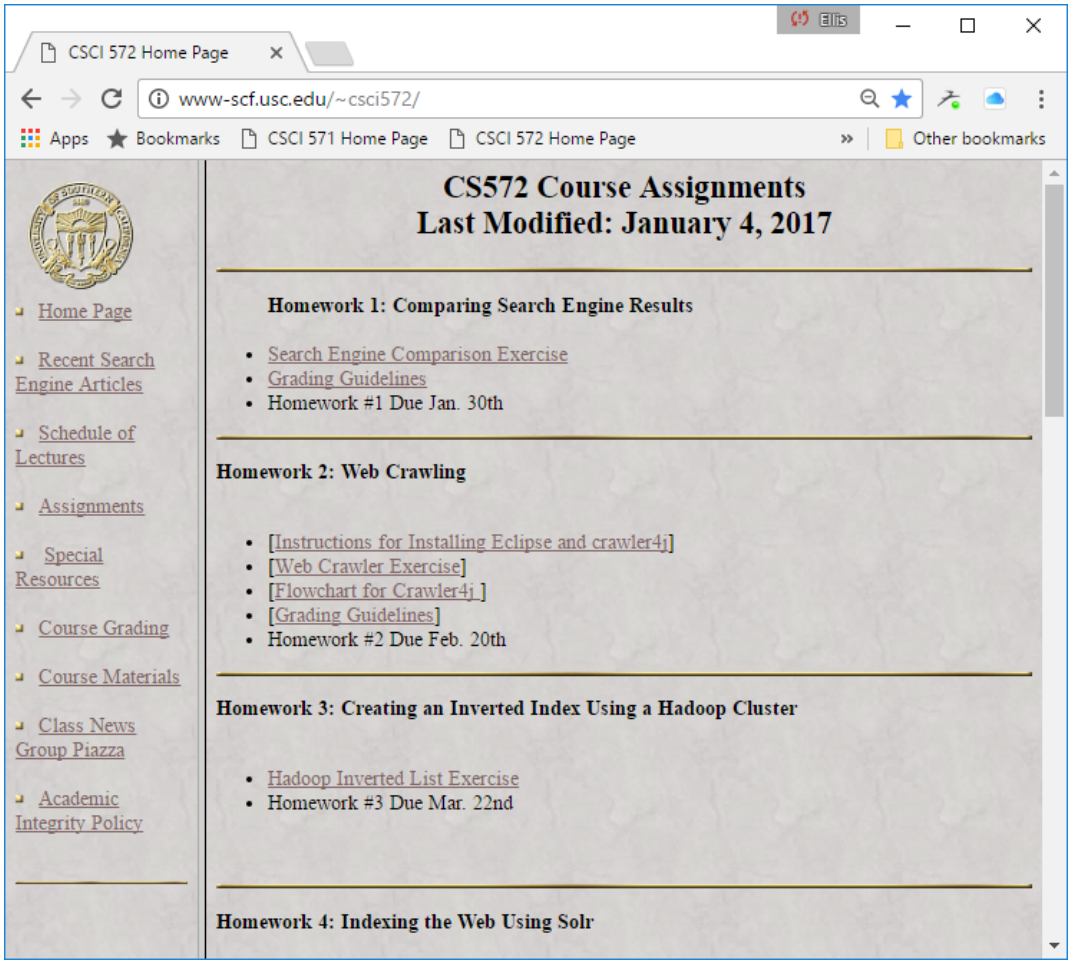
HW2 Assignment

- Involves
  1. Java programming
     - I assume all of you know how to program in Java!
  2. Eclipse Software Development Environment
  3. crawler4j, an open source java web crawler
  4. a crawl and analysis of a web site and an analysis of the crawl



What is Eclipse?

- Eclipse started as a proprietary IBM product (IBM Visual age for Smalltalk/Java)
  - Embracing the open source model IBM opened the product up
- Open Source
  - It is a general purpose open platform that facilitates and encourages the development of third party plug-ins
- Best known as an Integrated Development Environment (IDE)
  - Provides tools for coding, building, running and debugging applications
- Originally designed for Java, now supports many other languages
  - Good support for C, C++
  - Python, PHP, Ruby, etc…

Prerequisites for Running Eclipse

- Eclipse is written in Java and will thus need an installed JRE (Java Runtime Environment) or JDK (Java Development Kit) in which to execute
  - JDK recommended

Obtaining Eclipse

- Eclipse can be downloaded from…
  - http://www.eclipse.org/downloads/packages/
- Eclipse comes bundled as a zip file (Windows) or a tarball (all other operating systems)

Installing Eclipse

- Simply unwrap the zip file to some directory where you want to store the executables
- The document

"Instructions for Installing Eclipse and crawler4j"

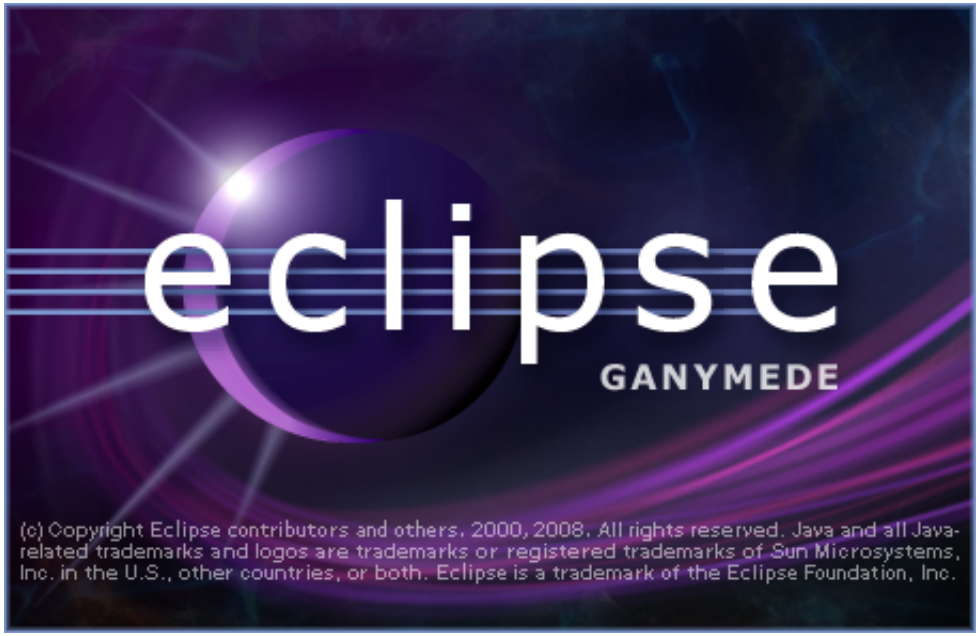
- located at

http://www-scf.usc.edu/~csci572/2017Fall/hw2/Crawler4jinstallation.pdf

describes the installation for Windows and Macs

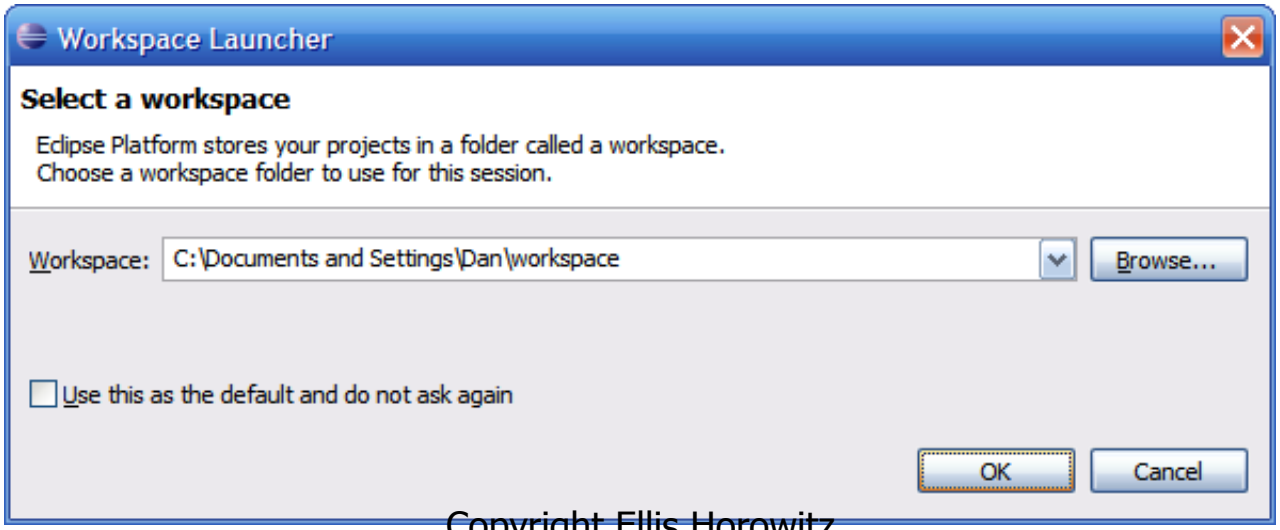
Launching Eclipse

- Once you have the environment setup, go ahead and launch eclipse
- You should see a splash screen such as the one below…

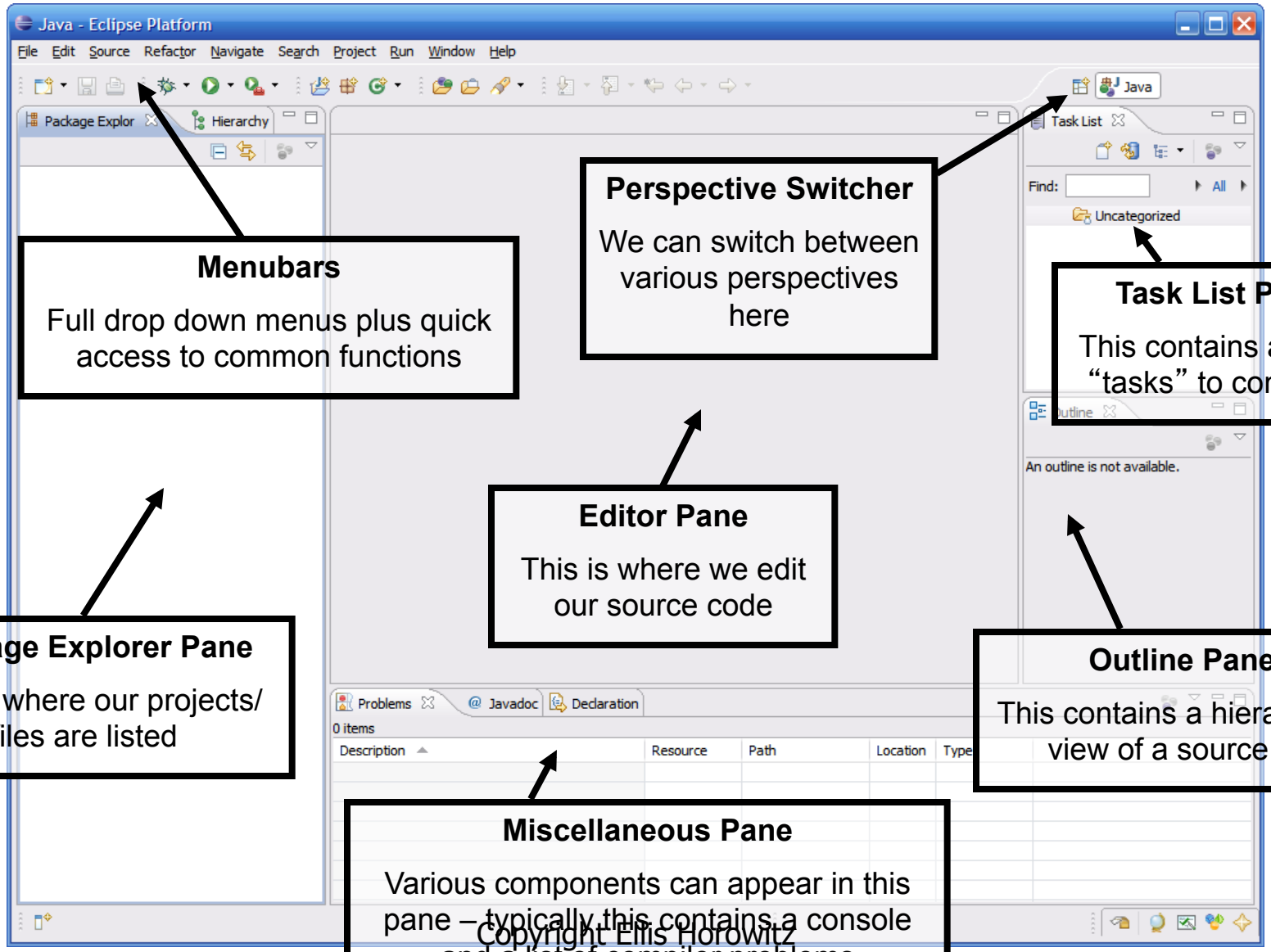


Selecting a Workspace

- In Eclipse, all of your code will live under a *workspace*
- A *workspace* is nothing more than a location where we will store the source code and where Eclipse will write out preferences
- Eclipse allows you to have multiple workspaces – each tailored in its own way
- Choose a location where you want to store your files, then click OK



