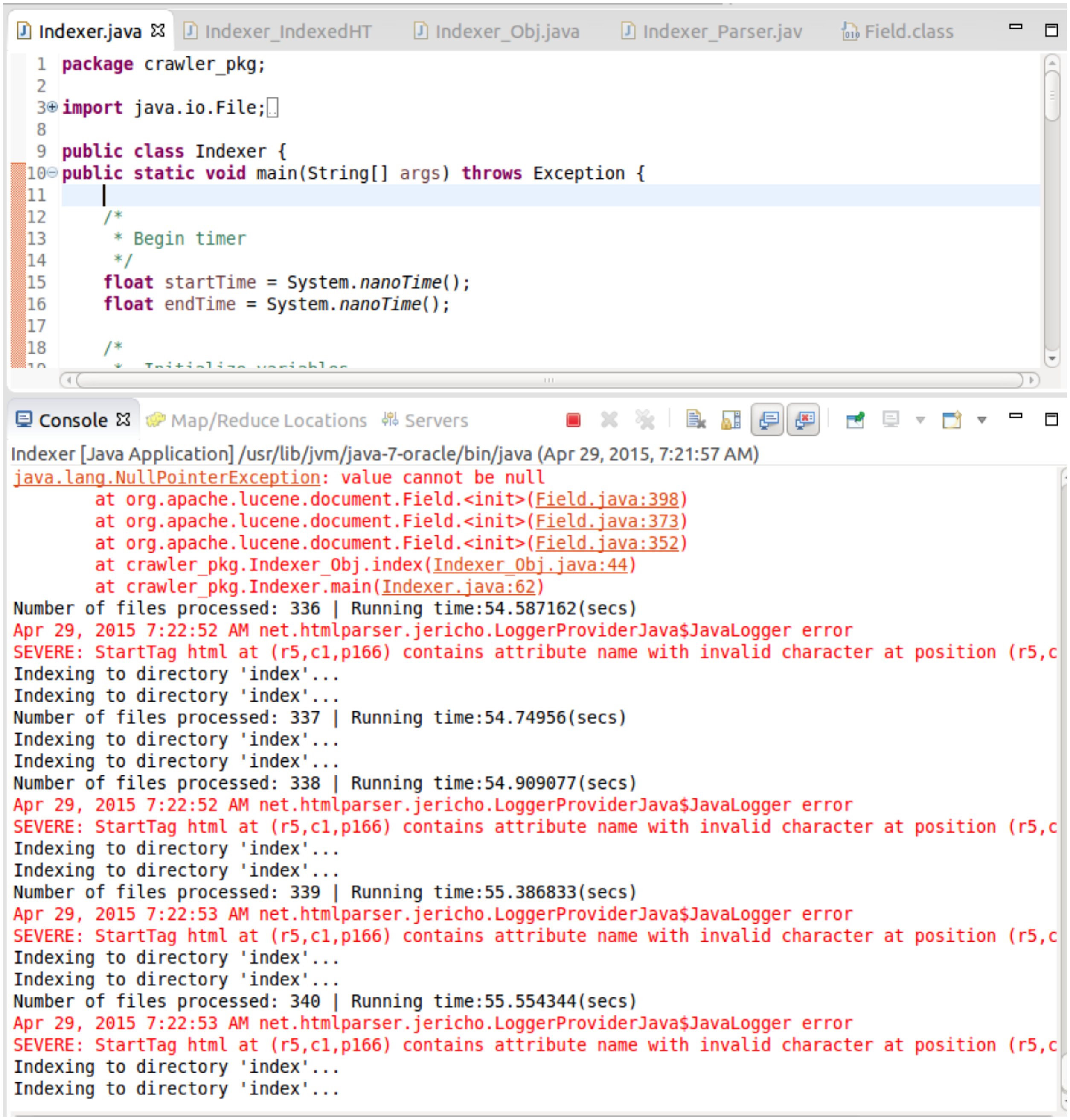
```
        at org.apache.lucene.document.Field.<init>(Field.java:352)
        at crawler_pkg.Indexer_Obj.index(Indexer_Obj.java:44)
        at crawler_pkg.Indexer_Parser.extractText(Indexer_Parser.java:93)
        at crawler_pkg.Indexer.main(Indexer.java:61)
Indexing to directory 'index'...
