Apache Nutch 1.7 Installation (Windows 7)

Xingquan Zhu (Jan. 2015, version 1.1)

**Attention:**

* This installation instruction was generated based on Nutch 1.7. You can use a recent Nutch version if you want.
* If you are using Mac computer, you can bypass the Cygwin installation part, so the installation is likely easier.

Apache Nutch Installation in this writing will require following software:

1. Cygwin (<http://www.cygwin.com/>)
2. Java SKD (http://www.oracle.com/technetwork/java/javase/downloads/index.html)
3. Apache Nutch (<http://nutch.apache.org/>)
4. Apache Ant (<http://ant.apache.org/>)

A very useful Nutch tutorial:

<http://wiki.apache.org/nutch/NutchTutorial>

Please note that we will not install Apache Solr (the search engine part) at this time, and will only install web crawler.

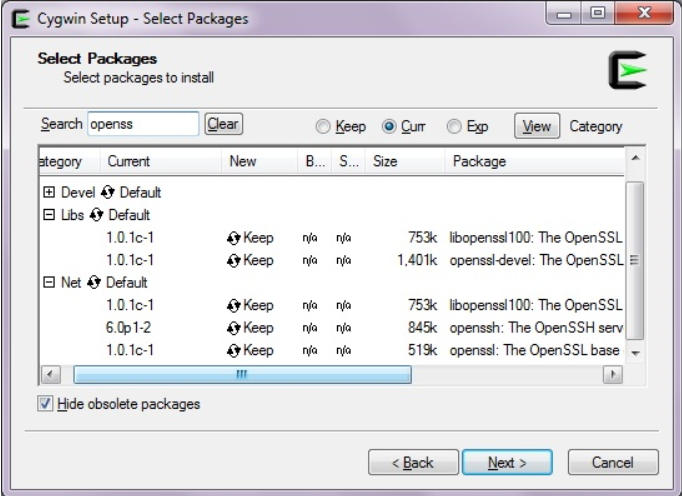
**Part 1: Cygwin Installation and Configuration**

**A:**

1. Download 'setup.exe' from Cygwin website, or you can click setup and install online

<http://www.cygwin.com/>

1. When you see the “package selection screen” window, similar to the following, please make sure that you select three packages:OpenSSL, OpenSSH, tcp\_wrappers, and diffutils.



1. click through until you come to the plugin selection window

3.1 - Make sure that the installation directory is 'C:\cygwin'

1. Once installed, go to Start -> All Programs -> Cygwin, right-click on the subsequent shortcut and select the option to 'Run as Administrator'
2. In your cygwin window, type in the following commands (one by one)

$ chmod +r /etc/passwd

$ chmod u+w /etc/passwd

$ chmod +r /etc/group

$ chmod u+w /etc/group

$ chmod 755 /var

$ touch /var/log/sshd.log

$ chmod 664 /var/log/sshd.log

1. Type 'ssh-host-config' and hit Enter

Should privilege separation be used? NO

Name of service to install: sshd

Do you want to install sshd as a service? YES

Enter the value of CYGWIN for the daemon: <LEAVE BLANK, JUST HIT ENTER>

Do you want to use a different name? (default is 'cyg\_server'): NO

Please enter the password for user 'cyg\_server': <leave blank, if possible. But if the system does not take blank password, just input a password you can remember>

Reenter: <leave blank or repeat password>

1. Once you receive message “The service has been installed under sshd\_server account. To start the service, call net start sshd…. Host configuration finished. Have fun”. It means that Cygwin has been installed properly
2. If you are experiencing any problems in the above process, you can always type ‘ssh-host-config’ again and overwrite the existing configuration and continue. Please choose to use “Should privilege separation be used: Yes”.
3. To start the ssh service, type in 'net start sshd' in your cygwin window. When you log in next time, this will automatically run.

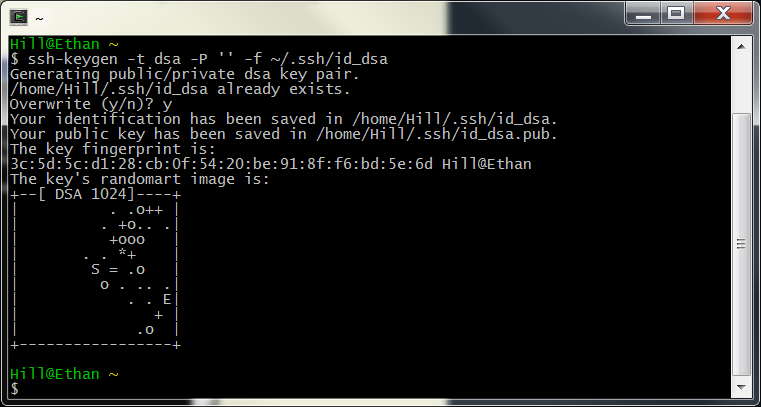
1. To test, you can type “ssh localhost” in your Cygwin window. You are expected to be prompted for nothing, but if the system ask for password (the password is your windows user password), then you need to continue step B.

**B: Configure Cygwin to avoid password**

In this step, we will generate key-based authentication that can be used so that ssh doesn’t require the use of a password every time it’s invoked. In order to do this, execute the following command in the Cygwin terminal:

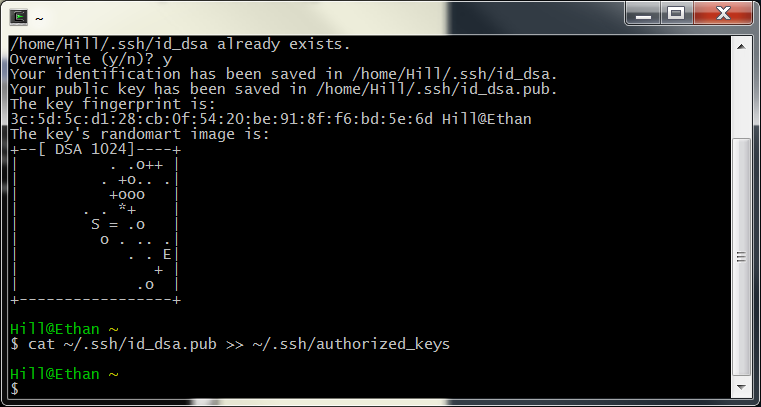
(generate public key)

