```
#!/bin/bash
rm -f "$2/tableau.html";  #effacer le résultat des essais précédents
rm -rf ./CORPUS-SEG;
mkdir ./CORPUS-SEG; #créer le répertoire pour les corpus chinois segmentés
#les trois "écho" suivants sont pour clarifier où se trouve les listes d'URLs et le fichier HTML, ainsi que le motif cherché
echo "Les URLs sont dans le répertoire : $1";
echo "Le tableau HTML créé est dans le répertoire : $2";
echo "Le motif cherché est : $3";
#créer un variable pour le motif, ainsi qu'un fichier texte pour l'utilisation de minigrep plus tard
motif=$3;
echo "MOTIF=$3" > ./minigrep/para-motif.txt;
#---------------commencer à construire le fichier html-----------------------------
echo "<!DOCTYPE HTML><html><head><meta charset=\"UTF-8\"/><title>TABLEAU D'URLs</title></head><body>" >> "$2/tableau.html";  #construire le début de html selon la structure
numerotableau=1 #pour numéroter les tableaux et compter les fichiers d'URLs traités
for fichier in $(ls $1) #Maintenant, on va parcourir tous les fichiers dans ce répertoire
do
echo "$1/$fichier";  #Montrer les fichiers utilisés dans le terminal bash
compteur=1; #dans chaque fichier, compter les lignes/URLs
#----------créer le tableau-----------------------------------
echo "<table border=\"2\" align=\"center\" width=\"80%\"><caption>Tableau $numerotableau</caption>" >> "$2/tableau.html";
echo "<tr bgcolor=\"grey\">
<th>N°</th>
<th>URL</th>
<th>Code http</th>
<th>Encodage</th>
<th>Page aspirée</th>
<th>Dump</th>
<th>Filtrage txt</th>
<th>Filtrage html</th>
<th>Index</th>
<th>Bitexte</th>
<th>Fréquence motif</th></tr>" >> "$2/tableau.html"
#créer la première ligne du tableau pour donner le nom de chaque colonne
#----------Maitenant, on entre dans le fichier, et on lit ligne par ligne----------
for ligne in $(cat "$1/$fichier")
        do
        echo "-----------------------------------------------";
        echo "Traitement de l'URL : $ligne";
        echo "-----------------------------------------------";
        coderetourhttp=$(curl -SIL -o tmp.txt -w %{http_code} $ligne); #récupérer le code retour de http pour vérifier si la connextion se passe bien
        if [[ $coderetourhttp == 200 ]] #qui signifie tout se passe bien
                then
                curl -L $ligne > ./PAGES-ASPIREES/$numerotableau-$compteur.html; #aspirer la page en format HTML
                encodage1=$(curl -SIL -o tmp1.txt -w %{content_type} $ligne | cut -d '=' -f2 | tr "[a-z]" "[A-Z]" | tr -d "\r");  #récupérer l'encodage de cette page
                #-----------loop sert à trouver le vrai encodage---------
                if [[ $encodage1 == "UTF-8" ]]
                        then encodage=$encodage1
                        else
                        encodage=$(egrep -oi "charset=\"?[^\"]+\"" ./PAGES-ASPIREES/$numerotableau-$compteur.html|cut -d"=" -f2 |tr -d "\""| tr "[a-z]" "[A-Z]"| head -n 1);
                        #parce que pour certaine page, l'encodage n'apparaît pas dans le content_type, le précédent récupère TEXT/HTML même si le page est en UTF-8, donc pour elles, je cherche le vrai encodage dans le code des pages aspirées; en plus, pour certaines pages, il existe plusieurs "charset" dans le script, donc j'utilise head pour prendre seulement le premier.
