NutchGotchas

The following acts as a comprehensive list of Nutch "Gotchas" which should act as a suitable prerequisite source of implicit information currently existing in the Nutch Codebase and in its general usage.

Developing Nutch: Gotchas

Developing Nutch Gotchas should be driven purely by community opinion and consensus that it is necessary to make implicit information explicit in an attempt to create an easier working environment for Nutch users at all levels. The list below has been compiled as a repository of information which emerged during discussions on various lists. As with many areas of the Nutch wiki, this list exists as a non static resource and all Nutch users are invited to edit based upon experience and community consensus.

Current Gotchas and using them:

No agents listed in 'http.agent.name' property

Since 1.3 Nutch is called from either of the runtime dirs (runtime/local and runtime/deploy). The conf files should be modified in runtime/local/conf, not in $NUTCH_HOME/conf.

Nutch-1016: Strip UTF-8 non-character codepoints

This JIRA issue affects the indexer and relates to the stripping of UTF-8 non-character codepoints which exist within some documents and was initially discovered during large crawls. When indexing to Solr this will yield the following exception:

```
EVERE: java.lang.RuntimeException: [was class java.io.CharConversionException] Invalid UTF-8 character 0xffff
at char #1142033, byte #1155068)
        at com.ctc.wstx.util.ExceptionUtil.throwRuntimeException(ExceptionUtil.java:18)
        at com.ctc.wstx.sr.StreamScanner.throwLazyError(StreamScanner.java:731)
```

The fix (committed by Markus) for the SolrWriter class passes the value of the content field to a method to strip away non-characters, effectively avoiding the runtime exception. Various patches are available here

Removal of crawl-urlfilter.txt

As of the release of Nutch 1.3, crawl-urlfilter.txt has been removed purposefully as it did not add anything to the other url filters (automaton | regex) in terms of functionality. By default the urlfilters contain (+.) which was what the crawl-urlfilter used to do.

Confusion about "solrUrl is not set, indexing will be skipped..." log message

This relates to the removal of the Nutch Lucene legacy dependence to support indexing with Solr, and the road map to enable various other indexing implementations. We have two options for passing the indexing command to Nutch.

- During the crawl command, as explained here.
- or during the later stage of sending an individual solrindex command to Solr as explained here.

DiskErrorException while fetching

Questions like this one arise fairly regularly on the user@ list

```
Hello,

I am getting some exception while fetching:

2011-07-10 23:25:21,427 WARN  mapred.LocalJobRunner - job_local_0001
org.apache.hadoop.util.DiskChecker$DiskErrorException: Could not find
taskTracker/jobcache/job_local_0001/attempt_local_0001_m_000000_0/output/spill0.out
in any of the configured local directories
       at org.apache.hadoop.fs.LocalDirAllocator$AllocatorPerContext.getLocalPathToRead(LocalDirAllocator.java:
389)
       at org.apache.hadoop.fs.LocalDirAllocator.getLocalPathToRead(LocalDirAllocator.java:138)
       at org.apache.hadoop.mapred.MapOutputFile.getSpillFile(MapOutputFile.java:94)
       at org.apache.hadoop.mapred.MapTask$MapOutputBuffer.mergeParts(MapTask.java:1443)
       at org.apache.hadoop.mapred.MapTask$MapOutputBuffer.flush(MapTask.java:1154)
       at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:359)
       at org.apache.hadoop.mapred.MapTask.run(MapTask.java:307)
       at org.apache.hadoop.mapred.LocalJobRunner$Job.run(LocalJobRunner.java:177)
2011-07-10 23:25:22,279 FATAL fetcher.Fetcher - Fetcher:
java.io.IOException: Job failed!
       at org.apache.hadoop.mapred.JobClient.runJob(JobClient.java:1252)
       at org.apache.nutch.fetcher.Fetcher.fetch(Fetcher.java:1107)
       at org.apache.nutch.fetcher.Fetcher.run(Fetcher.java:1145)
       at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:65)
       at org.apache.nutch.fetcher.Fetcher.main(Fetcher.java:1116)

What should I do? What happens if I restart the fetch job?
```

The answer we find addressed the situation is that you're most likely out of disk space in /tmp. Consider using another location, or possibly another partition for hadoop.tmp.dir (which can be set in nutch-site.xml as below) with plenty of room for large transient files or using a Hadoop cluster.

```
<property>
<name>hadoop.tmp.dir</name>
<value>/path/to/large/hadoop/tmp</value>
</property>
```

Difference between building Nutch with Maven/Ant&Ivy

If you use maven to build the project it will build the >= 1.3 version ex plugins. Ant/ivy builds 1.4 and the plugins.