1. **ssh-keygen –t dsa –P ‘’ –f ~/.ssh/id\_dsa**

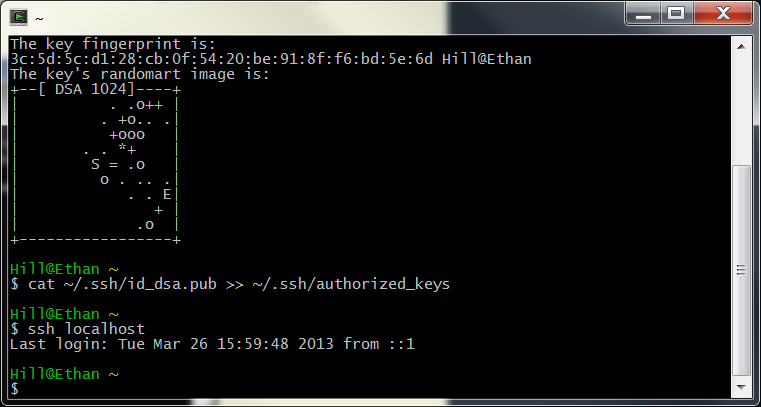


**Copy the key to the list of authorized keys for ssh**

1. **cat ~/.ssh/id\_dsa.pub >> ~/.ssh/authorized\_keys**



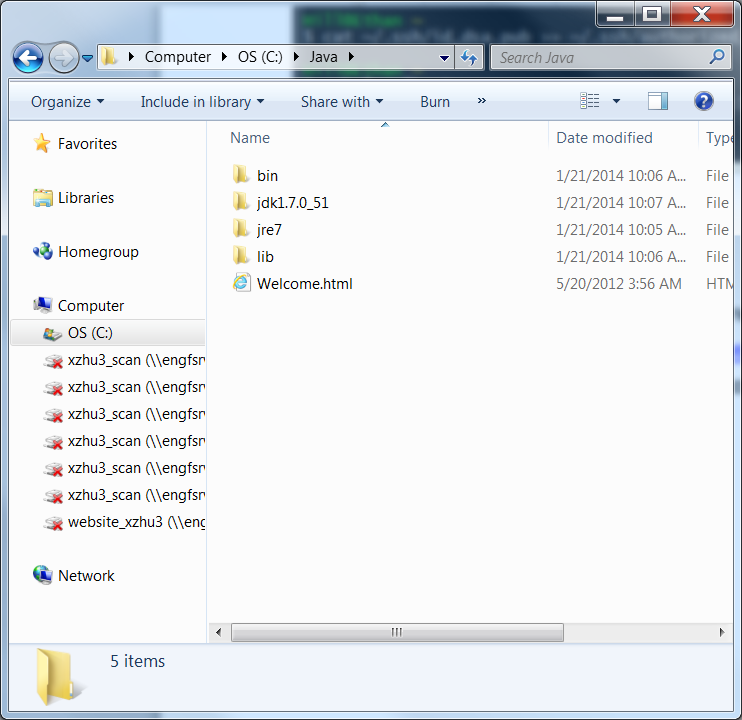
**Then when you run “ssh localhost”, you will NOT be greeted with any password.**



**Part II: Java SDK Installation:**

1. Please download java SDK from here (<http://www.oracle.com/technetwork/java/javase/downloads/index.html>), and install SDK to a new directory “C:/Java/” as follows (I used jdk.1.7.0\_51).

Please avoid installing SDK to “C:/Program Files/Java/…”, because the space key between “Program” and “Files” will cause error later on.

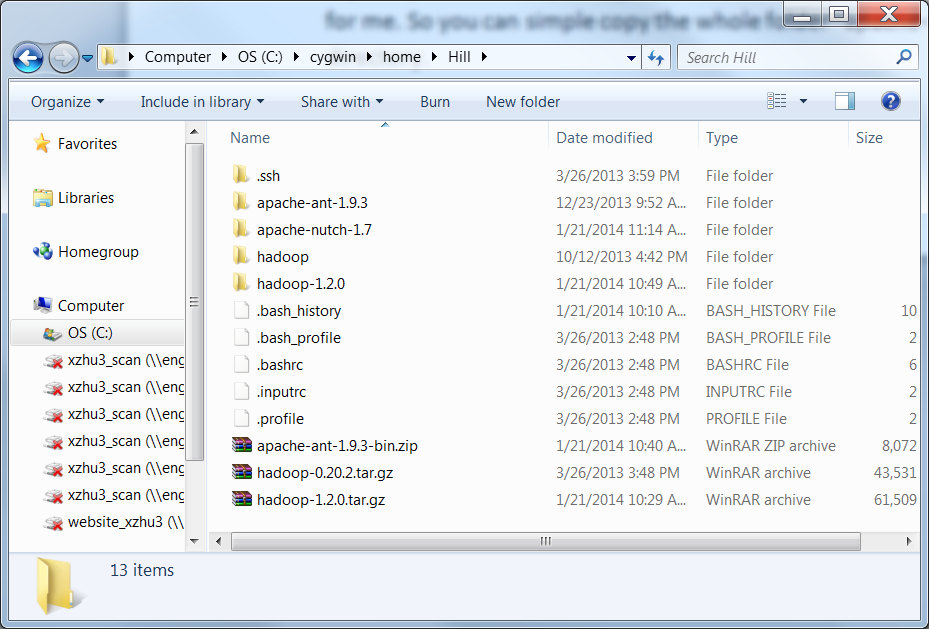


**Part III: Apache Nutch Installation**

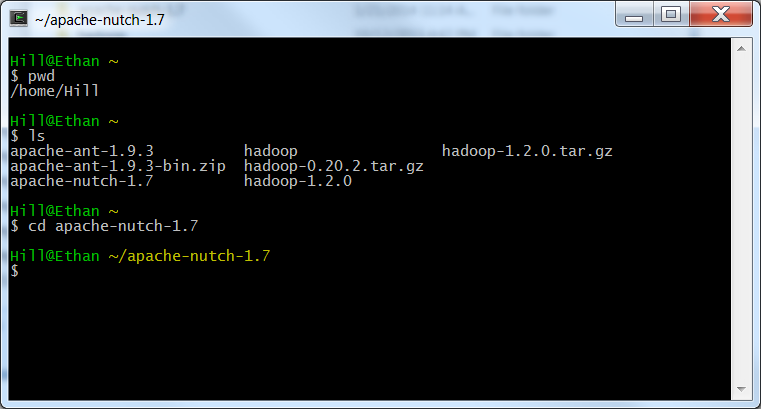
1. Please download apache-nutch-1.7.-bin.zip from here (<http://archive.apache.org/dist/nutch/1.7/>). You can download a recent Nutch version in your homework, but I validated the installation using 1.7 version in this report.
2. Upzip the downloaded zip file “[apache-nutch-1.7-bin.zip](http://www.apache.org/dyn/closer.cgi/nutch/1.7/apache-nutch-1.7-bin.zip)”, all files are now contained in a folder called “apache-nutch-1.7”.
3. Copy the whole folder to the following directory:

C:/Cygwin/home/your\_window\_ID/

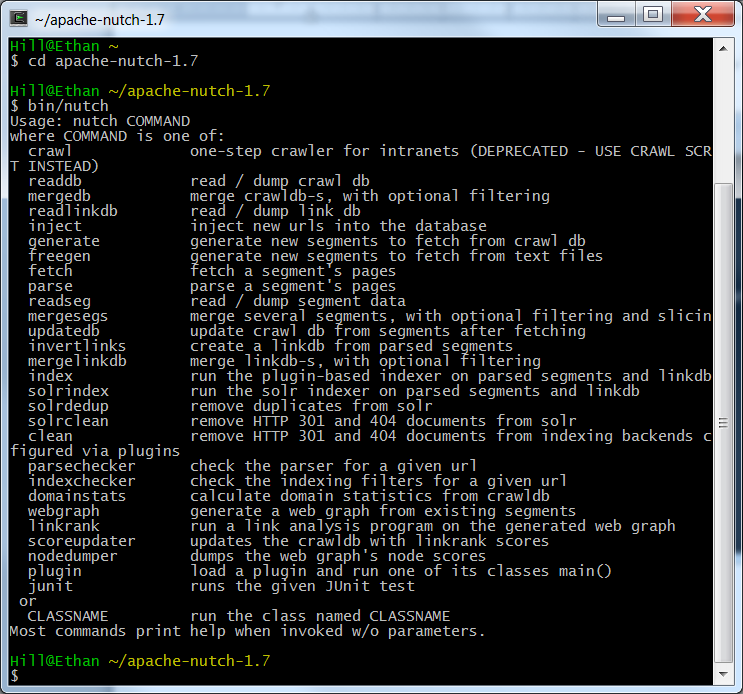
Where your\_window\_ID is the ID of your windows account. My account on my computer is “Hill”, so Cygwin will create a folder “c:/Cygwin/home/Hill” for me. So you can simple copy the whole folder “apache-nutch-1.7” to this directory.



