

```

#!/bin/bash

# $1 : ./URLS
# $2 : ./TABLEAUX
# $3 : regexp

#"((л|Л)актоз(|L)actose)"

rm -f "$2/tableau.html"; # on supprime le fichier qu'on veut reconstruire

echo "les URLS sont dans : $1" ;
echo "je stocke le tableau HTML dans : $2" ; # il n'y a pas besoin de définir les variable $1 et $2,
car elles sont prédéfinies !!
echo "voici le motif recherché : $3" ;

numerotableau=1 ;
motif=$3 ;

echo "<html><head><meta charset=\"utf-8\"/><title>TABLEAU D'URLS</title></head><body>"
>> "$2/tableau.html" ;

#####

# parcours du dossier contenant les fichiers URLS

for fichier in $(ls $1)
do
compteur=1
echo "$1/$fichier"
#####
# création du tableau #
#####
echo "<table border = \"2\" align=\"center\">" >> "$2/tableau.html" ;
echo "<tr bgcolor=\"yellow\"><td>N°</td><td>URL</td><td>Code http</td><td>encodage</
td><td>Page aspirée</td><td>Dump</td><td>Filtrage Txt</td><td>Filtrage Html</
td><td>Index</td><td>Bitexte</td><td>Fq Motif</td></tr>" >> "$2/tableau.html" ;
# nommer les colonnes

# parcours d'un fichier URL : lecture ligne à ligne des URLS

for ligne in $(cat "$1/$fichier")
do
echo "_____";
echo "Traitement de l'URL : $ligne" ;
echo "_____";
# 1 - on teste la connection vers l'URLSs
coderetour=$(curl -sIL -o retour.txt -w %{http_code} $ligne) ;
echo "CODE HTTP : $encodage" # curl -L pour que curl puisse trouver l'adresse, même
s'il est déplacé
# est-ce que la variable "coderetour" est égale à 200 ?
if [[ $coderetour == 200 ]]
then
# 2 - on essaie de récupérer l'encodage de la page associée à l'URL
encodage=$(curl -sIL -o extract -w %{content_type} $ligne | cut -f2 -d"=" | tr '[a-z]'
'[A-Z]' | tr -d '\r') ; # on transcode en majuscule pour ne pas avoir 'utf' et 'UTF' et on supprime
\r pour ne pas avoir des problèmes
echo "Encodage détecté par curl est : $encodage"

# 3 - et on aspire la page

```

```

curl -L -o "./PAGES-ASPIREES/$encodage-$numerotableau-$compteur.html"
"$ligne" ;

if [[ $encodage == "UTF-8" ]]
then
# on remplit le tableau
# 1 - on lynx la page aspirée
lynx -dump -nolist -assume_charset=$encodage -display-
charset=$encodage "./PAGES-ASPIREES/$encodage-$numerotableau-$compteur.html" > ./
DUMP-TEXT/$encodage-$numerotableau-$compteur.txt ;
# 2 - on cree le fichier contexte TXT via egrep
egrep -i -C2 "$motif" ./DUMP-TEXT/$encodage-$numerotableau-
$compteur.txt > ./CONTEXTES/$encodage-$numerotableau-$compteur.txt;
# 3 - la fréquence du motif
nbmotif=$(egrep -coi "$motif" ./DUMP-TEXT/$encodage-$numerotableau-
$compteur.txt);
# 4 - contexte HTML
perl ./minigrep/minigrepmultilingue.pl "utf-8" ./DUMP-TEXT/$encodage-
$numerotableau-$compteur.txt ./minigrep/parametre-motif.txt ;
mv resultat-extraction.html ./CONTEXTES/$encodage-$numerotableau-
$compteur.html ;
# 5 - index hierarchique
egrep -o "\w+" ./DUMP-TEXT/$encodage-$numerotableau-$compteur.txt |
sort | uniq -c | sort -r > ./DUMP-TEXT/index-$encodage-$numerotableau-$compteur.txt ;
