MergeCrawl

This script allows you to merge 2 Nutch crawls:

- Merge linkdb
- Merge crawldb
- Merge segments
- Update segments
- Index segments
- De-duplicate indexes
- Merge indexes
- Some stats

Tested with Nutch-0.8 release and Ubuntu-Dapper. Please report any bug you find on the mailing list.

Install

- copy this script in your NUTCH_HOME/bin

Usage:

- bin/mergecrawl newcrawl-path crawl1-path crawl2-path
- USE ABSOLUTE PATHS
- e.g. bin/merge_crawls.sh /home/ren/nutch/trunk/build/crawl /home/ren/nutch/trunk/build_f/crawl/ /home/ren/nutch/trunk/build_w/crawl/

```
#!/bin/bash

# Nutch merge crawls script.
# Based on recrawl script
#
# The script merges 2 or more nutch crawls into a single crawl
#
# USE ABSOLUTE PATHS for the script args
# e.g. bin/merge_crawls.sh /home/ren/nutch/trunk/build/crawl /home/ren/nutch/trunk/build_f/crawl/ /home/ren
/nutch/trunk/build_w/crawl/


if [ -n "$1" ]
then
  crawl_dir=$1
  if [ -d $1 ]; then
    echo "error: crawl already exists: '$1'"
    exit 1
  fi
else
  echo "Usage: bin/mergecrawl newcrawl-path crawl1-path crawl2-path, USE ABSOLUTE PATHS"
  exit 1
fi

if [ -n "$2" ]
then
  crawl_1=$2
else
  echo "Usage: bin/mergecrawl newcrawl-path crawl1-path crawl2-path, USE ABSOLUTE PATHS"
  exit 1
fi

if [ -n "$3" ]
then
  crawl_2=$3
else
  echo "Usage: bin/mergecrawl newcrawl-path crawl1-path crawl2-path, USE ABSOLUTE PATHS"
  exit 1
fi


#Sets the path to bin
nutch_dir=`dirname $0`

echo "Creating new crawl in: " $crawl_dir
mkdir $crawl_dir
```

```
webdb_dir=$crawl_dir/crawldb
segments_dir=$crawl_dir/segments
linkdb_dir=$crawl_dir/linkdb
index_dir=$crawl_dir/index

echo Merge linkdb
$nutch_dir/nutch mergelinkdb $linkdb_dir $crawl_1/linkdb $crawl_2/linkdb

echo Merge crawldb
$nutch_dir/nutch mergedb $webdb_dir $crawl_1/crawldb $crawl_2/crawldb

echo Merge segments
segments_1=`ls -d $crawl_1/segments/*`
 #echo 1 $segments_1
segments_2=`ls -d $crawl_2/segments/*`
 #echo 2 $segments_2
$nutch_dir/nutch mergesegs $segments_dir $segments_1 $segments_2


# From there, identical to recrawl.sh

echo Update segments
$nutch_dir/nutch invertlinks $linkdb_dir -dir $segments_dir

echo Index segments
new_indexes=$crawl_dir/newindexes
segment=`ls -d $segments_dir/* | tail -1`
$nutch_dir/nutch index $new_indexes $webdb_dir $linkdb_dir $segment

echo De-duplicate indexes
$nutch_dir/nutch dedup $new_indexes

echo Merge indexes
$nutch_dir/nutch merge $index_dir $new_indexes

echo Some stats
$nutch_dir/nutch readdb $webdb_dir -stats
```