bin/nutch readseg

Readseg is an alias for org.apache.nutch.segment.SegmentReader

This class is similar to readdb in that it dumps the contents of a segment. There are three ways we can use this class:

```
1st Usage: bin/nutch readseg -dump <segment_dir> <output> [general options]
```

-dump: Dumps content of a <segment_dir> as a text file to <output>.

[general options]: General options are provided below.

```
2nd Usage: bin/nutch readseg -list (<segment_dir1> ... | -dir <segments>) [general options]
```

-list: This arguement lists a synopsis of segments in specified directories, or all segments in a directory <segments>, and prints details of them to System.out.

<segment_dir1> ...: This should be a list of the paths for individual segment directories to process.

-dir <segments>: Should be a path to a directory that contains multiple segments.

[general options]: General options are provided below.

```
3rd Usage: bin/nutch readseg -get <segment_dir> <keyValue> [general options]
```

-get: This arguement gets a specified record from a segment, and prints it on System.out.

<segment_dir>: Path to the segment directory.

<keyValue>: This should be the value of the key (url) we wish to retreive specific information about. N.B. It is essential to put "double-quotes" around strings with spaces.

[general options]: General options are provided below.

- **-nocontent**: Pass this to ignore the content directory.
- **-nofetch**: To ignore the crawl_fetch directory.
- **-nogenerate**: To ignore the crawl_generate directory.
- **-noparse**: To ignore the crawl_parse directory.
- **-noparsedata**: To ignore the parse_data directory.
- **-noparsetext**: To ignore the parse_text directory.

CommandLineOptions