

```

1  #!/bin/bash
2  # 1. Lecture des paramètres dans le fichier PARAMETRES
3  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)
4  echo "Le dossier d'URLs : $DOSSIERURLS " ; #on commente les actions effectuées par
   le programme:
5  echo "Le fichier contenant le tableau : $fichier_tableau" ;
6  echo "Le motif est : $motif" ;
7  # on commence la création de notre tableau
8  compteur_tableau=1; #on crée un compteur pour chaque tableau (un par langue)
9  echo "<html><head><style>
10 a{color:red}          /*les liens en général*/
11 a:hover{color:black} /*quand on passe la souris sur un lien*/
12 a:active{color:black} /*quand on clique sur le lien*/
13 a:visited{color:grey;} /*quand le lien a déjà été visité*/
14 </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
15 for fichier in $(ls $DOSSIERURLS) #on peut également l'écrire: for fichier in `ls
   $DOSSIERURLS`
16 {
17     compteur=1; # on crée un compteur pour compter les URLs
18     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
19     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 recouvrera 11 colonne (cela sert simplement
   à améliorer l'apparence de notre tableau #tr c'est pour une ligne et td pour une
   cellule
20     echo "<tr>
21     <td align=\"center\"><b>N°</b></td>
22     <td align=\"center\"><b>Lien</b></td>
23     <td align=\"center\"><b>CODE CURL</b>
24     <td align=\"center\"><b>Statut CURL</b></td>
25     <td align=\"center\"><b>Page Aspirée</b></td>
26     <td align=\"center\"><b>Encodage Initial</b></td>
27     <td align=\"center\"><b>DUMP initial</b></td>
28     <td align=\"center\"><b>DUMP UTF-8</b></td>
29     <td align=\"center\"><b>CONTEXTE UTF-8</b></td>
30     <td align=\"center\"><b>CONTEXTE HTML UTF-8</b></td>
31     <td align=\"center\"><b>Fréquence du motif</b></td>
32     </tr>" >> $fichier_tableau ; #on vient de créer la structure de notre tableau en
   html
33     # on va maintenant procéder au traitement de chacune des URLs
34     for line in `cat $DOSSIERURLS/$fichier` # ou: for line in $(cat
   $DOSSIERURLS/fichier) #le cat sert à afficher
35     {
36         # ==> ASPIRATION DE LA PAGE
37         echo "TELECHARGEMENT de $line vers
   ./PAGES-ASPIREES/$compteur_tableau-$compteur.html" ;
38         # 1. RECUPERATION DU HEADER HTTP
39         status1=$(curl -sI $line | head -n 1);
40         # 2. RECUPERATION DU CODE RETOUR HTTP ET DE LA PAGE
41         status2=$(curl --silent --output
   ./PAGES-ASPIREES/"$compteur_tableau-$compteur".html --write-out "%{http_code}"
   $line);
42         echo "STATUT CURL : $status2" ;
43
44         # on va détecter l'encodage de nos fichiers
45         echo "DETECTION encodage de $line ";
46         encodage=$(curl -sI $line | egrep -i "charset=" | cut -f2 -d= | tr -d "\n" | tr
   -d "\r" | tr "[:upper:]" "[:lower:]");
47         echo "ENCODAGE $line : <$encodage>" ;
48         if [[ $encodage == "utf-8" ]]
49         then
50             echo "DUMP de $line via lynx" ;
51             lynx -dump -nolist -assume_charset=$encodage -display_charset=$encodage $line
   > ./DUMP-TEXT/$compteur_tableau-$compteur.txt ;
52
53             # ajouter ici l'extraction de contexte autour des mots choisis
54

