Upgrading from 0.8.x to 0.9

Upgrading nutch from 0.8.x to 0.9.0-dev (current svn trunk)

Hadoop 0.7.x deprecates the use of UTF8 class for storing String-s, instead providing Text. As a part of this upgrade all places in Nutch that used UTF8 have been changed to use Text - this change would have to be done sooner or later.

However, this means that all previously created data is no longer compatible with the new tools, which expect data to use Text class instead of UTF8 class.

Now, quickly before you panic ... there is an upgrade path to save your precious data, please read on. 😉

If your data is easily re-created from scratch, I recommend doing this, it might be quicker.

Otherwise, please follow these steps to upgrade your data to the new formats (note: this does not require re-fetching):

CrawlDb upgrade

Use the new tool `convdb <old_db> <new_db> [-withMetadata]` to convert your existing [CrawlDb] from old format using `<UTF8, [CrawlDatum]>` to the new format using `<Text, [CrawlDatum]>`. Optionally, you can also replace all UTF8 metadata keys to use Text (normally not needed).

Segment upgrade

At the moment you can only upgrade non-parsed segments. Please follow these steps:

- if you ran your fetcher in parsing mode, or if you already parsed the segments, then for each segment you need to first remove directories containing parsed content:

```
for i in segments/2007* ; do
   (cd $i && rm -rf crawl_parse parse_data parse_text)
done
```

- then you can use 'mergesegs' command, which was modified to perform the conversion of remaining segment parts, e.g. like this:

```
mkdir converted
for i in segments/2007* ; do
   nutch mergesegs converted $i
done
```

(Of course, you can use this opportunity to actually merge some segments and/or re-slice them - the above command creates exactly one converted segment for one input segment).

- finally, you will need to re-parse converted segments:

```
for i in converted/* ; do
   nutch parse $i
done
```

LinkDb upgrade

There is no option to upgrade your linkdb - but you can easily re-create it from parsed segments, using 'nutch invertlinks' command. Be sure to remove the old linkdb first! (otherwise the tool will attempt to merge new data with old data and things will explode).

Index upgrade

Theoretically, if you rename your converted segments to have exactly the same names as the old segments, you shouldn't need to rebuild your indexes. However, this part of the upgrade process wasn't tested - so to be safe it's better to re-index converted segments.

Your suggestions, comments, success (or horror) stories are appreciated. Good luck!