```
#!/bin/bash
rm -f "$2/tableau.html" ;
# S'il y existe déjà un tel fichier, on le supprime
echo "Les URLs sont stockés dans : $1";
echo "Le tableau va être crée dans : $2";
motif=$3;
echo "MOTIF=$3" > ./minigrepmultilingue/motif-regexp.txt
# $1 est le premier argument fournit à la commande bash (./URLS)
# $2 est le deuxième argument fournit à la commande bash (./TABLEAUX)
echo "<html><head><meta charset=\"utf-8\" /><title>Mon beau tableau
d'URLs</title></head><body>" > "$2/tableau.html";
# Il faut mettre le nom du fichier entre guillemets "$2/tableau.html"
compteur_fichier=1;
# Ce compteur désigne le numéro du fichier contenant des liens url (j'en ai deux)
#...
for fichier in $(ls $1)
# Attention, pas de \: après for !
# On commence une boucle : pour chaque fichier dans le répertoire $1,
# c'est à dire pour chaque fichier dans le répertoire ./URLS
do
compteur_ligne=1   # Compteur qui va compter les lignes du fichier, chaque ligne
                                                        # correspond à un lien url
echo "<table border=\"2\" align=\"center\" width=\"80%\"><caption>Tableau
$compteur_fichier</caption>" >> "$2/tableau.html" ;
echo "<tr bgcolor=\"grey\">
<th>N.</th>
<th>Lien URL</th>
<th>Code http</th>
<th>Encodage</th>
<th>Page aspirée</th>
<th>Dump txt</th>
<th>Contexte txt</th>
<th>Contexte html</th>
<th>Index</th>
<th>Bigrammes</th>
<th>Nb motif</th>
</tr>" >> "$2/tableau.html"  ;
echo "$1/$fichier";   # Affiche sur stdout le fichier d'url qui va être traité
for ligne in $(cat "$1/$fichier")    # On commence à lire ligne par ligne le fichier
          do
          echo "traitement d'URL : $ligne";
          # J'attribue une variable correspondant au code de la connexion avec la page
          http_code=$(curl -SIL -o tmp.txt -w %{http_code} $ligne) ;
          echo "CODE HTTP : $http_code";

   if [[ $http_code == 200 ]]
                    then
                    # Si ce code est 200, je continue -->
                    #
                    curl -L $ligne > ./PAGES-ASPIREES/$encodage-$compteur_fichier-
$compteur_ligne.html;
```

```
		# J'attribue une variable correspodant à l'encodage
		encodage=$(curl -SIL -o tmp1.txt -w %{content_type} $ligne | cut -f2 -d"=" | tr "[a-z]" "[A-Z]" | tr -d "\r" | sed -r "s/\"//g");
		echo "ENCODAGE DETECTE PAR CURL : $encodage";
		if [[ $encodage == "UTF-8" ]]
			then
			# dump de la page
			lynx -dump -nolist -assume_charset=$encodage -display_charset=$encodage "./PAGES-ASPIREES/$encodage-$compteur_fichier-$compteur_ligne.html" > ./DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt ;
			# contexte de la page en txt
			egrep -i -C2 "$motif" ./DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt > ./CONTEXTES/$encodage-$compteur_fichier-$compteur_ligne.txt ;
			# fréquence motif
			nbmotif=$(egrep -coi "$motif" ./DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt;) ;
			# contexte en html
			perl ./minigrepmultilingue/minigrepmultilingue.pl "utf-8" ./DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt ./minigrepmultilingue/motif-regexp.txt;
				#je rénomme et bouge le fichier pour qu'il ne soit pas écrasé
			mv ./resultat-extraction.html ./CONTEXTES/$encodage-$compteur_fichier-$compteur_ligne.html ;
