Useful scripts

Some useful scripts for running nutch in real life.

simple loop executes nutch commands:

```
# number of loops
LIMIT=20
# number of pages to fetch
MAXPAGES=50000
# you mail address
MAIL=you@domain.com

for ((a=1; a <= LIMIT ; a++))
 do
  echo '************** start new crawl loop '$a'**************'
  bin/nutch generate db segments -topN $MAXPAGES > gen_$a.log 2>&1
  s1=`ls -d segments/2* | tail -1`
  echo $s1
  bin/nutch fetch $s1 > fetch_$a.log 2>&1
  bin/nutch updatedb db $s1 > update_$a.log 2>&1
  bin/nutch analyze db 3 > analyse_$a.log 2>&1
  bin/nutch index $s1   > index_$a.log 2>&1
  bin/nutch dedup segments dedup$a.tmp  > dedup_$a.log 2>&1
  du -hs db segments | mail -s'nutch loop $a done' $MAIL
 done
exit 0
```

scripts for simple paralyzing nutch processes

The scripts can be used to distribute the nutch processes over a set of machines. It is possible to run fetching, indexing and "dedub" at the same time. However, since the web data base has a centralized architecture it is not possible to run the most time consuming tasks (segment generation, data base analysis and update) at the same time.

The scripts require that all machines share the same hard drive for example a NAS (network attached storage) but may usage of the nutch dfs would be an interesting alternative.

fetching script:

```
#!/bin/bash
# fetcher script by stefan groschupf sg(AT)http://www.media-style.com
file=segment.done
processFile=fetch.running
doneFile=fetch.done

while  [ 1 ]
do
for i in ./segments/*; do
  if [ -d "$i" ]; then

searchFile=$i/$file
    if [ -f $searchFile ]
    then
        echo "$searchFile file exist"
        rm $searchFile
        pFile=$i/$processFile
        echo "done" >$pFile

        # run command
        FileName=./logs/$(date +%y_%m_%d_%H_%M_%S)_$processFile.log
        echo $FileName

        echo "start command" >>$FileName 2>&1
        bin/nutch fetch $i >>$FileName 2>&1
        echo "command done" >>$FileName 2>&1
        rm $pFile
        dFile=$i/$doneFile
        echo "done" >$dFile

    fi
  fi
done
#echo "sleep for a while"
sleep 10
done
```

web db script:

```
#!/bin/bash
# webdb script by stefan groschupf sg(AT)http://www.media-style.com
file=fetch.done
processFile=webdb.running
doneFile=webdb.done

while  [ 1 ]
do
for i in ./segments/*; do
  if [ -d "$i" ]; then

searchFile=$i/$file
    if [ -f $searchFile ]
    then
        echo "$searchFile file exist"
        rm $searchFile
        pFile=$i/$processFile
        echo "done" > $pFile

        # run command
        FileName=./logs/$(date +%y_%m_%d_%H_%M_%S)_$processFile.log
        echo $FileName

        echo "start command" >>$FileName 2>&1

        bin/nutch updatedb db $i
        bin/nutch analyze db 2
        bin/nutch generate db segments -topN 1000000
        s3=`ls -d segments/2* | tail -1`
        echo $s3

        rm $pFile
        dFile=$i/$doneFile
        echo "done" >>$dFile
        echo "done" >> $s3/segment.done


    fi
  fi
done

#echo "sleep for a while"
sleep 10
done
```

indexing script:

```
#!/bin/bash
# index script by stefan groschupf sg(AT)http://www.media-style.com
file=webdb.done
processFile=indexer.running
doneFile=indexer.done

while  [ 1 ]
do
for i in ./segments/*; do
  if [ -d "$i" ]; then

searchFile=$i/$file
    if [ -f $searchFile ]
    then
        echo "$searchFile file exist"
        rm $searchFile
        pFile=$i/$processFile
        echo "done" > $pFile

        # run command
        FileName=./logs/$(date +%y_%m_%d_%H_%M_%S)_$processFile.log
        echo $FileName

        echo "start command" >>$FileName 2>&1
        bin/nutch index $i >>$FileName 2>&1
        echo "command done" >>$FileName 2>&1
        touch /Users/myUser/nutch/uiserver/jakarta-tomcat-5.5.4/webapps/ROOT/WEB-INF/web.xml
        rm $pFile
        dFile=$i/$doneFile
        echo "done" >>$dFile

    fi
  fi
done

#echo "sleep for a while"
sleep 10
done
```

dedub script:

```
#!/bin/bash
# dedub  script by stefan groschupf sg(AT)http://www.media-style.com
file=indexer.done
processFile=dedup.running
doneFile=dedup.done

while  [ 1 ]
do
for i in ./segments/*; do
  if [ -d "$i" ]; then

searchFile=$i/$file
    if [ -f $searchFile ]
    then
        echo "$searchFile file exist"
        rm $searchFile
        pFile=$i/$processFile
        echo "done" > $pFile

        # run command
        FileName=./logs/$(date +%y_%m_%d_%H_%M_%S)_$processFile.log
        echo $FileName

        echo "start command" >>$FileName 2>&1
        bin/nutch dedup segments dedup.tmp >>$FileName 2>&1
        rm dedup.tmp >>$FileName 2>&1
        echo "command done" >>$FileName 2>&1
        rm $pFile
        dFile=$i/$doneFile
        echo "done" >>$dFile

    fi
  fi
done

#echo "sleep for a while"
sleep 10
done
```