OldFeatures

Missing from the current Nutch documentation (Tutorial, FAQ) is a list of features. This wiki page could help, if someone who knows the answers can edit it.

(Please reformat this text and divide into feature lists, questions and questions & answers).

Features

- Fetching, parsing and indexation in parallel and/ou distributed
- Plugins
- Many formats: plain text, HTML, XML, ZIP, OpenDocument (OpenOffice.org), Microsoft Office (Word, Excel, Powerpoint), PDF, JavaScript, RSS, RTF, MP3 (ID3 tags)
- Ontology
- Clustering
- MapReduce ;
- Distributed filesystem (via Hadoop)
- Link-graph database
- NTLM authentication

Questions and Answers

*What kind of searches does Nutch support? (quoted, nested, truncation, wildcarding [and where], Boolean),

- "...." (phrase search?), + (what is this for?), - (negation) and fieldname:term. No "AND" or "OR". The and-logic is implied.
  *Is stemming an option?
- According to the Lucene in Action book: "Nutch does not use stemming or term aliasing of any kind. Search engines have not historically done much stemming, but it is a question that comes up regularly." – page 329
  *What kind of stemming does Nutch use? (and can you add exceptions/changes?)
- See previous answer 🙂
  *Does Nutch support Boolean operators? (can you use Google-like plus or minus or are you stuck with 1990s terms?)
- No
  *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?

- Guessing from the names of the available parser plugins, this is probably it. However, only the plain text and HTML are enabled by default. 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:
  - Plain Text (plugin: parse-text)
  - HTML (parse-html)
  - XML (parse-xml) uses XPath and namespaces to do the mapping between XML elements and Lucene fields.
  - JavaScript (for extracting links only?) (parse-js)
  - OpenOfice.org ODF (parse-oo) parses Open Office and Star Office documents.
  - Microsoft Power Point, the .ppt file (parse-mspowerpoint)
  - Microsoft Word, the .doc file (parse-msword)
  - Adobe PDF (parse-pdf)
  - RSS (parse-rss)
  - RTF (parse-rtf)
  - MP3 ❓ Is there any text in MP3? (parse-mp3) (JR: Sure, the mp3 itself contains the ID3v1 or ID3v2 tags which contain song information like
    title, artist, album, comments, etc. The useful information needed to search mp3s)
  - ZIP ❓ This seems to expand the zip of plain text files and return the concatenated text. (parse-zip)

Questions without Answers

*Does Nutch support weighted field searching, synonym support?

*What kinds of indexes does Nutch build? (multi-format indexing, incremental indexing, spell-check support, thesauri support, fielded searching, rank-by-reputation?)

*What post-coordination options are available? (hey Karen, what does this mean?)

*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?

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