**Solr + Nutch Tutorial**

**Requirements**

To follow along with this tutorial, you will need...

1. Unix environment, or Windows-[Cygwin](https://www.cygwin.com/) environment
2. Java Runtime/Development Environment (1.7 or greater). Some places you can get it are from [Oracle](http://www.oracle.com/technetwork/java/javase/downloads/index.html), [Open JDK](http://openjdk.java.net/), or [IBM](http://www.ibm.com/developerworks/java/jdk/).
   * Running **java -version** at the command line should indicate a version number starting with 1.7.
   * Gnu's GCJ is not supported and does not work with Solr.
3. A [Solr release](http://lucene.apache.org/solr/mirrors-solr-latest-redir.html).
4. Apache Ant: <http://ant.apache.org/>

## Install Solr

user:~solr$ **ls**

solr-4.10.2.zip

user:~solr$ **unzip -q solr-4.10.2.zip**

Solr can run in any Java Servlet Container of your choice, but to simplify this tutorial, the example index includes a small installation of Jetty.

To launch Jetty with the Solr WAR, and the example configs, just run the **start.jar**. But before that, please cd solr-4.10.2/example/solr/collection1/conf directory, and then replace file schema.xml with that posted in the [NSF-Polar-Cyberinfrastructure](https://github.com/NSF-Polar-Cyberinfrastructure/issue-1/blob/master/conf/acadis) github.

user:~solr$ **cd solr-4.10.2/example/**

user:~/solr/example$ **java -jar start.jar**

0 [main] INFO org.eclipse.jetty.server.Server – jetty-8.1.10.v20130312

21 [main] INFO org.eclipse.jetty.deploy.providers.ScanningAppProvider – Deployment monitor /Users/AngelaWang/Downloads/acadis/solr-4.10.2/example/contexts at interval 0

27 [main] INFO org.eclipse.jetty.deploy.DeploymentManager – Deployable added: /Users/AngelaWang/Downloads/acadis/solr-4.10.2/example/contexts/solr-jetty-context.xml

1066 [main] INFO org.eclipse.jetty.webapp.StandardDescriptorProcessor – NO JSP Support for /solr, did not find org.apache.jasper.servlet.JspServlet

...

2980 [searcherExecutor-6-thread-1] INFO org.apache.solr.core.SolrCore – [collection1] Registered new searcher Searcher@4251b296[collection1] main{StandardDirectoryReader(segments\_1:1:nrt)}

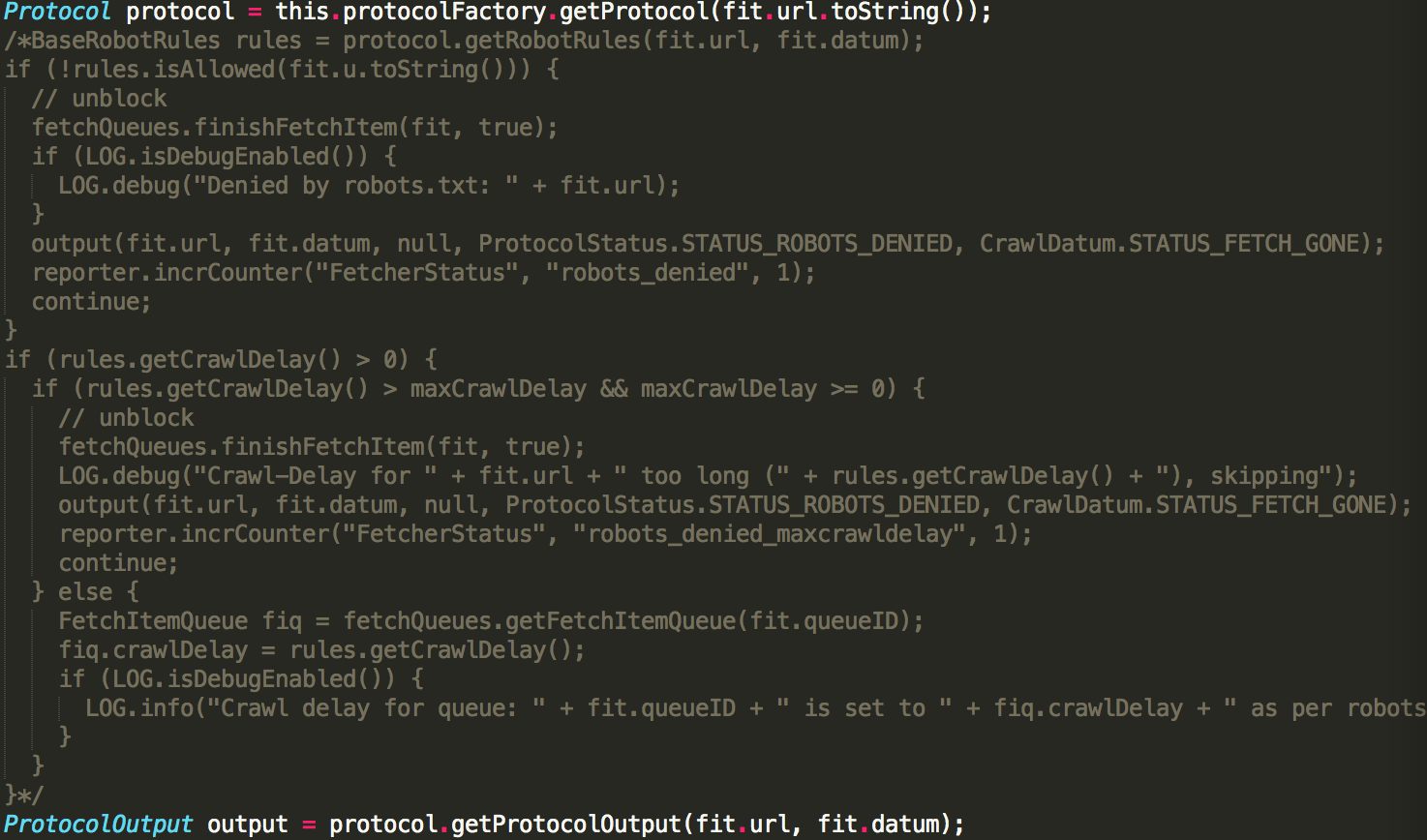
105400 [qtp733662211-12] INFO org.apache.solr.servlet.SolrDispatchFilter – [admin] webapp=null path=/admin/cores params={indexInfo=false&\_=1418277242649&wt=json} status=0 QTime=3