Eclipse IDE Components



Menubars

Full drop down menus plus quick access to common functions

Perspective Switcher

We can switch between various perspectives here

Task List Pane

This contains a list of "tasks" to complete

Editor Pane

This is where we edit our source code

Package Explorer Pane

This is where our projects/ files are listed

Outline Pane

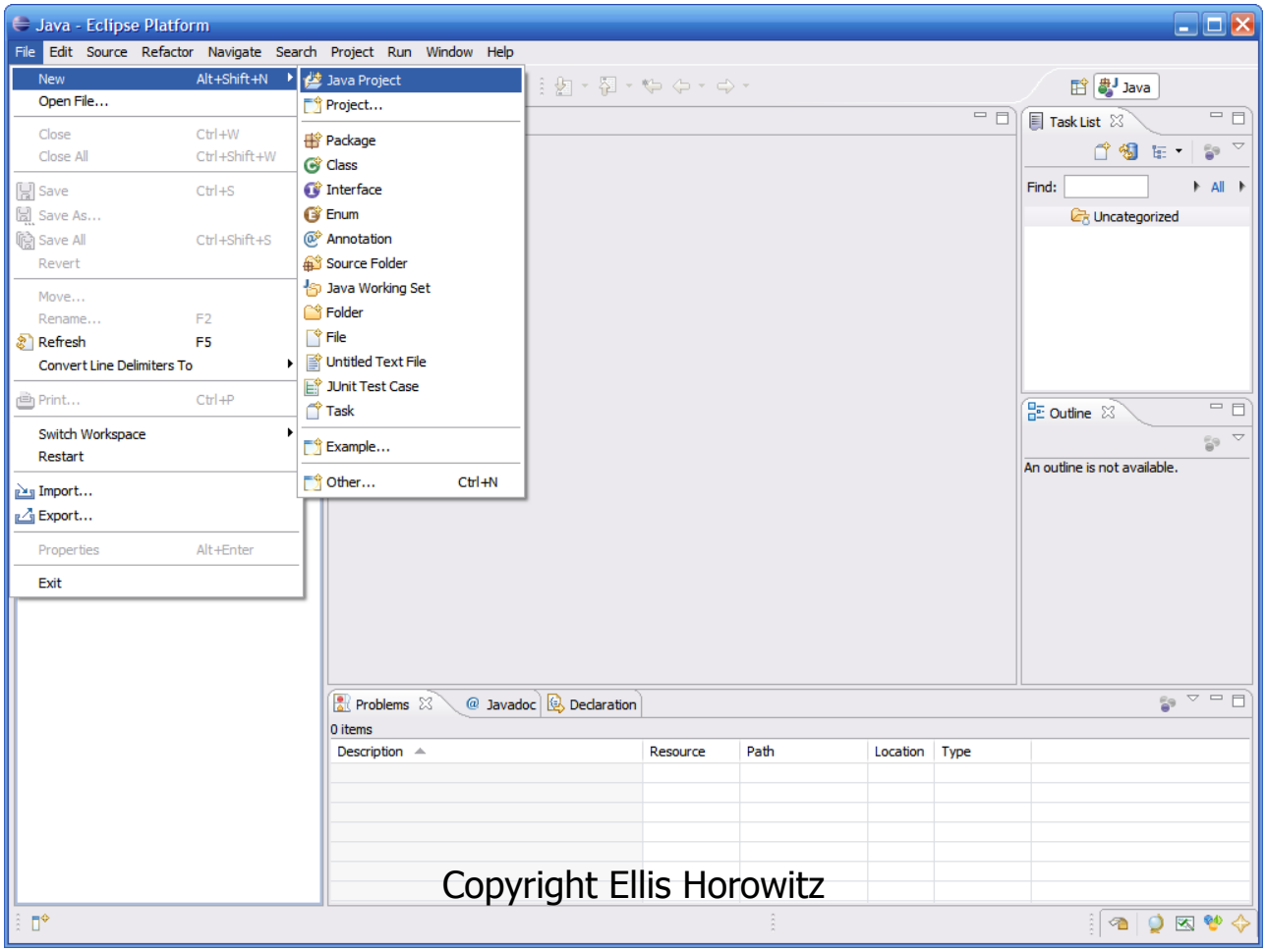
This contains a hierarchical view of a source file

Miscellaneous Pane

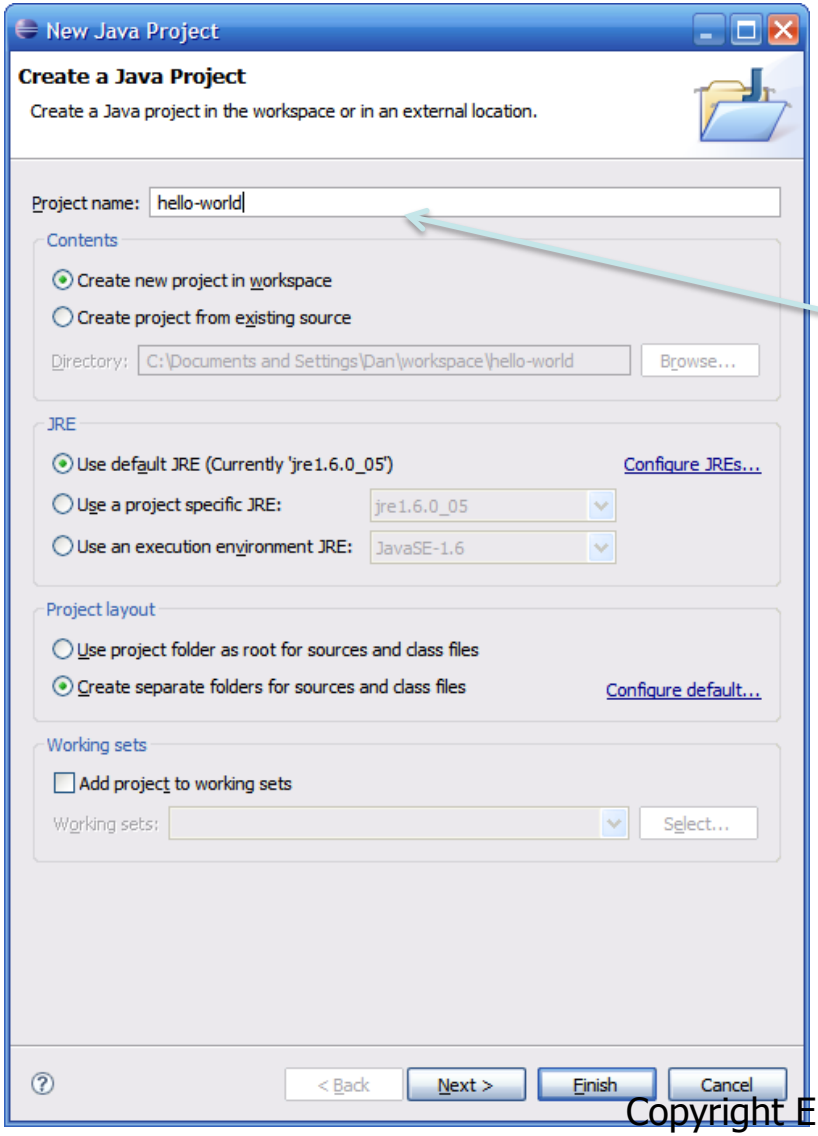
Various components can appear in this pane – typically this contains a console and a list of compiler problems

Creating a New Project

- All code in Eclipse needs to live under a project
- To create a project: File → New → Java Project



Creating a New Project (continued)

