```bash
#!/bin/bash
# 1. Lecture des paramètres dans le fichier PARAMETRES
read DOSSIERURLS;read fichier_tableau;read motif; #ici ont lit nos paramètres qui contiennent: le chemin vers 3 fichiers, cad le dossier avec les fichiers contenant les urls, le fichier qui contient le tableau html (on le crée s'il n'existe pas) dans le dossier TABLEAU et le motif qu'on recherche (le mot à analyser)
echo "Le dossier d'URLs : $DOSSIERURLS " ; #on commente les actions effectuées par le programme:
echo "Le fichier contenant le tableau : $fichier_tableau" ;
echo "Le motif est : $motif" ;
# on commence la création de notre tableau
compteur_tableau=1; #on crée un compteur pour chaque tableau (un par langue)
echo "<html><head><style>
a{color:red}            /*les liens en général*/
a:hover{color:black}    /*quand on passe la souris sur un lien*/
a:active{color:black}   /*quand on clique sur le lien*/
a:visited{color:grey;}  /*quand le lien a déja été visité*/
</style> <meta charset="UTF-8"></head><body>" > $fichier_tableau ; #on crée le début de la page html en écrivant du code html puis on l'insère dans notre document .html
for fichier in $(ls $DOSSIERURLS)  #on peut également l'écrire: for fichier in `ls $DOSSIERURLS`
{
    compteur=1; # on crée un compteur pour compter les URLs
    echo "<table  bordercolor=\"red\" align=\"center\" border=\"5\">">> $fichier_tableau ; #là on fait le bord de notre tableau, on modifie un peu son apparence avec des codes html
    echo "<tr><td colspan=\"11\" align=\"center\"> <p style=\"color:red;\" > Tableau n° $compteur_tableau</p></td></tr>" >> $fichier_tableau ; #on crée notre première ligne, et on indique qu'elle recouvra 11 colonne (cela sert simplement à améliorer l'apparence de notre tableau #tr c'est pour une ligne et td pour une cellule
    echo "<tr>
    <td align=\"center\"><b>N°</b></td>
    <td align=\"center\"><b>Lien</b></td>
    <td align=\"center\"><b>CODE CURL</b>
    <td align=\"center\"><b>Statut CURL</b></td>
    <td align=\"center\"><b>Page Aspirée</b></td>
    <td align=\"center\"><b>Encodage Initial</b></td>
    <td align=\"center\"><b>DUMP initial</b></td>
    <td align=\"center\"><b>DUMP UTF-8</b></td>
    <td align=\"center\"><b>CONTEXTE UTF-8</b></td>
    <td align=\"center\"><b>CONTEXTE HTML UTF-8</b></td>
    <td align=\"center\"><b>Fréquence du motif</b></td>
    </tr>" >> $fichier_tableau ; #on vient de créer la structure de notre tableau en html
        # on va maintenant procéder au traitement de chacune des URLs
    for line in `cat $DOSSIERURLS/$fichier` # ou: for line in $(cat $DOSSIERURLS/fichier) #le cat sert à afficher
    {
        # ==> ASPIRATION DE LA PAGE
    echo "TELECHARGEMENT de $line vers ./PAGES-ASPIREES/$compteur_tableau-$compteur.html" ;
            # 1. RECUPERATION DU HEADER HTTP
    status1=$(curl -sI $line | head -n 1);
            # 2. RECUPERATION DU CODE RETOUR HTTP ET DE LA PAGE
    status2=$(curl --silent --output ./PAGES-ASPIREES/"$compteur_tableau-$compteur".html --write-out "%{http_code}" $line);
    echo "STATUT CURL : $status2" ;

    # on va détecter l'encodage de nos fichiers
    echo "DETECTION encodage de $line ";
    encodage=$(curl -sI $line | egrep -i "charset=" | cut -f2 -d= | tr -d "\n" | tr -d "\r" | tr "[:upper:]" "[:lower:]");
    echo "ENCODAGE $line : <$encodage>" ;
    if [[ $encodage == "utf-8" ]]
    then
        echo "DUMP de $line via lynx" ;
        lynx -dump -nolist -assume_charset=$encodage -display_charset=$encodage $line > ./DUMP-TEXT/$compteur_tableau-$compteur.txt ;