# 6 - bigramme
egrep -o "\w+" ./DUMP-TEXT/$encodage-$numerotableau-$compteur.txt >
bi1.txt;
tail -n +2 bi1.txt > bi2.txt ;
paste bi1.txt bi2.txt > bi3.txt ;
cat bi3.txt | sort | uniq -c | sort -r > ./DUMP-TEXT/bigramme-$encodage-
$numerotableau-$compteur.txt ;
# 7 - on repmlit le tebleau
echo "<tr>
<td>$compteur</td>
<td><a href=\"$ligne\" target=\"_blank\">$ligne</a></td>
<td>Code_http:$coderetour</td>
<td>Encodage:$encodage</td>
<td><a href=\"../PAGES-ASPIREES/$encodage-$numerotableau-
$compteur.html\">$encodage-$numerotableau-$compteur.html</a></td>
<td><a href=\"../DUMP-TEXT/$encodage-$numerotableau-$compteur.txt\">$encodage-
$numerotableau-$compteur.txt</a></td>
<td><a href=\"../CONTEXTES/$encodage-$numerotableau-$compteur.txt\">$encodage-
$numerotableau-$compteur.txt</a></td>
<td><a href=\"../CONTEXTES/$encodage-$numerotableau-$compteur.html\">$encodage-
$numerotableau-$compteur.html</a></td>
<td><a href=\"../DUMP-TEXT/index-$encodage-$numerotableau-$compteur.txt\">index-
$encodage-$numerotableau-$compteur</a></td>
<td><a href=\"../DUMP-TEXT/bigramme-$encodage-$numerotableau-
$compteur.txt\">bigramme-$encodage-$numerotableau-$compteur</a></td>
<td>$nbmotif</td>
</tr>" >> "$2/tableau.html";

elif [[ $encodage == "CP1251" ]]
# extraire les information de CP1251 avant de le transormer en UTF-8
then
# on remplit le tableau
# 1 - on lynx la page aspirée
lynx -dump -nolist -assume_charset=windows-1251 -display-
charset=windows-1251 "./PAGES-ASPIREES/$encodage-$numerotableau-$compteur.html" > ./
DUMP-TEXT/$encodage-$numerotableau-$compteur.txt ;

```

```

# 2 - on cree le fichier contexte TXT via egrep
egrep -i -C2 "$motif" ./DUMP-TEXT/$encodage-$numerotableau-
$compteur.txt > ./CONTEXTES/$encodage-$numerotableau-$compteur.txt;
# 3 - la fréquence du motif
nbmotif=$(egrep -coi "$motif" ./DUMP-TEXT/$encodage-$numerotableau-
$compteur.txt);
# 4 - contexte HTML
perl ./minigrep/minigrepmultilingue.pl "utf-8" ./DUMP-TEXT/$encodage-
$numerotableau-$compteur.txt ./minigrep/parametre-motif.txt ;
mv resultat-extraction.html ./CONTEXTES/$encodage-$numerotableau-
$compteur.html ;
# 5 - index hierarchique
egrep -o "\w+" ./DUMP-TEXT/$encodage-$numerotableau-$compteur.txt |
sort | uniq -c | sort -r > ./DUMP-TEXT/index-$encodage-$numerotableau-$compteur.txt ;
# 6 - bigramme
egrep -o "\w+" ./DUMP-TEXT/$encodage-$numerotableau-$compteur.txt >
bi1.txt;
tail -n +2 bi1.txt > bi2.txt ;
paste bi1.txt bi2.txt > bi3.txt ;
cat bi3.txt | sort | uniq -c | sort -r > ./DUMP-TEXT/bigramme-$encodage-
$numerotableau-$compteur.txt ;
# 7 - on replit le tableau
echo "<tr>
<td>$compteur</td>
<td><a href=\"$ligne\" target=\"_blank\">$ligne</a></td>
<td>Code_http:$coderetour</td>
<td>Encodage:$encodage</td>
<td><a href=\"../PAGES-ASPIREES/$encodage-$numerotableau-
$compteur.html\">$encodage-$numerotableau-$compteur.html</a></td>