java.lang.NullPointerException: value cannot be null
        at org.apache.lucene.document.Field.<init>(Field.java:398)
        at org.apache.lucene.document.Field.<init>(Field.java:373)
        at org.apache.lucene.document.Field.<init>(Field.java:352)
        at crawler_pkg.Indexer_Obj.index(Indexer_Obj.java:44)
        at crawler_pkg.Indexer.main(Indexer.java:62)
Number of files processed: 63 | Running time:12.710052(secs)
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 64 | Running time:13.134594(secs)
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 65 | Running time:13.303808(secs)
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 66 | Running time:13.4715805(secs)
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 67 | Running time:13.647872(secs)
Apr 29, 2015 7:22:11 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
Indexing to directory 'index'...
```



```
package crawler_pkg;

import java.io.File;
...
public class Indexer {
public static void main(String[] args) throws Exception {

    /*
     * Begin timer
     */
    float startTime = System.nanoTime();
    float endTime = System.nanoTime();

    /*
     * Initialize variables
```

Indexer [Java Application] /usr/lib/jvm/java-7-oracle/bin/java (Apr 29, 2015, 7:21:57 AM)

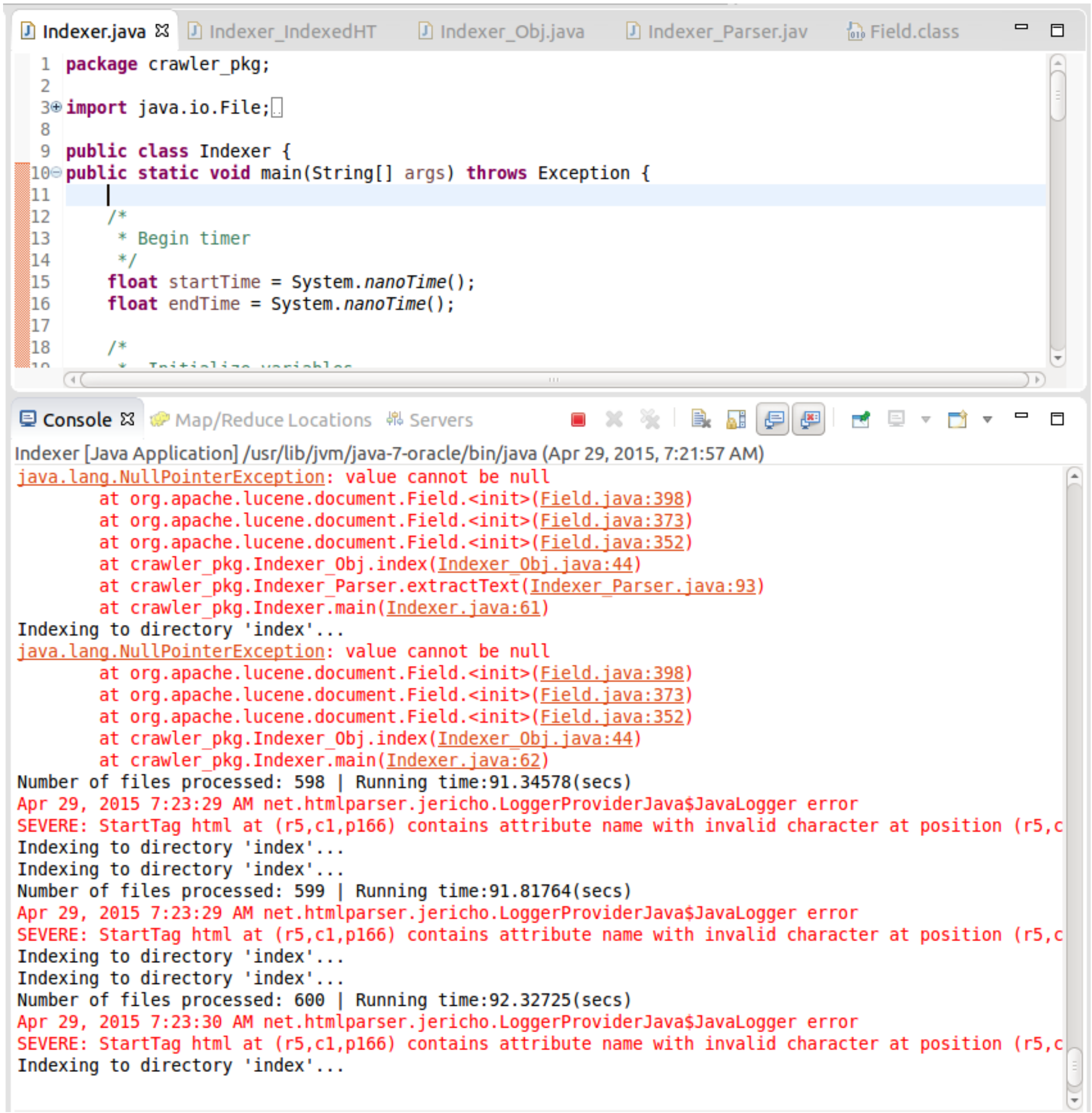
```
