

# bin/nutch readdb

Readdb is an alias for `org.apache.nutch.crawl.CrawlDbReader`

## Nutch 1.x

The `CrawlDbReader` implements all the read-only parts of accessing our web database. It provides us with a read utility for the `crawldb`.

Usage:

```
bin/nutch readdb <crawldb> (-stats | -dump <out_dir> | -topN <nnnn> <out_dir> [<min>] | -url <url>)
```

**<crawldb>**: The location of the `crawldb` directory we wish to read and obtain information from.

**-stats**: This prints the overall statistics to `System.out`.

**-dump <out\_dir>**: Enables us to dump the whole `crawldb` to a text file in any `<out_dir>` we wish to specify.

**[-regex <expr>]**: filter records with a regular expression

**[-status <status>]**: filter records by `CrawlDatum` status

**-topN <nnnn> <out\_dir> [<min>]**: This dumps the top `<nnnn>` urls sorted by score relevance to any `<out_dir>` we wish to specify. If the `[<min>]` parameter is passed in the command the reader will skip records with scores below this particular value. This can significantly improve retrieval performance of statistics or `crawldb` dump results.

**-url <url>**: This simply prints information of any particular `<url>` to `System.out`.

## Nutch 2.x

```
Usage: WebTableReader (-stats | -url [url] | -dump <out_dir> [-regex regex])
      [-crawlId <id>] [-content] [-headers] [-links] [-text]
      -crawlId <id>  - the id to prefix the schemas to operate on,
                      (default: storage.crawl.id)
      -stats [-sort] - print overall statistics to System.out
      [-sort]        - list status sorted by host
      -url <url>      - print information on <url> to System.out
      -dump <out_dir> [-regex regex] - dump the webtable to a text file in
                      <out_dir>
      -content        - dump also raw content
      -headers        - dump protocol headers
      -links          - dump links
      -text           - dump extracted text
      [-regex]        - filter on the URL of the webtable entry
```

[CommandLineOptions](#)