        # ajouter ici l'extraction de contexte autour des mots choisis

```

```
55          egrep -i $motif ./DUMP-TEXT/$compteur_tableau-$compteur.txt > ./CONTEXTES/
            $compteur_tableau-$compteur.txt ;
56          nombre_motif=$(egrep -coi $motif ./DUMP-TEXT/$compteur_tableau-$compteur.txt
            | wc -w);
57          perl ./PROGRAMMES/minigrepmultilingue-v2.2-regexp/minigrepmultilingue.pl
            "UTF-8" ./DUMP-TEXT/$compteur_tableau-$compteur.txt ./PROGRAMMES/
            minigrepmultilingue-v2.2-regexp/motif.txt ;
58          mv resultat-extraction.html ./CONTEXTES/$compteur_tableau-$compteur.html ;
59          echo "ECRITURE RESULTAT dans le tableau" ;
60          echo "<tr>
61          <td align=\"center\">$compteur</td>
62          <td align=\"center\"><a href=\"$line\">lien n°$compteur</a></td>
63          <td align=\"center\">$status2</td>
64          <td align=\"center\">$status1</td>
65          <td align=\"center\"><a
            href=\"../PAGES-ASPIREES/$compteur_tableau-$compteur.html\">P.A n°
            $compteur_tableau-$compteur</a></td>
66          <td align=\"center\">$encodage</td>
67          <td align=\"center\">&empty;</td>
68          <td align=\"center\"><a
            href=\"../DUMP-TEXT/$compteur_tableau-$compteur.txt\">DUMP n°
            $compteur_tableau-$compteur</a></td>
69          <td align=\"center\"><a
            href=\"../CONTEXTES/$compteur_tableau-$compteur.txt\">CONTEXTE n°
            $compteur_tableau-$compteur</a></td>
70          <td align=\"center\"><a
            href=\"../CONTEXTES/$compteur_tableau-$compteur.html\">CONTEXTE n°
            $compteur_tableau-$compteur</a></td>
71          <td align=\"center\">$nombre_motif</td>
72          </tr>" >> $fichier_tableau ;
73      else
74          if [[ $encodage != "" ]]
75              then
76                  enco_iconv=$(iconv -l |  egrep -o "[-A-Z0-9\_\:]+" |egrep -i
                    $encodage) ;
77                  # ici il faut s'assurer que l'encodage est bien connu de iconv !!!!
78                  if [[ $enco_iconv == "" ]]
79                      then
80                          # ici on ne fait rien donc on écrit une ligne avec le
                            symbole 'vide'
81                          echo "<tr>
82                          <td align=\"center\">$compteur</td>
83                          <td align=\"center\"><a href=\"$line\">lien
                            n°$compteur</a></td>
84                          <td align=\"center\">$status2</td>
85                          <td>$status1</td>
86                          <td align=\"center\"><a
                            href=\"../PAGES-ASPIREES/$compteur_tableau-$compteur.html\">PA
                             n° $compteur_tableau-$compteur</a></td>
87                          <td align=\"center\">$encodage<br/>via curl<br/>inconnu de
                            iconv</td><td align=\"center\">-</td>
88                          <td align=\"center\">&empty;</td>
89                          <td align=\"center\">&empty;</td>
90                          <td align=\"center\">&empty;</td>
91                          <td align=\"center\">&empty;</td>
92                          </tr>" >> $fichier_tableau ;
93                      else
94                          echo "DUMP (via $encodage) de $line via lynx" ;
95                          lynx -dump -nolist -assume_charset=$encodage -display_charset
                            =$encodage $line > ./DUMP-TEXT/$compteur_tableau-$compteur.
