# bin/nutch parse

Parse is an alias for org.apache.nutch.parse.ParseSegment

## Nutch 1.x

The class parses contents in one segment. It assumes, under the given segment, the existence of ./fetcher_output/, which is typically generated after a non-parsing fetcher run (i.e., fetcher is started with option -noParsing or as default 'false' boolean value as specified in nutch-default.xml).

Contents in one segment are parsed and saved in these steps:

**1.** ./fetcher_output/ and ./content/ are looped together (possibly by multiple ParserThreads), and content is parsed for each entry. The entry number and resultant ParserOutput are saved in ./parser.unsorted.

**2.** ./parser.unsorted is sorted by entry number, result saved as ./parser.sorted.

**3.** ./parser.sorted and ./fetcher_output/ are looped together. At each entry, ParserOutput is split into ParseDate and ParseText, which are saved in . /parse_data/ and ./parse_text/ respectively. Also updated is FetcherOutput with parsing status, which is saved in ./fetcher/.

In the end, ./fetcher/ should be identical to a directory produced as a result from the fetcher being run WITHOUT option -noParsing e.g. fetching and parsing in the same command. N.B. This is not suggested in a production environment.

By default, intermediates ./parser.unsorted and ./parser.sorted are removed at the end, unless option -noClean is used. However ./fetcher_output/ is kept intact.

Check Fetcher.java and FetcherOutput.java for further details.

```
Usage: bin/nutch parse <segment> [-noFilter] [-noNormalize]
     <segment>    - path to segment you wish to parse
     -noFilter    - optional flag to NOT filtering URLs
     -noNormalize - optional flag for NOT normalizing URLs
```

**<segmentdir>**: This should be the path to the segment directory containing our data for parsing.

## Nutch 2.x

```
Usage: ParserJob (<batchId> | -all) [-crawlId <id>] [-resume] [-force]
    <batchId>     - symbolic batch ID created by Generator
    -crawlId <id> - the id to prefix the schemas to operate on,
                    (default: storage.crawl.id)
    -all          - consider pages from all crawl jobs
    -resume       - resume a previous incomplete job
    -force        - force re-parsing even if a page is already parsed
```

CommandLineOptions