CommandLineOptions

Nutch Command Line Options of bin/nutch

The following is a **complete** list of Nutch command line options. That is to say that some or all of the options may not be available in the particular version of Nutch you are using. For version specific options please see the relevant check box, once you know that such a command exists for your particular Nutch distribution, you can navigate to the relevant wiki entry for a detailed descritpion of the tool.

The script bin/nutch is a helper which picks different java classes to "run".

The crawl script NUTCH-1087] [https://issues.apache.org/jira/browse/NUTCH-1087 replaces the bin/nutch crawl command used up to versions 1.7 and 2.2.1.

Note: Most commands print help when invoked w/o parameters.

See each entry for details of the command arguments and options.

command	function	version	
		1.x	2.x
bin/nutch readdb	Read / dump crawl db	X	X
bin/nutch mergedb	Merge crawldb-s, with optional filtering	X	
bin/nutch readlinkdb	Read / dump link db	X	
bin/nutch inject	Inject new urls into the database	X	X
bin/nutch generate	Generate new segments to fetch from crawldb	X	X
bin/nutch freegen	Generate new segments to fetch from text files	X	
bin/nutch fetch	Fetch a segment's pages	X	X
bin/nutch parse	Parse a segment's pages	X	X
bin/nutch readseg	Read / dump segment data	X	
bin/nutch mergesegs	Merges multiple segments, with optional filtering and slicing	X	
bin/nutch updatedb	Update crawldb (from segments if in 1.x) after fetching	X	X
bin/nutch updatehostdb	Update hostdb after fetching		X
bin/nutch invertlinks	Create a linkdb from parsed segments	X	
bin/nutch mergelinkdb	Merge's linkdb-s, with optional filtering	X	
bin/nutch elasticindex	Run the elastic search indexer on parsed batches		X
bin/nutch solrindex	Run the solr indexer on parsed segments and linkdb - DEPRECATED use the index command instead	X	X
bin/nutch solrdedup	Removes duplicate documents from solr - DEPRECATED use the dedup command instead	X	X
bin/nutch solrclean	Removes HTTP 301 and 404 documents from solr - DEPRECATED use the clean command instead	X	
bin/nutch index	Run the plugin-based indexer on parsed segments and linkdb	X	
bin/nutch dedup	Deduplicate entries in the crawldb and give them a special status	X	
bin/nutch clean	Remove HTTP 301 and 404 documents and duplicates from indexing backends configured via plugins	X	
bin/nutch parsechecker	Checks the parser for a given url	X	X
bin/nutch indexchecker	Checks the indexing filters for a given url	X	
bin/nutch normalizerchecker	Checks URL normalizers for given URLs	X	
bin/nutch domainstats	Calculates domain statistics from crawldb	X	
bin/nutch webgraph	Generates a web graph from existing segments	X	
bin/nutch linkrank	Runs a link analysis program on the generated web graph	X	
bin/nutch scoreupdater	Updates the crawldb with linkrank scores	X	
bin/nutch nodedumper	Dumps the web graph's node scores	X	

bin/nutch plugin	Loads a plugin and run one of its classes main()	X	X
bin/nutch nutchserver	run a (local) Nutch server on a user defined port	X	X
bin/nutch webapp	run a (local) Nutch WebApp GUI on port 8080		X
bin/nutch junit	Runs the given JUnit test	X	X
bin/nutch commoncrawldump	Dump out Nutch segments into Common Crawl data format	X	
bin/nutch CLASSNAME	run the class named CLASSNAME	X	X

Webgraph classes

- bin/nutch org.apache.nutch.scoring.webgraph.WebGraph
- bin/nutch org.apache.nutch.scoring.webgraph.Loops
- bin/nutch org.apache.nutch.scoring.webgraph.LinkRank
- bin/nutch org.apache.nutch.scoring.webgraph.ScoreUpdater
- bin/nutch org.apache.nutch.scoring.webgraph.NodeDumper
- bin/nutch org.apache.nutch.scoring.webgraph.NodeReader
- bin/nutch org.apache.nutch.scoring.webgraph.LoopReader
- bin/nutch org.apache.nutch.scoring.webgraph.LinkDumper

Useful Plugin Classes

- bin/nutch plugin urlnormalizer-regex org.apache.nutch.net.urlnormalizer.regex.RegexURLNormalizer

Other Classes

- bin/nutch org.apache.nutch.net.URLFilterChecker
- bin/nutch org.apache.nutch.net.URLNormalizerChecker
- bin/nutch org.apache.nutch.tools.CrawlDBScanner
- bin/nutch org.apache.nutch.protocol.RobotRulesParser

back to FrontPage