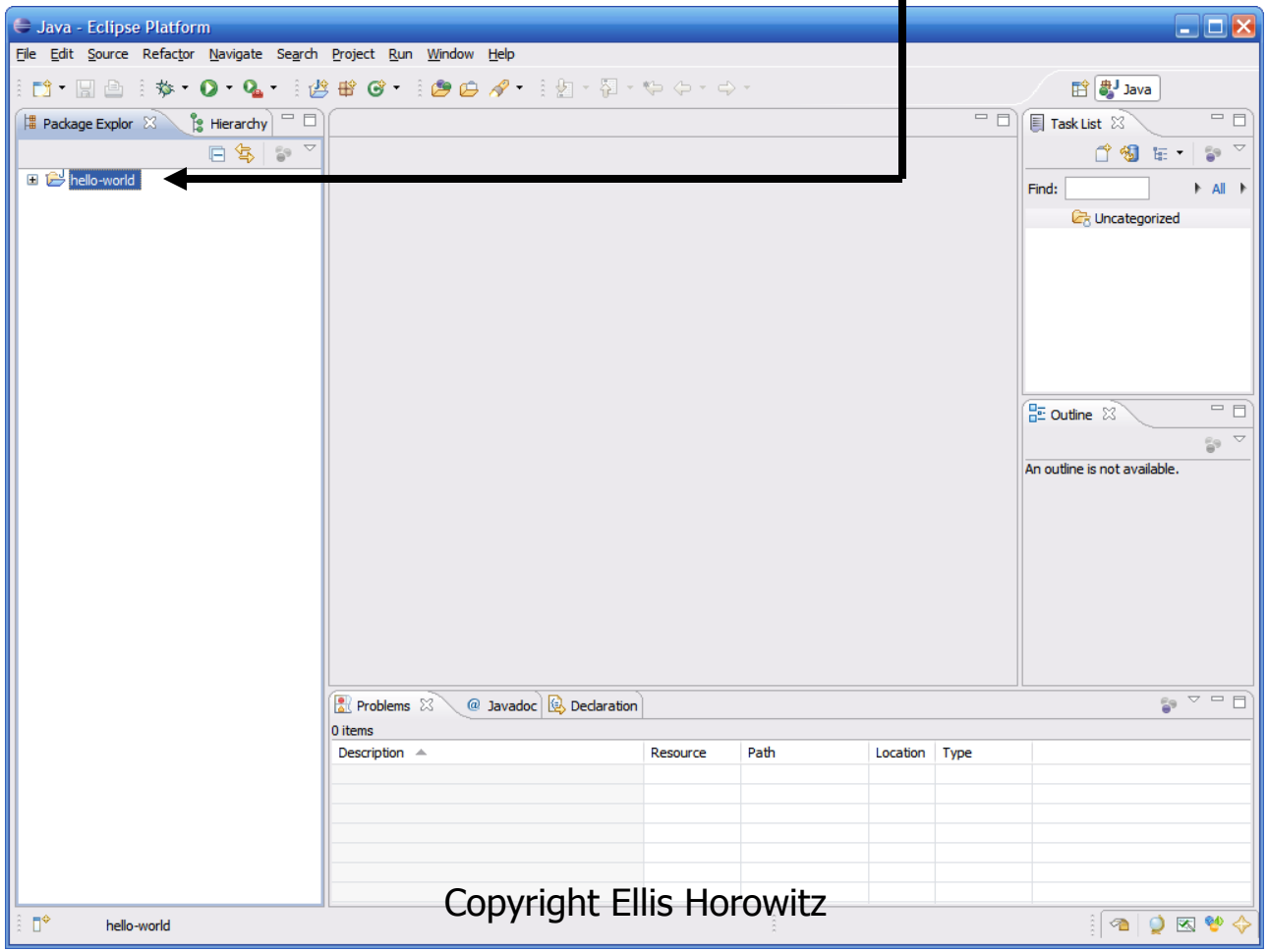


- Enter a name for the project, then click Finish

Hello-world Project

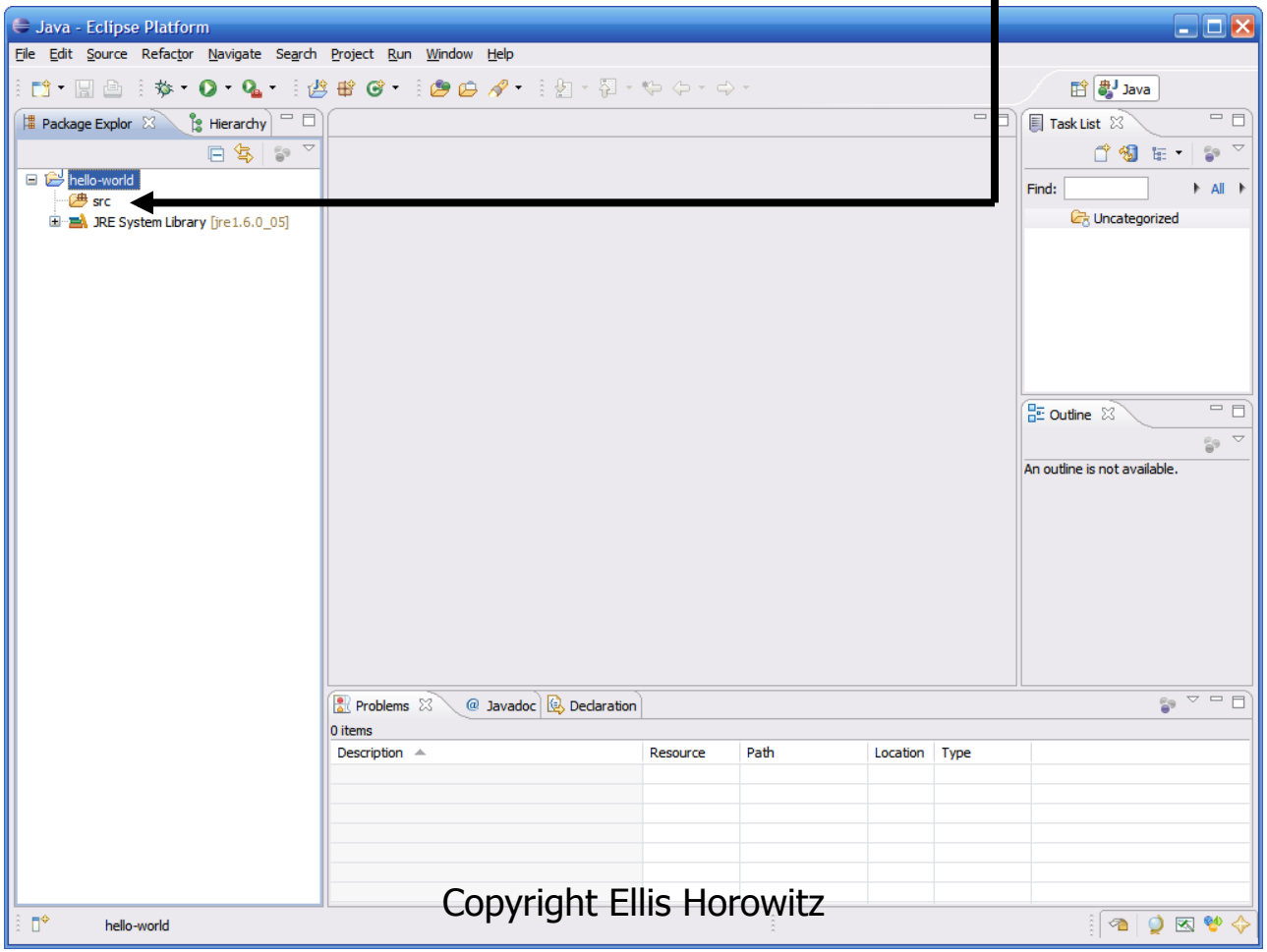
Creating a New Project (continued)

- The newly created project should then appear under the Package Explorer



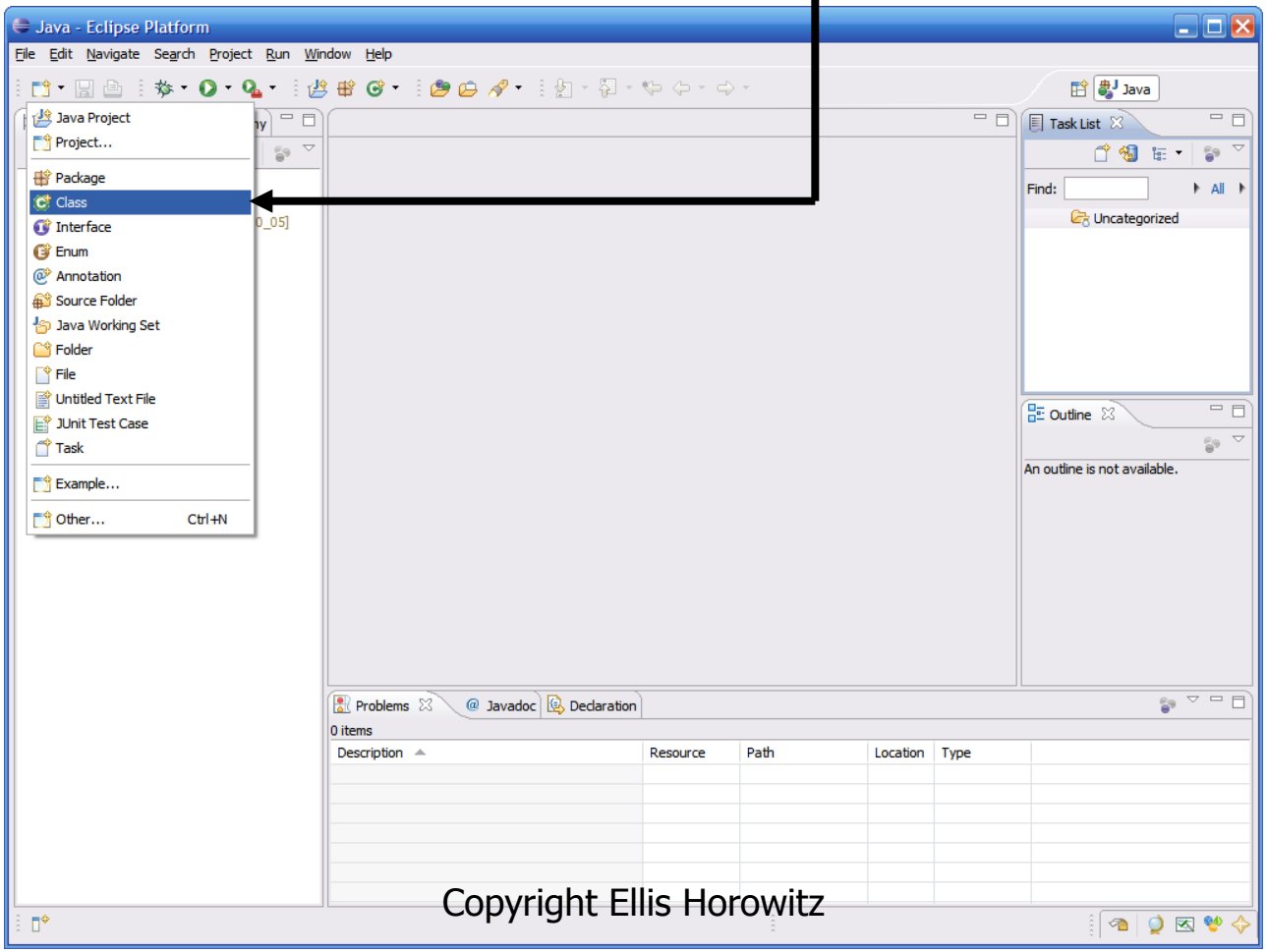
The src folder

- Eclipse automatically creates a folder to store your source code in called src

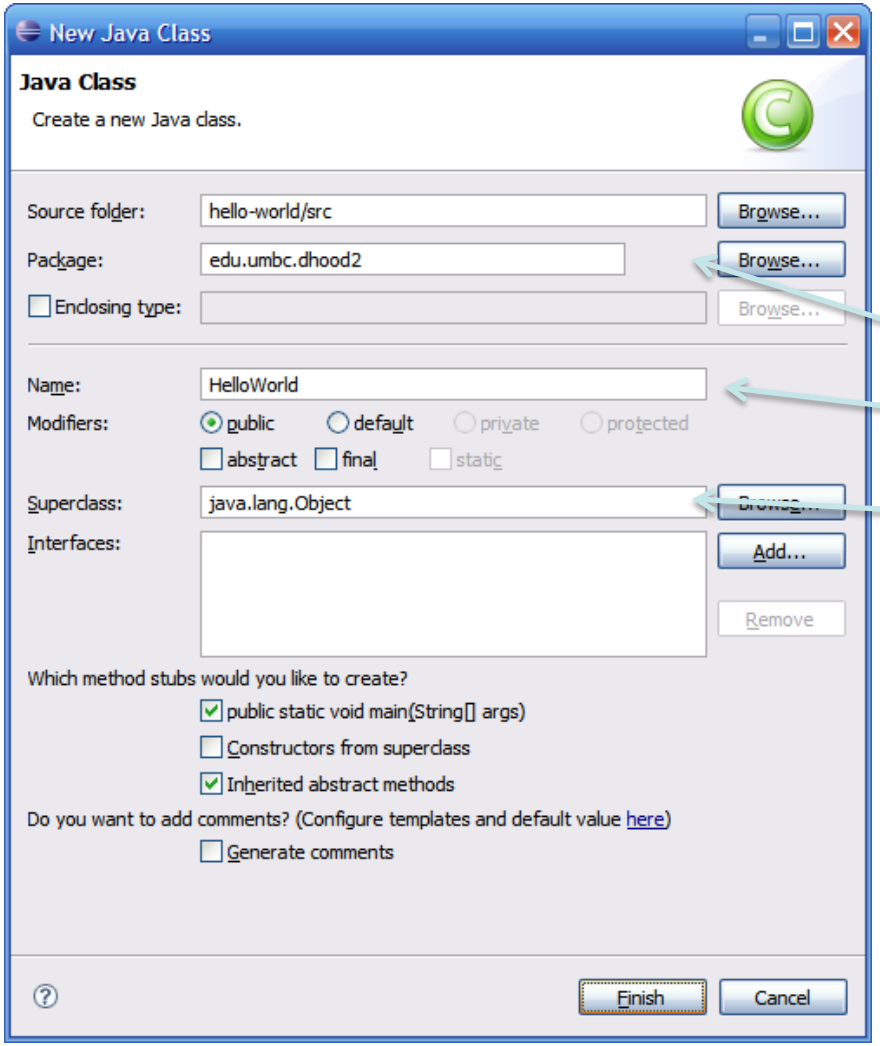


Creating a Class

- To create a class, simply click on the New button, then select Class



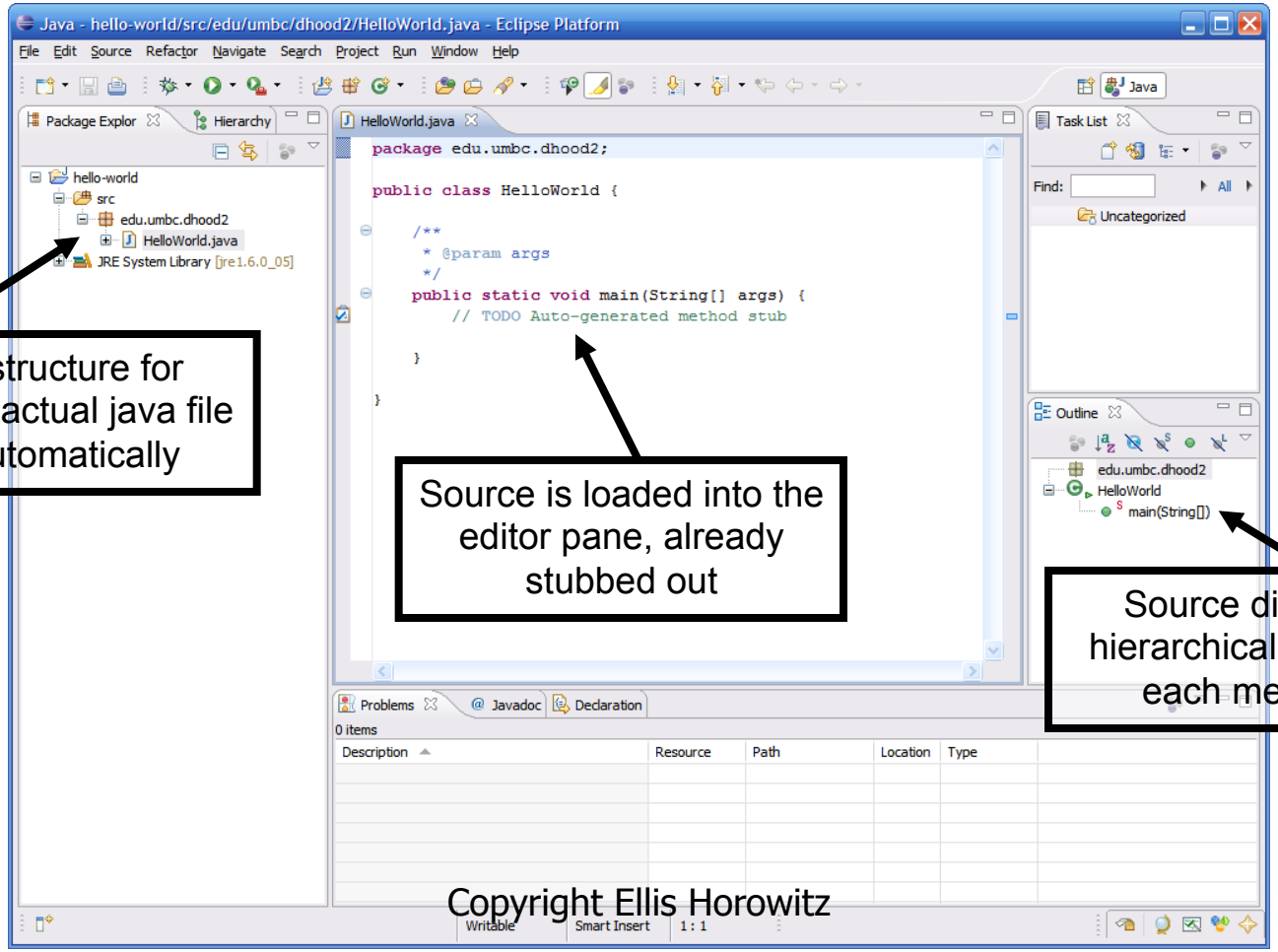
Creating a Class (continued)



