GettingNutchRunningWithUbuntu

Recently, and with a bit of effort, I got db1.spack up and running on nutch trunk. I decided to keep track of what I did to get db2.spack up and running, and contribute this tutorial.

Install Ubuntu

Here are some minimal steps:

- get either the Desktop or Server Version from http://www.ubuntu.com/download/
- boot and install
- familiarize yourself with: sudo su -

Add Nutch User

Let's add a nutch user to do our nutch stuff

```
# adduser nutch
```

java

I tried to get java from normal apt sources and I am guessing it is my Athlon that broke me. I broke down and got java from Sun (http://java.sun.com/j2se/1.5.0/download.jsp), the Download JDK 5.0 Update 4 link. I tried getting the 1.4.2 and it didn't work, but 1.5.0 worked.

```
root@db2:/opt# ./jdk-1_5_0_04-linux-amd64.bin
```

You might also want to follow the instructions for Debian-izing the Sun JDK: http://plugindoc.mozdev.org/faqs/distronotes/ubuntu-x86.html#java-sun

Let's put JAVA_HOME in our ~/.bash_profiles, and source said ~/.bash_profiles for root and nutch

```
# echo 'export JAVA_HOME=/opt/jdk1.5.0_04' >> ~/.bash_profile
# . ~/.bash_profile
nutch@db2:~$ echo 'export JAVA_HOME=/opt/jdk1.5.0_04' >> ~/.bash_profile
nutch@db2:~$ . ~/.bash_profile
```

apt

Add the Multiverse to your sources.list or use the GUI:

System -> Administration -> Synaptic Package Manager

Settings -> Repositories

With the new apt sources, let's update

```
# apt-get update
```

And get the packages we need.

```
# apt-get install subversion ant ant-optional lynx
```

subversion is used to get nutch, ant is used to build nutch and lynx is used to test nutch.

Build Nutch Code and Index

Let's change over to the nutch user

```
# su - nutch
```

Checkout the code AND the gora code

```
nutch@db2:~$ svn checkout http://svn.apache.org/repos/asf/nutch/trunk nutch
nutch@db2:~$ cd nutch
nutch@db2:~$ svn checkout https://svn.apache.org/repos/asf/incubator/gora/
```

Since this tutorial is for getting trunk to work, let's go there

```
nutch@db2:~ $ cd ~/nutch
```

We build with ant

```
nutch@db2:~/nutch $ ant
```

And build a war for tomcat and later searching

```
nutch@db2:~/nutch/trunk $ ant war
```

Follow the nutch tutorial (http://lucene.apache.org/nutch/tutorial.html) to build a index, or for a simple index:

If you are using the latest "trunk" stuff, the url seeding has been changed from a single file to a directory. Using trunk (after 0.7.2), put the urls in a file (here, called "nutch") in a DIRECTORY called "urls":

```
nutch@db2:~/nutch $ mkdir urls
nutch@db2:~/nutch $ echo 'http://lucene.apache.org/nutch/' > urls/nutch
```

Using 0.7.2 or before, just put urls in a FILE called "urls":

```
nutch@db2:~/nutch $ echo 'http://lucene.apache.org/nutch/' > urls
```

Then, in any case, you specify in the same fashion ("urls" below referring either to a dir or a file, depending on the version you're using):

```
nutch@db2:~/nutch $ perl -pi -e 's|MY.DOMAIN.NAME|lucene.apache.org/nutch|' \
  conf/crawl-urlfilter.txt
nutch@db2:~/nutch $ src/bin/nutch crawl urls -dir crawl.test -depth 3
```

See, perl can be useful 🙂

tomcat

Again, I tried apt without much luck, so I downloaded tomcat from Apache (http://jakarta.apache.org/site/downloads/downloads_tomcat-4.cgi).

As above, I put the java stuff in /opt

```
root@db2:/opt# tar -xzvf jakarta-tomcat-4.1.31.tar.gz
```

Out with the old and in with the new

```
# rm -rf /opt/jakarta-tomcat-4.1.31/webapps/ROOT*
# cp ~nutch/nutch/trunk/build/nutch-0.8-dev.war \
    /opt/jakarta-tomcat-4.1.31/webapps/ROOT.war
```

Let's move to where we put the index

```
# cd ~nutch/nutch/trunk/crawl.test
```

And start tomcat from there

```
# /opt/jakarta-tomcat-4.1.31/bin/catalina.sh start
```

Test

Connect to tomcat and perform a search.

```
$ lynx localhost:8080
```

I searched for 'nutch' and all was well! (you can use <TAB> to get to the search input in lynx)

Tutorial written by Earl Cahill, 2005