

bin/nutch index

Pluggable Indexing

The index command (running `org.apache.nutch.indexer.IndexingJob`) takes the content from one or multiple segments and passes it to all enabled [IndexWriter](#) plugins which send the documents to Solr, Elasticsearch, and various other index back-ends.

Nutch 1.x

Usage: `Indexer (<crawlDb> | -nocrawlDb) (<segment> ... | -dir <segments>) [general options]`

Index given segments using configured indexer plugins

The CrawlDb is optional but it is required to send deletion requests for duplicates and to read the proper document score/boost/weight passed to the indexers.

Required arguments:

<code><crawlDb></code>	path to CrawlDb, or
<code>-nocrawlDb</code>	flag to indicate that no CrawlDb shall be used
<code><segment> ...</code>	path(s) to segment, or
<code>-dir <segments></code>	path to segments/ directory, (all subdirectories are read as segments)

General options:

<code>-linkDb <linkDb></code>	use LinkDb to index anchor texts of incoming links
<code>-params k1=v1&k2=v2...</code>	parameters passed to indexer plugins (via property <code>indexer.additional.params</code>)
<code>-noCommit</code>	do not call the commit method of indexer plugins
<code>-deleteGone</code>	send deletion requests for 404s, redirects, duplicates
<code>-filter</code>	skip documents with URL rejected by configured URL filters
<code>-normalize</code>	normalize URLs before indexing
<code>-addBinaryContent</code>	index raw/binary content in field <code>`binaryContent`</code>
<code>-base64</code>	use Base64 encoding for binary content

Indexwriter plugins have to be enabled by the property `plugin.includes`. See [IndexWriter](#) how to configure these plugins.