			# quels sont les mots qui apparaissent le plus?
			egrep -o "\w+" ./DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt | sort | uniq -c | sort -gr > ./DUMP-TEXT/index-$encodage-$compteur_fichier-$compteur_ligne.txt ;
			# bigrammes
			egrep -o "\w+" ./DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt > bi1.txt;
			tail -n +2 bi1.txt > bi2.txt ;
			paste bi1.txt bi2.txt > bi3.txt ;
			cat bi3.txt | sort | uniq -c | sort -gr >  ./DUMP-TEXT/bigramme-$encodage-$compteur_fichier-$compteur_ligne.txt ;
			echo "<tr>
			<td>$compteur_ligne</td>
			<td><a href=\"$ligne\" target=\"_blank\">$ligne</a></td>
			<td>$http_code</td>
			<td>$encodage</td>
			<td><a href=\"../PAGES-ASPIREES/$encodage-$compteur_fichier-$compteur_ligne.html\"> $encodage-$compteur_fichier-$compteur_ligne.html</a></td>
			<td><a href=\"../DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt\">$encodage-$compteur_fichier-$compteur_ligne.txt</a></td>
			<td><a href=\"../CONTEXTES/$encodage-$compteur_fichier-$compteur_ligne.txt\">$encodage-$compteur_fichier-$compteur_ligne.txt</a></td>
			<td><a href=\"../CONTEXTES/$encodage-$compteur_fichier-$compteur_ligne.html\">$encodage-$compteur_fichier-$compteur_ligne.html</a></td>
			<td><a href=\"../DUMP-TEXT/index-$encodage-$compteur_fichier-$compteur_ligne.txt\">index-$encodage-$compteur_fichier-$compteur_ligne.txt</a></td>
			<td><a href=\"../DUMP-TEXT/bigramme-$encodage-$compteur_fichier-$compteur_ligne.txt\">bigramme-$encodage-$compteur_fichier-$compteur_ligne.txt</a></td>
			<td>$nbmotif</td>
```

```
                </tr>">>"$2/tableau.html";
                # si l'encodage n'est pas UTF-8
                else
                retouriconv=$(iconv -l | egrep -io "\b$encodage\b")
                if [[ $retouriconv != "" ]]
                        then
                        # je fais le dump de la page
                        lynx -dump -nolist -assume_charset=$encodage -
display_charset=$encodage "./PAGES-ASPIREES/$encodage-$compteur_fichier-
$compteur_ligne.html" > ./DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne-
$encodage.txt;
                        # je la transcode grâce à iconv
                        iconv -f $encodage -t UTF-8 ./DUMP-TEXT/$encodage-
$compteur_fichier-$compteur_ligne-$encodage.txt > ./DUMP-TEXT/$encodage-$compteur_fichier-
$compteur_ligne.txt ;
                        # je fais tout le tableau
                        # ----------------------------------------------------------------------------------------
--------------------------------
                        # contexte de la page en txt
                        egrep -i -C2 "$motif" ./DUMP-TEXT/$encodage-$compteur_fichier-
$compteur_ligne.txt > ./CONTEXTES/$encodage-$compteur_fichier-$compteur_ligne.txt ;
                        # fréquence motif
                        nbmotif=$(egrep -coi "$motif" ./DUMP-TEXT/$encodage-
$compteur_fichier-$compteur_ligne.txt;) ;
                        # contexte en html
                        perl ./minigrepmultilingue/minigrepmultilingue.pl "utf-8" ./DUMP-
TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt ./minigrepmultilingue/motif-regexp.txt;
                                #je rénomme et bouge le fichier pour qu'il ne soit pas écrasé
                        mv ./resultat-extraction.html ./CONTEXTES/$encodage-
$compteur_fichier-$compteur_ligne.html ;
                        # quels sont les mots qui apparaissent le plus?
                        egrep -o "\w+" ./DUMP-TEXT/$encodage-$compteur_fichier-
$compteur_ligne.txt | sort | uniq -c | sort -gr > ./DUMP-TEXT/index-$encodage-$compteur_fichier-
$compteur_ligne.txt ;
                        # bigrammes
                        egrep -o "\w+" ./DUMP-TEXT/$encodage-$compteur_fichier-
$compteur_ligne.txt > bi1.txt;
                        tail -n +2 bi1.txt > bi2.txt ;
                        paste bi1.txt bi2.txt > bi3.txt ;
                        cat bi3.txt | sort | uniq -c | sort -gr >  ./DUMP-TEXT/bigramme-
$encodage-$compteur_fichier-$compteur_ligne.txt ;
                        # tout ce qui est écrit dans ce tableau doit avoir le chemin relatif par
rapport au tableau (TABLEAUX)
                        echo "
                        <tr>
                        <td>$compteur_ligne</td>
                        <td><a href=\"$ligne\" target=\"_blank\">$ligne</a></td>
                        <td>$http_code</td>
                        <td>$encodage</td>
                        <td><a href=\"../PAGES-ASPIREES/$encodage-$compteur_fichier-
$compteur_ligne.html\"> $encodage-$compteur_fichier-$compteur_ligne.html</a></td>
```

```
                                <td><a href=\"../DUMP-TEXT/$encodage-$compteur_fichier-
$compteur_ligne.txt\">$encodage-$compteur_fichier-$compteur_ligne.txt</a></td>
                                <td><a href=\"../CONTEXTES/$encodage-$compteur_fichier-
$compteur_ligne.txt\">$encodage-$compteur_fichier-$compteur_ligne.txt</a></td>
                                <td><a href=\"../CONTEXTES/$encodage-$compteur_fichier-
$compteur_ligne.html\">$encodage-$compteur_fichier-$compteur_ligne.html</a></td>
                                <td><a href=\"../DUMP-TEXT/index-$encodage-$compteur_fichier-
$compteur_ligne.txt\">index-$encodage-$compteur_fichier-$compteur_ligne.txt</a></td>
                                <td><a href=\"../DUMP-TEXT/bigramme-$encodage-
$compteur_fichier-$compteur_ligne.txt\">bigramme-$encodage-$compteur_fichier-
$compteur_ligne.txt</a></td>
                                <td>$nbmotif</td>
                                </tr>">>"$2/tableau.html";
                                else
                                # je cherche l'encodage grâce à egrep
                                encodage2=$(egrep -io "charset=.+\b" "./PAGES-
ASPIREES/$encodage-$compteur_fichier-$compteur_ligne.html"| cut -f2 -d "=" | sed -r "s/'|\"//g" |
