JeromeCharron

Jerome Charron

<<MailTo(jerome.charron AT PASDEPOURRIELS gmail DOT com)>>

Activities

- http://motrech.free.fr/
  - A french mailing list on search engines (http://groups.yahoo.com/group/motrech)
  - A french blog on search engines (http://motrech.blogspot.com/)
- http://www.frutch.org/
  - A french mailing list on Nutch (http://groups.yahoo.com/group/frutch)
  - A french wiki on Nutch (http://www.frutch.org/wikini)

Nutch contributions

- MimeTypeUtil package (org.apache.nutch.util.mime)
  - **TODO**: Provide an content-type mapper (see ParserFactoryImprovementProposal requirements)
  - **TODO**: Replace the current XML descriptor by the Freedesktop shared-mime-info-spec one
- LanguageIdentifierPlugin
  - Some benchs LanguageIdentifierBenchs
  - Enhance the LanguageParseFilter by checking the validity of the parsed language string.
  - **TODO**: Enhance the LanguageParseFilter by correlating (instead of taking only the first information available) all the clues available : DublinCore / Meta-Http-Equiv / Content-Language and statistical content analysis.
  - **TODO**: Improve API :
    - returns an ordered list of candidate languages instead of just one.
    - See also Andrzej comments :
      - exporting a list of supported languages,
      - exporting an NGramProfile of the analyzed text,
      - allow processing of chunks of input.
- MultiLingualSupport proposal.
  - Framework for a multi-lingual analysis:
    - Analysis ExtensionPoint
    - AnalyzerFactory
  - LibLuceneAnalyzersPlugin packaged and committed
  - AnalysisFrPlugin (Lucene French Analyzer Wrapper) packaged and committed
  - AnalysisDePlugin (Lucene German Analyzer Wrapper) packaged and committed
  - **TODO**: Multilingual querying support
- ParserFactoryImprovementProposal
  - **TODO**: Use content-type/extension-id mapping instead of content-type/plugin-id
- PluginRepository enhancements:
  - Add ability to handle plugins inter-dependencies (ie, a plugin can specify it has a runtime dependency on another(s) plugin(s) using the <requires><import plugin="plugin-id"/></requires> directive in the plugin.xml plugin descriptor.
  - Add ability to automatically load (depending on config) the required plugins specified by plugins dependencies (circular dependencies checked).
- MarkupLanguageParserProposal
- Microformats HtmlParseFilter:
  - rel-tag (see microformats-reltag plugin)
  - **TODO** hreview
  - ...
- Nutch article on french wikipedia.
- URL Filters enhancements:
  - Add a *mini framework* plugin for regular expression based URL Filters (lib-regex-filter)
  - Add a regex url filter implementation based on dk.brics.automaton Finite-State Automata for Java.
  - See RegexURLFiltersBenchs for a comparison of urlfilter-regex and urlfilter-automaton plugins

---

CategoryHomepage