Web Document Index and Search

Code and Documentation Written By
Steven Fan

Overview of System

Both the indexer and the searcher files are written in Java and mainly uses the Lucene architecture for building indexes and searching through them. The indexer utilizes the jsoup library to extract information from HTML files. The searcher utilizes the Spring program to allow configuration with HTML to build an interactive web searcher. The following documentation will document the source code itself. For convenience, Java files *Index.java* and *Search.java* are stored in the *files/* directory separate of the Java package hierarchy of the program.

Index.java

Initialize Code Block

```
System.out.print("Input index directory: ");
Scanner scanner = new Scanner(System.in);
String path = scanner.nextLine();
Analyzer analyzer = new StandardAnalyzer();
Directory directory = FSDirectory.open(Paths.get(path));
IndexWriterConfig config = new IndexWriterConfig(analyzer);
IndexWriter indexWriter = new IndexWriter(directory, config);
```

This code section initializes the Lucene class variables. The analyzer is selected as the *StandardAnalyzer* which parses the input based on the standard, such as taking into account capitalization, stop words, etc. The index will be written to the index directory folder provided by the user.

Extraction and Indexing

Lines 31 through 115 of the code consists mainly of extraction of HTML information using *jsoup* libraries. The program extracts data based on the UTF-8 code. The code block extracts an HTML’s title, metadata, header, URL, and content. The program searches for HTML files in the data directory folder provided by the user.

```
Document doc = new Document();
try
{
	doc.add(new TextField("title", title, Field.Store.YES));
	doc.add(new TextField("metadata", metadata, Field.Store.YES));
	doc.add(new TextField("header", header, Field.Store.YES));
```

```
        doc.add(new StringField("url", url, Field.Store.YES));
        doc.add(new TextField("content", content, Field.Store.YES));
        indexWriter.addDocument(doc);
}
```

This code block adds the extracted information from the HTML file into a document where the *indexWriter* can then create an index based on this document.

The extraction and indexing of HTML files is done through a loop, where the loop terminates when the expected file is not found. The program searches for HTML files in the data directory folder and expects the HTML files to be named exactly like the data directory folder followed by a number. For example, if the user states the data directory folder is named *data/*, then the first HTML file to parse must be *data/data0.html*. The program iterates through each number. Thus, after extracting and parsing *data/data0.html*, the program attempts to extract and parse *data/data1.html*. The syntax of these HTML file names are based on the HTML files generated by the web crawler in Part A.

Search.java

Class and Function Start

```
@RestController
@CrossOrigin("*")
public class Search
{
        @RequestMapping(value = "/search", method = RequestMethod.GET)
        public @ResponseBody String search(@RequestParam(required = false,
                defaultValue = "") String query) throws IOException,\
                ParseException
        {
```

The *@RestController* and *@CrossOrigin* designates types as controllers and allows sending data back to the origin connection (HTML search engine interface). The *@RequestMapping* allow execution of the program’s *search()* function by appending “/search” after the URL of the server the program is running on. The program expects an HTTP GET signal from the client program. The search program also accepts parameters which will become the query for the searching of the index as denoted by *@RequestParam*.

Initialize Code Block

```
Analyzer analyzer = new StandardAnalyzer();
Directory directory = FSDirectory.open(Paths.get("index/"));
DirectoryReader directoryReader = DirectoryReader.open(directory);
```

```
IndexSearcher indexSearcher = new IndexSearcher(directoryReader);
String[] fields = {"title", "metadata", "header", "url", "content"};
Map<String, Float> boosts = new HashMap<>();
boosts.put(fields[0], 1.0f);
boosts.put(fields[1], 0.5f);
boosts.put(fields[2], 0.5f);
boosts.put(fields[3], 0.25f);
boosts.put(fields[4], 0.25f);
MultiFieldQueryParser parser = new MultiFieldQueryParser(fields, analyzer,\
      boosts);
```

This code block initializes the Lucene class variables. Once again, the *StandardAnalyzer* is selected as the default analyzer to use. The directory opens the index directory folder, which is expected to be *index/* in this version of the code. The code also initializes score weights for the features of the index documents. The program thus reads the query and utilizes the analyzer and score weights in order to determine the document scoring and fetches the documents that best fit the query.

Scoring and Fetching

```
Query parsed = parser.parse(query);
ScoreDoc[] score = indexSearcher.search(parsed, 10).scoreDocs;
```

This code segment simply scores the documents based on the query and feature weights and then fetches the top ten documents based on their score.