- This brings up the new class wizard
- From here you can specify the following...
  - Package
  - Class name
  - Superclass
  - Whether or not to include a main
  - Etc…
- Fill in necessary information then click Finish to continue

The Created Class

- As you can see a number of things have now happened…



Directory structure for package and actual java file created automatically

Source is loaded into the editor pane, already stubbed out

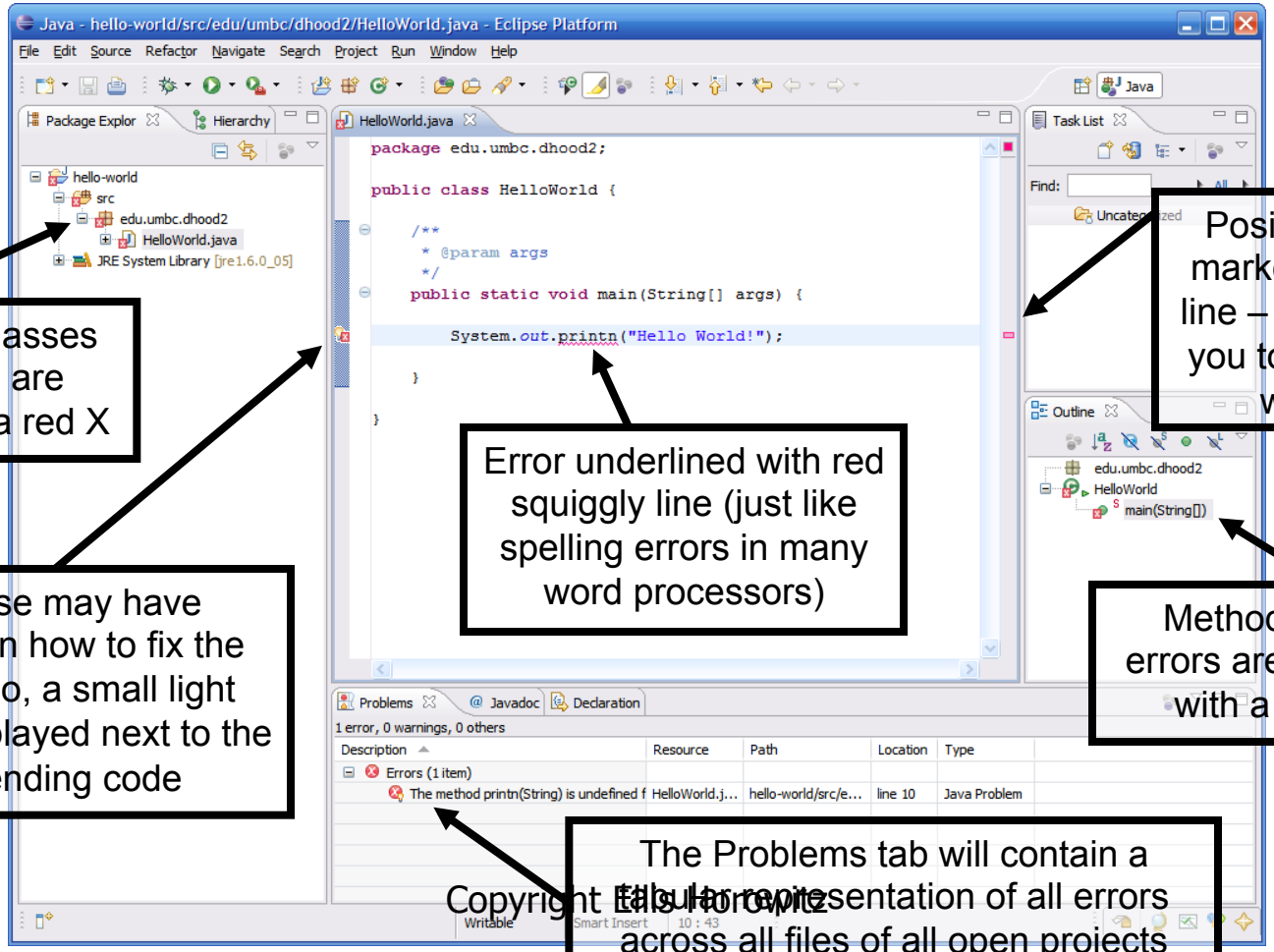
Source displayed in a hierarchical fashion listing each method name

Compiling Source Code

- One important feature of Eclipse is that it automatically compiles your code in the background
- This means that errors can be corrected when made
  - We all know that iterative development is the best approach to developing code, but going to shell to do a compile can interrupt the normal course of development
  - You no longer need to go to the command prompt and compile code directly

Example Compilation Error

- This code contains a typo in the println statement…



Packages/Classes with errors are marked with a red X

Position in file is marked with a red line – 1 click allows you to jump to line with error

Error underlined with red squiggly line (just like spelling errors in many word processors)

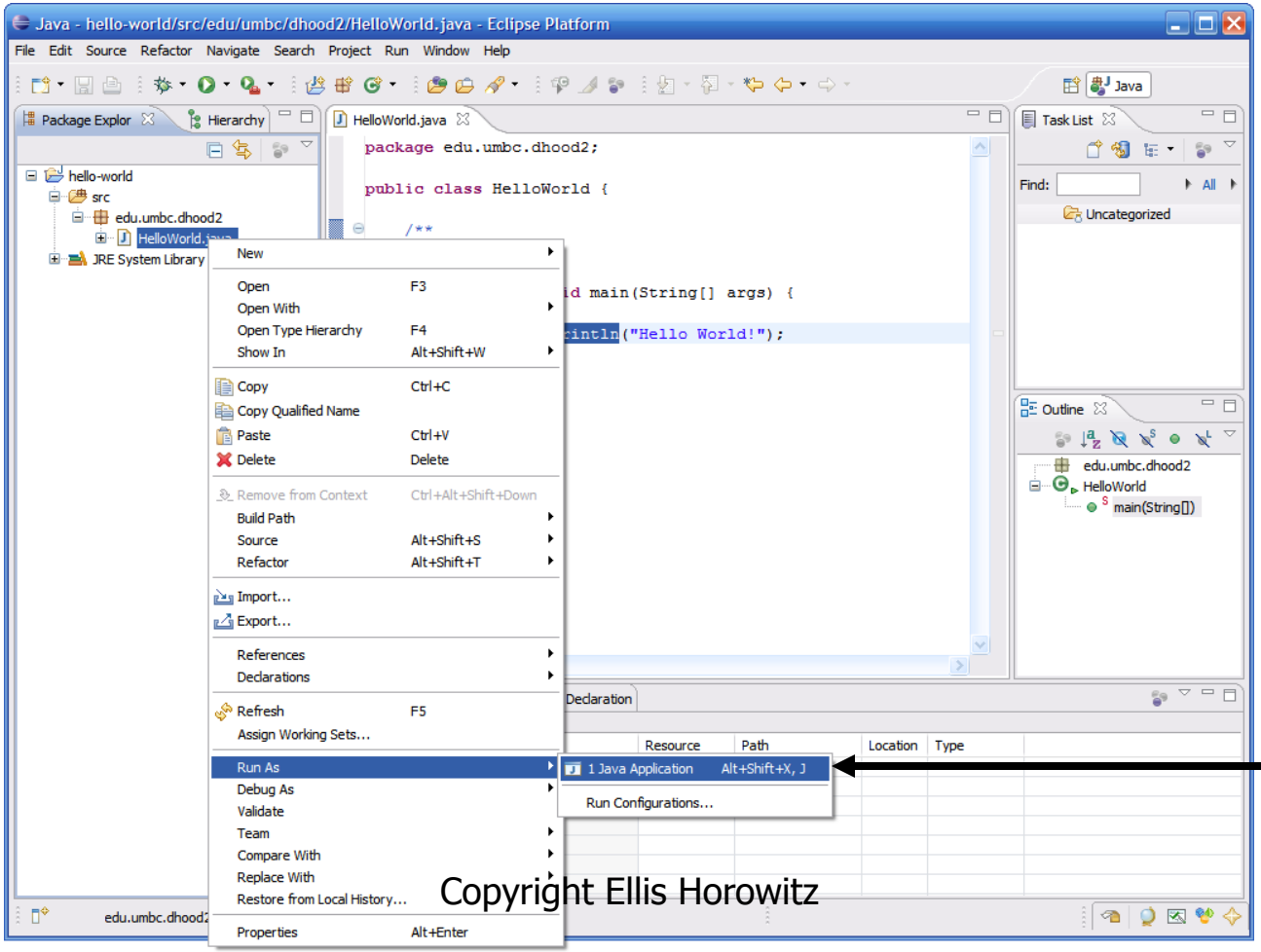
Often Eclipse may have suggestions on how to fix the problem – if so, a small light bulb will be displayed next to the line of offending code

Methods with errors are marked with a red X

The Problems tab will contain a tabular representation of all errors across all files of all open projects

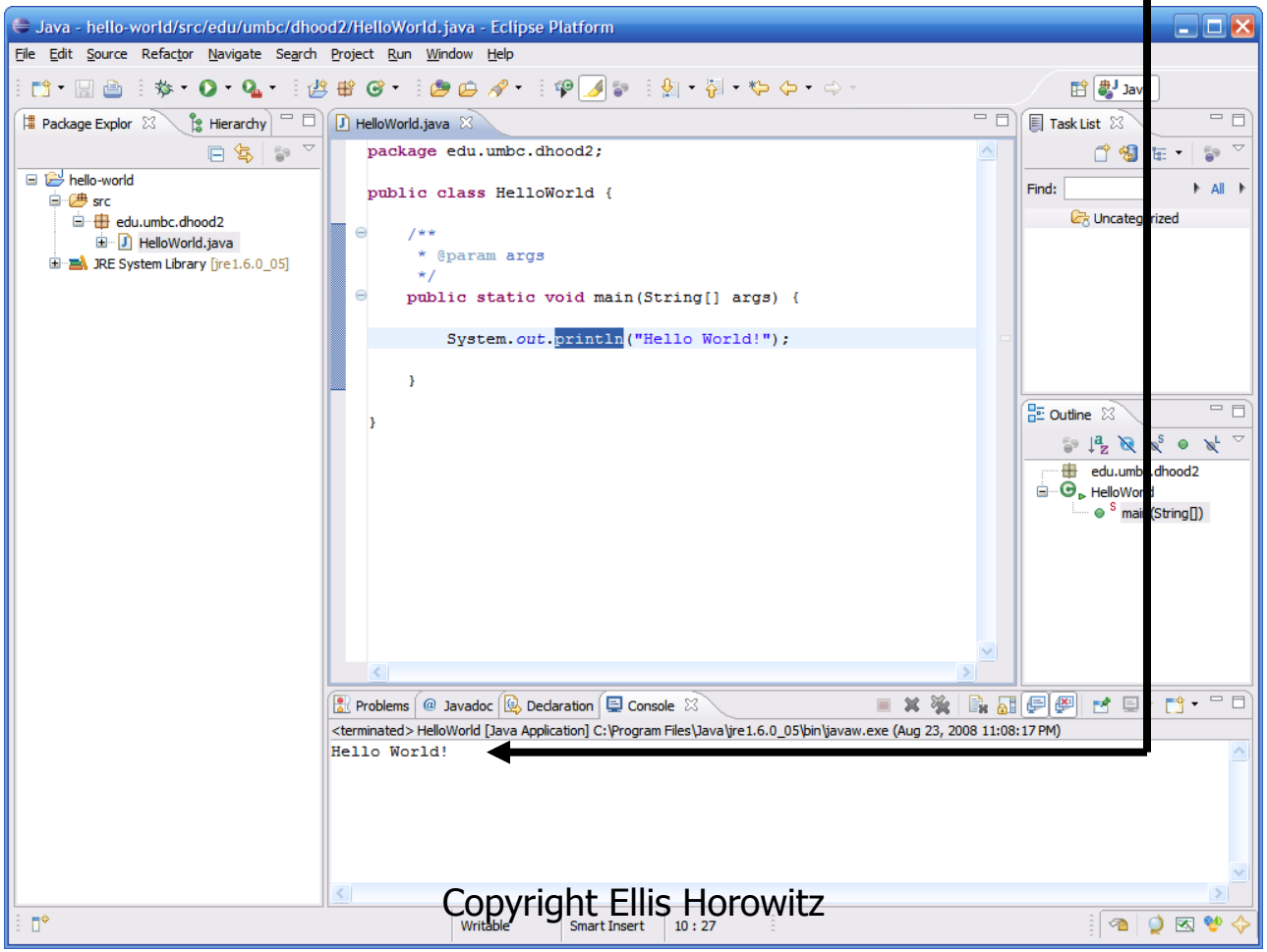
Running Code

- An easy way to run code is to right click on the class and select Run As → Java Application



