FabioGiavazzi HowtoGettingNutchRunningonWindows

Howto to setup nutch on a Windows Server 2008 R2 Enterprise(64-bit) and crawl samba shares.

First of all you need to download the following software:

Java 1.6 (or newer version):

http://www.oracle.com/technetwork/java/javase/downloads/index.html

Tomcat 7:

http://tomcat.apache.org/download-70.cgi

Cygwin:

http://www.cygwin.com/

Nutch-1.2 (or newer version):

ftp://mirror.switch.ch/mirror/apache/dist//nutch/ (apache-nutch-1.2-bin.zip)

Step 1:

Install Cygwin, (run cygwin.exe) follow the setup-assistant.

Setp 2:

Install Java (run jdk-6u24-windows-i586.exe) and set JAVA_HOME in_ Start -> Computer -> Properties -> Advanced system settings -> Advanced -> Environment Variables..._

- http://img576.imageshack.us/img576/2909/picture1fmi.png

(Use 32-bit Version of Java, there are some troubles with the 64-bit version and the os!)

Step 3:

Install Tomcat, (run apache-tomcat-7.0.11.exe).

After installation, Tomcat should start the service automatically. When the service is not running, start it manually by clicking on Configure Tomcat and then Start:

http://img228.imageshack.us/img228/4713/picture2nwn.png

Now go to http://localhost:8080 in your browser and check if you see the following screen:

http://img689.imageshack.us/img689/4540/picture002lw.png

Step 4:

For crawling samba share, you first have to setup the networkdrive:

(In this example it's ipa-data1)

http://img806.imageshack.us/img806/3530/picture3fl.png

Step 5:

Unzip the apache-nutch-1.2-bin.zip to any directory you like, I prefer C:\:

http://img850.imageshack.us/img850/4992/picture4c.png

Now go to the nutch-1.2 directory and create an_ urls_ folder.

In this folder, you create a text file with any name you like (e.g. files). Now edit it and paste your file urls:

http://img560.imageshack.us/img560/8753/picture6l.png

You have to type file:///, otherwise it won't work.

Step 6:

Go to the nutch-1.2\conf directory and edit the nutch-default.xml:

http://img834.imageshack.us/img834/8493/picture7vq.png

Here we have to change the property plugin-includes and set the limit for file content to -1 for unlimited file length. Take a look at the changes:

http://img851.imageshack.us/img851/9772/picture8b.png

Change the value protocol-http to protocol-file in plugin-includes (Don't change the other default values):

http://img269.imageshack.us/img269/3586/picture9ks.png

To specifie that nutch only crawls your specified links in the folder urls, you have to disable this property with set it to false:

http://img22.imageshack.us/img22/5789/picture10xc.png

Step 7:

Go to nutch-1.2\conf\ and edit the file crawl-urlfilter.txt:

http://img%38%35%36.imageshack.us/img%38%35%36/%36%36%32%36/picture%31%32m.png

Change -(file|ftp|mailto) to -(http|ftp|mailto)

Disable skip URLs with slash-delimited & accept hosts in MY.DOMAIN.NAME

Change skip everything else to accept everything else

Step 8:

Edit the file nutch-1.2\conf\nutch-site.xml, paste some default properties:

<configuration>

<property>

- <name>http.agent.name</name> <value>test</value>
- <description>test </description>

</property>

<property>

- <name>http.agent.description</name>
- <value>Nutch</value>
- <description>Nutch </description>

</property>

<property>

- <name>http.agent.url</name>
- <value>http://test.url </value>
- <description>http://test.url </description>

</property>

<property>

- <name>http.agent.email</name>
- <value> test@test.ch </value>
- <description> test@test.ch </description>
- </property>
- </configuration> .

Step 9:

Open cygwin.exe and run the crawl, just use this command:

(First, navigate to the nutch-1.2 directory with cd /cygdrive/c/nutch-1.2)

http://img%36%38%33.imageshack.us/img%36%38%33/%32%36%30%36/picture%31%33an.png

Options which u can use:

• -dir *dir* names the directory to put the crawl in

• -threads *threads* determines the number of threads that will fetch in parallel

• -depth *depth* indicates the link depth from the root page that should be crawled

Step10:

To use the Tomcat manager you have to edit the tomcat-users.xml in C:\Program Files (x86)\Apache Software Foundation\Tomcat 7.0\conf\:

http://img%37.imageshack.us/img%37/%31%30%37%30/picture%31%34ap.png

Add a new user and new role, like this:

http://img%32%30%32.imageshack.us/img%32%30%32/%37%35%36%33/picture%31%35t.png

Save the settings and restart Tomcat (Take a look at Step 3).

Step 11:

Go to http://localhost:8080/manager/html in the browser (login with the user in Step 10).

In the WAR file to deploy section, select the \nutch-1.2\nutch-1.2.war file to upload:

http://img%31%36.imageshack.us/img%31%36/%36%31%37%35/picture%31%36bl.png

Then you will see the /nutch-1.2 in the list, start it.

Go to C:\Program Files (x86)\Apache Software Foundation\Tomcat 7.0\webapps\ and you will see that there is a folder called nutch-1.2.

Step 12:

Navigate to C:\Program Files (x86)\Apache Software Foundation\Tomcat 7.0\webapps\nutch-1.2\WEB-INF\classes\ and edit the nutch-site.xml:

<configuration>

- <property>
  - <name>searcher.dir</name> <value>your_crawl_folder (like C:\nutch-1.2\crawl)</value>
    </property>

</configuration>

After that, restart Tomcat.

Step 13:

Go to http://localhost:8080/nutch-1.2 and you should see the following:

http://img%38%36%30.imageshack.us/img%38%36%30/%33%35%34%30/picture%31%37r.png

Now you can search for your files!

Don't forget, that you have to set up the networkdrives on every system, to enable editing files directly over nutch!

http://img%32%31.imageshack.us/img%32%31/%39%34%35/picture%31%38p.png

Enjoy!