                            txt ; #on extrait l'encodage
96                          iconv -f $encodage -t utf-8 ./DUMP-TEXT/$compteur_tableau-
                            $compteur.txt > ./DUMP-TEXT/$compteur_tableau-$compteur-
                            utf8.txt ; #on change l'encodage de celui dans lequel il
                            était (-from) vers de l'utf-8 (-to)
97                          egrep -i $motif ./DUMP-TEXT/$compteur_tableau-$compteur-
                            utf8.txt > ./CONTEXTES/$compteur_tableau-$compteur.txt ;
98                          nombre_motif=$(egrep -coi $motif
                            ./DUMP-TEXT/$compteur_tableau-$compteur-utf8.txt);
99                          perl ./PROGRAMMES/minigrepmultilingue-v2.2-regexp/
                            minigrepmultilingue.pl "utf-8" ./DUMP-TEXT/$compteur_tableau-
                            $compteur-utf8.txt ./PROGRAMMES/
                            minigrepmultilingue-v2.2-regexp/motif.txt ; # utilise minigrep
```

```
100                         mv resultat-extraction.html ./CONTEXTES/$compteur_tableau-
                            $compteur.html ;
101                         echo "ECRITURE RESULTAT dans le tableau" ;
102                         echo "<tr>
103                         <td align=\"center\">$compteur</td>
104                         <td align=\"center\"><a href=\"$line\">lien
                            n°$compteur</a></td>
105                         <td align=\"center\">$status2</td>
106                         <td>$status1</td>
107                         <td align=\"center\"><a
                            href=\"../PAGES-ASPIREES/$compteur_tableau-$compteur.html\">PA
                             n° $compteur_tableau-$compteur</a></td>
108                         <td align=\"center\">$encodage<br/>via curl</td><td
                            align=\"center\"><a
                            href=\"../DUMP-TEXT/$compteur_tableau-$compteur.txt\">DUMP
                            n° $compteur_tableau-$compteur</a></td>
109                         <td align=\"center\"><a
                            href=\"../DUMP-TEXT/$compteur_tableau-$compteur-utf8.txt\">DUM
                            P n° $compteur_tableau-$compteur</a></td>
110                         <td align=\"center\"><a
                            href=\"../CONTEXTES/$compteur_tableau-$compteur.txt\">CONTEXTE
                             n° $compteur_tableau-$compteur</a></td>
111                         <td align=\"center\"><a
                            href=\"../CONTEXTES/$compteur_tableau-$compteur.html\">CONTEXT
                            E n° $compteur_tableau-$compteur</a></td>
112                         <td align=\"center\">$nombre_motif</td></tr>" >>
                            $fichier_tableau ;
113                     fi    #sert à fermer le if
114                 else
115                     cherche_charset=$(egrep -i -o "meta(.*)?charset"
                        ./PAGES-ASPIREES/"$compteur_tableau-$compteur".html);#on cherche
                        l'encodage(le charset) dans les pages aspirées. Comme leur nom est
                        composé du numero du compteur du tableau pouis du compteur de lignes
                        on met cette variable pour automatiser la recherche pour toutes nos
                        pages.