        at org.apache.lucene.document.Field.<init>(Field.java:398)
        at org.apache.lucene.document.Field.<init>(Field.java:373)
        at org.apache.lucene.document.Field.<init>(Field.java:352)
        at crawler_pkg.Indexer_Obj.index(Indexer_Obj.java:44)
        at crawler_pkg.Indexer_Parser.extractText(Indexer_Parser.java:93)
        at crawler_pkg.Indexer.main(Indexer.java:61)
Indexing to directory 'index'...
java.lang.NullPointerException: value cannot be null
        at org.apache.lucene.document.Field.<init>(Field.java:398)
        at org.apache.lucene.document.Field.<init>(Field.java:373)
        at org.apache.lucene.document.Field.<init>(Field.java:352)
        at crawler_pkg.Indexer_Obj.index(Indexer_Obj.java:44)
        at crawler_pkg.Indexer.main(Indexer.java:62)
Number of files processed: 198 | Running time:31.67119(secs)
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 199 | Running time:31.84001(secs)
Apr 29, 2015 7:22:29 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 200 | Running time:32.000706(secs)
Apr 29, 2015 7:22:29 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 201 | Running time:32.152092(secs)
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 202 | Running time:32.754894(secs)
Indexing to directory 'index'...
```



```java
package crawler_pkg;

import java.io.File;