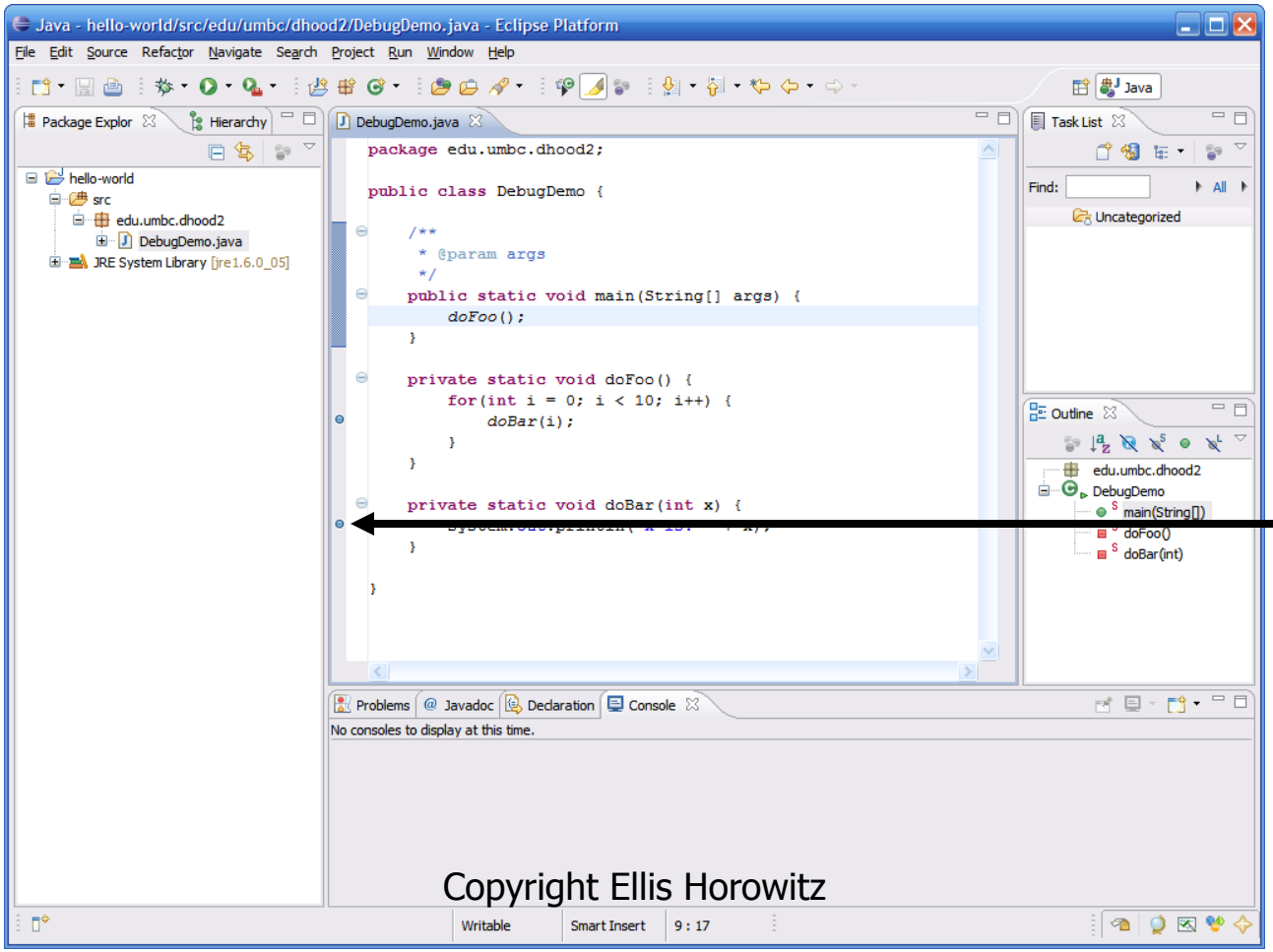
Running Code (continued)

- The output of running the code can be seen in the Console tab in the bottom pane



Debugging Code

- Eclipse comes with a pretty good built-in debugger
- You can set break points in your code by double clicking in the left hand margin – break points are represented by these blue bubbles



Tools for Surface Web Crawling

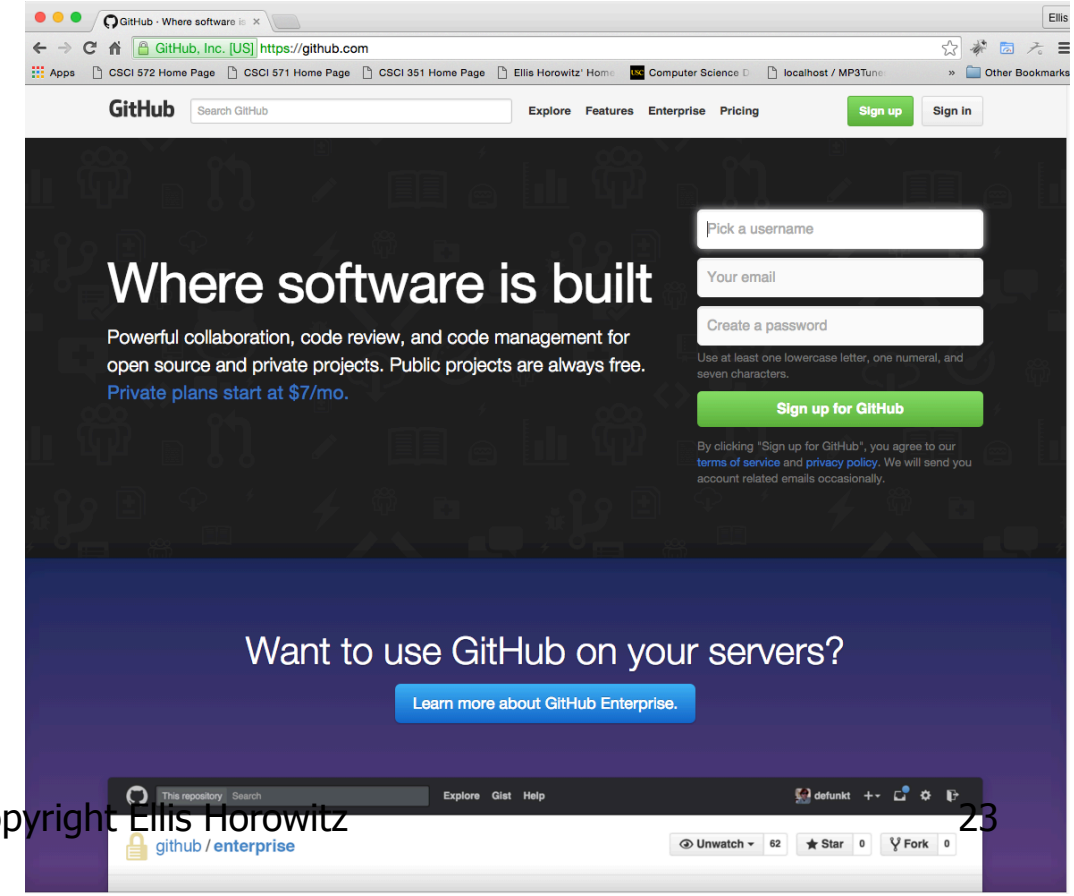
- **Command line**
  - wget, pre-installed in Ubuntu
    - get a single page
    - wget http://www.example.com/index.html
    - support http, ftp etc., e.g.
    - wget ftp://ftp.gnu.org/pub/gnu/wget/wget-latest.tar.gz
  - curl, OSX pre-installed
- **Simple crawling APIs**
  - Java, crawler4j
  - Python, scrapy: http://scrapy.org
- **Large-scale crawling**
  - Heritrix, crawler for archive.org
  - Nutch, Apache Software Foundation

How To Get a Web Page in Java

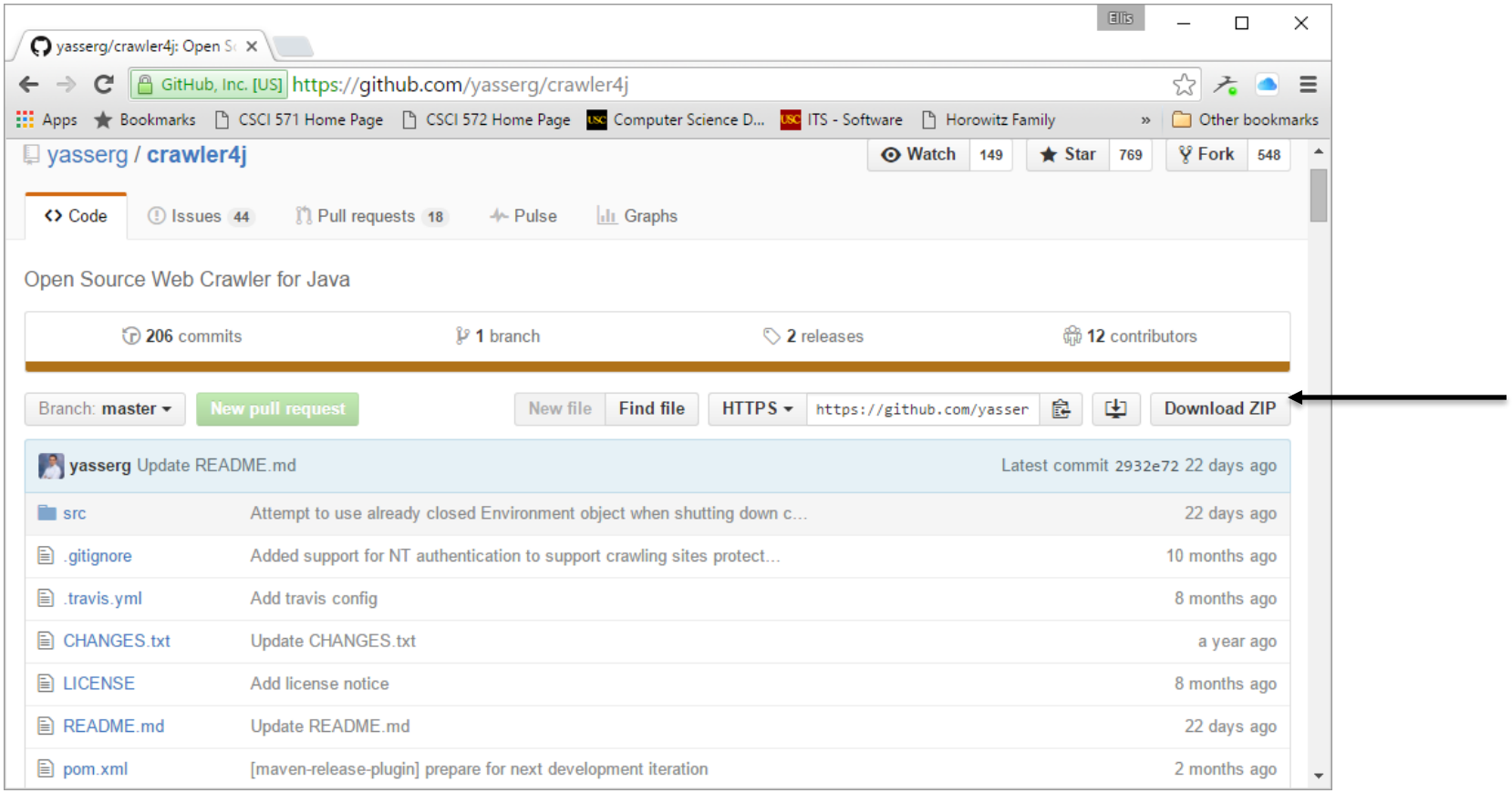
```
import java . net .*;
import java . io .*;
public class URLReader {
public static void main(String [] args) throws
Exception { } }
  URL oracle = new URL(”http://www.oracle.com/”);
  BufferedReader in = new BufferedReader (
  new InputStreamReader(oracle.openStream()));
  String inputLine ;
  while (( inputLine = in . readLine ()) != null)
     System . out . println ( inputLine );
     in . close ();
   }
}
```

Instructions for Installing Crawler4j

- download crawler4j from github
  - **GitHub** is a web-based repository hosting service for software. Originally the Git system offered distributed revision control and source code management (SCM) functionality, but on the command line; GitHub offers a web interface and some additional features.
  - As of Dec 2016, GitHub reports having over 24 million users and over 35 million repositories

