ErrorMessages

Error messages, reasons and solutions

Please feel free to add error messages, reasons and solutions!

Please report bugs to the mailing list!

General

Java.io.IOException: No input directories specified in: NutchConf: nutch-default.xml , mapred-default.xml

The crawl tool expects as its first parameter the folder name where the seeding urls file is located so for example if your urls.txt is located in /nutch/seeds the crawl command would look like: crawl seed -dir /user/nutchuser...

Exception: java.net.SocketException: Invalid argument or cannot assign requested address on Fedora Core 3 or 4

It seems you have installed IPV6 on your machine.

To solve this problem, add the following java param to the java instantiation in bin/nutch:

JAVA_IPV4=-Djava.net.preferIPv4Stack=true

run it exec "$JAVA" $JAVA_HEAP_MAX $NUTCH_OPTS $JAVA_IPV4 -classpath "$CLASSPATH" $CLASS "$@"

FileNotFoundException: 1

delay 1 fails crawltest and subdirectories are created; also ant compiles no probs; ROOT.war is installed and runs; urls file exists. Adding ./ or full path as x below changes nothing. Server runs squid on 80 and real Apache 1.3 on 81. Catalina is on 8080 and is up and running.

/x/nutch/nutch-0.7 # bin/nutch crawl /x/nutch/nutch-0.7/urls -dir /x/nutch/nutch-0.7/crawl.test -threads 2 -delay 1 -depth 3
run java in /usr/local/java/j2sdk1.4.2
050827 032536 parsing file:/x/nutch/nutch-0.7/conf/nutch-default.xml
050827 032536 parsing file:/x/nutch/nutch-0.7/conf/crawl-tool.xml
050827 032536 parsing file:/x/nutch/nutch-0.7/conf/nutch-site.xml
050827 032537 No FS indicated, using default:local
050827 032537 crawl started in: /x/nutch/nutch-0.7/crawl.test
050827 032537 rootUrlFile = 1
050827 032537 threads = 2
050827 032537 depth = 3
050827 032537 Created webdb at LocalFS,/x/nutch/nutch-0.7/crawl.test/db
Exception in thread "main" java.io.FileNotFoundException: 1 (No such file or directory)
at java.io.FileInputStream.open(Native Method)
at java.io.FileInputStream.<init>(FileInputStream.java:106)
at java.io.FileReader.<init>(FileReader.java:55)
at org.apache.nutch.db.WebDBInjector.injectURLFile(WebDBInjector.java:372)
at org.apache.nutch.db.WebDBInjector.main(WebDBInjector.java:535)
at org.apache.nutch.tools.CrawlTool.main(CrawlTool.java:134)

crawl test exists

ls -R crawl.test/
crawl.test/:

- .. db

crawl.test/db:

- .. dbreadlock dbwritelock webdb

crawl.test/db/webdb:

- .. linksByMD5 linksByURL pagesByMD5 pagesByURL

crawl.test/db/webdb/linksByMD5:

- .. data index

crawl.test/db/webdb/linksByURL:

- .. data index

crawl.test/db/webdb/pagesByMD5:

- .. data index

crawl.test/db/webdb/pagesByURL:

- .. data index

---

export NUTCH_JAVA_HOME is set and working..

It always fails with above error, while omitting the delay tag seems to work 🙂 ... I tried putting the -delay tag at several places above, it always fails

nutch 0.7 Apache Tomcat/5.0.19 jdsk 1.4.2-b28 Sun Microsystems Inc. Linux (Suse 8.2 1.5 years old but updated) Linux Kernel 2.4.21 i386

Well its working without the delay tag but I can't release it on other sites with no delay tag. What am I doing wrong?

Fetching Errors

Why do I get error "123456 104934 fetch of http://mydomain/index.html failed with: net.nutch.net.protocols.http.HttpError: HTTP Error: 401" when crawling?

- An HTTP 401 error is returned from a remote webserver when you not authorized to view the page. Currently nutch does not support HTTP authentication but it will be trivial to add when the new HTTPClient fetcher code is committed.

/etc/host.conf: line 1: cannot specify more then 4 services

- Please have a look at http://sources.redhat.com/ml/bug-glibc/2002-07/msg00269.html

While fetching I get UnknownHostException for known hosts

Make sure your DNS server is working and/or it can handle the load of requests.

Updating Errors

Until updating my DB I got a OutOfMemoryException or a 'to many files open' error.

- The problems is that nutch opens more files then your OS allows to open. You can check the limits of your machine with "ulimit -a". In case you run nutch as superuser you can set the limit of open files for the actual session with "ulimit -n 65536". To change this limit permanently please read: http://bbcr.uwaterloo.ca/~brecht/servers/openfiles.html

Indexing Errors

While indexing documents, I get the following error:

050529 011245 fetch okay, but can't parse myfile, reason: Content truncated at 65536 bytes. Parser can't handle incomplete msword file.

What is happening?

- By default, the size of the documents downloaded by Nutch is limited (to 65536 bytes). To allow Nutch to download larger files (via HTTP), modify nutch-site.xml and add an entry like this:

```
<property>
  <name>http.content.limit</name>
  <value>150000</value>
</property>
```

- If you do not want to limit the size of downloaded documents, set http.content.limit to a negative value:

```
<property>
  <name>http.content.limit</name>
  <value>-1</value>
</property>
```

Searching Errors

Tomcat reports root cause: java.lang.OutOfMemoryError and does not find anything.

- Try to give java / tomcat some more memory. Add to catalina.sh (linux): JAVA_OPTS=-Xmx256m

Installation Errors

See GettingNutchRunningWithUbuntu for some help.

Nutch on Debian (cont)

What is mentioned here

http://nutch.sourceforge.net/cgi-bin/twiki/view/Main/GettingNutchRunningOnDebian

java.lang.NoClassDefFoundError: org/apache/coyote/http11/Http11Processor$1
at org.apache.coyote.http11.Http11Processor.prepareResponse(Http11Processor.java:1513)

can be avoided with permission java.io.FilePermission "*", "read,write,execute,delete";

pityfully the cache anchor option doesn't work still

java.security.AccessControlException: access denied (java.util.PropertyPermission * read,write)
at java.security.AccessControlContext.checkPermission(AccessControlContext.java:264)

this happens independent of putting

permission java.io.FilePermission "*", "read,write,execute,delete";

in

/etc/tomcat4/policy.d/04webapps.policy

😞

so if you are then entirely fed up trying to find what's up ... because bad stack trace + idiotic and unpenetrable security settings are selfdefeating..

you enter permission java.security.AllPermission;

in /etc/tomcat4/policy.d/04webapps.policy

and the thing works ... (but I am not even contemplating what security holes I have opened here :|)

Setup on a SUSE 8.1 system was no problem btw ...