public class Indexer {
public static void main(String[] args) throws Exception {
    
    /*
     * Begin timer
     */
    float startTime = System.nanoTime();
    float endTime = System.nanoTime();
    
    /*
     * Initialize variables
```

Indexer [Java Application] /usr/lib/jvm/java-7-oracle/bin/java (Apr 29, 2015, 7:21:57 AM)

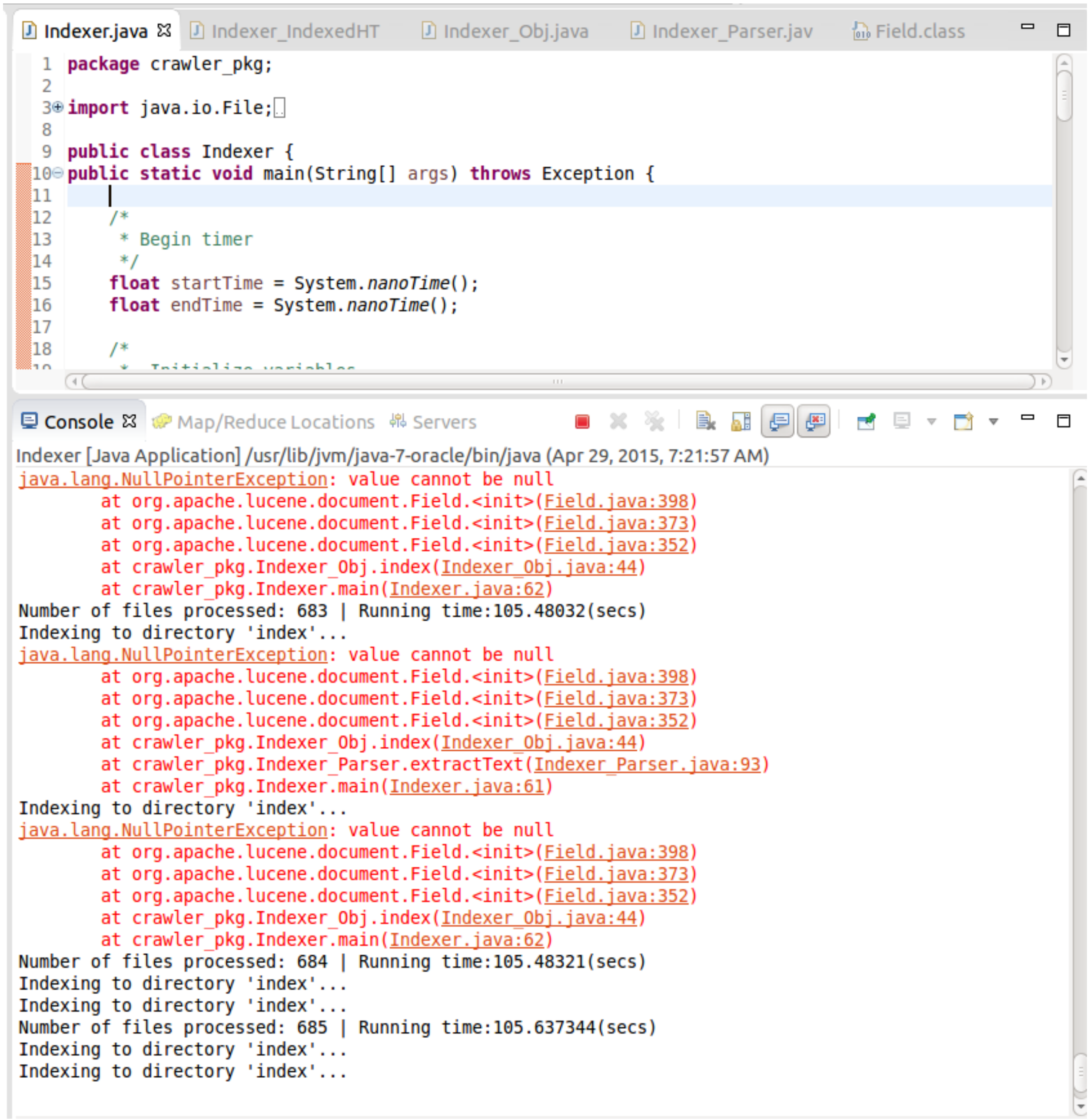
```
java.lang.NullPointerException: value cannot be null
        at org.apache.lucene.document.Field.<init>(Field.java:398)
        at org.apache.lucene.document.Field.<init>(Field.java:373)
        at org.apache.lucene.document.Field.<init>(Field.java:352)
        at crawler_pkg.Indexer_Obj.index(Indexer_Obj.java:44)
        at crawler_pkg.Indexer.main(Indexer.java:62)
Number of files processed: 336 | Running time:54.587162(secs)
Apr 29, 2015 7:22:52 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 337 | Running time:54.74956(secs)
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 338 | Running time:54.909077(secs)
Apr 29, 2015 7:22:52 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 339 | Running time:55.386833(secs)
Apr 29, 2015 7:22:53 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 340 | Running time:55.554344(secs)
Apr 29, 2015 7:22:53 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
Indexing to directory 'index'...
