Requirements for installing Nutch

1. Java 1.4.x, either from Sun or IBM on Linux is preferred. Set NUTCH_JAVA_HOME to the root of your JVM installation.
2. Apache's Tomcat 4.x.
3. On Win32, cygwin, for shell support.
4. Edit the file conf/crawl-urlfilter.txt and replace MY.DOMAIN.NAME with the name of the domain you wish to crawl. For example, if you wished to limit the crawl to the apache.org domain, the line should read: (This will include any url in the domain apache.org.)

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

(Reference: “Nutch Tutorial”, http://lucene.apache.org/nutch/tutorial.html)

Step 1: Perform crawl

```
./nutch crawl ../urls -dir ../crawled/ -depth 1
```

where “urls” file contains one url (http://lucene.apache.org/nutch/) for demo purpose and “crawled” directory is the directory where crawled content will be stored.

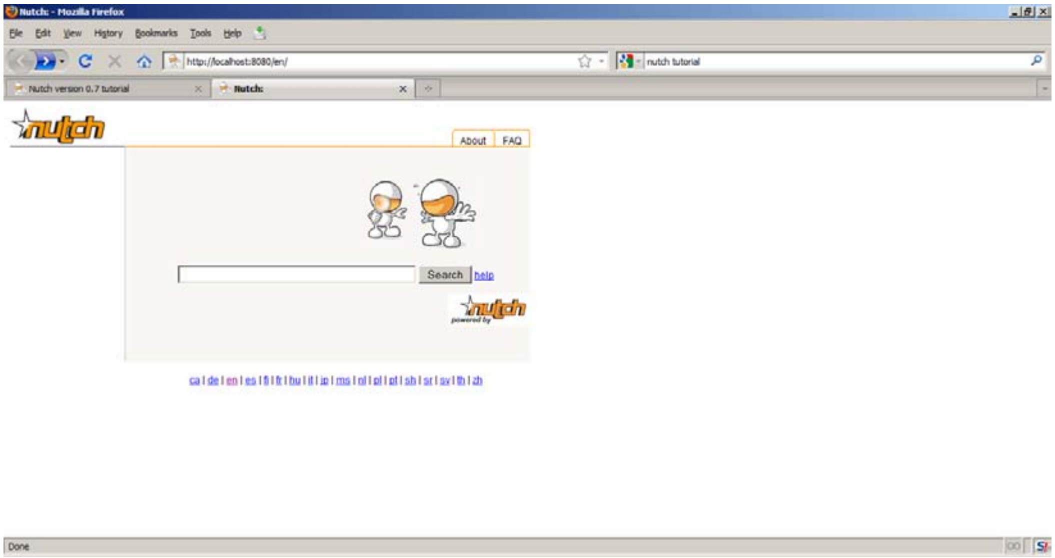
Step 2: Start tomcat server

```
catalina.sh start
```

Step 3: Now open the following URL in a browser to access Nutch search interface

```
http://localhost:8080/
```

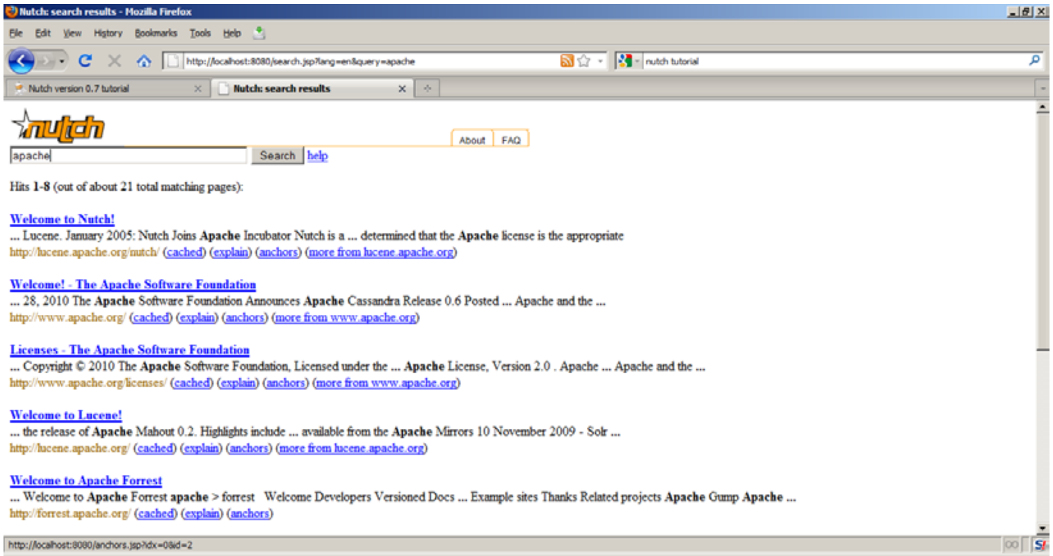
The snapshot of this interface is as shown below:



http://localhost:8080/en/

Nutch: - Mozilla Firefox
File Edit View History Bookmarks Tools Help
nutch tutorial
Nutch version 0.7 tutorial
Nutch:
nutch
About
FAQ
Search
help
nutch
powered by
ca | de | en | es | fi | fr | hu | it | jp | ms | nl | pl | pt | sh | sr | sv | th | zh
Done

The below snapshot shows the query results for the keyword “apache”:


Nutch: search results - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
http://localhost:8080/search.jsp?lang=en&query=apache
nutch tutorial
Nutch version 0.7 tutorial
Nutch: search results
nutch
About
FAQ
apache
Search
help

Hits **1-8** (out of about **21** total matching pages):

Welcome to Nutch!
... Lucene. January 2005: Nutch Joins **Apache** Incubator Nutch is a ... determined that the **Apache** license is the appropriate
http://lucene.apache.org/nutch/ (cached) (explain) (anchors) (more from lucene.apache.org)

Welcome! - The Apache Software Foundation
... 28, 2010 The **Apache** Software Foundation Announces **Apache** Cassandra Release 0.6 Posted ... Apache and the ...
http://www.apache.org/ (cached) (explain) (anchors) (more from www.apache.org)

Licenses - The Apache Software Foundation
... Copyright © 2010 The **Apache** Software Foundation, Licensed under the ... **Apache** License, Version 2.0 . Apache ... Apache and the ...
http://www.apache.org/licenses/ (cached) (explain) (anchors) (more from www.apache.org)

Welcome to Lucene!
... the release of **Apache** Mahout 0.2. Highlights include ... available from the **Apache** Mirrors 10 November 2009 - Solr ...
http://lucene.apache.org/ (cached) (explain) (anchors) (more from lucene.apache.org)

Welcome to Apache Forrest
... Welcome to **Apache** Forrest **apache** > forrest Welcome Developers Versioned Docs ... Example sites Thanks Related projects **Apache** Gump **Apache** ...
http://forrest.apache.org/ (cached) (explain) (anchors)

http://localhost:8080/anchors.jsp?idx=0&id=2