This will start up the Jetty application server on port 8983, and use your terminal to display the logging information from Solr.

You can see that the Solr is running by loading <http://localhost:8983/solr/> in your web browser. This is the main starting point for Administering Solr.

## Install Nutch - Set up Nutch from a source distribution

* Download a [Nutch source package](http://nutch.apache.org/downloads.html), e.g., apache-nutch-1.9-src.zip
* Unzip
* cd apache-nutch-1.9/src/java/org/apache/nutch/fetcher
* vim Fetcher.java, in FetcherThread inner class, comment out the following lines to disable robot check.



* cd apache-nutch-1.9/
* Run ant in this folder
* Now there is a directory runtime/local which contains a ready to use Nutch installation.

When the source distribution is used ${NUTCH\_RUNTIME\_HOME} refers to apache-nutch-1.9/runtime/local/. Note that

* config files should be modified in apache-nutch-1.9/runtime/local/conf/
* ant clean will remove this directory (keep copies of modified config files)

## Verify your Nutch installation

* run "bin/nutch" - You can confirm a correct installation if you see something similar to the following:
* Usage: nutch COMMAND

where COMMAND is one of:

readdb read / dump crawl db

mergedb merge crawldb-s, with optional filtering

readlinkdb read / dump link db

inject nject new urls into the database

generate generate new segments to fetch from crawl db

freegen generate new segments to fetch from text files

fetch fetch a segment's pages

## Crawl the ACADIS website

Nutch requires two configuration changes before a website can be crawled:

1. Customize your crawl properties, where at a minimum, you provide a name for your crawler for external servers to recognize
2. Set a seed list of URLs to crawl

### 4.1 Customize your crawl properties

* Default crawl properties can be viewed and edited within conf/nutch-default.xml - where most of these can be used without modification
* The file conf/nutch-site.xml serves as a place to add your own custom crawl properties that overwrite conf/nutch-default.xml.

## cd conf directory, and then replace nutch-site.xml, regex-normalize.xml, regex-urlfilter.txt and solrindex-mapping.xml files with those posted in the [NSF-Polar-Cyberinfrastructure](https://github.com/NSF-Polar-Cyberinfrastructure/issue-1/blob/master/conf/acadis) github.

### 4.2 Create a URL seed list

* A URL seed list includes a list of websites, one-per-line, which nutch will look to crawl
* The file conf/regex-urlfilter.txt will provide Regular Expressions that allow nutch to filter and narrow the types of web resources to crawl and download

#### Create a URL seed list

* cd bin
* mkdir -p urls
* cd urls
* touch seed.txt to create a text file seed.txt under urls/ with the following content (one URL per line for each site you want Nutch to crawl).
* https://www.aoncadis.org/home.htm

### Using the crawl script

* Usage: ./crawl <seedDir> <crawlDir> <solrURL> <numberOfRounds>

Example: ./crawl urls/ AcadisCrawl/ http://localhost:8983/solr/ 30

Typically one starts testing one's configuration by crawling at shallow depths, sharply limiting the number of rounds, and watching the output to check that desired pages are fetched and undesirable pages are not. Once one is confident of the configuration, then an appropriate depth for a full crawl is around 10.