Org.apache.nutch.net.BasicUrlNormalizer

BasicUrlNormalizer Notes

The Basic URL Normalizer class manipulates an URL in several ways.

1. Trims white space from the end of the URL. (java.lang.String.trim())
2. may lower case protocol. (java.net.URL)
3. if protocol is http or ftp:
    a. lower cases host.
    b. removes port if default.
    c. adds trailing slash if no file specified.
    d. removes any refrence text
    e. removes any relative paths

For example:

`http://wiKI.apache.ORG:80/somedirectory/../DevelopmentCommandLineOptions`

would be rewriten:

`http://wiki.apache.org/DevelopmentCommandLineOptions`

Notes

Other then trimming trailing white space and the normalization performed by java.net.URL no protocols other then http and ftp are further normalized.