1. Now double click Cygwin on your screen, and enter “apache-nutch-1.7” folder as follows



1. Now, type “bin/nutch” (without double quote), you will find some output as follows

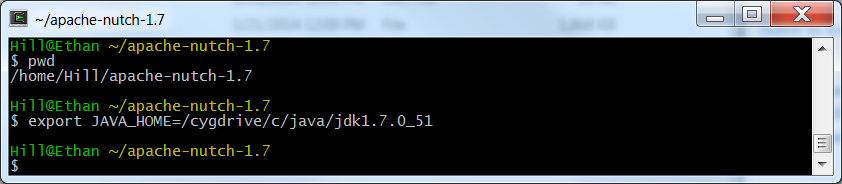


1. If you are seeing “permission denied”, you may type “chmod +x bin/nutch”, or follow suggestions listed here

<http://wiki.apache.org/nutch/NutchTutorial>

1. Set JAVA\_HOME

Please type: “export JAVA\_HOME=/cygdrive/c/java/jdk1.7.0\_51” (without double quote), showing as follows



1. Modify http.agent.name in the nutch-site.xml file. Please find “nutch-site.xml” file under “apache-nutch-1.7/conf/nutch-site.xml”, and change content as:

<?xml version="1.0"?>

<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

<property>

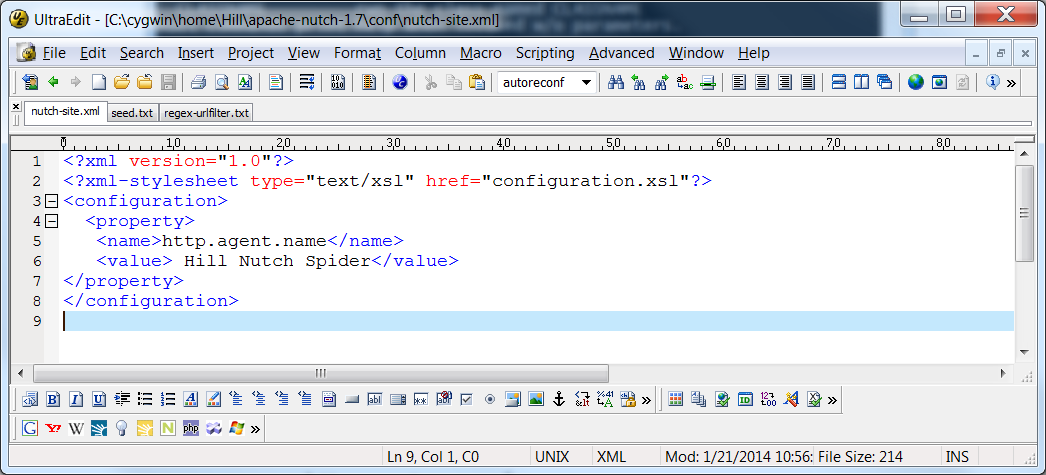
<name>http.agent.name</name>

<value> Hill Nutch Spider</value>

</property>

</configuration>

An example is showing as follows:



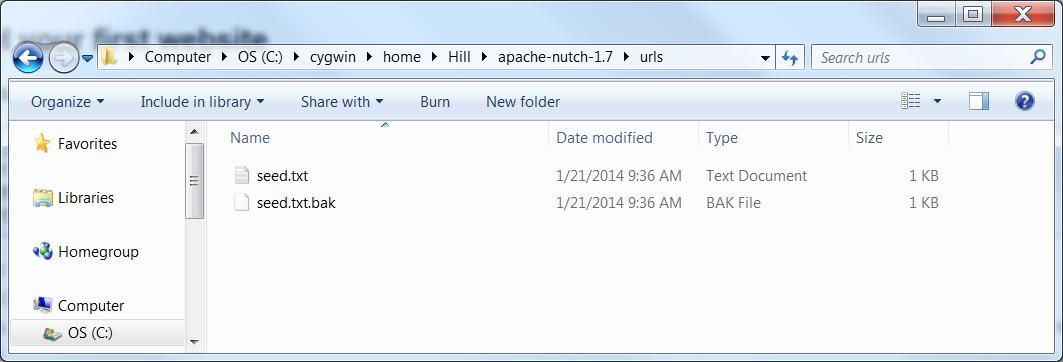
1. **Please type following commands in Cygwin (under apache-nutch-1.7 folder)**

**mkdir –p urls**

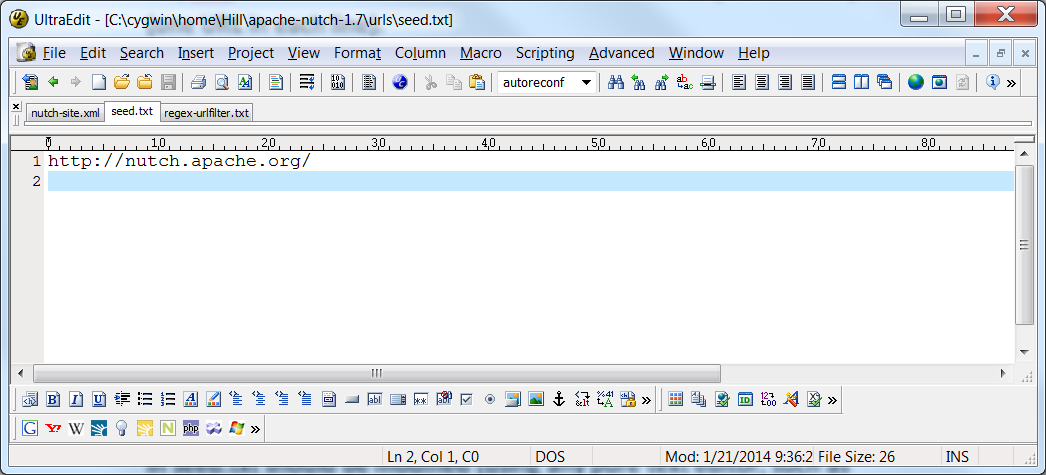
**cd urls**

**touch seed.txt**

**now you will find “seed.txt” file, which is located at “apache-nutch-1.7/urls/seed.txt”. This file will contain all seed URLs you want to provide (one URL in each line).**



In my case, I used “<http://nutch.apache.org>” as a seed url, so the content in seed.txt should be modified (using any pure text editor, such as notepad) as follows:



1. Now open “regex-urlfilter.txt” which is located at “apache-nutch-1.7/conf/regex-urlfilter.txt”.

Please replace

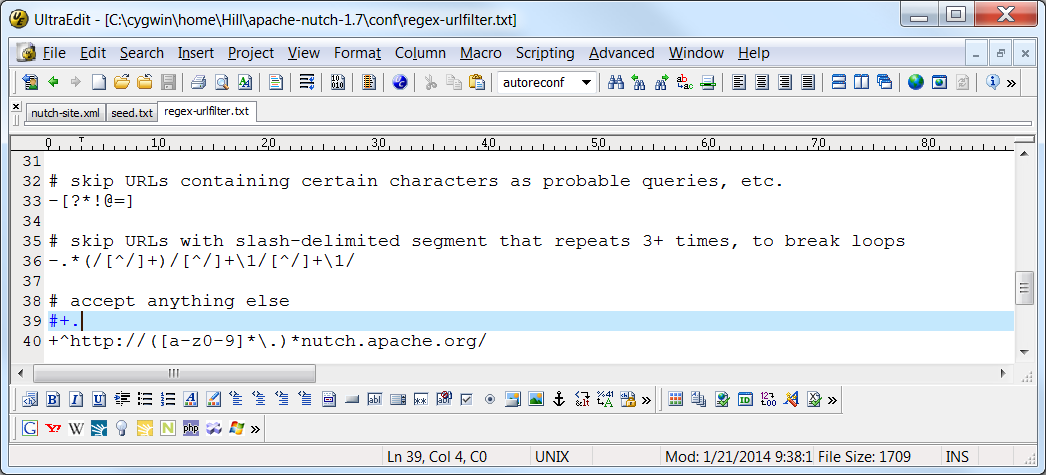
# accept anything else

+.

As following:

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

A screenshot is showing below: (please note “#” is a comment symbol.)



