GettingNutchRunningWithFedoraCore

This is based on GettingNutchRunningWithRedHatApplicationServer. To make this easier to start we are using the yum command line as an example.

Repositories we need

- Core (for sure)
- Extras
- maybe Jpackage

Packages to Install

This is a primary list from the Redhat server

```
yum install ant ant-apache-regexp axis jaf jakarta-commons-beanutils jakarta-commons-collections jakarta-
commons-daemon jakarta-commons-dbcp jakarta-commons-digester jakarta-commons-discovery jakarta-commons-el
jakarta-commons-fileupload jakarta-commons-httpclient jakarta-commons-launcher jakarta-commons-logging jakarta-
commons-modeler jakarta-commons-pool jakarta-commons-validator jakarta-regexp jakarta-taglibs-standard  jakarta-
taglibs-standard-javadoc javamail jta jta-javadoc junit
libgcj34 log4j mx4j oro regexp servletapi4 servletapi5 struts11 tomcat5 tomcat5-admin-webapps tomcat5-webapps
tyrex wsdl4j xalan
xerces xml-commons xml-commons-apis xml-commons-resolver
```

Installing for dependencies:

```
 bcel                    i386       5.1-8jpp.1       core             983 k
 eclipse-ecj             i386       1:3.2.1-4.fc6    core             7.9 M
 gcc-java                i386       4.1.1-30         core             2.8 M
 geronimo-specs          i386       1.0-0.M2.2jpp.12  core             230 k
 jakarta-oro             i386       2.0.8-3jpp.1     core             173 k
 java-1.4.2-gcj-compat-devel  i386       1.4.2.0-40jpp.110  core              49 k
 libgcj-devel            i386       4.1.1-30         core             1.4 M
 mx4j                    i386       1:3.0.1-6jpp.4   core             2.5 M
 regexp                  i386       1.4-2jpp.2       core              91 k
 wsdl4j                  i386       1.5.2-4jpp.1     core             388 k
 zlib-devel              i386       1.2.3-3          core
```

Yum Install Errors:

- No Match for argument: jta-javadoc

Install Java

- Install Linux RPM in self-extracting file

Download and Testing

- DownloadingNutch: downloaded nutch-0.8.tar.gz

```
tar xzf nutch-08.tar.gz
cd nutch-0.8

{{{
export JAVA_HOME=/usr/java/jdk1.5.0_08/
bin/nutch
```

- Test using NutchTutorial

1. make a new dir `urls`
2. add an url in a new file 'urls/nutch'

3. add/edit `conf/crawl-urlfilter.txt' (under # accept hosts in MY.DOMAIN.NAME)

```
bin/nutch crawl urls -dir crawl -depth 3 -topN 50
```

Check logs/hadoop.log for success.

Instead oft catalina.sh you starting the tomcat5 service by running:

```
/sbin/service tomcat5 start
```

You find tomcats log in /var/log/tomcat5/catalina.out

<<< FrontPage