Lines 53 through 95 of the code block simply extracts the document data from the list of top ten documents and generates snippets for each document. The data is then passed into a *String* variable and returned, thus sending the data back to the origin connection.

engine.html

The purpose of the *engine.html* file is to create an interactive interface to send queries and receive documents. The HTML file displays snippets of the returned documents and also provides a hyperlink to the original web page that was fetched from the web crawler. The HTML file allows a cleaner and simpler method in sending and receiving data as a web search program. Otherwise, the user will have to utilize the console in order to execute searches. The back end of the web search interface is built utilizing Spring and Maven. Spring and Maven aid in setting up a local host server to which the *Search.java* file is run. An external browser, such as *engine.html*, which acts as the front end, can then communicate with *Search.java* and search documents. The *engine.html* file is built using HTML, CSS, Javascript, and AngularJS.

Limitations

The index and search programs do contain limitations. One limitation is the rigidness of the scoring, in which once the boosting of the score has been set, the architecture cannot be changed without reindexing the entirety of the collection. Another limitation is the inability of the search algorithm to establish semantic meaning on the language of the query. For example, the query “white house” cannot connect the semantic meaning of the two words together to indicate the meaning of the actual White House in Washington D.C. In extension, the program cannot derive definitions of the word. It simply fetches strings that match the terms of the query. For example, it cannot distinguish that “money” and “capital” are synonyms. Furthermore, it cannot distinguish whether the word “bank” refers to an establishment where money is stored or a slope next to a body of water. Another limitation of the program is its inability to query snippets of URLs, especially with the incorporation of non-alphanumeric symbols.

Instructions

The program is built using Eclipse. It is recommended to run the indexer on Eclipse due to the hierarchy nature of Java. Since *Index.java* uses Lucene and jsoup, the necessary JAR files where the classes are stored in are necessary. The JAR files are stored in the *class/* directory. Before running *Index.java*, the data collection of the HTML documents must be present in the root directory (directory that contains *bin/*, *class/*, *files/*, *src/*, *mvnw.cmd*, etc.) in its designated data directory folder. The data directory folder and the HTML files within it must share the same name format mentioned above. Due to the size of the collection, I have omitted it in my upload. Running the *Index.java* file will build the indexes within the index directory folder designated by the user.

To run the search program, the user can simply run *Search.java* and communicate through the console. However, that is not as neat as a web interface. In order to run Maven, your environment variable JAVA_HOME must match the location of your Java installation. Once all initial setups are finished, run the local host server by executing the command in the root directory:

```
./mvn spring-boot:run
```

or

```
mvnw.cmd spring-boot:run
```

depending on the terminal used

Note: GitHub has altered line break types in the files of the repository. This may affect the compilation or execution of the program in unintended ways.

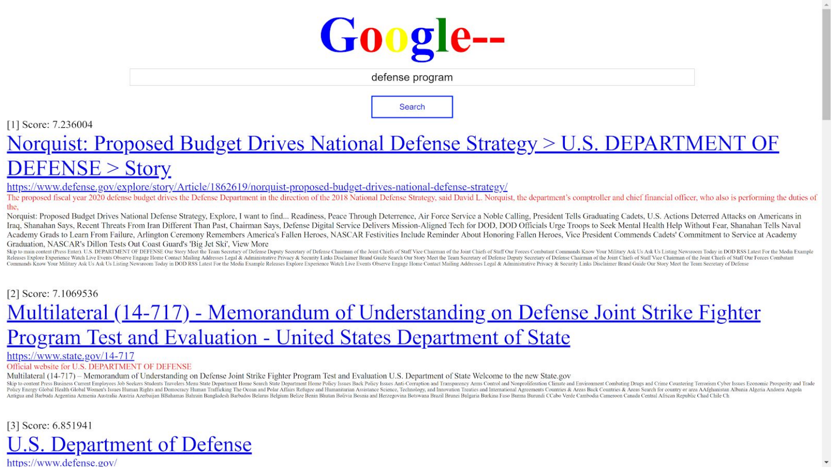
Once the server is running, simply open the *engine.html* located in *files/* to open a web browser. The user can then write a query in the search box and press search in order to execute a search through the index.

