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:

        <crawldb>        path to CrawlDb, or
        -nocrawldb       flag to indicate that no CrawlDb shall be used

        <segment> ...    path(s) to segment, or
        -dir <segments> path to segments/ directory,
                         (all subdirectories are read as segments)

General options:

        -linkdb <linkdb>         use LinkDb to index anchor texts of incoming links
        -params k1=v1&k2=v2...   parameters passed to indexer plugins
                                 (via property indexer.additional.params)

        -noCommit        do not call the commit method of indexer plugins
        -deleteGone      send deletion requests for 404s, redirects, duplicates
        -filter          skip documents with URL rejected by configured URL filters
        -normalize       normalize URLs before indexing
        -addBinaryContent        index raw/binary content in field `binaryContent`
        -base64          use Base64 encoding for binary content
```

Indexwriter plugins have to be enabled by the property plugin.includes. See IndexWriter how to configure these plugins.