1. **Now type following command in Cygwin (under “**apache-nutch-1.7**”)**

bin/nutch crawl urls -dir crawl -depth 3 -topN 5

If all the above settings are correct, you will observed the programming is running with many command line (as follows)

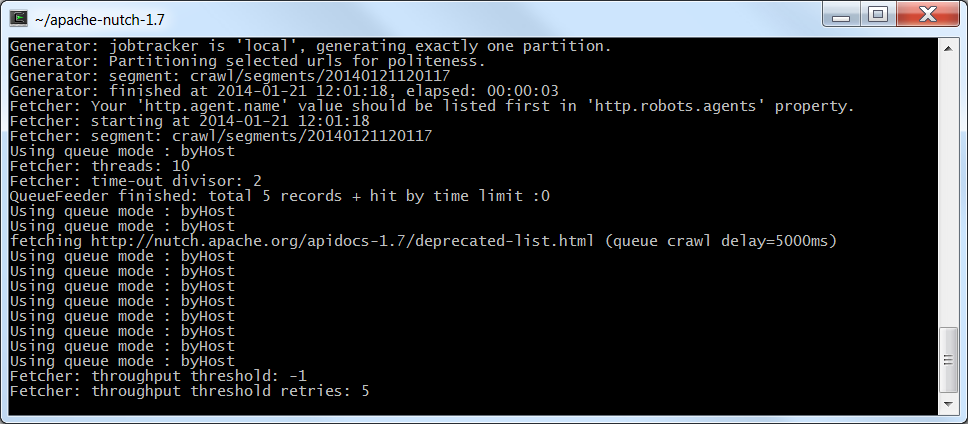
Meaning of some parameters

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

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

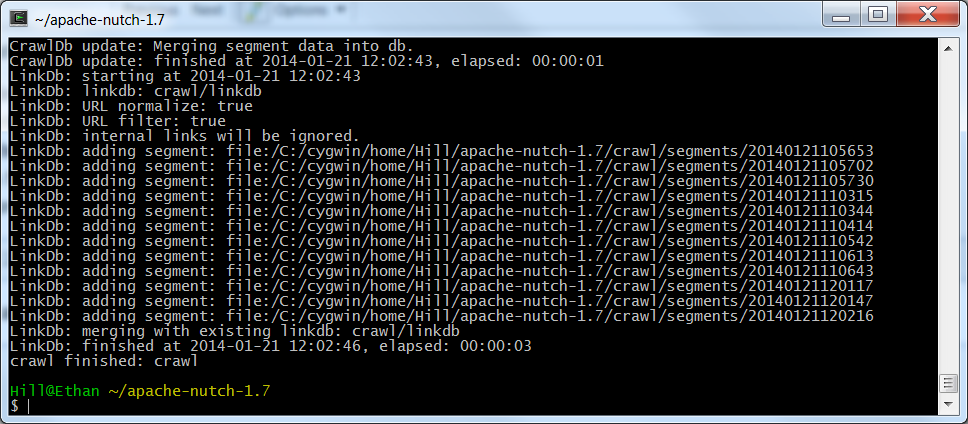
**-depth** *depth* indicates the link depth from the root page that should be crawled.

**-topN** *N* determines the maximum number of pages that will be retrieved at each level up to the depth.



Once crawling is finished, you will find a window showing:

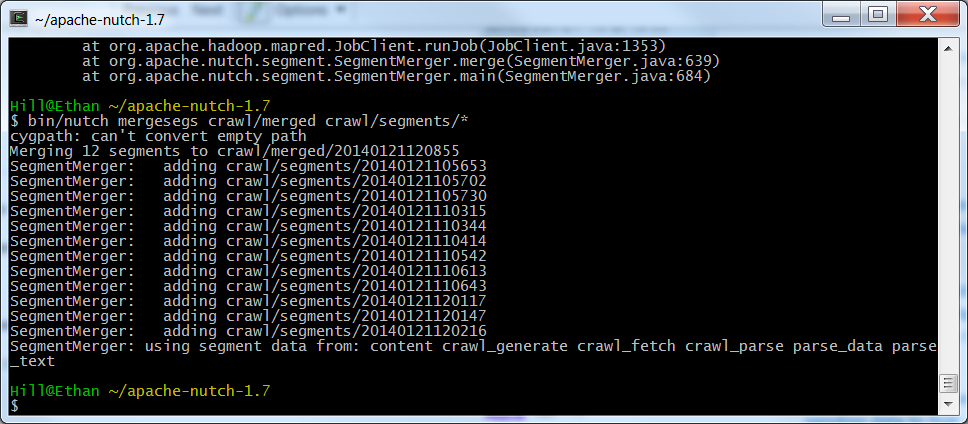
Which means that all results are saved in /crawl/segments/… folders



**The results in the above folders are indexed (binary encoded). In order to read the original files downloaded from the server. We will combine all results in the segments.**

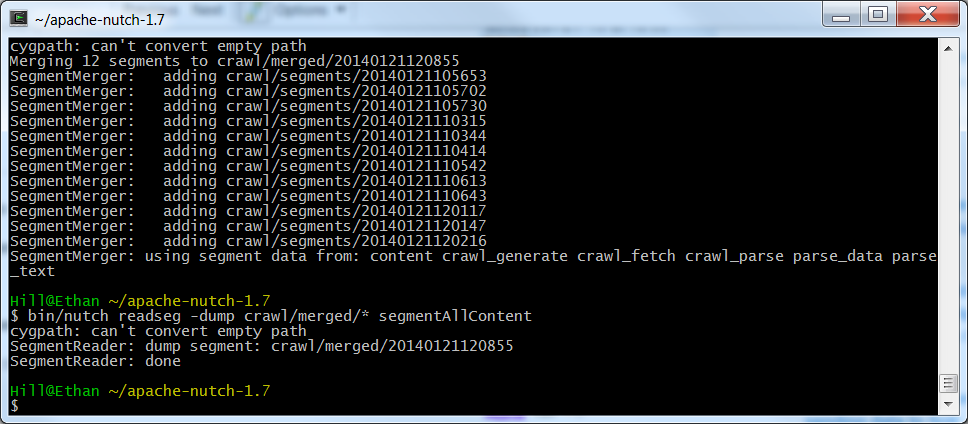
1. **Enter the following two command lines (one by one) in the Cygwin (under “apache-nutch-1.7”)**

