bin/nutch webgraph

WebGraph is an alias for org.apache.nutch.scoring.webgraph.WebGraph

This class Creates three databases, one for inlinks, one for outlinks, and a node database that holds the number of in and outlinks to a url and the current score for the url.

The score is set by an analysis program such as LinkRank. The WebGraph is an update-able database. Outlinks are stored by their fetch time or by the current system time if no fetch time is available. Only the most recent version of outlinks for a given url is stored. As more crawls are executed and the Web Graph updated, newer Outlinks will replace older Outlinks. This allows the WebGraph to adapt to changes in the link structure of the web.

The Inlink database is created from the Outlink database and is regenerated when the WebGraph is updated. The Node database is created from both the Inlink and Outlink databases. Because the Node database is overwritten when the WebGraph is updated and because the Node database holds current scores for urls it is recommended that a crawl-cyle (one or more full crawls) should be fully complete before the WebGraph is updated and some type of analysis, such as LinkRank, is run to update scores in the Node database in a stable fashion.

Usage:

```
bin/nutch webgraph (-segment <segment> | -segmentDir <segmentDir>) -webgraphdb <webgraphdb> [-filter -normalize]
| -help
```

**-segment **: The location of a segment(s) we wish to read and obtain information from.

-segmentDir <segmentDir>: The location of a segment(s) directory we wish to read and obtain information from.

-webgraphdb <webgraphdb>: The web graph database to use.

[-filter]: Whether to use URLFilters on the URLs in the segment(s).

[-normalize]: Whether to use URLNormalizers on the URLs in the segment(s).

-help: Prints the above output message, even if this is not included help message is still displayed to cmdout.

<<< CommandLineOptions