README

Web Crawler and Inverted Indexer

This project is a web crawler and inverted indexer that extracts information from web pages and creates an inverted index for efficient keyword-based searching.

Prerequisites

To compile and run this project, you need to have the following dependencies installed. Clone the repo, and the files below will automatically be included

- Java Development Kit (JDK)
- htmlparser.jar
- jsoup-1.17.2.jar
- jdbm-1.0.jar

Compilation

To compile the project and move the class files to the appropriate destination, use the following command:

```
javac -cp htmlparser.jar:jsoup-1.17.2.jar:jdbm-1.0.jar:. *.java
```

Then, copy class files to the appropriate folder in order to crawl pages

```
cp *.class ./PROJECT
```

Also, move all class files to the appropriate folder for the webapp

```
mv *.class ./apache-tomcat-10.1.20/webapps/comp4321/WEB-INF/classes/PROJECT
```

Running the Cralwer and Inverter

To run the program, use the following command

```
java -cp htmlparser.jar:jsoup-1.17.2.jar:jdbm-1.0.jar:. PROJECT.Inverter "https://www.cse.ust.hk/~kwtleung
```

Moving the database files to the correct location:

```
mv *.lg *.db ./apache-tomcat-10.1.20/webapps/comp4321/WEB-INF/database/
```

Optional: Running other java files for testing (e.g: SearchEngine)

```
java -cp htmlparser.jar:jsoup-1.17.2.jar:jdbm-1.0.jar:. PROJECT.SearchEngine
```

Using the Web interface:

Firstly, add environment variables.

- Set CATALINA_HOME to {path to this project}/apache-tomcat-10.1.20/
- Set JAVA_HOME to {Path to your JDK}

Next, change directory to the correct folder to start apache tomcat, and run the startup.sh file

```
cd ./apache-tomcat-10.1.20/bin
./startup.sh
```

Head over to your browser: http://localhost:8080/comp4321/

You may then perform the searching

Shutting down Apache Tomcat

You should then shutdown apache-tomcat when you are done

```
./shutdown.sh
```

License