bin/nutch mergelinkdb

Mergelinkdb is an alias for org.apache.nutch.crawl.LinkDbMerger

This tool merges several LinkDb-s into one, optionally filtering URLs through the current URLFilters, to skip prohibited URLs and links.

It is possible to use this tool just for filtering - in that case only one LinkDb should be specified in the arguments passed as parameters. If more than one LinkDb contains information about the same URL, all inlinks are accumulated, but only at most <code>db.max.inlinks</code> inlinks will ever be added.

If activated, URLFilters will be applied to both the target URLs and to any incoming link URL. If a target URL is prohibited, all inlinks to that target will be removed, including the target URL. If some of incoming links are prohibited, only they will be removed, and they won't count when checking the above-mentioned maximum limit.

Usage:

```
bin/nutch mergelinkdb <output_linkdb> <linkdb1> [<linkdb2> <linkdb3> ...] [-normalize] [-filter]
```

<output_linkdb>: This should be the path the the output linkdb to create from merging various linkdbs.

<linkdb1>: This corresponds to a the path for a single linkdb directory, (single input LinkDb is ok) OR

[<linkdb2> <linkdb3> ...: A list of paths to linkdb directories to create a merged linkdb from.

[-normalize]: We pass this if wish to use URLNormalizer on both fromUrls and toUrls in linkdb(s) (usually not needed).

[-filter]: This parameter uses current URLFilters on both fromUrls and toUrls in linkdb(s).

CommandLineOptions