<td><a href=\"../DUMP-TEXT/$encodage-$numerotableau-$compteur.txt\">$encodage-
$numerotableau-$compteur.txt</a></td>
<td><a href=\"../CONTEXTES/$encodage-$numerotableau-$compteur.txt\">$encodage-
$numerotableau-$compteur.txt</a></td>
<td><a href=\"../CONTEXTES/$encodage-$numerotableau-$compteur.html\">$encodage-
$numerotableau-$compteur.html</a></td>
<td><a href=\"../DUMP-TEXT/index-$encodage-$numerotableau-$compteur.txt\">index-
$encodage-$numerotableau-$compteur</a></td>
<td><a href=\"../DUMP-TEXT/bigramme-$encodage-$numerotableau-
$compteur.txt\">bigramme-$encodage-$numerotableau-$compteur</a></td>
<td>$nbmotif</td>
</tr>" >> "$2/tableau.html";

# changer d'encodage

echo "Encodage détecté est windows-1251"
lynx -dump -nolist -assume_charset=windows-1251 -display-
charset=windows-1251 "../PAGES-ASPIREES/$encodage-$numerotableau-$compteur.html" > ./
DUMP-TEXT/UTF-8-$numerotableau-$compteur.txt
iconv -f "CP1251" -t "UTF-8" ./DUMP-TEXT/$encodage-$numerotableau-
$compteur.txt > ./DUMP-TEXT/UTF-8-$numerotableau-$compteur.txt ;
egrep -i -C2 "$motif" ./DUMP-TEXT/UTF-8-$numerotableau-$compteur.txt
> ./CONTEXTES/UTF-8-$numerotableau-$compteur.txt;

# 3 - la fréquence du motif
nbmotif=$(egrep -coi "$motif" ./DUMP-TEXT/UTF-8-$numerotableau-
$compteur.txt);
# 4 - contexte HTML
perl ./minigrep/minigrepmultilingue.pl "utf-8" ./DUMP-TEXT/UTF-8-
$numerotableau-$compteur.txt ./minigrep/parametre-motif.txt ;

```

```

mv resultat-extraction.html ./CONTEXTES/UTF-8-$numerotableau-
$compteur.txt ;
# 5 - index hierarchique
egrep -o "\w+" ./DUMP-TEXT/UTF-8-$numerotableau-$compteur.txt | sort |
uniq -c | sort -r > ./DUMP-TEXT/index-UTF-8-$numerotableau-$compteur.txt ;
# 6 - bigramme
egrep -o "\w+" ./DUMP-TEXT/UTF-8-$numerotableau-$compteur.txt >
bi1.txt;
tail -n +2 bi1.txt > bi2.txt ;
paste bi1.txt bi2.txt > bi3.txt ;
cat bi3.txt | sort | uniq -c | sort -r > ./DUMP-TEXT/bigramme-UTF-8-
$numerotableau-$compteur-utf-8.txt ;
# 7 - on repplit le tebleau
echo "<tr>
<td>$compteur</td>
<td><a href=\"$ligne\" target=\"_blank\">$ligne</a></td>
<td>Code_http:$coderetour</td>
<td>Encodage:$encodage</td>
<td><a href=\"../PAGES-ASPIREES/UTF-8-$numerotableau-
$compteur.html\">UTF-8-$numerotableau-$compteur.html</a></td>
<td><a href=\"../DUMP-TEXT/UTF-8-$numerotableau-$compteur.txt\">UTF-8-
$numerotableau-$compteur.txt</a></td>
<td><a href=\"../CONTEXTES/UTF-8-$numerotableau-$compteur.txt\">UTF-8-
$numerotableau-$compteur.txt</a></td>
<td><a href=\"../CONTEXTES/UTF-8-$numerotableau-$compteur.html\">UTF-8-
$numerotableau-$compteur.html</a></td>
<td><a href=\"../DUMP-TEXT/index-UTF-8--$numerotableau-$compteur.txt\">index-
UTF-8-$numerotableau-$compteur</a></td>
<td><a href=\"../DUMP-TEXT/bigramme-UTF-8-$numerotableau-
$compteur.txt\">bigramme-UTF-8-$numerotableau-$compteur</a></td>
<td>$nbmotif</td>
</tr>" >> "$2/tableau.html";

else