```

```
                        fi
                #----------------------------------------------------------
                if [[ $encodage == "UTF-8" ]] #pour vérifier l'encodage de cette page, si en UTF-8,
on peut obtenir le texte dump directement par page aspirée avec lynx
                        then
                        #obtenir le texte dump par les pages aspirées
                        lynx -dump -nolist -assume_charset=$encodage
-display_charset=$encodage "./PAGES-ASPIREES/$numerotableau-$compteur.html" > ./DUMP-
TEXT/$numerotableau-$compteur.txt;
                        #créer le fichier txt comporte le contexte (lignes avant et après) de motif
(egrep)
                        egrep -i -C2 "$motif" ./DUMP-TEXT/$numerotableau-$compteur.txt > ./
CONTEXTES/$numerotableau-$compteur.txt;
                        #compter le nombre de motif
                        nombre_motif=$(egrep -coi $motif ./DUMP-TEXT/$numerotableau-
$compteur.txt);
                        #faire le page html qui montre le contexte en utilisant minigrep
                        perl ./minigrep/minigrepmultilingue.pl "utf-8" ./DUMP-TEXT/
$numerotableau-$compteur.txt ./minigrep/para-motif.txt;
                        #renommer et mettre le  résultat dans le répertoire correct
                        mv ./resultat-extraction.html ./CONTEXTES/$numerotableau-
$compteur.html;
                        if [[ $fichier == "URL-chinois.txt" ]]
                                then
                                #--------pour segmenter le corpus chinois---------
                                python3 ./PROGRAMMES/seg.py ./DUMP-TEXT/$numerotableau-
$compteur.txt;
                                #renommer et réorienter le fichier crée dans le bon répertoire
                                mv ./fichier_segmente.txt ./CORPUS-SEG/$numerotableau-
$compteur.txt;
                                #-------------------------------------------------
                                #pour faire l'index hierarchique
                                egrep -o "\w+" ./CORPUS-SEG/$numerotableau-$compteur.txt |
sort | uniq -c| sort -r > ./DUMP-TEXT/index-$numerotableau-$compteur.txt;
                                #pour faire le bigramme
                                egrep -o "\w+" ./CORPUS-SEG/$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-
$numerotableau-$compteur.txt
                                #---------maintenant, on traite les URL anglais/...----------------------
                                else
                                #pour faire l'intex hierarchique
                                egrep -o "\w+" ./DUMP-TEXT/$numerotableau-$compteur.txt | sort |
uniq -c| sort -r > ./DUMP-TEXT/index-$numerotableau-$compteur.txt;
                                #pour faire le bigramme
                                egrep -o "\w+" ./DUMP-TEXT/$numerotableau-$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-
$numerotableau-$compteur.txt
                        fi
                        #remplir le tableau avec toutes les infos de ce URL
                        echo "<tr>
                        <td>$compteur</td>
                        <td><a href=\"$ligne\" target=\"_blank\">$ligne</a></td>
                        <td>$coderetourhttp</td>
                        <td>$encodage</td>
```

```
                        <td><a href=\"../PAGES-ASPIREES/$numerotableau-
$compteur.html\">$numerotableau-$compteur.html</a></td>
                        <td><a href=\"../DUMP-TEXT/$numerotableau-
$compteur.txt\">$numerotableau-$compteur.txt</a></td>
                        <td><a href=\"../CONTEXTES/$numerotableau-
$compteur.txt\">$numerotableau-$compteur.txt</a></td>
                        <td><a href=\"../CONTEXTES/$numerotableau-
$compteur.html\">$numerotableau-$compteur.html</a></td>
                        <td><a href=\"../DUMP-TEXT/index-$numerotableau-
$compteur.txt\">index-$numerotableau-$compteur.txt</a></td>
                        <td><a href=\"../DUMP-TEXT/bigramme-$numerotableau-
$compteur.txt\">bigramme-$numerotableau-$compteur.txt</a></td>
                        <td>$nombre_motif</td>
                        </tr>" >> "$2/tableau.html";
                        else #le vrai encodage n'est pas UTF-8
                        #pour tester si encodage est connu par iconv
                        testeur=$(iconv -l | egrep -i $encodage)
                        if [[ $testeur != "" ]] #si l'encodage de la page est connu par iconv
                                then
                                #obtenir le texte dump en encodage non UTF-8
                                lynx -dump -nolist -assume_charset=$encodage
-display_charset=$encodage "./PAGES-ASPIREES/$numerotableau-$compteur.html" > ./DUMP-
TEXT/$numerotableau-$compteur-$encodage.txt;
                                #convertir son encodage à UTF-8 avec iconv
                                #iconv ./DUMP-TEXT/$numerotableau-$compteur-$encodage.txt -f
$encodage -t UTF-8 -o ./DUMP-TEXT/$numerotableau-$compteur.txt;########c'est faux
                                iconv -c -f $encodage -t UTF-8  ./DUMP-TEXT/$numerotableau-
$compteur-$encodage.txt  > ./DUMP-TEXT/$numerotableau-$compteur.txt;
                                #traitement sur le motif (contexte, fréquence)