bin/nutch mergesegs crawl/merged crawl/segments/\*

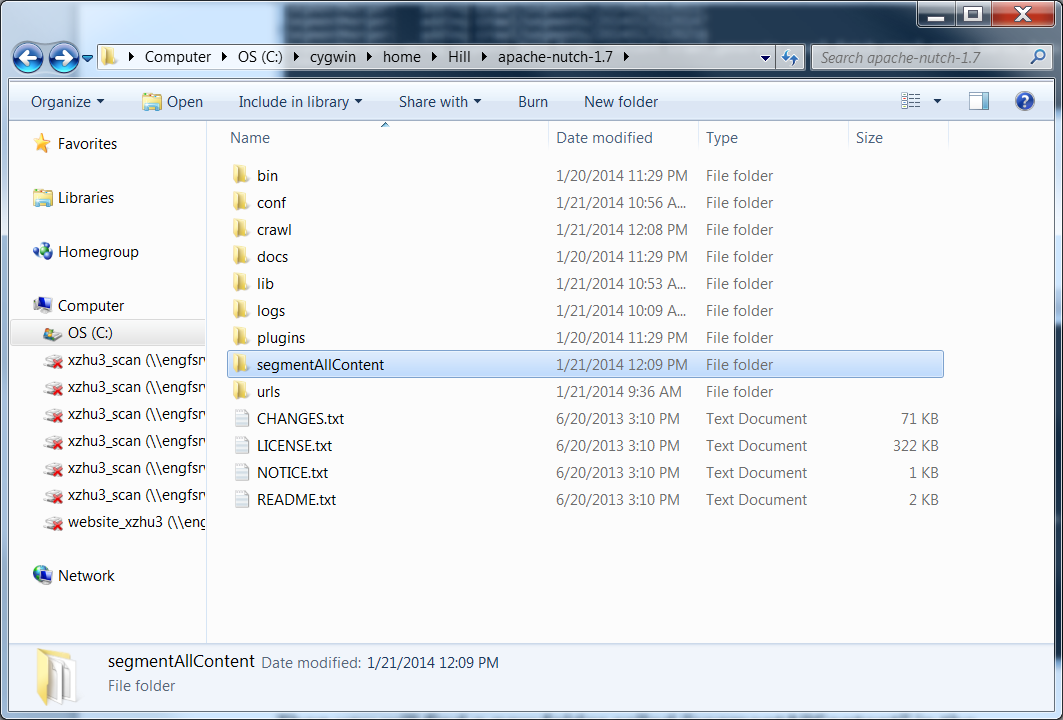


**Then type the second command**

bin/nutch readseg -dump crawl/merged/\* segmentAllContent



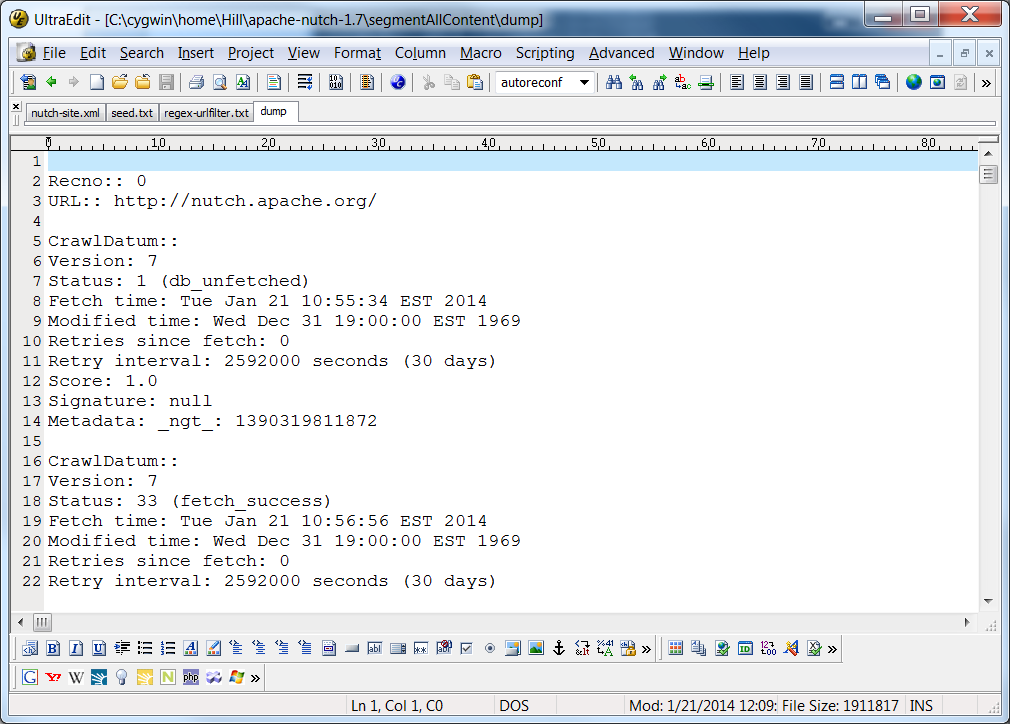
**Then you will find a new folder called “segmentAllContent” in the “apache-nutch-1.7” folder.**



**The original files crawled from the servers are included in the “dump” file.**

**I am attaching a screenshot for your information.**

**Then your first web crawling task is done!**



**Trouble Shooting:**

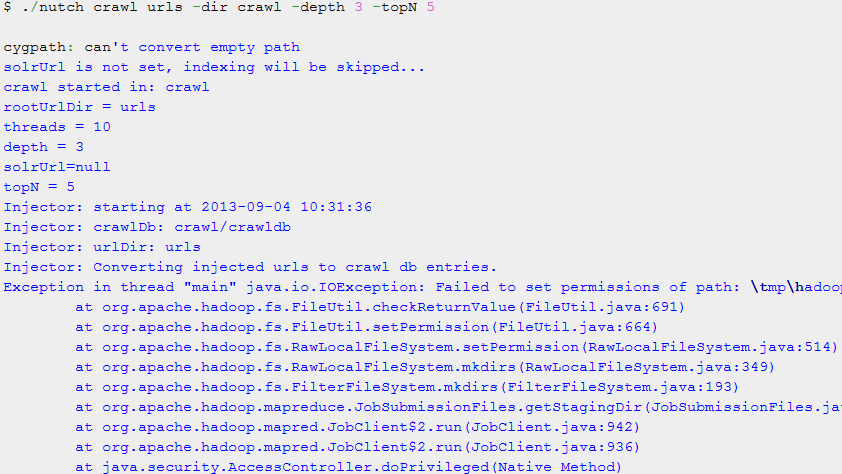
**The following parts provide tips to solve some error messages:**

1. **Entering error message: “**Exception in thread "main" java.io.IOException: Failed to set permissions of path: **\t**mp**\h**adoop**”**

**After you type**

bin/nutch crawl urls -dir crawl -depth 3 -topN 5

**and encounter error messages showing as follows.**

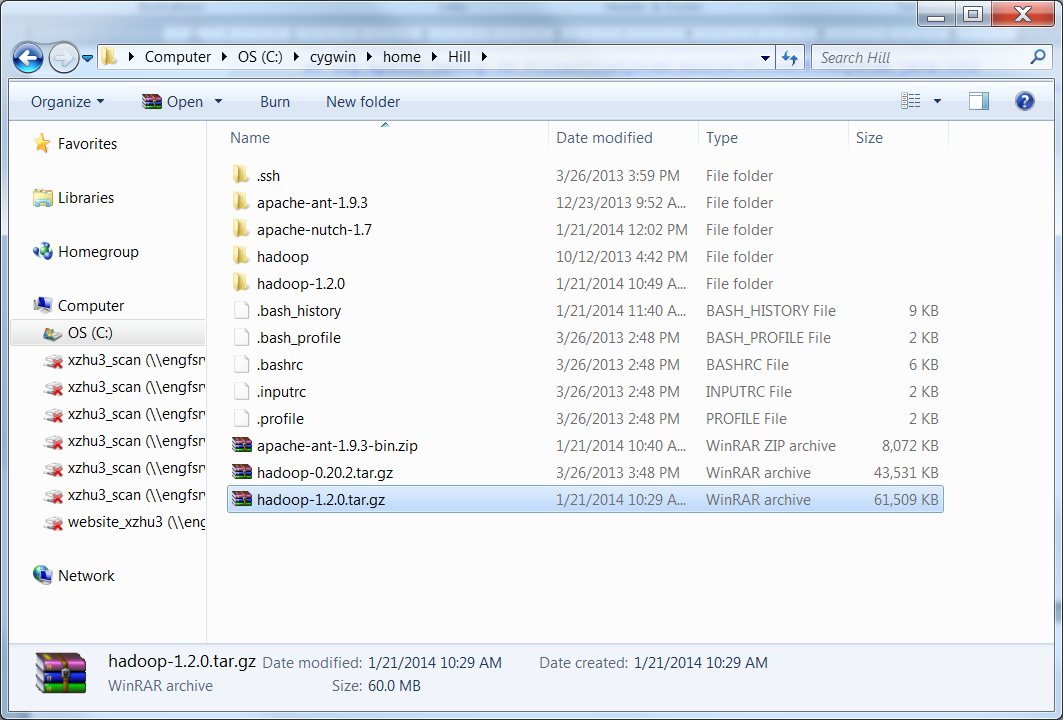
****

**This problem was caused by windows permission setting (the problem should not exist in a Linux system).**

**Solutions:**

1. **Please download hadoop-1.2.0.tar.gz from here (**[**http://archive.apache.org/dist/hadoop/common/hadoop-1.2.0/**](http://archive.apache.org/dist/hadoop/common/hadoop-1.2.0/)**)**