116                     if [[ $cherche_charset != "" ]] #si on ne trouve rien
117                         then #alors
118                             encodage=$(egrep -i -o
                                "meta(.*)charset[^=]*?=[^\"]*?\"?[^\"]+?\""
                                ./PAGES-ASPIREES/$compteur_tableau-$compteur.html | egrep -i
                                -o "charset[^=]*?= *?\"?[^\"]+?\"" | cut -f2 -d= | sed
                                "s/\"//g" | sed "s/>//g" | sed "s/ //g" | sed "s/\///g" |
                                sort -u | tr [A-Z] [a-z] ); #on crée une variable qui
                                s'appelle encodage dans laquelle on
119                             echo "ENCODAGE EXTRAIT DE LA PAGE ASPIREE : $encodage" ;
120                             if [[ $encodage == "utf-8" ]] #si l'encodage est en utf-8
121                                 then #alors
122                                     echo "DUMP de $line via lynx" ;
123                                     lynx -dump -nolist -assume_charset=$encodage -
                                        display_charset=$encodage $line > ./DUMP-TEXT/
                                        $compteur_tableau-$compteur.txt ; #
124                                     egrep -i $motif ./DUMP-TEXT/$compteur_tableau-
                                        $compteur.txt > ./CONTEXTES/$compteur_tableau-
                                        $compteur.txt ;
125                                     nombre_motif=$(egrep -coi $motif
                                        ./DUMP-TEXT/$compteur_tableau-$compteur.txt);
126                                     perl ./PROGRAMMES/minigrepmultilingue-v2.2-regexp/
                                        minigrepmultilingue.pl "utf-8" ./DUMP-TEXT/
                                        $compteur_tableau-$compteur.txt ./PROGRAMMES/
                                        minigrepmultilingue-v2.2-regexp/motif.txt ;#ne pas
                                        oublier de mettre le chemin de motif.txt ! Sinon ça
                                        ne marche pas
127                                     mv resultat-extraction.html ./CONTEXTES/
                                        $compteur_tableau-$compteur.html ;
128                                     echo "ECRITURE RESULTAT dans le tableau" ;
129                                     echo "<tr>
130                                     <td align=\"center\">$compteur</td>
131                                     <td align=\"center\"><a href=\"$line\">lien
                                        n°$compteur</a></td>
132                                     <td align=\"center\">$status2</td>
133                                     <td align=\"center\">$status1</td>
134                                     <td align=\"center\"><a
                                        href=\"../PAGES-ASPIREES/$compteur_tableau-$compteur.h
```

```
tml\">P.A n° $compteur_tableau-$compteur</a></td>
<td align=\"center\">$encodage<br/>avec charset</td><td align=\"center\">&empty;</td>
<td align=\"center\"><a href=\"../DUMP-TEXT/$compteur_tableau-$compteur.txt\">DUMP n° $compteur_tableau-$compteur</a></td>
<td align=\"center\"><a href=\"../CONTEXTES/$compteur_tableau-$compteur.txt\">CONTEXTE n° $compteur_tableau-$compteur</a></td>
<td align=\"center\"><a href=\"../CONTEXTES/$compteur_tableau-$compteur.html\">CONTEXTE n° $compteur_tableau-$compteur</a></td>
<td align=\"center\">$nombre_motif</td>
</tr>" >> $fichier_tableau ;
else #sinon
    enco_iconv=$(iconv -l |  egrep -o "[-A-Z0-9\_\:]+" | egrep -i $encodage) ;
    if [[ $enco_iconv == "" ]]
        then
            echo "<tr>
            <td align=\"center\">$compteur</td>
            <td align=\"center\"><a href=\"$line\">lien n°$compteur</a></td>
            <td align=\"center\">$status2</td><td>$status1</td>
            <td align=\"center\"><a href=\"../PAGES-ASPIREES/$compteur_tableau-$compteur.html\">PA n° $compteur_tableau-$compteur</a></td>
            <td align=\"center\">$encodage<br/><br/>avec charset<br/>inconnu de iconv</td>
            <td align=\"center\"><a href=\"../DUMP-TEXT/$compteur_tableau-$compteur.txt\">DUMP n° $compteur_tableau-$compteur</a></td>
            <td align=\"center\">&empty;</td>
