Crawl

Introduction

This is a script to crawl an Intranet as well as the web. It does not crawl using the 'bin/nutch crawl' command or 'Crawl' class present in Nutch. Therefore the filters present in 'conf/crawl-urlfilter.txt ' has no effect on this script. The filters for this script must be set in 'regex-urlfilter.txt'.

Steps

The complete job of this script has been divided broadly into 8 steps.

1. Inject URLs 2. Generate, Fetch, Parse, Update Loop 3. Merge Segments 4. Invert Links 5. Index 6. Dedup 7. Merge Indexes 8. Load new indexes

Modes of Execution

The script can be executed in two modes:-

- Normal Mode
- Safe Mode

Normal Mode

If the script is executed with the command 'bin/runbot', it will delete all the directories such as fetched segments, generated indexes, etc, so as to save space.

Caution: This also means that if something has gone wrong during the crawl and the resultant crawl DB is corrupt or incomplete, there is no way any recovery action can be taken.

Safe Mode

Alternatively, the script can be executed in safe mode as 'bin/runbot safe' which will prevent deletion of these directories. All important temporary directories would be backed up with the prefix BACKUP. e.g. crawl/BACKUPsegments, crawl/BACKUPindexes, crawl/BACKUPindex. If errors occur, you can take recovery action because the directories haven't been deleted. You can then manually merge the segments, generate indexes, etc. from the directories and reload the index.

Normal Mode vs. Safe Mode

Ideally, you should run the script in safe mode a couple of times, to make sure the crawl is running fine. If you are sure, that everything will go fine, you need not run it in safe mode.

Tinkering

Adjust the variables, 'depth', 'threads', 'adddays' and 'topN' as per your needs. Delete or comment out the statement for 'topN' assignment if you do not wish to set a 'topN' value.

NUTCH_HOME

If you are not executing the script as 'bin/runbot' from Nutch directory, you should either set the environment variable 'NUTCH_HOME' or edit the following in the script:-

```
if [ -z "$NUTCH_HOME" ]
then
  NUTCH_HOME=.
```

Set 'NUTCH_HOME' to the path of the Nutch directory (if you are not setting it as an environment variable, since if environment variable is set, the above assignment is ignored).

CATALINA_HOME

'CATALINA_HOME' points to the Tomcat installation directory. You must either set this as an environment variable or set it by editing the following lines in the script:-

```
if [ -z "$CATALINA_HOME" ]
then
  CATALINA_HOME=/opt/apache-tomcat-6.0.10
```

Similar to the previous section, if this variable is set in the environment, then the above assignment is ignored.

Can it re-crawl?

The author has used this script to re-crawl a couple of times. However, no real world testing has been done for re-crawling. Therefore, you may try to use the script for re-crawl. If it works fine or it doesn't work properly for re-crawl, please let us know.

Script

```
#!/bin/sh

# runbot script to run the Nutch bot for crawling and re-crawling.
# Usage: bin/runbot [safe]
#        If executed in 'safe' mode, it doesn't delete the temporary
#        directories generated during crawl. This might be helpful for
#        analysis and recovery in case a crawl fails.
#
# Author: Susam Pal

depth=2
threads=5
adddays=5
topN=15 #Comment this statement if you don't want to set topN value

# Arguments for rm and mv
RMARGS="-rf"
MVARGS="--verbose"

# Parse arguments
if [ "$1" == "safe" ]
then
  safe=yes
fi

if [ -z "$NUTCH_HOME" ]
then
  NUTCH_HOME=.
  echo runbot: $0 could not find environment variable NUTCH_HOME
  echo runbot: NUTCH_HOME=$NUTCH_HOME has been set by the script
else
  echo runbot: $0 found environment variable NUTCH_HOME=$NUTCH_HOME
fi

if [ -z "$CATALINA_HOME" ]
then
  CATALINA_HOME=/opt/apache-tomcat-6.0.10
  echo runbot: $0 could not find environment variable NUTCH_HOME
  echo runbot: CATALINA_HOME=$CATALINA_HOME has been set by the script
else
  echo runbot: $0 found environment variable CATALINA_HOME=$CATALINA_HOME
fi

if [ -n "$topN" ]
then
  topN="-topN $topN"
else
  topN=""
fi

steps=8
echo "----- Inject (Step 1 of $steps) -----"
$NUTCH_HOME/bin/nutch inject crawl/crawldb urls
```

```
echo "----- Generate, Fetch, Parse, Update (Step 2 of $steps) -----"
for((i=0; i < $depth; i++))
do
  echo "--- Beginning crawl at depth `expr $i + 1` of $depth ---"
  $NUTCH_HOME/bin/nutch generate crawl/crawldb crawl/segments $topN \
      -adddays $adddays
  if [ $? -ne 0 ]
  then
    echo "runbot: Stopping at depth $depth. No more URLs to fetch."
    break
  fi
  segment=`ls -d crawl/segments/* | tail -1`

  $NUTCH_HOME/bin/nutch fetch $segment -threads $threads
  if [ $? -ne 0 ]
  then
    echo "runbot: fetch $segment at depth `expr $i + 1` failed."
    echo "runbot: Deleting segment $segment."
    rm $RMARGS $segment
    continue
  fi

  $NUTCH_HOME/bin/nutch updatedb crawl/crawldb $segment
done

echo "----- Merge Segments (Step 3 of $steps) -----"
$NUTCH_HOME/bin/nutch mergesegs crawl/MERGEDsegments crawl/segments/*
if [ "$safe" != "yes" ]
then
  rm $RMARGS crawl/segments
else
  rm $RMARGS crawl/BACKUPsegments
  mv $MVARGS crawl/segments crawl/BACKUPsegments
fi

mv $MVARGS crawl/MERGEDsegments crawl/segments

echo "----- Invert Links (Step 4 of $steps) -----"
$NUTCH_HOME/bin/nutch invertlinks crawl/linkdb crawl/segments/*

echo "----- Index (Step 5 of $steps) -----"
$NUTCH_HOME/bin/nutch index crawl/NEWindexes crawl/crawldb crawl/linkdb \
    crawl/segments/*

echo "----- Dedup (Step 6 of $steps) -----"
$NUTCH_HOME/bin/nutch dedup crawl/NEWindexes

echo "----- Merge Indexes (Step 7 of $steps) -----"
$NUTCH_HOME/bin/nutch merge crawl/NEWindex crawl/NEWindexes

echo "----- Loading New Index (Step 8 of $steps) -----"
${CATALINA_HOME}/bin/shutdown.sh

if [ "$safe" != "yes" ]
then
  rm $RMARGS crawl/NEWindexes
  rm $RMARGS crawl/index
else
  rm $RMARGS crawl/BACKUPindexes
  rm $RMARGS crawl/BACKUPindex
  mv $MVARGS crawl/NEWindexes crawl/BACKUPindexes
  mv $MVARGS crawl/index crawl/BACKUPindex
fi

mv $MVARGS crawl/NEWindex crawl/index

${CATALINA_HOME}/bin/startup.sh

echo "runbot: FINISHED: Crawl completed!"
echo ""
```