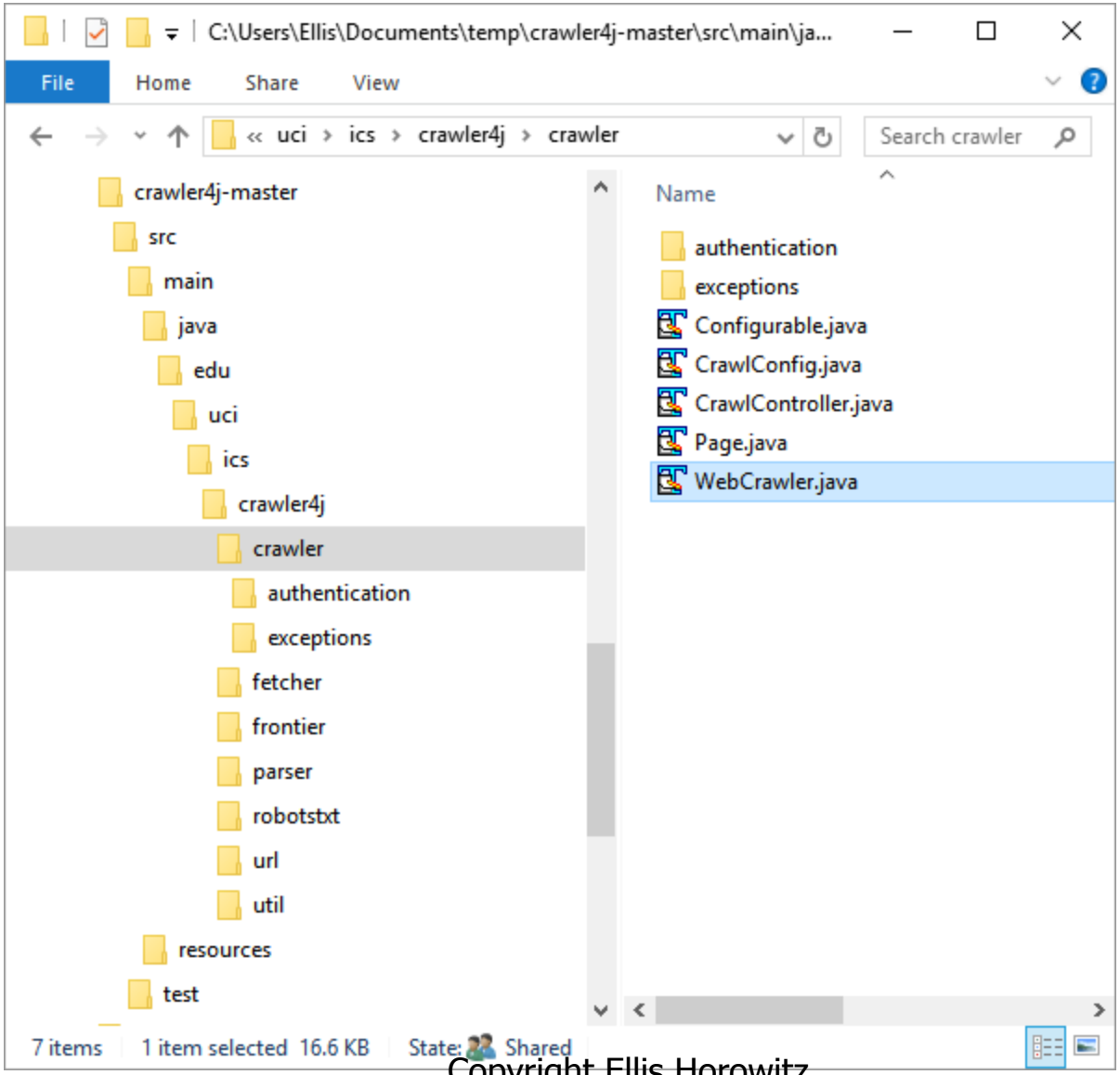


Downloading Crawler4j from GitHub



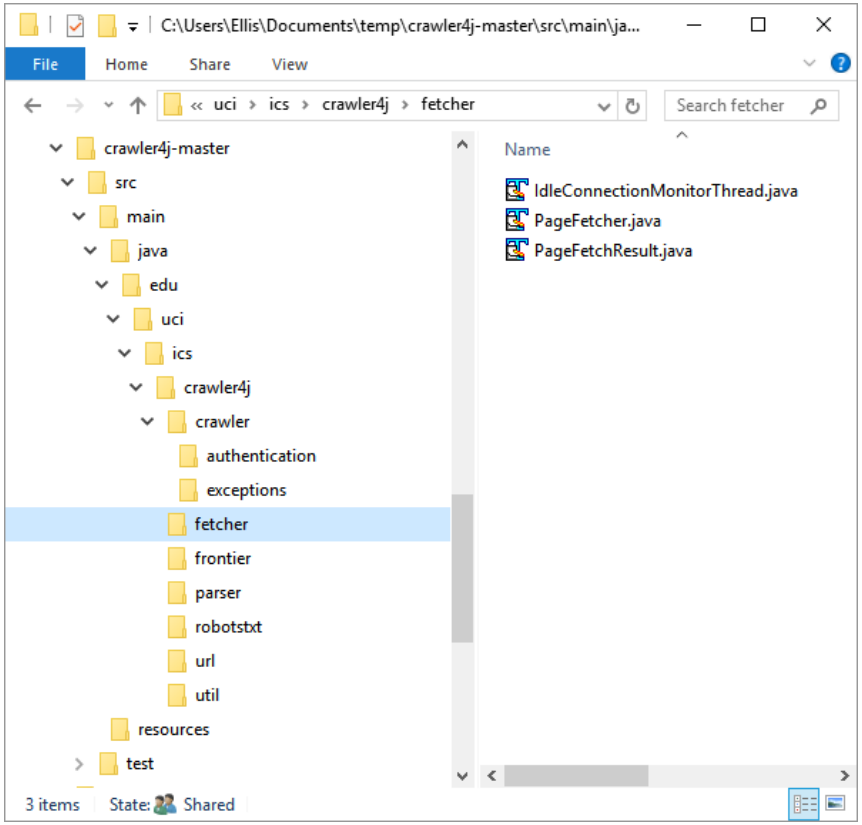
See especially the README file page at
https://github.com/yasserg/crawler4j/blob/master/README.md

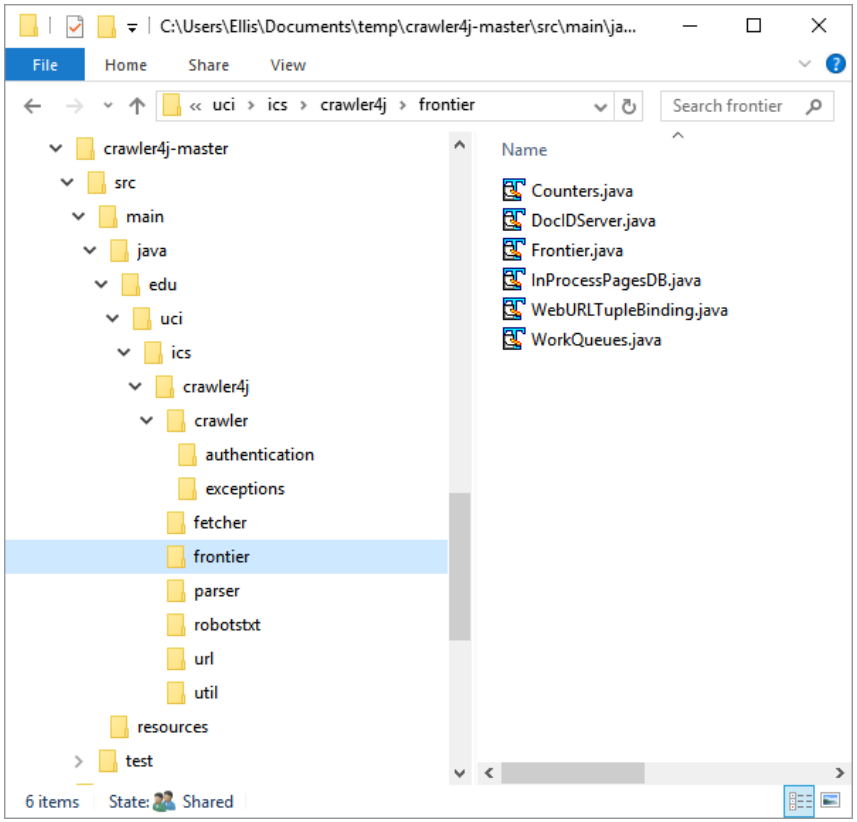
Crawler4j Source Code



Crawler folder, a good place to start; look especially at WebCrawler.java

Crawler4j Source code is Logically Organized into folders

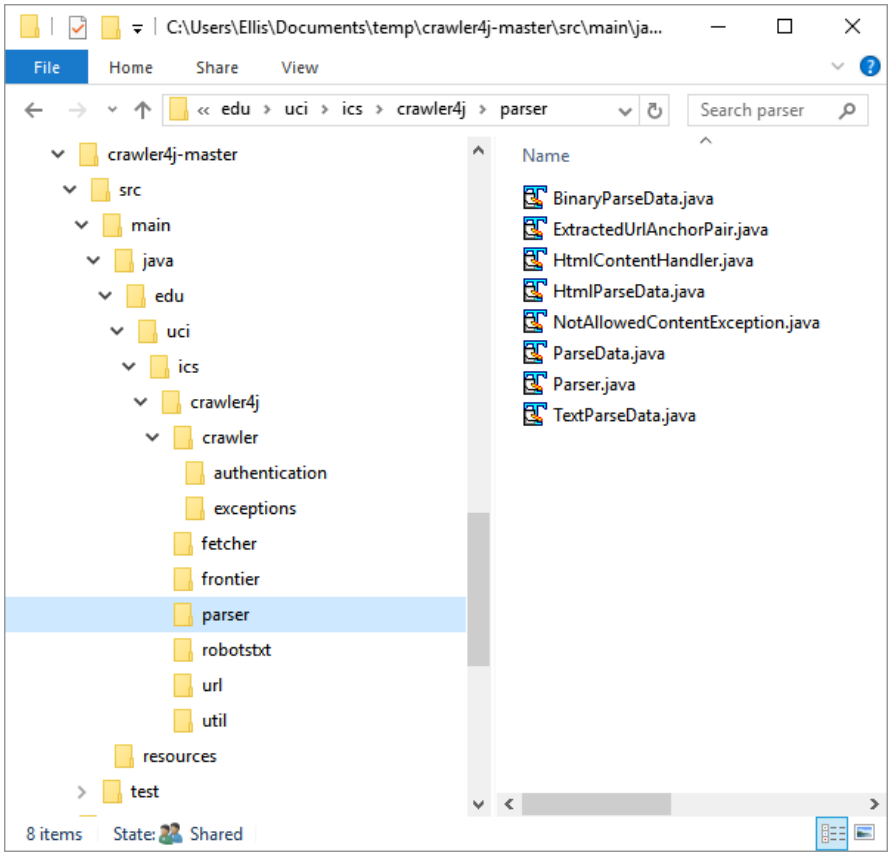


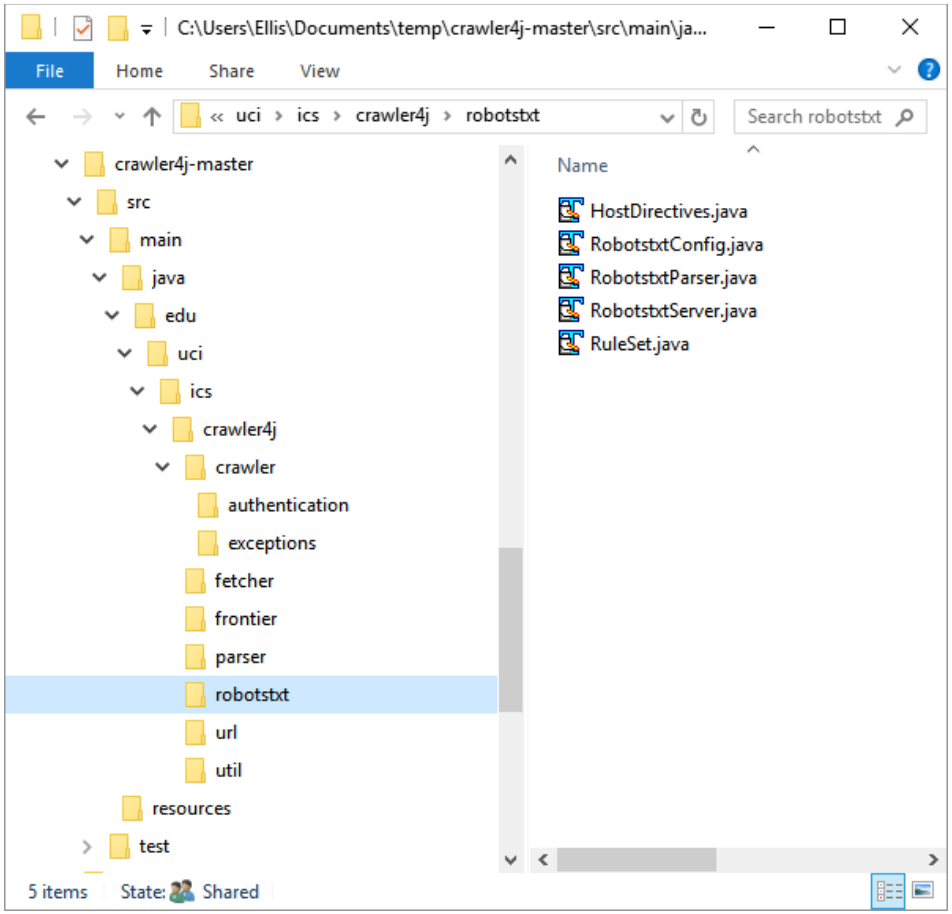


Fetcher Code handles:
- schemes: http, https
- politeness delay;
- redirects;
- max-size settings;
- expired connections

Frontier Code handles:
- statistics database;
- previously seen URLs
- queue of pending URLs