```



```
package crawler_pkg;

import java.io.File;

public class Indexer {
public static void main(String[] args) throws Exception {

    /*
     * Begin timer
     */
    float startTime = System.nanoTime();
    float endTime = System.nanoTime();

    /*
     * Initialize variables
```

Indexer [Java Application] /usr/lib/jvm/java-7-oracle/bin/java (Apr 29, 2015, 7:21:57 AM)

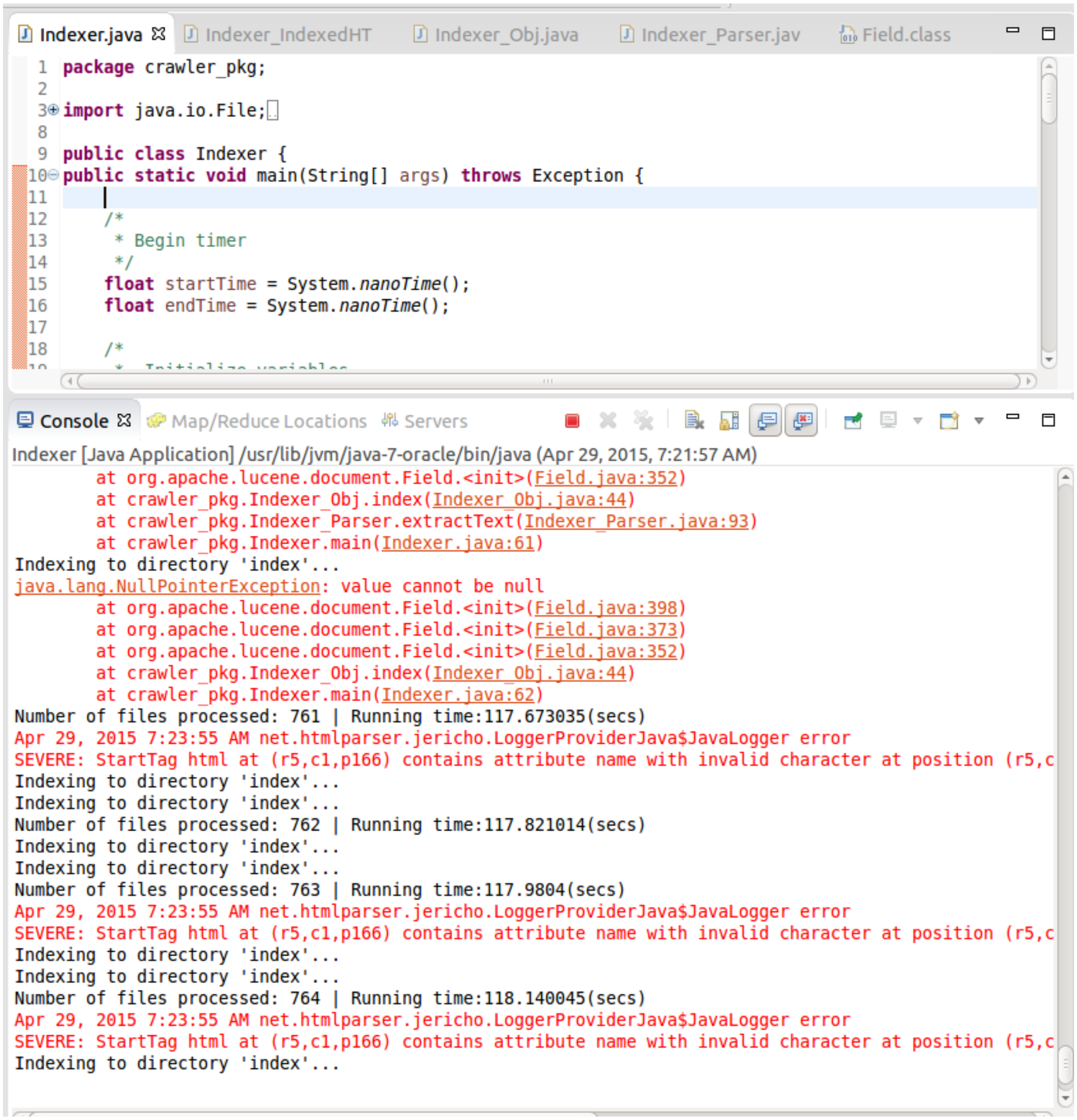
```
java.lang.NullPointerException: value cannot be null
	at org.apache.lucene.document.Field.<init>(Field.java:398)
	at org.apache.lucene.document.Field.<init>(Field.java:373)
	at org.apache.lucene.document.Field.<init>(Field.java:352)
	at crawler_pkg.Indexer_Obj.index(Indexer_Obj.java:44)
	at crawler_pkg.Indexer_Parser.extractText(Indexer_Parser.java:93)
	at crawler_pkg.Indexer.main(Indexer.java:61)
Indexing to directory 'index'...