tr "[a-z]" "[A-Z]" | tr -d "\r"  ) ;
                                echo "ENCODAGE DETECTE PAR EGREP : $encodage2";
                                retouriconv2=$(iconv -l | egrep -i "\b$encodage2\b") ;
                                if [[ $retouriconv2 != "" ]]
                                        then
                                        if [[ $encodage2 == "UTF-8" ]]
                                                then
                                                # dump de la page
                                                lynx -dump -nolist -assume_charset=$encodage -
display_charset=$encodage  "./PAGES-ASPIREES/$encodage-$compteur_fichier-
$compteur_ligne.html" > ./DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt ;
                                                # contexte de la page en txt
                                                egrep -i -C2 "$motif" ./DUMP-TEXT/$encodage-
$compteur_fichier-$compteur_ligne.txt > ./CONTEXTES/$encodage-$compteur_fichier-
$compteur_ligne.txt ;
                                                # fréquence motif
                                                nbmotif=$(egrep -coi "$motif" ./DUMP-
TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt;) ;
                                                # contexte en html
                                                perl ./minigrepmultilingue/minigrepmultilingue.pl
"utf-8" ./DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt
./minigrepmultilingue/motif-regexp.txt;
                                                        #je rénomme et bouge le fichier pour qu'il ne
soit pas écrasé
                                                mv ./resultat-extraction.html
./CONTEXTES/$encodage-$compteur_fichier-$compteur_ligne.html ;
                                                # quels sont les mots qui apparaissent le plus?
                                                egrep -o "\w+" ./DUMP-TEXT/$encodage-
$compteur_fichier-$compteur_ligne.txt | sort | uniq -c | sort -gr > ./DUMP-TEXT/index-$encodage-
$compteur_fichier-$compteur_ligne.txt ;
                                                # bigrammes
                                                egrep -o "\w+" ./DUMP-TEXT/$encodage-
$compteur_fichier-$compteur_ligne.txt > bi1.txt;
                                                tail -n +2 bi1.txt > bi2.txt ;
```

```
                                        paste bi1.txt bi2.txt > bi3.txt ;
                                        cat bi3.txt | sort | uniq -c | sort -gr >  ./DUMP-
TEXT/bigramme-$encodage-$compteur_fichier-$compteur_ligne.txt ;
                                        echo "<tr>
                                        <td>$compteur_ligne</td>
                                        <td><a href=\"$ligne\"
target=\"_blank\">$ligne</a></td>
                                        <td>$http_code</td>
                                        <td>$encodage2</td>
                                        <td><a href=\"../PAGES-ASPIREES/$encodage-
$compteur_fichier-$compteur_ligne.html\"> $encodage-$compteur_fichier-
$compteur_ligne.html</a></td>
                                        <td><a href=\"../DUMP-TEXT/$encodage-
$compteur_fichier-$compteur_ligne.txt\">$encodage-$compteur_fichier-
$compteur_ligne.txt</a></td>
                                        <td><a href=\"../CONTEXTES/$encodage-
$compteur_fichier-$compteur_ligne.txt\">$encodage-$compteur_fichier-
$compteur_ligne.txt</a></td>
                                        <td><a href=\"../CONTEXTES/$encodage-
$compteur_fichier-$compteur_ligne.html\">$encodage-$compteur_fichier-
$compteur_ligne.html</a></td>
                                        <td><a href=\"../DUMP-TEXT/index-
$compteur_fichier-$compteur_ligne.txt\">index-$encodage-$compteur_fichier-
$compteur_ligne.txt</a></td>
                                        <td><a href=\"../DUMP-TEXT/bigramme-$encodage-
$compteur_fichier-$compteur_ligne.txt\">bigramme-$encodage-$compteur_fichier-
$compteur_ligne.txt</a></td>
                                        <td>$nbmotif</td>
                                        </tr>">>"$2/tableau.html";
                                fi
                                # je dump
                                lynx -dump -nolist -assume_charset=$encodage2 -
display_charset=$encodage2 "./PAGES-ASPIREES/$encodage-$compteur_fichier-
$compteur_ligne.html" > ./DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne-
$encodage2.txt;
                                # je la transcode grâce à iconv
                                iconv -f $encodage2 -t UTF-8 ./DUMP-TEXT/$encodage-
$compteur_fichier-$compteur_ligne-$encodage2.txt > ./DUMP-TEXT/$encodage-$compteur_fichier-
$compteur_ligne.txt ;
                                # je fais tout le tableau
                                # ----------------------------------------
                                # contexte de la page en txt
                                egrep -i -C2 "$motif" ./DUMP-TEXT/$encodage-
$compteur_fichier-$compteur_ligne.txt > ./CONTEXTES/$encodage-$compteur_fichier-
$compteur_ligne.txt ;
                                # fréquence motif
                                nbmotif=$(egrep -coi "$motif" ./DUMP-TEXT/$encodage-
$compteur_fichier-$compteur_ligne.txt;) ;
                                # contexte en html
```

```
                                        perl ./minigrepmultilingue/minigrepmultilingue.pl "utf-8" ./DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt ./minigrepmultilingue/motif-regexp.txt;
                                                #je rénomme et bouge le fichier pour qu'il ne soit pas écrasé
                                        mv ./resultat-extraction.html ./CONTEXTES/$encodage-$compteur_fichier-$compteur_ligne.html ;
