bin/nutch inject

Inject is an alias for org.apache.nutch.crawl.Injector

This class takes a flat file of URLs and adds them to the of pages to be crawled. It is useful for bootstrapping the system. The URL files contain one URL per line, optionally followed by custom metadata separated by tabs with the metadata key separated from the corresponding value by '='.

Note that some metadata keys are reserved:

nutch.score: allows to set a custom score for a specific URL

nutch.fetchInterval: allows to set a custom fetch interval for a specific URL

userType: this can be any metadata field which you then assign a value. In the example here we use userType to refer to the nature of Nutch as an open source project.

e.g. http://www.xyz.org/ nutch.score=10 nutch.fetchInterval=2592000 userType=open_source

Nutch 1.x

```
bin/nutch inject [-D...] <crawldb> <url_dir> [-overwrite|-update] [-noFilter] [-noNormalize] [-filterNormalizeAll]

  <crawldb>     Path to a crawldb directory. If not present, a new one would be created.
  <url_dir>     Path to URL file or directory with URL file(s) containing URLs to be injected.
                A URL file should have one URL per line, optionally followed by custom metadata.
                Blank lines or lines starting with a '#' would be ignored. Custom metadata must
                be of form 'key=value' and separated by tabs.
                Below are reserved metadata keys:

                        nutch.score: A custom score for a url
                        nutch.fetchInterval: A custom fetch interval for a url
                        nutch.fetchInterval.fixed: A custom fetch interval for a url that is not changed by
AdaptiveFetchSchedule

                Example:
                 http://www.apache.org/
                 http://www.nutch.org/ \t nutch.score=10 \t nutch.fetchInterval=2592000 \t userType=open_source

 -overwrite     Overwite existing crawldb records by the injected records. Has precedence over 'update'
 -update        Update existing crawldb records with the injected records. Old metadata is preserved

 -nonormalize   Do not normalize URLs before injecting
 -nofilter      Do not apply URL filters to injected URLs
 -filterNormalizeAll
                Normalize and filter all URLs including the URLs of existing CrawlDb records

 -D...          set or overwrite configuration property (property=value)
 -Ddb.update.purge.404=true
                remove URLs with status gone (404) from CrawlDb
```

<crawldb>: The directory containing the crawldb

<url_dir>: The directory containing our seed list (referred to above as 'flat file'), usually a text document containing URLs, one URL per line.

The injector uses the following configurations (see https://issues.apache.org/jira/browse/NUTCH-1405)

* db.injector.overwrite = [true|false] : replace the entries in the crawldb with the corresponding ones from the seed data. Will set the status to UNFETCHED.

* db.injector.update = [true|false] : Keeps the existing entries in the crawldb but replaces the score and fetch interval with the values found for the corresponding entries in the seed data. Any metadata found for the seed entry are added. The status remains what it was in the original version of the crawldb, e.g. FETCHED.

Nutch 2.x

```
Usage: InjectorJob <url_dir> [-crawlId <id>]
```

CommandLineOptions