NonDefaultIntranetCrawlingOptions

Options for intranet crawling that are not enabled by default

Here are some options you might want to add to your conf/nutch-site.xml configuration file if you plan on crawling your local network intranet. You will notice that some plugins are not enabled by default but accurately reflect the type of data present on the typical enterprise intranet.

Enable additional parser plugins

```
<property>
<name>plugin.includes</name>
<value>protocol-http|urlfilter-regex|parse-(html|tika|zip|js|swf|feed)|index-(basic|more)|scoring-
opic|urlnormalizer-(pass|regex|basic)</value>
</property>
```

This will enable the parser plugins for html, zip, javascript, swf and rss/atom feed. Text, pdf, excel, powerpoint, word and various other document formats are also parsed by the tika implementation. Additional parsers can be specified in conf/parse-plugins.xml. If you have additional document types you wish to parse and they are listed in the parse-plugins file, just add them to the list. For more information please see PluginCentral

Increase the file size fetch limit

```
<property>
<name>http.content.limit</name>
<value>2097152</value>
</property>
```

This will increase the default file size fetching limit to 2 megabytes. If your documents are larger (such as PDFs) then increase the number appropriately.