bin/nutch dedup

The **dedup** command is an alias for the class **org.apache.nutch.crawl.DeduplicationJob** and is available since Nutch 1.8 (Never ported to Nutch 2.x).

This command takes a path to a crawldb as parameter and finds duplicates based on the signature. If several entries share the same signature, the one with the highest score is kept. If the scores are the same, then the fetch time is used to determine which one to keep with the most recent one being kept. If their fetch times are the same we keep the one with the shortest URL. The entries which are not kept have their status changed to STATUS_DB_DUPLICATE, this is then used by the Cleaning and Indexing jobs to delete the corresponding documents in the backends (SOLR, Elasticsearch).

The dedup command replaces the SOLR dedup command which was limited to SOLR and was a lot less efficient.