# l'encodage n'est pas UTF-8
# est-ce qu'il est connu par iconv ?
retouriconv=$(iconv -l | egrep -o -i "\b$encodage\b") ;

# l'output n'est pas vide

if [[ $retouriconv != "" ]]
then
# on remplit le tableau
# 1 - on lynx la page aspirée
lynx -dump -nolist -assume_charset=$encodage -display-
charset=$encodage ./PAGES-ASPIREES/$encodage-$numerotableau-$compteur.html > ./
DUMP-TEXT/$encodage-$numerotableau-$compteur.txt ;
# 2 - on cree le fichier contexte TXT via egrep
egrep -i -C2 "$motif" ./DUMP-TEXT/$encodage-
$numerotableau-$compteur.txt > ./CONTEXTES/$encodage-$numerotableau-$compteur.txt;
# 3 - la fréquence du motif
nbmotif=$(egrep -coi "$motif" ./DUMP-TEXT/$encodage-
$numerotableau-$compteur.txt);
# 4 - contexte HTML
perl ./minigrep/minigrepmultilingue.pl "utf-8" ./DUMP-TEXT/
$encodage-$numerotableau-$compteur.txt ./minigrep/parametre-motif.txt ;

```

```

mv resultat-extraction.html ./CONTEXTES/$encodage-
$numerotableau-$compteur.html ;
# 5 - index hierarchique
egrep -o "\w+" ./DUMP-TEXT/$encodage-$numerotableau-
$compteur.txt | sort | uniq -c | sort -r > ./DUMP-TEXT/index-$encodage-$numerotableau-
$compteur.txt ;
# 6 - bigramme
egrep -o "\w+" ./DUMP-TEXT/$encodage-$numerotableau-
$compteur.txt > bi1.txt;
tail -n +2 bi1.txt > bi2.txt ;
paste bi1.txt bi2.txt > bi3.txt ;
cat bi3.txt | sort | uniq -c | sort -r > ./DUMP-TEXT/bigramme-
$encodage-$numerotableau-$compteur.txt ;
# 7 - on replit le tableau
echo "<tr>
<td>$compteur</td>
<td><a href=\"$ligne\" target=\"_blank\">$ligne</a></td>
<td>Code_http:$coderetour</td>
<td>Encodage:$encodage</td>
<td><a href=\"../PAGES-ASPIREES/$encodage-$numerotableau-
$compteur.html\">$encodage-$numerotableau-$compteur.html</a></td>
<td><a href=\"../DUMP-TEXT/$encodage-$numerotableau-
$compteur.txt\">$encodage-$numerotableau-$compteur.txt</a></td>
<td><a href=\"../CONTEXTES/$encodage-$numerotableau-
$compteur.txt\">$encodage-$numerotableau-$compteur.txt</a></td>
<td><a href=\"../CONTEXTES/$encodage-$numerotableau-
$compteur.html\">$encodage-$numerotableau-$compteur.html</a></td>
<td><a href=\"../DUMP-TEXT/index-$encodage-$numerotableau-
$compteur.txt\">index-$encodage-$numerotableau-$compteur</a></td>
<td><a href=\"../DUMP-TEXT/bigramme-$encodage-$numerotableau-
$compteur.txt\">bigramme-$encodage-$numerotableau-$compteur</a></td>
<td>$nbmotif</td>
</tr>" >> "$2/tableau.html";

# changer d'encodage

echo "Encodage est $retouriconv est connu par iconv"

lynx -dump -nolist -assume_charset=$encodage -display-