The user can simply write a normal query, or search through a specific document feature as shown in this example:

```
title:Trade
```

Note that in order for the search program to work, an index must have been built. Currently, *Search.java* is hardcoded to fetch indexes in the *index/* directory. Due to the size of the index, I have omitted it in my upload.

Screenshots



defense program

[1] Score: 7.236004

Norquist: Proposed Budget Drives National Defense Strategy > U.S. DEPARTMENT OF DEFENSE > Story

https://www.defense.gov/explore/story/Article/1862619/norquist-proposed-budget-drives-national-defense-strategy/

The proposed fiscal year 2020 defense budget drives the Defense Department in the direction of the 2018 National Defense Strategy, said David L. Norquist, the department's comptroller and chief financial officer, who also is performing the duties of the,

Norquist: Proposed Budget Drives National Defense Strategy, Explore, I want to find... Readiness, Peace Through Deterrence, Air Force Service a Noble Calling, President Tells Graduating Cadets, U.S. Actions Deterred Attacks on Americans in Iraq, Shanahan Says, Recent Threats From Iran Different Than Past, Chairman Says, Defense Digital Service Delivers Mission-Aligned Tech for DOD, DOD Officials Urge Troops to Seek Mental Health Help Without Fear, Shanahan Tells Naval Academy Grads to Learn From Failure, Arlington Ceremony Remembers America's Fallen Heroes, NASCAR Festivities Include Reminder About Honoring Fallen Heroes, Vice President Commends Cadets' Commitment to Service at Academy Graduation, NASCAR's Dillon Tests Out Coast Guard's 'Big Jet Ski', View More

Skip to main content (Press Enter). U.S. DEPARTMENT OF DEFENSE Our Story Meet the Team Secretary of Defense Deputy Secretary of Defense Chairman of the Joint Chiefs of Staff Vice Chairman of the Joint Chiefs of Staff Our Forces Combatant Commands Know Your Military Ask Us Ask Us Listing Newsroom Today in DOD RSS Latest For the Media Example Releases Explore Experience Watch Live Events Observe Engage Home Contact Mailing Addresses Legal & Administrative Privacy & Security Links Disclaimer Brand Guide Search Our Story Meet the Team Secretary of Defense Deputy Secretary of Defense Chairman of the Joint Chiefs of Staff Vice Chairman of the Joint Chiefs of Staff Our Forces Combatant Commands Know Your Military Ask Us Ask Us Listing Newsroom Today in DOD RSS Latest For the Media Example Releases Explore Experience Watch Live Events Observe Engage Home Contact Mailing Addresses Legal & Administrative Privacy & Security Links Disclaimer Brand Guide Our Story Meet the Team Secretary of Defense

[2] Score: 7.1069536

Multilateral (14-717) - Memorandum of Understanding on Defense Joint Strike Fighter Program Test and Evaluation - United States Department of State

https://www.state.gov/14-717

Official website for U.S. DEPARTMENT OF DEFENSE

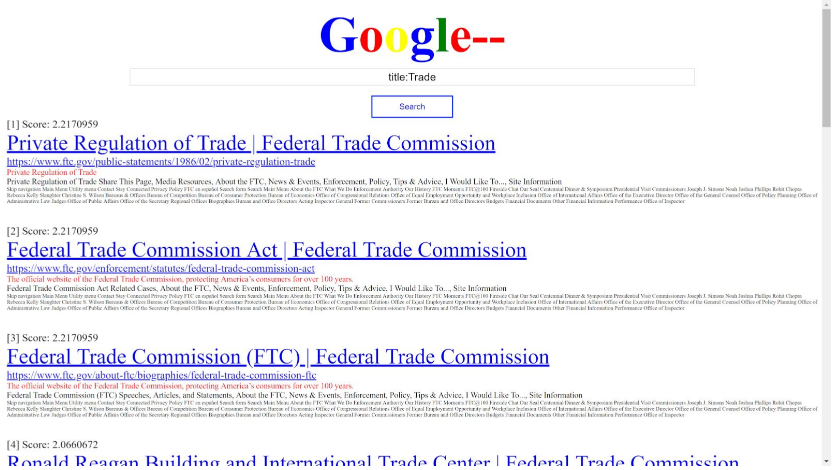
Multilateral (14-717) – Memorandum of Understanding on Defense Joint Strike Fighter Program Test and Evaluation U.S. Department of State Welcome to the new State.gov

