DebugTool

Based on some conversations on list:

We've gathered some requirements for a Debug Tool, that could be useful in allowing users to know precisely what decisions that Nutch is making while it navigates the URL space. So far, here's what we have from Ken Krugler, primarily, and those others (Markus Jelsma, Chris Mattmann, Lewis John McGibbney) participating in the above referenced thread:

It should be possible to generate information that would have answered all of the "is it X" questions that came up during a user's crawl. E.g.

1. which URLs were put on the fetch list, versus skipped.
2. which fetched documents were truncated.

- The code currently has primitive logging for all parse plugins to log verification of truncation to stdout. What more could we do here? It is a common problem so would be good to improve this area.

1. which URLs in a parsed page were skipped, due to the max outlinks per page limit.
2. which URLs got filtered by regex, prefix, suffix, domain filters
3. exclusions by robots directives
    - robots.txt
    - outlinks skipped by meta nofollow
4. URLs mapped to another URL
    - URL normalization
    - redirects

Please add more requirements and discussion here.