

# Program 1

Hello,

I wrote this program in java. It doesn't need any libraries, dependencies, or build tools. Compile the Crawler.java and run the program. I compiled and ran the program on Linux (with java 14) and Windows (with java 15) machines without issues.

When I tested the program using the css labs, I couldn't compile my code. They have a very old version of java.

This program should be able to work with java 11, but I couldn't test that because I don't have java 11.

To run the program:

```
javac Crawler.java
java Crawler http://courses.washington.edu/css502/dimpsey/ 30
```

I represent hops by a vertical bar, "|". I don't count visiting the starting url as a hop.

I parse pages with regex. I assume, among other things, anchors will end with "".

I keep track of the pages I visit by using a hashset. I backtrack with a stack.

I handle redirections (3xx) manually so I can print them, but I don't count them as hops. If links with trailing slashes redirect to one without or vice versa, then my program only counts that as one hop. If both links return 200, then each will be considered a different webpage (since they are).


I attempt requests thrice for server error responses (5xx).

The program terminates when numHops is reached, there are no more pages to visit, or the last visited page had no embedded urls. If you want to change the program so it doesn't terminate if the last page had no embedded urls, please look at the comments for response status 2xx.

## Output

The output for `java Crawler http://courses.washington.edu/css502/dimpsey/ 50` is in output.html. You can't open it with your browser because it has relative links for css and js files.

Here is a screenshot as well.



```
PS D:\Desktop\CSS 436\Programs\program1> java Crawler http://courses.washington.edu/css502/dimpsey/ 50
<!-- Crawling the web:
Status: 200| URL:http://courses.washington.edu/css502/dimpsey/
|Status: 301| URL:http://faculty.washington.edu/dimpsey
|Status: 200| URL:http://faculty.washington.edu/dimpsey/
||Status: 200| URL:http://faculty.washington.edu/dimpsey/about/
|||Status: 200| URL:http://faculty.washington.edu/dimpsey/software-shipped/
||||Status: 200| URL:http://faculty.washington.edu/dimpsey/home-2/education/
|||||Status: 200| URL:http://faculty.washington.edu/dimpsey/home-2/papers/

Finished Crawling
The last page is printed below:  ----->
<!DOCTYPE html>
<html lang="en-US">
<head>
<meta charset="UTF-8">
<meta name="viewport" content="width=device-width, initial-scale=1">
<title>Papers &#8211; Robert Dimpsey</title>
<link rel="profile" href="http://gmpg.org/xfn/11">
<link rel="pingback" href="http://faculty.washington.edu/dimpsey/wordpress/xmlrpc.php">
<title>Papers &#8211; Robert Dimpsey</title>
<link rel="dns-prefetch" href="//fonts.googleapis.com/" />
<link rel="dns-prefetch" href="//s.w.org/" />
<link rel="alternate" type="application/rss+xml" title="Robert Dimpsey &#8211; Feed" href="http://faculty.washington.edu/dimpsey/feed/" />
<link rel="alternate" type="application/rss+xml" title="Robert Dimpsey &#8211; Comments Feed" href="http://faculty.washington.edu/dimpsey/comments/feed/" />
<script type="text/javascript">
```