Skip to content Press Business Current Employees Job Seekers Students Travelers Menu State Department Home Search State Department Home Policy Issues Back Policy Issues Anti-Corruption and Transparency Arms Control and Nonproliferation Climate and Environment Combating Drugs and Crime Countering Terrorism Cyber Issues Economic Prosperity and Trade Policy Energy Global Health Global Women's Issues Human Rights and Democracy Human Trafficking The Ocean and Polar Affairs Refugee and Humanitarian Assistance Science, Technology, and Innovation Treaties and International Agreements Countries & Areas Back Countries & Areas Search for country or area AAfghanistan Albania Algeria Andorra Angola Antigua and Barbuda Argentina Armenia Australia Austria Azerbaijan BBahamas Bahrain Bangladesh Barbados Belarus Belgium Belize Benin Bhutan Bolivia Bosnia and Herzegovina Botswana Brazil Brunei Bulgaria Burkina Faso Burma Burundi CCabo Verde Cambodia Cameroon Canada Central African Republic Chad Chile Ch

[3] Score: 6.851941

U.S. Department of Defense

https://www.defense.gov/



title:Trade

[1] Score: 2.2170959

Private Regulation of Trade | Federal Trade Commission

https://www.ftc.gov/public-statements/1986/02/private-regulation-trade

Private Regulation of Trade

Private Regulation of Trade Share This Page, Media Resources, About the FTC, News & Events, Enforcement, Policy, Tips & Advice, I Would Like To..., Site Information

Skip navigation Main Menu Utility menu Contact Stay Connected Privacy Policy FTC en español Search form Search Main Menu About the FTC What We Do Enforcement Authority Our History FTC Moments FTC@100 Fireside Chat Our Seal Centennial Dinner & Symposium Presidential Visit Commissioners Joseph J. Simons Noah Joshua Phillips Rohit Chopra Rebecca Kelly Slaughter Christine S. Wilson Bureaus & Offices Bureau of Competition Bureau of Consumer Protection Bureau of Economics Office of Congressional Relations Office of Equal Employment Opportunity and Workplace Inclusion Office of International Affairs Office of the Executive Director Office of the General Counsel Office of Policy Planning Office of Administrative Law Judges Office of Public Affairs Office of the Secretary Regional Offices Biographies Bureau and Office Directors Acting Inspector General Former Commissioners Former Bureau and Office Directors Budgets Financial Documents Other Financial Information Performance Office of Inspector

[2] Score: 2.2170959

Federal Trade Commission Act | Federal Trade Commission

https://www.ftc.gov/enforcement/statutes/federal-trade-commission-act

The official website of the Federal Trade Commission, protecting America's consumers for over 100 years.

Federal Trade Commission Act Related Cases, About the FTC, News & Events, Enforcement, Policy, Tips & Advice, I Would Like To..., Site Information

Skip navigation Main Menu Utility menu Contact Stay Connected Privacy Policy FTC en español Search form Search Main Menu About the FTC What We Do Enforcement Authority Our History FTC Moments FTC@100 Fireside Chat Our Seal Centennial Dinner & Symposium Presidential Visit Commissioners Joseph J. Simons Noah Joshua Phillips Rohit Chopra Rebecca Kelly Slaughter Christine S. Wilson Bureaus & Offices Bureau of Competition Bureau of Consumer Protection Bureau of Economics Office of Congressional Relations Office of Equal Employment Opportunity and Workplace Inclusion Office of International Affairs Office of the Executive Director Office of the General Counsel Office of Policy Planning Office of Administrative Law Judges Office of Public Affairs Office of the Secretary Regional Offices Biographies Bureau and Office Directors Acting Inspector General Former Commissioners Former Bureau and Office Directors Budgets Financial Documents Other Financial Information Performance Office of Inspector