charset=$encodage "./PAGES-ASPIREES/$encodage-$numerotableau-$compteur.html" > ./
DUMP-TEXT/UTF-8-$numerotableau-$compteur.txt
iconv -f "$encodage" -t "UTF-8" ./DUMP-TEXT/$encodage-
$numerotableau-$compteur.txt > ./DUMP-TEXT/UTF-8-$numerotableau-$compteur.txt ;
egrep -i -C2 "$motif" ./DUMP-TEXT/UTF-8-$numerotableau-
$compteur.txt > ./CONTEXTES/UTF-8-$numerotableau-$compteur.txt;

# 3 - la fréquence du motif
nbmotif=$(egrep -coi "$motif" ./DUMP-TEXT/UTF-8-
$numerotableau-$compteur.txt) ;
# 4 - contexte HTML
perl ./minigrep/minigrepmultilingue.pl "utf-8" ./DUMP-TEXT/
UTF-8-$numerotableau-$compteur.txt ./minigrep/parametre-motif.txt ;
mv resultat-extraction.html ./CONTEXTES/UTF-8-
$numerotableau-$compteur.txt ;
# 5 - index hierarchique
egrep -o "\w+" ./DUMP-TEXT/UTF-8-$numerotableau-
$compteur.txt | sort | uniq -c | sort -r > ./DUMP-TEXT/index-UTF-8-$numerotableau-$compteur.txt
;
# 6 - bigramme

```

```

egrep -o "\w+" ./DUMP-TEXT/UTF-8-$numerotableau-
$compteur.txt > bi1.txt;
tail -n +2 bi1.txt > bi2.txt ;
paste bi1.txt bi2.txt > bi3.txt ;
cat bi3.txt | sort | uniq -c | sort -r > ./DUMP-TEXT/bigramme-
UTF-8-$numerotableau-$compteur-utf-8.txt ;
# 7 - on replit le tebleau
echo "<tr>
<td>$compteur</td>
<td><a href=\"$ligne\" target=\"_blank\">$ligne</a></td>
<td>Code_http:$coderetour</td>
<td>Encodage:$encodage</td>
<td><a href=\"../PAGES-ASPIREES/UTF-8-$numerotableau-
$compteur.html\">UTF-8-$numerotableau-$compteur.html</a></td>
<td><a href=\"../DUMP-TEXT/UTF-8-$numerotableau-
$compteur.txt\">UTF-8-$numerotableau-$compteur.txt</a></td>
<td><a href=\"../CONTEXTES/UTF-8-$numerotableau-
$compteur.txt\">UTF-8-$numerotableau-$compteur.txt</a></td>
<td><a href=\"../CONTEXTES/UTF-8-$numerotableau-
$compteur.html\">UTF-8-$numerotableau-$compteur.html</a></td>
<td><a href=\"../DUMP-TEXT/index-UTF-8--$numerotableau-
$compteur.txt\">index-UTF-8-$numerotableau-$compteur</a></td>
<td><a href=\"../DUMP-TEXT/bigramme-UTF-8-$numerotableau-
$compteur.txt\">bigramme-UTF-8-$numerotableau-$compteur</a></td>
<td>$nbmotif</td>
</tr>" >> "$2/tableau.html";

else

# l'encodage n'est pas reconnu pas iconv, donc, on essaye de le
chercher manuellement dans le fichier

echo "Je cherche l'encodage avec egrep" ;

encodgrep=$(egrep -i -o "charset=(\"|\")?[\^\"]{1,3};)+(\"|\")" ./PAGES-
ASPIREES/$encodage-$numerotableau-$compteur.html | tr -d "\"|\'" | tr '[a-z]' '[A-Z]') ;

echo "Encodage trouvé est $encodgrep" ;
echo "Je cherche si iconv connaît $encodgrep" ;

iconvgrep=$(iconv -l | egrep -o -i "\b$encodgrep\b") ;