                                        # quels sont les mots qui apparaissent le plus?
                                        egrep -o "\w+" ./DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt | sort | uniq -c | sort -gr > ./DUMP-TEXT/index-$encodage-$compteur_fichier-$compteur_ligne.txt ;
                                        # bigrammes
                                        egrep -o "\w+" ./DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt > bi1.txt;
                                        tail -n +2 bi1.txt > bi2.txt ;
                                        paste bi1.txt bi2.txt > bi3.txt ;
                                        cat bi3.txt | sort | uniq -c | sort -gr >  ./DUMP-TEXT/bigramme-$encodage-$compteur_fichier-$compteur_ligne.txt ;
                                        echo "<tr>
                                        <td>$compteur_ligne</td>
                                        <td><a href=\"$ligne\" target=\"_blank\">$ligne</a></td>
                                        <td>$http_code</td>
                                        <td>$encodage2</td>
                                        <td><a href=\"../PAGES-ASPIREES/$encodage-$compteur_fichier-$compteur_ligne.html\"> $encodage-$compteur_fichier-$compteur_ligne.html</a></td>
                                        <td><a href=\"../DUMP-TEXT/$encodage-$compteur_fichier-$compteur_ligne.txt\"> $encodage-$compteur_fichier-$compteur_ligne.txt</a></td>
                                        <td><a href=\"../CONTEXTES/$encodage-$compteur_fichier-$compteur_ligne.txt\"> $encodage-$compteur_fichier-$compteur_ligne.txt</a></td>
                                        <td><a href=\"../CONTEXTES/$encodage-$compteur_fichier-$compteur_ligne.html\"> $encodage-$compteur_fichier-$compteur_ligne.html</a></td>
                                        <td><a href=\"../DUMP-TEXT/index-$encodage-$compteur_fichier-$compteur_ligne.txt\">index-$encodage-$compteur_fichier-$compteur_ligne.txt</a></td>
                                        <td><a href=\"../DUMP-TEXT/bigramme-$encodage-$compteur_fichier-$compteur_ligne.txt\">bigramme-$encodage-$compteur_fichier-$compteur_ligne.txt</a></td>
                                        <td>$nbmotif</td>
                                        </tr>">>"$2/tableau.html";
                                fi
#-----------------------------------
                        fi
                        echo "<tr>
                        <td>$compteur_ligne</td>
                        <td><a href=\"$ligne\" target=\"_blank\">$ligne</a></td>
                        <td>$http_code</td>
                        <td>$encodage</td><td><a href=\"../PAGES-ASPIREES/$encodage-$compteur_fichier-$compteur_ligne\">$encodage-$compteur_fichier-$compteur_ligne</a></td>
                        <td>-</td>
                        <td>-</td>
```

```
                        <td>-</td>
                        <td>-</td>
                        <td>-</td>
                        <td>-</td>
                        </tr>" >> "$2/tableau.html";
                        fi
                else
                echo "<tr><td>$compteur_ligne</td><td><a href=\"$ligne\"
target=\"_blank\">$ligne</a></td>" >>"$2/tableau.html";
                #           code 200/autre                      cellule vide
                echo "<td>$http_code</td><td>-</td>" >>"$2/tableau.html";
                #               cellule vide
                echo "<td>-</td></tr>">>"$2/tableau.html";
                fi
                compteur_ligne=$((compteur_ligne+1));
                done
# Une fois la table terminée, on incremente le compteur_fichier, car nouveau fichier --> nouvelle
table
# on refait la boucle depuis for fichier in $(ls $1)
compteur_fichier=$((compteur_fichier+1));
        done
echo "</body></html>" >> "$2/tableau.html"
```