[3] Score: 2.2170959

Federal Trade Commission (FTC) | Federal Trade Commission

https://www.ftc.gov/about-ftc/biographies/federal-trade-commission-ftc

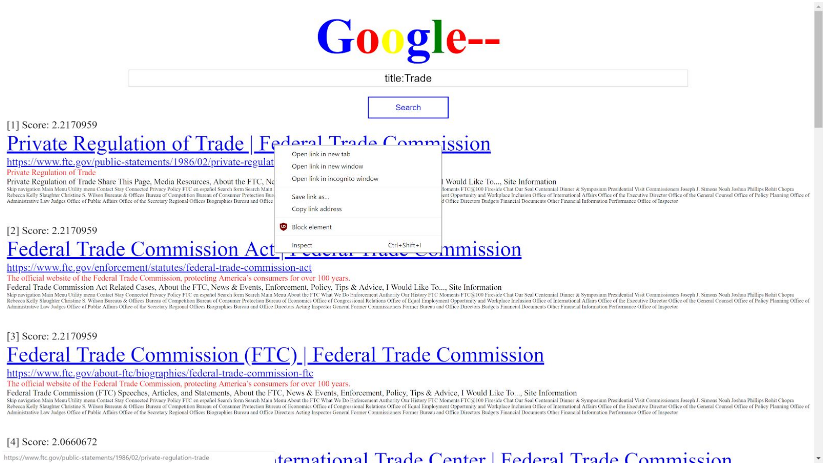
The official website of the Federal Trade Commission, protecting America's consumers for over 100 years.

Federal Trade Commission (FTC) Speeches, Articles, and Statements, About the FTC, News & Events, Enforcement, Policy, Tips & Advice, I Would Like To..., Site Information

Skip navigation Main Menu Utility menu Contact Stay Connected Privacy Policy FTC en español Search form Search Main Menu About the FTC What We Do Enforcement Authority Our History FTC Moments FTC@100 Fireside Chat Our Seal Centennial Dinner & Symposium Presidential Visit Commissioners Joseph J. Simons Noah Joshua Phillips Rohit Chopra Rebecca Kelly Slaughter Christine S. Wilson Bureaus & Offices Bureau of Competition Bureau of Consumer Protection Bureau of Economics Office of Congressional Relations Office of Equal Employment Opportunity and Workplace Inclusion Office of International Affairs Office of the Executive Director Office of the General Counsel Office of Policy Planning Office of Administrative Law Judges Office of Public Affairs Office of the Secretary Regional Offices Biographies Bureau and Office Directors Acting Inspector General Former Commissioners Former Bureau and Office Directors Budgets Financial Documents Other Financial Information Performance Office of Inspector

[4] Score: 2.0660672

Ronald Reagan Building and International Trade Center | Federal Trade Commission



[1] Score: 2.2170959

Private Regulation of Trade | Federal Trade Commission

https://www.ftc.gov/public-statements/1986/02/private-regulat

Private Regulation of Trade

Private Regulation of Trade Share This Page, Media Resources, About the FTC, Ne I Would Like To..., Site Information

Skip navigation Main Menu Utility menu Contact Stay Connected Privacy Policy FTC en español Search form Search Main Moments FTC@100 Fireside Chat Our Seal Centennial Dinner & Symposium Presidential Visit Commissioners Joseph J. Simons Noah Joshua Phillips Rohit Chopra Rebecca Kelly Slaughter Christine S. Wilson Bureaus & Offices Bureau of Competition Bureau of Consumer Protection Bur ent Opportunity and Workplace Inclusion Office of International Affairs Office of the Executive Director Office of the General Counsel Office of Policy Planning Office of Administrative Law Judges Office of Public Affairs Office of the Secretary Regional Offices Biographies Bureau and Office d Office Directors Budgets Financial Documents Other Financial Information Performance Office of Inspector

[2] Score: 2.2170959

Federal Trade Commission Act | Federal Trade Commission

https://www.ftc.gov/enforcement/statutes/federal-trade-commission-act

The official website of the Federal Trade Commission, protecting America's consumers for over 100 years.

Federal Trade Commission Act Related Cases, About the FTC, News & Events, Enforcement, Policy, Tips & Advice, I Would Like To..., Site Information

Skip navigation Main Menu Utility menu Contact Stay Connected Privacy Policy FTC en español Search form Search Main Menu About the FTC What We Do Enforcement Authority Our History FTC Moments FTC@100 Fireside Chat Our Seal Centennial Dinner & Symposium Presidential Visit Commissioners Joseph J. Simons Noah Joshua Phillips Rohit Chopra Rebecca Kelly Slaughter Christine S. Wilson Bureaus & Offices Bureau of Competition Bureau of Consumer Protection Bureau of Economics Office of Congressional Relations Office of Equal Employment Opportunity and Workplace Inclusion Office of International Affairs Office of the Executive Director Office of the General Counsel Office of Policy Planning Office of Administrative Law Judges Office of Public Affairs Office of the Secretary Regional Offices Biographies Bureau and Office Directors Acting Inspector General Former Commissioners Former Bureau and Office Directors Budgets Financial Documents Other Financial Information Performance Office of Inspector