            <td align=\"center\">&empty;</td>
            <td align=\"center\">&empty;</td>
            <td align=\"center\">&empty;</td>
            </tr>" >> $fichier_tableau ;
        else
            lynx -dump -nolist -assume_charset=$encodage -display_charset=$encodage $line > ./DUMP-TEXT/$compteur_tableau-$compteur.txt ;
            iconv -f $encodage -t utf-8 ./DUMP-TEXT/$compteur_tableau-$compteur.txt > ./DUMP-TEXT/$compteur_tableau-$compteur-utf8.txt #on converti l'encodage du fichier en utf-8
            egrep -i $motif ./DUMP-TEXT/$compteur_tableau-$compteur-utf8.txt > ./CONTEXTES/$compteur_tableau-$compteur.txt ;
            nombre_motif=$(egrep -coi $motif ./DUMP-TEXT/$compteur_tableau-$compteur-utf8.txt);
            perl ./PROGRAMMES/minigrepmultilingue-v2.2-regexp/minigrepmultilingue.pl "utf-8" ./DUMP-TEXT/$compteur_tableau-$compteur-utf8.txt ./PROGRAMMES/minigrepmultilingue-v2.2-regexp/motif.txt;
            mv resultat-extraction.html ./CONTEXTES/$compteur_tableau-$compteur.html ;
            echo "ECRITURE RESULTAT dans le tableau" ;
            echo "<tr>
            <td align=\"center\">$compteur</td><td align=\"center\"><a href=\"$line\">lien n°$compteur</a></td>
            <td align=\"center\">$status2</td><td>$status1</td>
            <td align=\"center\"><a href=\"../PAGES-ASPIREES/$compteur_tableau-$co
```

```
                                        mpteur.html\">PA n°
                                        $compteur_tableau-$compteur</a></td>
169                                     <td align=\"center\">$encodage<br/>avec
                                        charset</td>
170                                     <td align=\"center\"><a
                                        href=\"../DUMP-TEXT/$compteur_tableau-$compteu
                                        r.txt\">DUMP n°
                                        $compteur_tableau-$compteur</a></td>
171                                     <td align=\"center\"><a
                                        href=\"../DUMP-TEXT/$compteur_tableau-$compteu
                                        r-utf8.txt\">DUMP n°
                                        $compteur_tableau-$compteur</a></td>
172                                     <td><a
                                        href=\"../CONTEXTES/$compteur_tableau-$compteu
                                        r.txt\">CONTEXTE n°
                                        $compteur_tableau-$compteur</a></td>
173                                     <td align=\"center\"><a
                                        href=\"../CONTEXTES/$compteur_tableau-$compteu
                                        r.html\">CONTEXTE n°
                                        $compteur_tableau-$compteur</a></td>
174                                     <td align=\"center\">$nombre_motif</td>
175                                     </tr>" >> $fichier_tableau ;
176                                 fi
177                             fi
178                 else #sinon
179                     echo "<tr>
180                     <td align=\"center\">$compteur</td>
181                     <td align=\"center\"><a href=\"$line\">lien n°$compteur</a></td>
182                     <td align=\"center\">$status2</td>
183                     <td align=\"center\">$status1</td>
184                     <td align=\"center\"><a
                        href=\"../PAGES-ASPIREES/$compteur_tableau-$compteur.html\">PA
                        n° $compteur_tableau-$compteur</a></td>
185                     <td align=\"center\">Aucun encodage extrait...</td>
186                     <td align=\"center\">&empty;</td>
187                     <td align=\"center\">&empty;</td>
188                     <td>&empty;</td>
189                     <td>&empty;</td>
190                     <td>&empty;</td>
191                     </tr>" >> $fichier_tableau ;
192                 fi
193             fi
194         fi
195         let "compteur+=1"; #on rajoute 1 à notre compteur
196         }
197         echo "</table>" >> $fichier_tableau ; #on ferme notre table html
198         let "compteur_tableau=compteur_tableau+1";
199     }
200     echo "</body></html>" >> $fichier_tableau ; #on ferme nos balises html ouvertes plus
        haut, lorsque nous avons crée la page
```