GettingNutchRunningWithUtf8

How to Configure App Servers to Pass non-ASCII Characters?

Nutch GUI uses the GET method to pass the query strings to the server. Tomcat 4 and 5 need to be configured to enable passing of non-ASCII characters.

Note that this note describes how to make Tomcat pass non-ASCII characters. Nutch, in its "factory set" configuration, handle only limited characters. Especially, it will not handle Chinese/Japanese/Korean text properly. (Each CJK character is treated as if it were a word by itself.) German special chars are also wrongly displayed (ö, ä, ü).

Tomcat 4 and Tomcat 5

Tomcat changed its "factory set" configuration to allow only the ISO 8859-1 encoding to be used in the GET method. See http://issues.apache.org/bugzilla/show_bug.cgi?id=29900 for the rationale for the change.

To enable passing of UTF-8 characters, edit $TOMCAT/conf/server.xml. Locate the <Connector> tag for the web (look for "8080") and insert this parameter assignment:
`URIEncoding="UTF-8"`
as explained in Tomcat 5 FAQ at http://tomcat.apache.org/faq/connectors.html#utf8

Contributors, please add special configurations needed for other app servers below.