                                egrep -i -C2 "$motif" ./DUMP-TEXT/$numerotableau-$compteur.txt
> ./CONTEXTES/$numerotableau-$compteur.txt;
                                nombre_motif=$(egrep -coi $motif ./DUMP-TEXT/$numerotableau-
$compteur.txt);
                                #contexte html
                                perl ./minigrep/minigrepmultilingue.pl "utf-8" ./DUMP-TEXT/
$numerotableau-$compteur.txt ./minigrep/para-motif.txt;
                                mv ./resultat-extraction.html ./CONTEXTES/$numerotableau-
$compteur.html;
                                if [[ $fichier == "URL-chinois.txt" ]]
                                        then
                                        #--------pour segmenter le corpus chinois---------
                                        python3 ./PROGRAMMES/seg.py ./DUMP-TEXT/
$numerotableau-$compteur.txt;
                                        #renommer et réorienter le fichier crée dans le bon répertoire
                                        mv ./fichier_segmente.txt ./CORPUS-SEG/$numerotableau-
$compteur.txt;
                                        #-------------------------------------------------
                                        #pour faire l'index hierarchique
                                        egrep -o "\w+" ./CORPUS-SEG/$numerotableau-
$compteur.txt | sort | uniq -c| sort -r > ./DUMP-TEXT/index-$numerotableau-$compteur.txt;
                                        #pour faire le bigramme
                                        egrep -o "\w+" ./CORPUS-SEG/$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-
$numerotableau-$compteur.txt
                                        #---------maintenant, on traite les URL
anglais/...----------------------
                                        else
```

```
                                        #pour faire l'intex hierarchique
                                        egrep -o "\w+" ./DUMP-TEXT/$numerotableau-$compteur.txt
| sort | uniq -c| sort -r > ./DUMP-TEXT/index-$numerotableau-$compteur.txt;
                                        #pour faire le bigramme
                                        egrep -o "\w+" ./DUMP-TEXT/$numerotableau-$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-
$numerotableau-$compteur.txt
                                fi
                                echo "<tr>
                                <td>$compteur</td>
                                <td><a href=\"$ligne\" target=\"_blank\">$ligne</a></td>
                                <td>$coderetourhttp</td>
                                <td>$encodage</td>
                                <td><a href=\"../PAGES-ASPIREES/$numerotableau-
$compteur.html\">$numerotableau-$compteur.html</a></td>
                                <td><a href=\"../DUMP-TEXT/$numerotableau-
$compteur.txt\">$numerotableau-$compteur.txt</a></td>
                                <td><a href=\"../CONTEXTES/$numerotableau-
$compteur.txt\">$numerotableau-$compteur.txt</a></td>
                                <td><a href=\"../CONTEXTES/$numerotableau-
$compteur.html\">$numerotableau-$compteur.html</a></td>
                                <td><a href=\"../DUMP-TEXT/index-$numerotableau-
$compteur.txt\">index-$numerotableau-$compteur.txt</a></td>
                                <td><a href=\"../DUMP-TEXT/bigramme-$numerotableau-
$compteur.txt\">bigramme-$numerotableau-$compteur.txt</a></td>
                                <td>$nombre_motif</td>
                                </tr>" >> "$2/tableau.html";
                                else #si l'encodage est inconnu par iconv
                                ##########selon le prof, mais pourquoi?
                                ## ici il faut aller extraire le charset dans la page HTML et s'assurer
que c'est un charset connu de iconv aussi
                                echo "<tr>
                                <td>$compteur</td>
                                <td><a href=\"$ligne\">$ligne</a></td>
                                <td>$coderetourhttp</td>
                                <td>$encodage</td>
                                <td><a href=\"../PAGES-ASPIREES/$numerotableau-
$compteur.html\">$numerotableau-$compteur.html</a></td>
                                <td>encodage inconnu</td>
                                <td>-</td>
                                <td>-</td>
                                <td>-</td>
                                <td>-</td>
                                <td>-</td>
                                </tr>" >> "$2/tableau.html";
                                fi
                        #-------------------------------------------------------
                        fi
                else #le code retour n'est pas 200, donc on ne peut rien faire, on laisse les cases
vides dans cette ligne
                echo "<tr>
                <td>$compteur</td>
                <td><a href=\"$ligne\">$ligne</a></td>
                <td>$coderetourhttp</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));  #N'oubliez pas parenthèses deux fois
        done
echo "</table><br />" >> "$2/tableau.html" #fermer le tableau et faire le saut de ligne
numerotableau=$((numerotableau+1))
done
echo "</body></html>" >> "$2/tableau.html"; #finir le fichier html
```