Crawler4j Routines are Named According to their Function

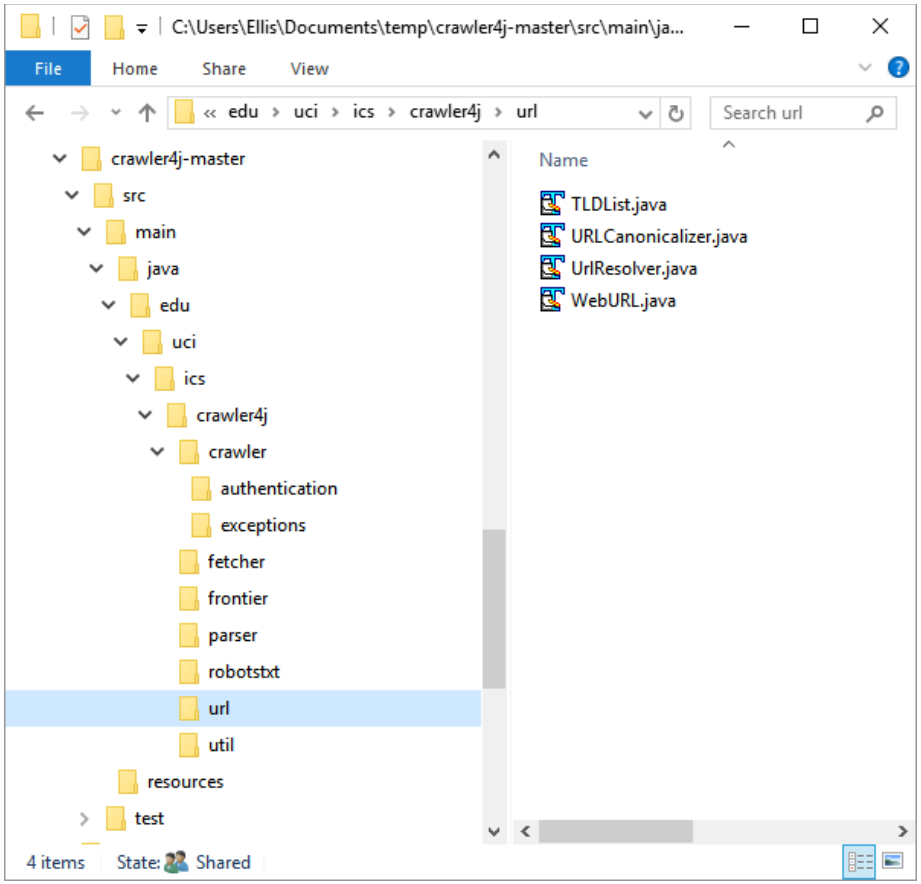


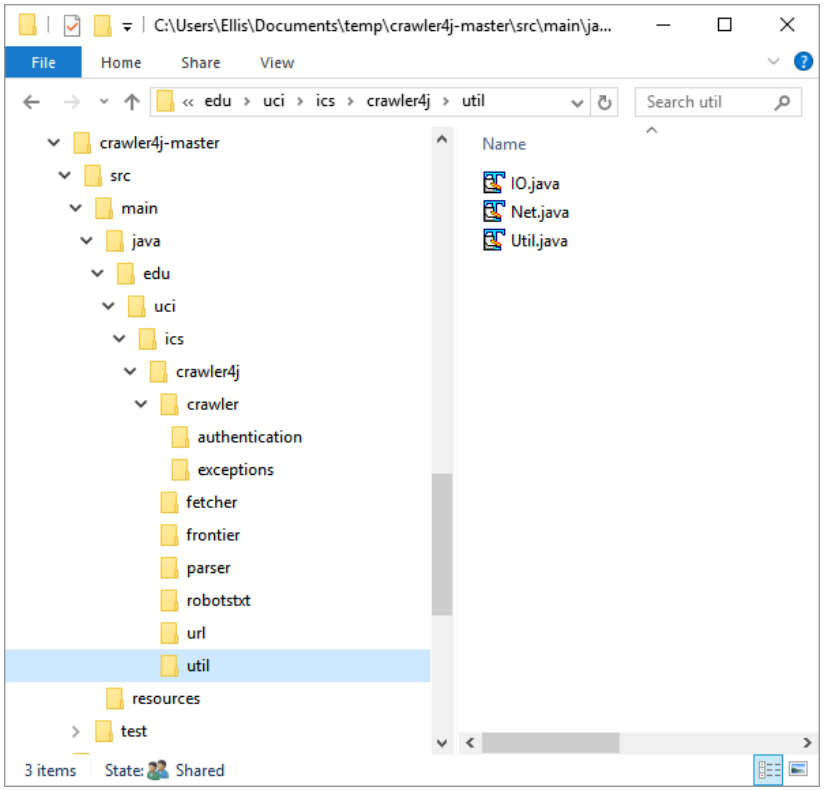


Parser Code handles:
- binary data
- html pages
- extracting links

Robots.txt Code handles:
- fetching and re-fetching robots.txt
- caching robots.txt files
- interpreting commands
- working with Page Fetcher

More crawler4j Source code

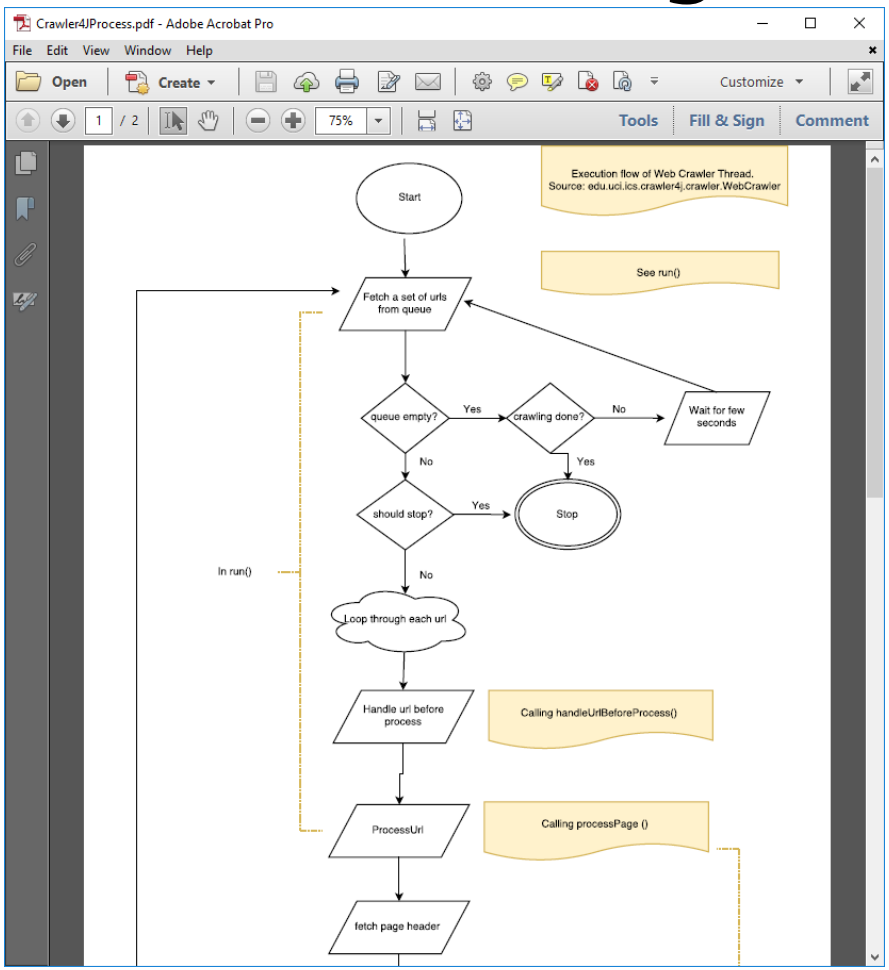


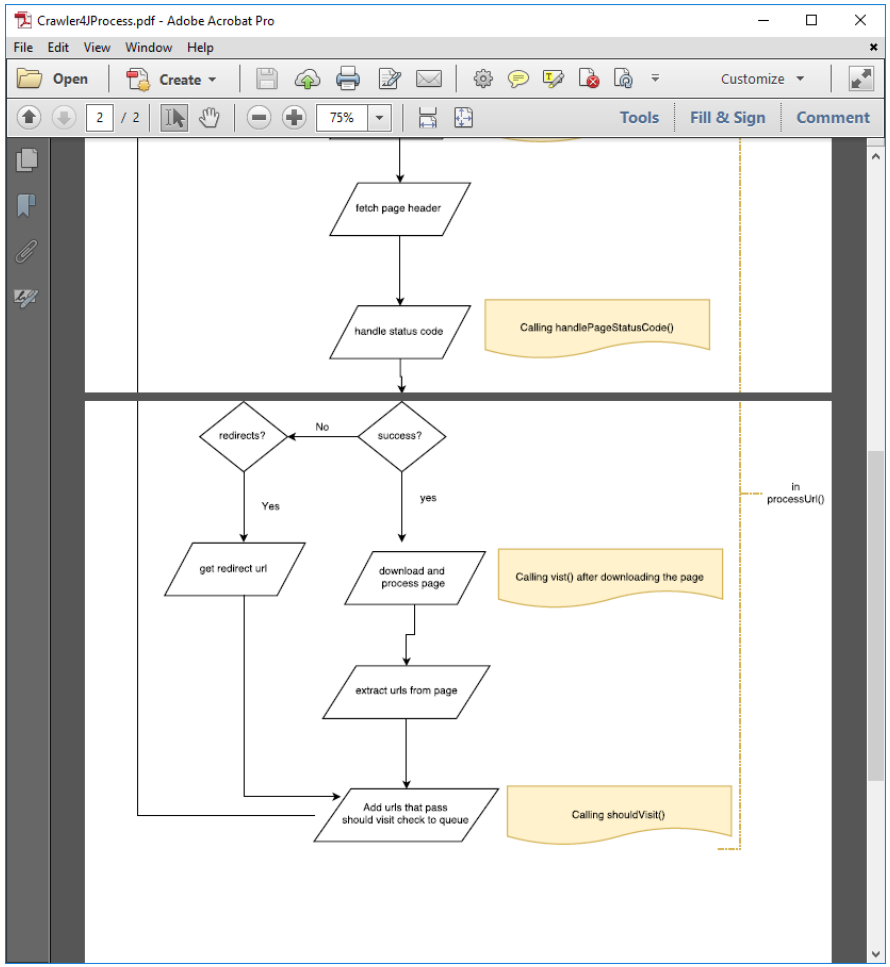


utility routines

URL resolver and canonicalizer handles:
- checking against list of TLDs
- normalizes URL, removes . or .., etc
- alters name/value pairs
- converts #nn values
- evaluates <base>

Logic Flowchart





http://www-scf.usc.edu/~csci572/2017Fall/hw2/Crawler4JProcess.pdf

Configuring the Crawler and Seeding it

```
public class Controller {
   public static void main(String[] args) throws Exception {
      String crawlStorageFolder = "/data/crawl";
      int numberOfCrawlers = 7;
      CrawlConfig config = new CrawlConfig();
      config.setCrawlStorageFolder(crawlStorageFolder);
      /*  Instantiate the controller for this crawl.*/
      PageFetcher pageFetcher = new PageFetcher(config);
      RobotstxtConfig robotstxtConfig = new RobotstxtConfig();
      RobotstxtServer robotstxtServer = new RobotstxtServer(robotstxtConfig, pageFetcher);
      CrawlController controller = new CrawlController(config, pageFetcher, robotstxtServer);
      /*  For each crawl, you need to add some seed urls. These are the first
       * URLs that are fetched and then the crawler starts following links
       * which are found in these pages  */
      controller.addSeed("http://www.cnn.com/”);
      /*  Start the crawl. This is a blocking operation, meaning that your code
       * will reach the line after this only when crawling is finished.  */
      controller.start(MyCrawler.class, numberOfCrawlers);
   }
}
```

folder to store downloads;

#crawlers

set up pagefetcher and robots.txt handler

crawling www.cnn.com

Defining Which Pages to Crawl

```
public class MyCrawler extends WebCrawler {
    private final static Pattern FILTERS =
Pattern.compile(".*(\\.(css|js|gif|jpg"  + "|png|mp3|mp3|zip|gz))$");
   /** This method receives two parameters. The first parameter is the page
     * in which we have discovered this new url and the second parameter is
     * the new url. You should implement this function to specify whether
     * the given url should be crawled or not (based on your crawling logic).
     * In this example, we are instructing the crawler to ignore urls that
     * have css, js, git, ... extensions and to only accept urls that start
     * with "http://www.cnn.com/". In this case, we didn't need the
     * referring Page parameter to make the decision.   */
     @Override
     public boolean shouldVisit(Page referringPage, WebURL url) {
         String href = url.getURL().toLowerCase();
         return !FILTERS.matcher(href).matches()
                 && href.startsWith("http://www.cnn.com/");
     }
```

see next slide

Matching URLs

- ".*(\\.(css|js|gif|jpg" + "|png|mp3|mp4|zip|gz))$"

- A regular expression, specified as a string, must first be compiled into an instance of this class.
- a Matcher object that can match arbitrary character sequences against the regular expression
- See https://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html

- In the above there are two strings concatenated by plus; consider the simpler form:
- ".*(\\.(css|js|zip|gz))$"