# est-ce que c'est de l'utf-8 ?

if [[ $iconvgrep == "UTF-8" ]]

then

lynx -dump -nolist -assume_charset=$encodage
-display_charset=$encodage ./PAGES-ASPIREES/$encodage-$numerotableau-$compteur.html"
> ./DUMP-TEXT/$encodage-$numerotableau-$compteur.txt ;
# 2 - on cree le fichier
contexte TXT via egrep
egrep -i -C2 "$motif" ./DUMP-TEXT/$encodage-
$numerotableau-$compteur.txt > ./CONTEXTES/$encodage-$numerotableau-$compteur.txt;
# 3 - la fréquence du motif
nbmotif=$(egrep -coi "$motif" ./DUMP-TEXT/
$encodage-$numerotableau-$compteur.txt);
# 4 - contexte HTML

```

```

perl ./minigrep/minigrepmultilingue.pl "utf-8" ./DUMP-
TEXT/$encodage-$numerotableau-$compteur.txt ./minigrep/parametre-motif.txt ;
mv resultat-extraction.html ./CONTEXTES/
$encodage-$numerotableau-$compteur.html ;
# 5 - index hierarchique
egrep -o "\w+" ./DUMP-TEXT/$encodage-
$numerotableau-$compteur.txt | sort | uniq -c | sort -r > ./DUMP-TEXT/index-$encodage-
$numerotableau-$compteur.txt ;

# 6 - bigramme
egrep -o "\w+" ./DUMP-TEXT/$encodage-
$numerotableau-$compteur.txt > bi1.txt;
tail -n +2 bi1.txt > bi2.txt ;
paste bi1.txt bi2.txt > bi3.txt ;
cat bi3.txt | sort | uniq -c | sort -r > ./DUMP-TEXT/
bigramme-$encodage-$numerotableau-$compteur.txt ;
# 7 - on rempli le tableau
echo "<tr>
<td>$compteur</td>
<td><a href=\"$ligne\" target=\"_blank\">$ligne</
a></td>
<td>Code_http:$coderetour</td>
<td>Encodage:$encodage</td>
<td><a href=\"../PAGES-ASPIREES/$encodage-
$numerotableau-$compteur.html\">$encodage-$numerotableau-$compteur.html</a></td>
<td><a href=\"../DUMP-TEXT/$encodage-
$numerotableau-$compteur.txt\">$encodage-$numerotableau-$compteur.txt</a></td>
<td><a href=\"../CONTEXTES/$encodage-
$numerotableau-$compteur.txt\">$encodage-$numerotableau-$compteur.txt</a></td>
<td><a href=\"../CONTEXTES/$encodage-
$numerotableau-$compteur.html\">$encodage-$numerotableau-$compteur.html</a></td>
<td><a href=\"../DUMP-TEXT/index-$encodage-
$numerotableau-$compteur.txt\">index-$encodage-$numerotableau-$compteur</a></td>
<td><a href=\"../DUMP-TEXT/bigramme-$encodage-
$numerotableau-$compteur.txt\">bigramme-$encodage-$numerotableau-$compteur</a></td>
<td>$nbmotif</td>
</tr>" >> "$2/tableau.html";

# c'est pas de l'utf-8 :

else

if [[ $iconvgrep != "" ]]

then

echo "iconv ne connait pas l'encodage $encodgrep" ;
echo "le traitement du fichier $fichier s'arrête ici" ;

echo "<tr>
<td>$compteur</td>
<td><a href=\"$ligne\"
target=\"_blank\">$ligne</a></td>
<td>Code_http:$coderetour</
td>
<td>Encodage:$encodage</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>

```

```
<td>"---"</td>
<td>"---"</td>
</tr>" >> "$2/tableau.html";
```

```
fi
```

```
fi
```

```
fi
```

```
        # echo "<tr><td>$compteur :</td><td><a href=\"\$ligne\"
target=\"_blank\">$ligne</a></td><td>Code_http:$coderetour</td><td>Encodage:$encodage/
td><td><a href=\"../PAGES-ASPIREES/$encodage-$numerotableau-
$compteur.html\">$encodage-$numerotableau-$compteur.html</a></td><td>-</td></tr>" >>
"$2/tableau.html" ;
```

```
        fi
```

```
    else
```

```
        # code est différent de 200, donc les colonnes restent vides
```

```
        # la sortie : ligne de tableau html sans sortie "-"
```

```
        echo "<tr><td>$compteur</td><td><a href=\"\$ligne\" target=\"_blank\">$ligne</a></
td><td>Code_http:$coderetour</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-
</td><td>-</td><td>-</td></tr>" >> "$2/tableau.html";
```

```
        fi
```

```
        compteur=$((compteur+1)) ;
```

```
    done
```

```
echo "</table><br />" >> "$2/tableau.html" ;
```

```
# on incrémente le compteur de tableau
```

```
numerotableau=$((numerotableau+1))
```

```
done
```

```
echo "</body></html>" >> "$2/tableau.html" ;
```