bin/nutch mergedb

Mergedb is an alias for org.apache.nutch.crawl.CrawlDbMerger

This tool merges several crawldb's into one, optionally filtering URLs through the current URLFilters, to skip prohibited pages. It is possible to use this tool just for filtering - in that case only one crawldb should be specified in arguments. If more than one crawldb contains information about the same URL, only the most recent version is retained, as determined by the value of org.apache.nutch.crawl.CrawlDatum#getFetchTime(). However, all metadata information from all versions is accumulated, with newer values taking precedence over older values.

Usage:

```
bin/nutch mergedb <output_crawldb> <crawldb1> [<crawldb2> <crawldb3> ...] [-normalize] [-filter]
```

<output_crawldb>: This allows us to specify a name for the new merged output crawldb.

<crawldb1>: Only one crawldb parameter is used if we only wish to filter URLs through the current URLFilters. This enables us to filter unwanted pages from the crawldb.

[<crawldb2> <crawldb3> ...]]: Two or more crawldb arguements can be passed if we wish to undertake a merging of crawldb's. More information regarding dulication of URLs and URL metadata etc can be found above.

[-normalize]: If we know/think that URLs require to be normalized prior to being merged we pass this parameter. This uses the URLNormalizer on urls in the crawldb(s), although this is usually not required.

[-filter]: Enables is to filter URLs through the current URLFilters. This can be used in conjuction with a single crawldb arguement.

CommandLineOptions