- . matches any character
- * matches zero or more of preceding character
- \. matches a literal dot
- $ anchors the pattern at the end of the string

Parsing the Downloaded Page

```
/**   This function is called when a page is fetched and ready
     * to be processed by your program.  */
    @Override
    public void visit(Page page) {
        String url = page.getWebURL().getURL();
        System.out.println("URL: " + url);
        if (page.getParseData() instanceof HtmlParseData) {
            HtmlParseData htmlParseData = (HtmlParseData) page.getParseData();
            String text = htmlParseData.getText();
            String html = htmlParseData.getHtml();
            Set<WebURL> links = htmlParseData.getOutgoingUrls();
            System.out.println("Text length: " + text.length());
            System.out.println("Html length: " + html.length());
            System.out.println("Number of outgoing links: " + links.size());
        }
   }
```

The Actual Exercise

- *the URLs it attempts to fetch*, **fetch.csv.** The number of rows should be no more than 20,000 as that is our pre-set limit.
- *the files it successfully downloads*, **visit.csv;** clearly the number of rows will be less than the number of rows in fetch.csv
- *all of the URLs that were discovered* and processed in some way; **urls.csv**. This file could be much larger than 20,000 rows as it will have numerous repeated URLs

Things to Save

- Fetch statistics:
  - # fetches attempted:
    The total number of URLs that the crawler attempted to fetch. This is usually equal to the MAXPAGES setting if the crawler reached that limit; less if the website is smaller than that.
  - # fetches succeeded:
    The number of URLs that were successfully downloaded in their entirety, i.e. returning a HTTP status code of 2XX.
  - # fetches failed or aborted:
    The number of fetches that failed for whatever reason, including, but not limited to: HTTP redirections (3XX), client errors (4XX), server errors (5XX) and other network-related errors.
-

Outgoing URLs

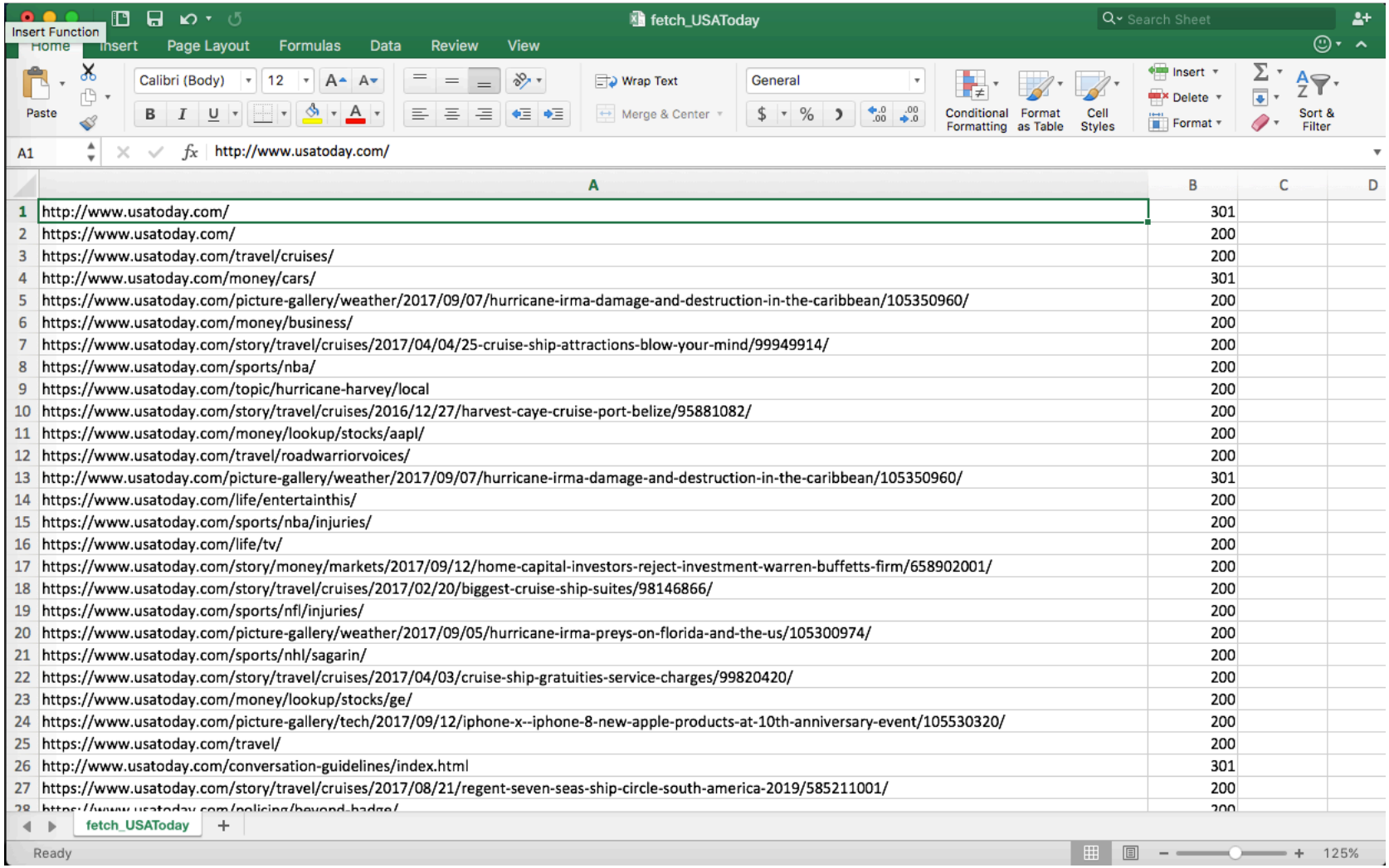
- Outgoing URLs: statistics about URLs extracted from visited HTML pages
  - Total URLs extracted:
    The grand total number of URLs extracted from all visited pages
  - # unique URLs extracted:
    The number of *unique* URLs encountered by the crawler
  - # unique URLs within the news web site:
    The number of *unique* URLs encountered that are associated with the news website,
    i.e. the URL begins with the given root URL of the news website.
  - # unique URLs outside the news website:
    The number of *unique* URLs encountered that were *not* from the website.

Sample Crawl Report for USA Today News Using 20,000 as the Download Limit

CrawlReport_USAToday.txt

```
Fetch statistics:
# Fetches attempted: 19980
# Fetches succeeded: 18326
# Fetches failed or aborted: 1654

Outgoing URLs:
# Total URLs extracted: 1328788
# Unique URLs extracted: 365062
# Unique URLs within a news website: 152480
# Unique URLs outside a news website: 212582

Status codes:
200 : 18326
301 : 1468
302 : 128
404 : 5
500 : 53

Content types:
text/html : 18326

Sizes:
<1KB : 1
1KB ~ <10KB : 36
10KB ~ <100KB : 898
100KB ~ <1MB : 17269
>1MB : 122
```

Sample Fetch File for USA Today



	A	B	C	D
1	http://www.usatoday.com/	301		
2	https://www.usatoday.com/	200		
3	https://www.usatoday.com/travel/cruises/	200		
4	http://www.usatoday.com/money/cars/	301		
5	https://www.usatoday.com/picture-gallery/weather/2017/09/07/hurricane-irma-damage-and-destruction-in-the-caribbean/105350960/	200		
6	https://www.usatoday.com/money/business/	200		
7	https://www.usatoday.com/story/travel/cruises/2017/04/04/25-cruise-ship-attractions-blow-your-mind/99949914/	200		
8	https://www.usatoday.com/sports/nba/	200		
9	https://www.usatoday.com/topic/hurricane-harvey/local	200		
10	https://www.usatoday.com/story/travel/cruises/2016/12/27/harvest-caye-cruise-port-belize/95881082/	200		
11	https://www.usatoday.com/money/lookup/stocks/aapl/	200		
12	https://www.usatoday.com/travel/roadwarriorvoices/	200		
13	http://www.usatoday.com/picture-gallery/weather/2017/09/07/hurricane-irma-damage-and-destruction-in-the-caribbean/105350960/	301		
14	https://www.usatoday.com/life/entertainthis/	200		
15	https://www.usatoday.com/sports/nba/injuries/	200		
16	https://www.usatoday.com/life/tv/	200		
17	https://www.usatoday.com/story/money/markets/2017/09/12/home-capital-investors-reject-investment-warren-buffetts-firm/658902001/	200		
18	https://www.usatoday.com/story/travel/cruises/2017/02/20/biggest-cruise-ship-suites/98146866/	200		
19	https://www.usatoday.com/sports/nfl/injuries/	200		
20	https://www.usatoday.com/picture-gallery/weather/2017/09/05/hurricane-irma-preys-on-florida-and-the-us/105300974/	200		
21	https://www.usatoday.com/sports/nhl/sagarin/	200		
22	https://www.usatoday.com/story/travel/cruises/2017/04/03/cruise-ship-gratuities-service-charges/99820420/	200		
23	https://www.usatoday.com/money/lookup/stocks/ge/	200		
24	https://www.usatoday.com/picture-gallery/tech/2017/09/12/iphone-x--iphone-8-new-apple-products-at-10th-anniversary-event/105530320/	200		
25	https://www.usatoday.com/travel/	200		
26	http://www.usatoday.com/conversation-guidelines/index.html	301		
27	https://www.usatoday.com/story/travel/cruises/2017/08/21/regent-seven-seas-ship-circle-south-america-2019/585211001/	200		
28	https://www.usatoday.com/policing/beyond-badge/	200		

Sample Visit File for USA Today

visit_USAToday

A1 https://www.usatoday.com/

	A	B	C	D
1	https://www.usatoday.com/	382590	259	text/html
2	https://www.usatoday.com/travel/cruises/	296808	236	text/html
3	https://www.usatoday.com/picture-gallery/weather/2017/09/07/hurricane-irma-damage-and-destruction-in-the-caribbean/105350960/	245521	105	text/html
4	https://www.usatoday.com/money/business/	271990	210	text/html
5	https://www.usatoday.com/story/travel/cruises/2017/04/04/25-cruise-ship-attractions-blow-your-mind/99949914/	530478	234	text/html
6	https://www.usatoday.com/sports/nba/	582209	418	text/html
7	https://www.usatoday.com/topic/hurricane-harvey/local	224644	197	text/html
8	https://www.usatoday.com/story/travel/cruises/2016/12/27/harvest-caye-cruise-port-belize/95881082/	401517	170	text/html
9	https://www.usatoday.com/money/lookup/stocks/aapl/	141887	137	text/html
10	https://www.usatoday.com/travel/roadwarriorvoices/	256391	203	text/html
11	https://www.usatoday.com/life/entertainthis/	274231	219	text/html
12	https://www.usatoday.com/sports/nba/injuries/	222900	292	text/html
13	https://www.usatoday.com/life/tv/	268628	211	text/html
14	https://www.usatoday.com/story/money/markets/2017/09/12/home-capital-investors-reject-investment-warren-buffetts-firm/658902001/	165958	56	text/html
15	https://www.usatoday.com/story/travel/cruises/2017/02/20/biggest-cruise-ship-suites/98146866/	817187	356	text/html
16	https://www.usatoday.com/sports/nfl/injuries/	586027	318	text/html
17	https://www.usatoday.com/picture-gallery/weather/2017/09/05/hurricane-irma-preys-on-florida-and-the-us/105300974/	431773	200	text/html
18	https://www.usatoday.com/sports/nhl/sagarin/	257099	302	text/html
19	https://www.usatoday.com/story/travel/cruises/2017/04/03/cruise-ship-gratuities-service-charges/99820420/	662824	302	text/html
20	https://www.usatoday.com/money/lookup/stocks/ge/	141887	137	text/html
21	https://www.usatoday.com/picture-gallery/tech/2017/09/12/iphone-x--iphone-8-new-apple-products-at-10th-anniversary-event/105530320/	238718	100	text/html
22	https://www.usatoday.com/travel/	426522	262	text/html
23	https://www.usatoday.com/story/travel/cruises/2017/08/21/regent-seven-seas-ship-circle-south-america-2019/585211001/	363323	146	text/html
24	https://www.usatoday.com/policing/beyond-badge/	888171	345	text/html
25	https://www.usatoday.com/topic/efb4d49e-13af-44db-89ac-a3d51e74851d/small-business/	295782	250	text/html
26	https://www.usatoday.com/sports/wnba/scores/	207600	295	text/html
27	https://www.usatoday.com/sports/motor-sports/	302519	319	text/html
28	https://www.usatoday.com/topic/c75c4482-ce20-4c69-8461-de49880dee08/small-business-central/	304103	254	text/html
29	https://www.usatoday.com/story/news/nation-now/2017/09/12/fire-damages-robert-e-lee-elementary-tampa/660355001/	150870	55	text/html

What to Submit

- Compress all of the above into a single zip archive and name it:

  crawl.zip
- Use only standard zip format. Do **NOT** use other formats such as zipx, rar, ace, etc. For example the zip file might contain the following three files:

1. CrawlReport_usatoday.txt,
2. fetch_usatoday.csv
3. visit_usatoday.csv

- To submit your file electronically to the csci572 account enter the following command from your UNIX prompt:
- $ submit -user csci572 -tag hw2 crawl.zip