**Put the zip file under “**C:/Cygwin/home/your\_window\_ID/**” as following**

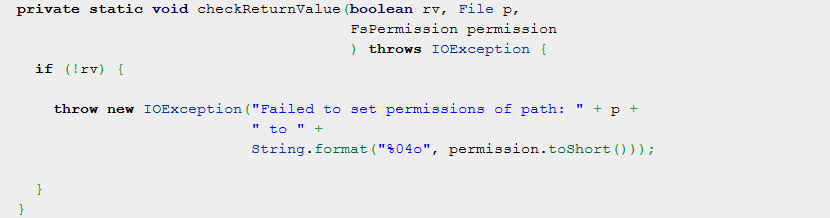


**(please ignore “hadoop-0.20.2.tar.gz”, which was used in another class. You do NOT need to download this file).**

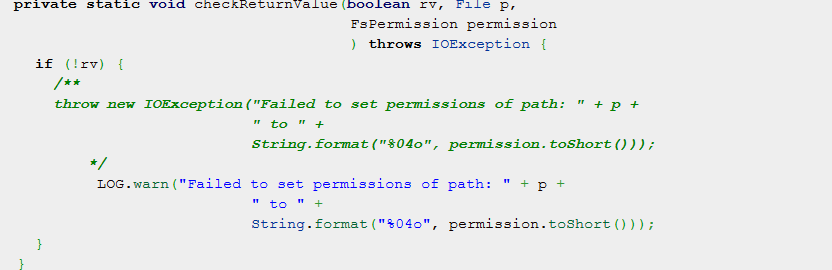
1. **Unzip hadoop-1.2.0.tar.gz. Then you will find a new folder “hadoop-1.2.0”, showing in the above window.**
2. **Now in the “hadoop-1.2.0” folder, find the following file**

hadoop-1.2.0\src\core\org\apache\hadoop\fs\FileUtil.java

**Please search the FileUtil.java file, and find “**Failed to set permissions of path**”, and modify your file from the original form:**