java.lang.NullPointerException: value cannot be null
	at org.apache.lucene.document.Field.<init>(Field.java:398)
	at org.apache.lucene.document.Field.<init>(Field.java:373)
	at org.apache.lucene.document.Field.<init>(Field.java:352)
	at crawler_pkg.Indexer_Obj.index(Indexer_Obj.java:44)
	at crawler_pkg.Indexer.main(Indexer.java:62)
Number of files processed: 598 | Running time:91.34578(secs)
Apr 29, 2015 7:23:29 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 599 | Running time:91.81764(secs)
Apr 29, 2015 7:23:29 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 600 | Running time:92.32725(secs)
Apr 29, 2015 7:23:30 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
```



```
package crawler_pkg;

import java.io.File;

public class Indexer {
public static void main(String[] args) throws Exception {

    /*
     * Begin timer
     */
    float startTime = System.nanoTime();
    float endTime = System.nanoTime();

    /*
     * Initialize variables
```

Indexer [Java Application] /usr/lib/jvm/java-7-oracle/bin/java (Apr 29, 2015, 7:21:57 AM)

```
java.lang.NullPointerException: value cannot be null
	at org.apache.lucene.document.Field.<init>(Field.java:398)
	at org.apache.lucene.document.Field.<init>(Field.java:373)
	at org.apache.lucene.document.Field.<init>(Field.java:352)
	at crawler_pkg.Indexer_Obj.index(Indexer_Obj.java:44)
	at crawler_pkg.Indexer.main(Indexer.java:62)
Number of files processed: 683 | Running time:105.48032(secs)
Indexing to directory 'index'...
java.lang.NullPointerException: value cannot be null
	at org.apache.lucene.document.Field.<init>(Field.java:398)
	at org.apache.lucene.document.Field.<init>(Field.java:373)
	at org.apache.lucene.document.Field.<init>(Field.java:352)
	at crawler_pkg.Indexer_Obj.index(Indexer_Obj.java:44)
	at crawler_pkg.Indexer_Parser.extractText(Indexer_Parser.java:93)
	at crawler_pkg.Indexer.main(Indexer.java:61)
Indexing to directory 'index'...
java.lang.NullPointerException: value cannot be null
	at org.apache.lucene.document.Field.<init>(Field.java:398)
	at org.apache.lucene.document.Field.<init>(Field.java:373)
	at org.apache.lucene.document.Field.<init>(Field.java:352)
	at crawler_pkg.Indexer_Obj.index(Indexer_Obj.java:44)
	at crawler_pkg.Indexer.main(Indexer.java:62)
Number of files processed: 684 | Running time:105.48321(secs)
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 685 | Running time:105.637344(secs)
Indexing to directory 'index'...
Indexing to directory 'index'...
```



```
package crawler_pkg;

import java.io.File;

public class Indexer {
public static void main(String[] args) throws Exception {

    /*
     * Begin timer
     */
    float startTime = System.nanoTime();
    float endTime = System.nanoTime();

    /*
     *  Initialize variables
```

```
Indexer [Java Application] /usr/lib/jvm/java-7-oracle/bin/java (Apr 29, 2015, 7:21:57 AM)
        at org.apache.lucene.document.Field.<init>(Field.java:352)
        at crawler_pkg.Indexer_Obj.index(Indexer_Obj.java:44)
        at crawler_pkg.Indexer_Parser.extractText(Indexer_Parser.java:93)
        at crawler_pkg.Indexer.main(Indexer.java:61)
Indexing to directory 'index'...
java.lang.NullPointerException: value cannot be null
        at org.apache.lucene.document.Field.<init>(Field.java:398)
        at org.apache.lucene.document.Field.<init>(Field.java:373)
        at org.apache.lucene.document.Field.<init>(Field.java:352)
        at crawler_pkg.Indexer_Obj.index(Indexer_Obj.java:44)
        at crawler_pkg.Indexer.main(Indexer.java:62)
Number of files processed: 761 | Running time:117.673035(secs)
Apr 29, 2015 7:23:55 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 762 | Running time:117.821014(secs)
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 763 | Running time:117.9804(secs)
Apr 29, 2015 7:23:55 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
Indexing to directory 'index'...
Number of files processed: 764 | Running time:118.140045(secs)
Apr 29, 2015 7:23:55 AM net.htmlparser.jericho.LoggerProviderJava$JavaLogger error
SEVERE: StartTag html at (r5,c1,p166) contains attribute name with invalid character at position (r5,c
Indexing to directory 'index'...
```

THE END. THANK YOU.