[3] Score: 2.2170959

Federal Trade Commission (FTC) | Federal Trade Commission

https://www.ftc.gov/about-ftc/biographies/federal-trade-commission-ftc

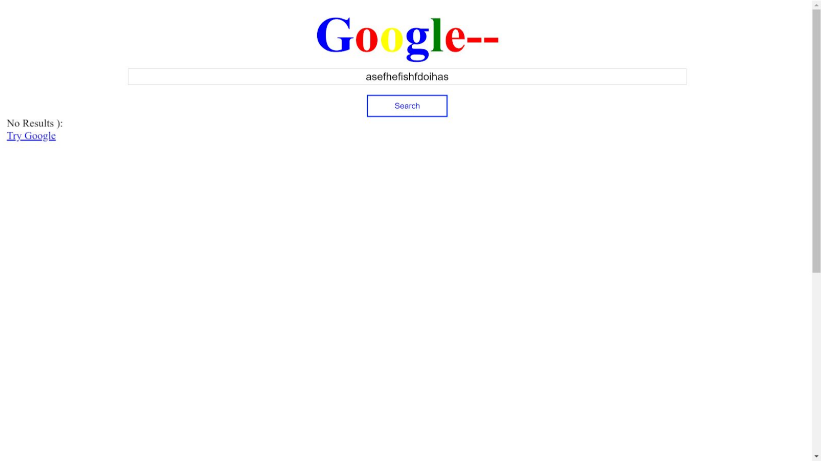
The official website of the Federal Trade Commission, protecting America's consumers for over 100 years.

Federal Trade Commission (FTC) Speeches, Articles, and Statements, About the FTC, News & Events, Enforcement, Policy, Tips & Advice, I Would Like To..., Site Information

Skip navigation Main Menu Utility menu Contact Stay Connected Privacy Policy FTC en español Search form Search Main Menu About the FTC What We Do Enforcement Authority Our History FTC Moments FTC@100 Fireside Chat Our Seal Centennial Dinner & Symposium Presidential Visit Commissioners Joseph J. Simons Noah Joshua Phillips Rohit Chopra Rebecca Kelly Slaughter Christine S. Wilson Bureaus & Offices Bureau of Competition Bureau of Consumer Protection Bureau of Economics Office of Congressional Relations Office of Equal Employment Opportunity and Workplace Inclusion Office of International Affairs Office of the Executive Director Office of the General Counsel Office of Policy Planning Office of Administrative Law Judges Office of Public Affairs Office of the Secretary Regional Offices Biographies Bureau and Office Directors Acting Inspector General Former Commissioners Former Bureau and Office Directors Budgets Financial Documents Other Financial Information Performance Office of Inspector