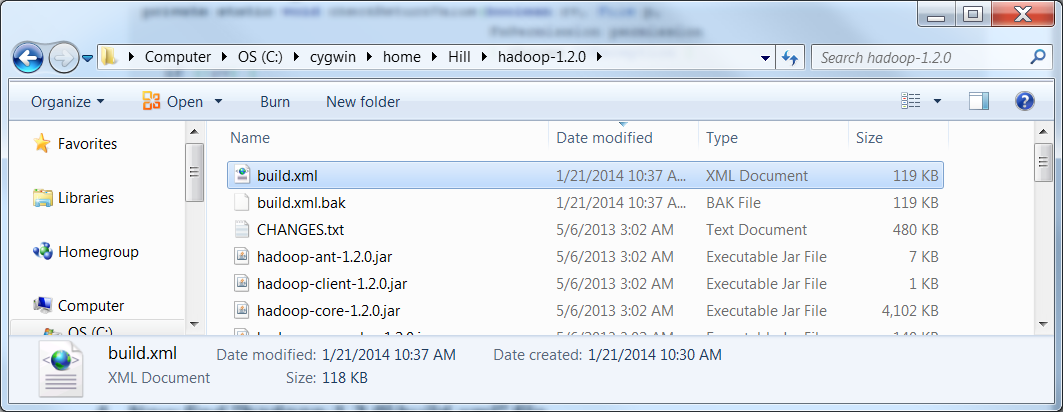
****

**To the following form:**

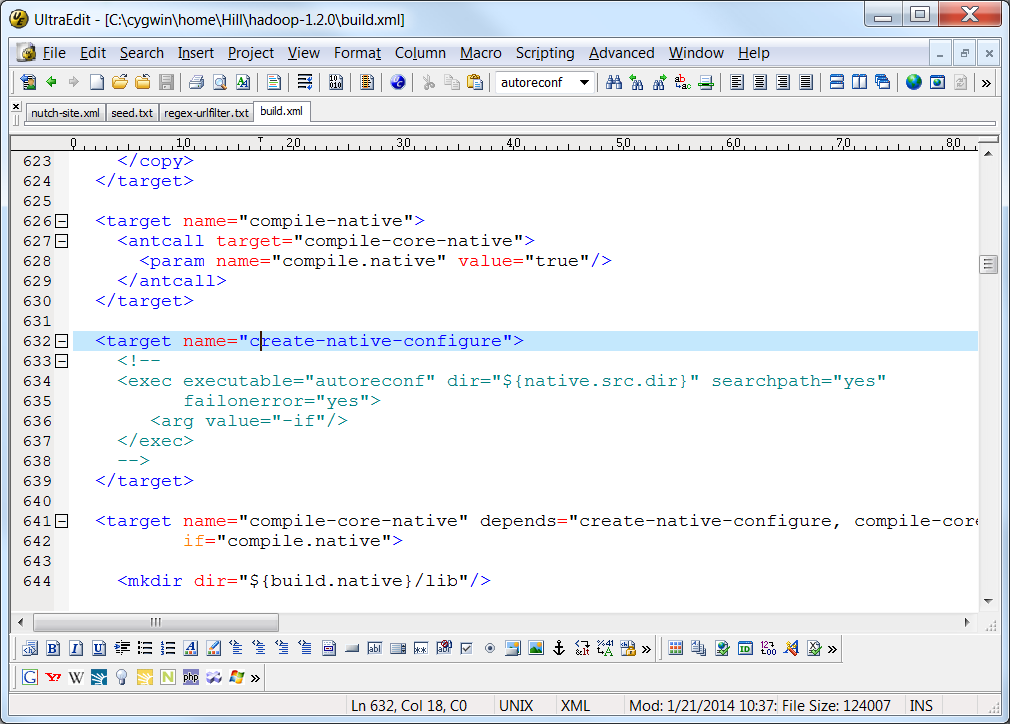
****

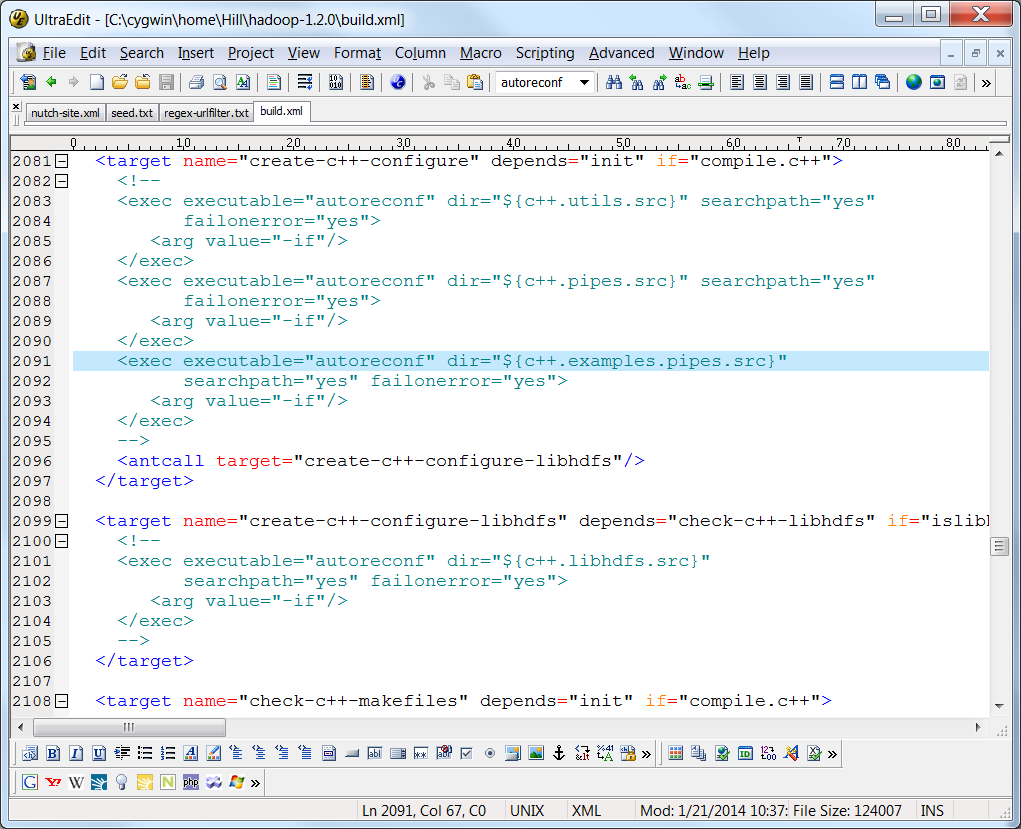
Save and exist FileUtil.java

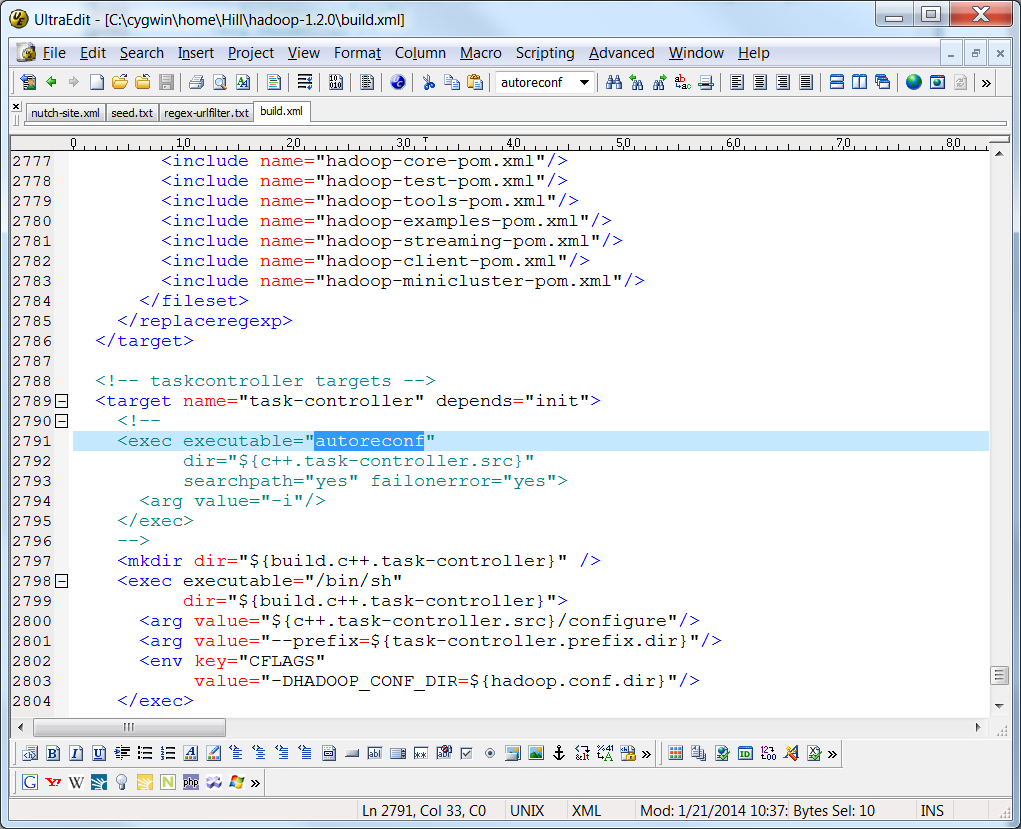
1. **Now find “hadoop-1.2.0\build.xml” file**



1. **Use a text editor (such as notepad) to open build.xml file.**
2. **There are SIX places with executable=”autoreconf”. Please find them and comment all those SIX places as follows:**



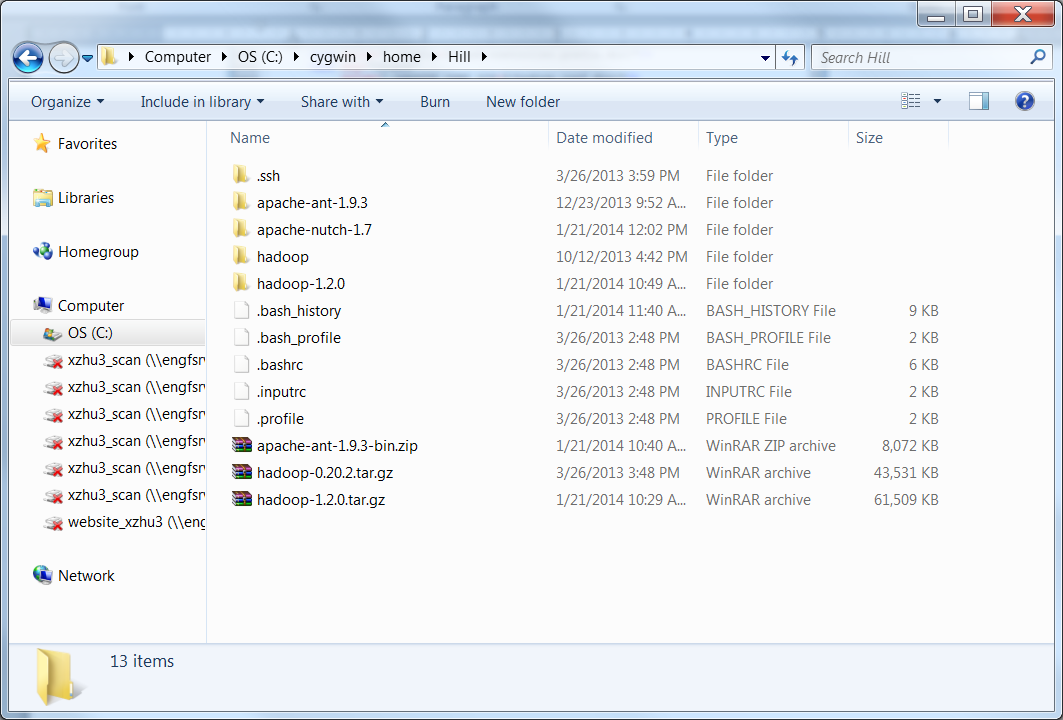




1. **Now we need to recompile hadoop-1.2.0 by using Apache ant. Then you will need following step**
   1. **Download apache ant (apache-ant-1.9.3-bin.zip) from there**

[**http://archive.apache.org/dist/ant/binaries/**](http://archive.apache.org/dist/ant/binaries/)

* 1. **Save apache-ant-1.9.3-bin.zip to** C:/Cygwin/home/your\_window\_ID/**” as following**



* 1. **Unzip “apache-ant-1.9.3-bin.zip”, you will find a new folder “apache-ant-1.9.3”.**
  2. **In “cygwin”, enter folder “apache-nutch-1.7”, and type following commands:**

export ANT\_HOME=/cygdrive/c/cygwin/home/Hill/apache-ant-1.9.3

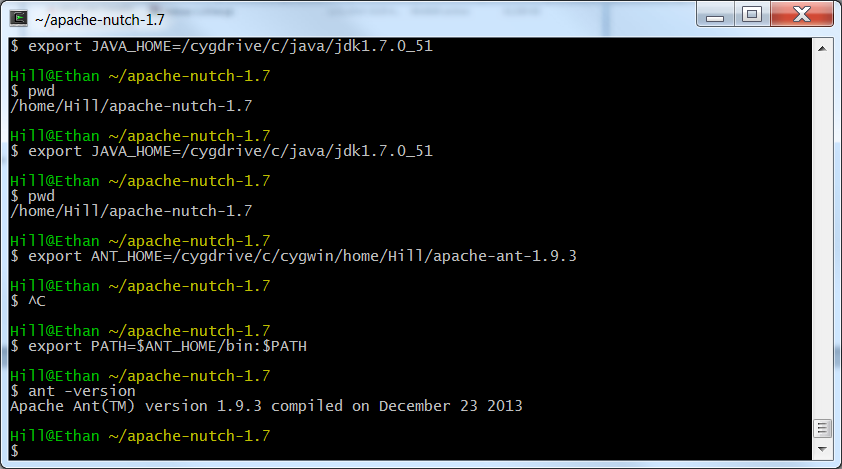
Then type:

export PATH=$ANT\_HOME/bin:$PATH

After that, you can type

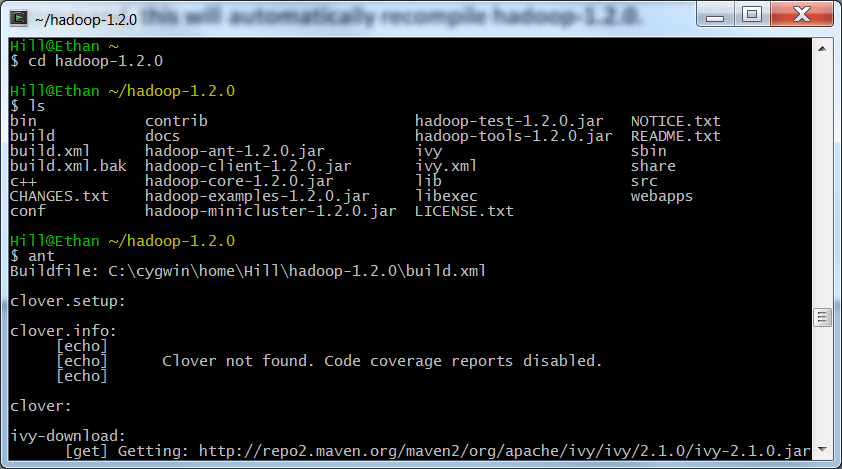
ant -version

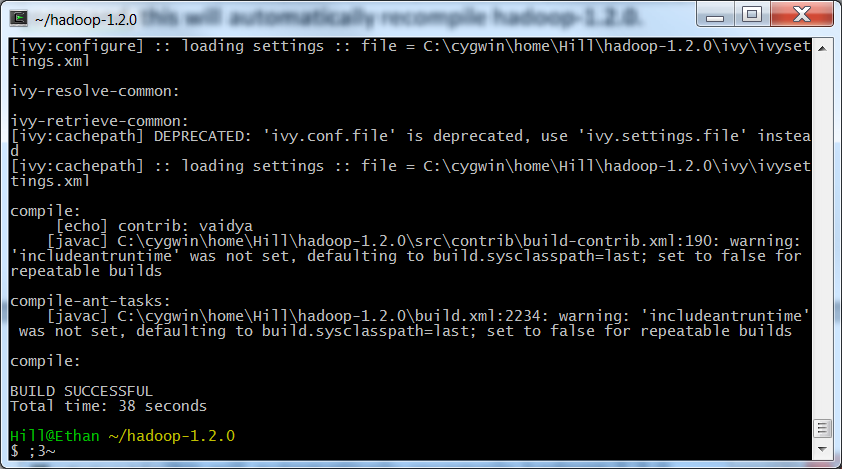
One you get the results showing as follows, it means that ant was properly configured.



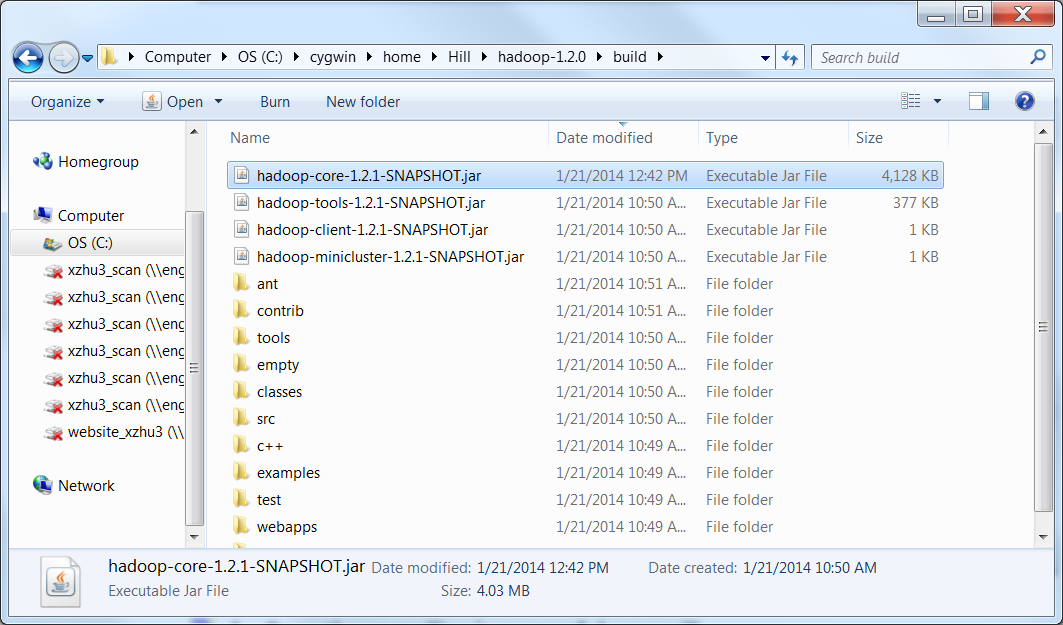
* 1. **Now in cygwin, enter “hadoop-1.2.0”, and then type “ant” command, this will automatically recompile hadoop-1.2.0.**

**Then, you will see some results as following:**

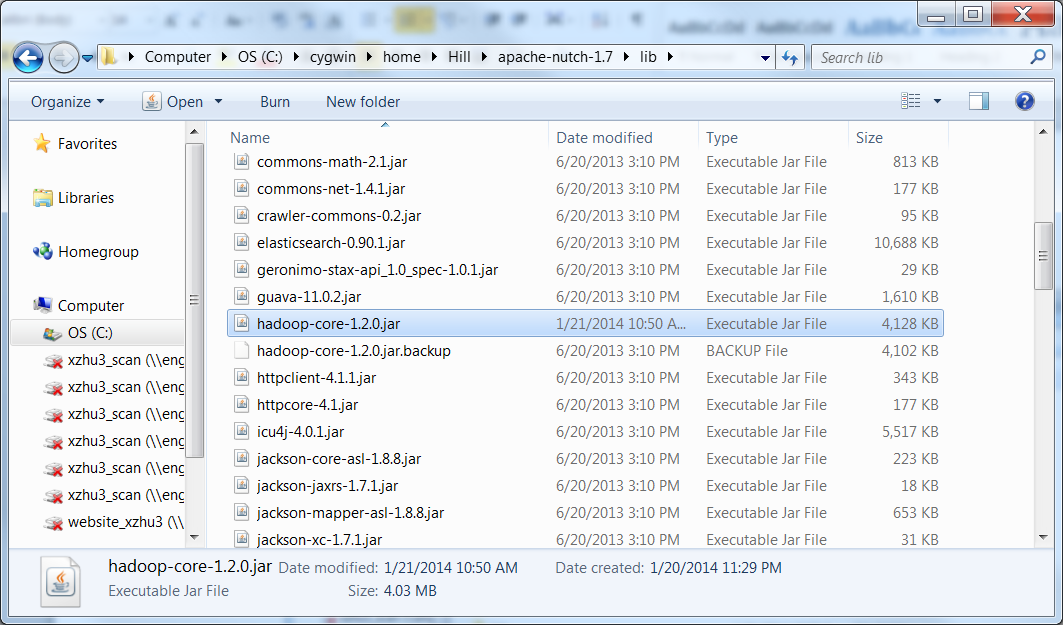




* 1. **Once the compiling is successful, you will have a new “hadoop-core-1.2.1.SNAPSHOT.jar”, in the “hadoop-1.2.0/build” folder, as following:**



* 1. **Copy “hadoop-core-1.2.1.SNAPSHOT.jar” and past the file to “apache-nutch-1.7/lib”, and rename the file as “hadoop-core-1.2.0.jar” (you will need to delete the original hadoop-core-1.2.0.jar or rename is, as showing in the following window).**



**H: then you can return to “apache-Nutch-1.7” (in cygwin), and type the following command. The error messages should disappear.**

bin/nutch crawl urls -dir crawl -depth 3 -topN 5