GettingNutchRunningWithWindows

Since Nutch is written in Java, it is possible to get Nutch working in a Windows environment, provided that the correct software is installed.

Note: If you're just interested in a basic installation on Windows and are not interested in knowing the details of how it is done, you might want check and see if theWhelanLabs SearchEngine Manager fits your needs. It is a free installer for Nutch on Windows.

The following documents describe how I got it working on Windows XP Pro running Tomcat 5.28. Edit: page updated with my experience installing on Windows Server 2003.

Required Software

Java

You will need to have Java 1.4.2 (or Java 1.5 for Nutch 0.8.x or higher) installed.

This also works with Java 6, Nutch 0.9, and Tomcat 6. Just the Java 6 JRE is necessary, unless you want to build nutch from sources yourself.

Cygwin

You'll need cygwin to run the shell commands since there are no separate scripts for NT cmd (the NT cmd shell does not nest environments recursively). Mks ksh does not work correctly with the scripts. Make sure you have installed the utility 'uname' in cygwin.

See also GettingNutchRunningOnCygwin for more details about configuring cygwin when using nutch.

Tomcat

You'll need Tomcat 4.* or higher running on your machine. I know of no reason to not go with the latest release (Tomcat 6 at time of last writing).

Setup

Download

Download the release and extract on your hard disk in a directory that *does not* contain a space in it (e.g., `c:\nutch-0.9`). If the directory does contain a space (e.g., `c:\my programs\nutch-0.9`), the Nutch scripts will not work properly.

Create an empty text file (use any name you wish) in your nutch directory (e.g., `urls`) and add the URLs of the sites you want to crawl.

Add your URLs to the `crawl-urlfilter.txt` (e.g., `C:\nutch-0.9\conf\crawl-urlfilter.txt`). An entry could look like this:

```
+^http://([a-z0-9]*\.)*apache.org/
```

Load up cygwin and navigate to your `nutch` directory. When cygwin launches, you'll usually find yourself in your user folder (e.g. `C:\Documents and Settings\username`).

If your workstation needs to go through a Windows Authentication Proxy to get to the Internet (this is not common), then you can use an application such as the NTLM Authorization Proxy Server to get through it. You'll then need to edit the `nutch-site.xml` file to point to the port opened by the app.

Intranet Crawling

Follow the tutorial instructions to begin the crawl by entering commands in cygwin. Nutch will create a crawl directory and a log file.

For example, if you enter the following command from the root of your Nutch install:

```
bin/nutch crawl urls -dir crawl -depth 3 >& crawl.log
```

then a folder called `crawl` is created in your `nutch` directory, along with the crawl.log file. Use this log file to debug any errors you might have.

You'll need to delete or move the crawl directory before starting the crawl off again unless you specify another path on the command above.

Analyzing Additional Resource Types

From the Features:

Edit `conf/nutch-site.xml` and change the value of `plugin.includes` to include the plugins for the document types that you want Nutch to handle.

Example: to add parsing for PDF, MS Office, and OpenOffice documents, you'll have something like:

```
<property>
  <name>plugin.includes</name>
  <value>protocol-http|urlfilter-regex|parse-(text|html|js|msexcel|mspowerpoint|msword|oo|pdf|swf|zip)|
index-basic|query-(basic|site|url)|summary-basic|scoring-opic|
urlnormalizer-(pass|regex|basic)</value>
</property>
```

Web Interface for Search

In your Environment Variables settings, add `NUTCH_JAVA_HOME` and the location of your JVM (e.g. `C:\j2sdk1.4.2_09`) as a new Environment Variable.

Open up a web browser and navigate to the Tomcat webapps manager (e.g. `http://localhost:8080/manager/html`) and upload the nutch WAR file to the context.

If you are going to run nutch in the root context *and* a root context already exists, undeploy it. Otherwise, skip to the Alternative, below.

Create a context fragment file so that the root url points to your nutch webapp. Navigate to your [tomcat_home]/conf/Catalina/localhost/ and put it there. Create a new xml file (name it the same as the webapp?) e.g. nutch-0.9.xml and add something like the following line to it.

```
<Context path="" debug="5" privileged="true" docBase="nutch-0.7.1"/>
```

Alternative: if you want to run other web applications alongside nutch, copy or rename the `nutch-0.9.0.war` to whatever you'd like the subdirectory URL to be. Deploy the renamed version using the Tomcat Web Application Manager.

For example, to use nutch via `http://localhost/search/`, rename the nutch `.war` file to `search.war` and then deploy `search.war`.

Set Your Searcher Directory

Next, navigate to your nutch webapp folder then `WEB-INF/classes`. Edit the `nutch-site.xml` file and add the following to it (make sure you don't have two sets of <configuration></configuration> tags!):

```
<configuration>
  <property>
    <name>searcher.dir</name>
    <value>your_crawl_folder_here</value>
  </property>
</configuration>
```

For example, if your nutch directory resides at `C:\nutch-0.9.0` and you specified `crawl` as the directory after the `-dir` command, then enter `C:\nutch-0.9.0\crawl` instead of `your_crawl_folder_here`.

Reload

Reload the Application. Use the Tomcat Manager and simply click the "Reload" command for nutch, or restart Tomcat using the windows services tool.

Open up a browser and enter the url `http://localhost:8080`. The nutch search page should appear. As long as you've defined the correct location of your nutch index directory (as shown above), clicking search should yield results.