Features

This page act's as an up-to-date resource for features included in the most current stable release of Nutch (at time of writing this is 1.3).

Features

- Fetching and parsing are done separately by default, this reduces the risk of an error corrupting the fetch parse stage of a crawl with Nutch.
- Plugins have been overhauled as a direct result of removal of legacy Lucene dependency for indexing and search.
- The number of plugins for processing various document types being shipped with Nutch has been refined. Plain text, XML, OpenDocument (OpenOffice.org), Microsoft Office (Word, Excel, Powerpoint), PDF, RTF, MP3 (ID3 tags) are all now parsed by the **Tika** plugin. The only parser plugins shipped with Nutch now are Feed (RSS/Atom), HTML, Ext, JavaScript, SWF, Tika & ZIP.
- MapReduce ;
- Distributed filesystem (via Hadoop)
- Link-graph database
- NTLM authentication

Questions and Answers

*How does the search engine handle punctuation and special characters? (and what's configurable?)

- They are treated like a space.
  *Which document formats are supported?

- This is directly linked to the available parser plugins mentioned above, however only some are enabled by default as most of the parsing is now delegated to Tika in an attempt to clean up the Nutch codebase. Edit conf/nutch-site.xml and change the value of plugin.includes property to include the plugins for the document types that you want Nutch to handle. Additionally have a look at conf/parse-plugins.xml for more details of plugin implementations. To recap:
  - Plain Text (plugin: tika)
  - HTML/XHTML+XML (parse-html/tika)
  - XML (parse-Tika/feed) uses XPath and namespaces to do the mapping between XML elements and index fields.
  - JavaScript (for extracting links only?) (parse-js)
  - OpenOfice.org ODF (parse-tika) parses Open Office and Star Office documents.
  - Microsoft Power Point, the .ppt file (parse-tika)
  - Microsoft Word, the .doc file (parse-tika)
  - Adobe PDF (parse-tika)
  - RSS (parse-feed/tika)
  - RTF (parse-tika)
  - MP3 (parse-tika) The mp3 itself contains the ID3v1 or ID3v2 tags which contain metadata song information like title, artist, album, comments, etc. The useful information needed to search mp3s
  - ZIP (parse-zip) This seems to expand the zip of plain text files and return the concatenated text.

Questions without Answers

*What post-coordination options are available?

*How easy is Nutch to configure?

*How transparent is its configuration to a working organization: does it require geeky command line stuff, or can a knowledgable manager enter a web or software interface to view or modify settings?

- Does Nutch support deduping?
- Can one tinker with relevance algoritms?
- Are there ranking overrides?