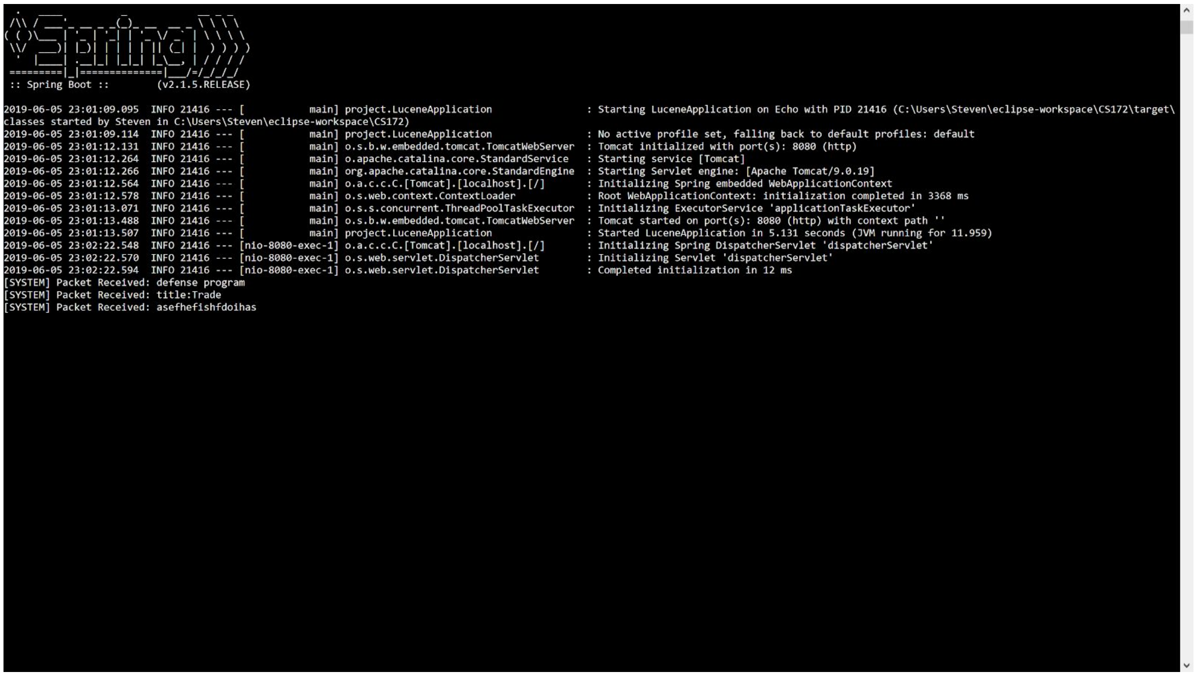
[4] Score: 2.0660672

ternational Trade Center | Federal Trade Commission



No Results):

Try Google



```
  .   ____          _            __ _ _
 /\\ / ___'_ __ _ _(_)_ __  __ _ \ \ \ \
( ( )\___ | '_ | '_| | '_ \/ _` | \ \ \ \
 \\/  ___)| |_)| | | | | || (_| |  ) ) ) )
  '  |____| .__|_| |_|_| |_\__, | / / / /
 =========|_|==============|___/=/_/_/_/
 :: Spring Boot ::        (v2.1.5.RELEASE)

2019-06-05 23:01:09.095  INFO 21416 --- [           main] project.LuceneApplication                : Starting LuceneApplication on Echo with PID 21416 (C:\Users\Steven\eclipse-workspace\CS172\target\
classes started by Steven in C:\Users\Steven\eclipse-workspace\CS172)
2019-06-05 23:01:09.114  INFO 21416 --- [           main] project.LuceneApplication                : No active profile set, falling back to default profiles: default
2019-06-05 23:01:12.131  INFO 21416 --- [           main] o.s.b.w.embedded.tomcat.TomcatWebServer  : Tomcat initialized with port(s): 8080 (http)
2019-06-05 23:01:12.264  INFO 21416 --- [           main] o.apache.catalina.core.StandardService   : Starting service [Tomcat]
2019-06-05 23:01:12.266  INFO 21416 --- [           main] org.apache.catalina.core.StandardEngine  : Starting Servlet engine: [Apache Tomcat/9.0.19]
2019-06-05 23:01:12.564  INFO 21416 --- [           main] o.a.c.c.C.[Tomcat].[localhost].[/]       : Initializing Spring embedded WebApplicationContext
2019-06-05 23:01:12.578  INFO 21416 --- [           main] o.s.web.context.ContextLoader            : Root WebApplicationContext: initialization completed in 3368 ms
2019-06-05 23:01:13.071  INFO 21416 --- [           main] o.s.s.concurrent.ThreadPoolTaskExecutor  : Initializing ExecutorService 'applicationTaskExecutor'
2019-06-05 23:01:13.488  INFO 21416 --- [           main] o.s.b.w.embedded.tomcat.TomcatWebServer  : Tomcat started on port(s): 8080 (http) with context path ''
2019-06-05 23:01:13.507  INFO 21416 --- [           main] project.LuceneApplication                : Started LuceneApplication in 5.131 seconds (JVM running for 11.959)
2019-06-05 23:02:22.548  INFO 21416 --- [nio-8080-exec-1] o.a.c.c.C.[Tomcat].[localhost].[/]       : Initializing Spring DispatcherServlet 'dispatcherServlet'
2019-06-05 23:02:22.570  INFO 21416 --- [nio-8080-exec-1] o.s.web.servlet.DispatcherServlet        : Initializing Servlet 'dispatcherServlet'
2019-06-05 23:02:22.594  INFO 21416 --- [nio-8080-exec-1] o.s.web.servlet.DispatcherServlet        : Completed initialization in 12 ms
[SYSTEM] Packet Received: defense program
[SYSTEM] Packet Received: title:Trade
[SYSTEM] Packet Received: asefhefishfdoihas
```

Note: Google retains all rights to the namesake Google. The name Google-- is used to portray this program's inferiority to the real deal.