```

```

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

```

```

100 mv resultat-extraction.html ./CONTEXTES/$compteur_tableau-
101 $compteur.html ;
102 echo "ECRITURE RESULTAT dans le tableau" ;
103 echo "<tr>
104 <td align=\"center\">$compteur</td>
105 <td align=\"center\"><a href=\"$line\">lien
106 n°$compteur</a></td>
107 <td align=\"center\">$status2</td>
108 <td>$status1</td>
109 <td align=\"center\"><a
110 href=\"../PAGES-ASPIREES/$compteur_tableau-$compteur.html\">PA
111 n° $compteur_tableau-$compteur</a></td>
112 <td align=\"center\">$encodage<br/>via curl</td><td
113 align=\"center\"><a
114 href=\"../DUMP-TEXT/$compteur_tableau-$compteur.txt\">DUMP
115 n° $compteur_tableau-$compteur</a></td>
116 <td align=\"center\"><a
117 href=\"../DUMP-TEXT/$compteur_tableau-$compteur-utf8.txt\">DUM
118 P n° $compteur_tableau-$compteur</a></td>
119 <td align=\"center\"><a
120 href=\"../CONTEXTES/$compteur_tableau-$compteur.txt\">CONTEXTE
121 n° $compteur_tableau-$compteur</a></td>
122 <td align=\"center\"><a
123 href=\"../CONTEXTES/$compteur_tableau-$compteur.html\">CONTEXT
124 E n° $compteur_tableau-$compteur</a></td>
125 <td align=\"center\">$nombre_motif</td></tr>" >>
126 $fichier_tableau ;
127 fi #sert à fermer le if
128 else
129 recherche_charset=$(egrep -i -o "meta(.*)?charset"
130 ./PAGES-ASPIREES/"$compteur_tableau-$compteur".html);#on cherche
131 l'encodage(le charset) dans les pages aspirées. Comme leur nom est
132 composé du numero du compteur du tableau puis du compteur de lignes
133 on met cette variable pour automatiser la recherche pour toutes nos
134 pages.
135 if [[ $recherche_charset != "" ]] #si on ne trouve rien
136 then #alors
137 encodage=$(egrep -i -o
138 "meta(.*)charset[^\=]*?=[^\"]*?\"?[\^"]+?\""
139 ./PAGES-ASPIREES/$compteur_tableau-$compteur.html | egrep -i
140 -o "charset[^\=]*?=[^\"]*?\"?[\^"]+?\"" | cut -f2 -d= | sed
141 "s/\"//g" | sed "s/>//g" | sed "s/ //g" | sed "s/\\\\//g" |
142 sort -u | tr [A-Z] [a-z] ); #on crée une variable qui
143 s'appelle encodage dans laquelle on
144 echo "ENCODAGE EXTRAIT DE LA PAGE ASPIREE : $encodage" ;
145 if [[ $encodage == "utf-8" ]] #si l'encodage est en utf-8
146 then #alors
147 echo "DUMP de $line via lynx" ;
148 lynx -dump -nolist -assume_charset=$encodage -
149 display_charset=$encodage $line > ./DUMP-TEXT/
150 $compteur_tableau-$compteur.txt ; #
151 egrep -i $motif ./DUMP-TEXT/$compteur_tableau-
152 $compteur.txt > ./CONTEXTES/$compteur_tableau-
153 $compteur.txt ;
154 nombre_motif=$(egrep -coi $motif
155 ./DUMP-TEXT/$compteur_tableau-$compteur.txt);
156 perl ./PROGRAMMES/minigrepmultilingue-v2.2-regexp/
157 minigrepmultilingue.pl "utf-8" ./DUMP-TEXT/
158 $compteur_tableau-$compteur.txt ./PROGRAMMES/
159 minigrepmultilingue-v2.2-regexp/motif.txt ;#ne pas
160 oublier de mettre le chemin de motif.txt ! Sinon ça
161 ne marche pas
162 mv resultat-extraction.html ./CONTEXTES/
163 $compteur_tableau-$compteur.html ;
164 echo "ECRITURE RESULTAT dans le tableau" ;
165 echo "<tr>
166 <td align=\"center\">$compteur</td>
167 <td align=\"center\"><a href=\"$line\">lien
168 n°$compteur</a></td>
169 <td align=\"center\">$status2</td>
170 <td align=\"center\">$status1</td>
171 <td align=\"center\"><a
172 href=\"../PAGES-ASPIREES/$compteur_tableau-$compteur.h

```

```

135 tml\">P.A n° $compteur_tableau-$compteur</a></td>
<td align=\"center\">$encodage<br/>avec
charset</td><td align=\"center\">&empty;</td>
136 <td align=\"center\"><a
href=\"../DUMP-TEXT/$compteur_tableau-$compteur.txt\">
DUMP n° $compteur_tableau-$compteur</a></td>
137 <td align=\"center\"><a
href=\"../CONTEXTES/$compteur_tableau-$compteur.txt\">
CONTEXTTE n° $compteur_tableau-$compteur</a></td>
138 <td align=\"center\"><a
href=\"../CONTEXTES/$compteur_tableau-$compteur.html\">
CONTEXTTE n° $compteur_tableau-$compteur</a></td>
139 <td align=\"center\">$nombre_motif</td>
140 </tr>" >> $fichier_tableau ;
141 else #sinon
142 enco_iconv=$(iconv -l | egrep -o "[-A-Z0-9\_\\:]+\" |
egrep -i $encodage) ;
143 if [[ $enco_iconv == "" ]]
144 then
145 echo "<tr>
146 <td align=\"center\">$compteur</td>
147 <td align=\"center\"><a href=\"$line\">lien
n°$compteur</a></td>
148 <td
align=\"center\">$status2</td><td>$status1</td>
>
149 <td align=\"center\"><a
href=\"../PAGES-ASPIREES/$compteur_tableau-$co
mpteur.html\">PA n°
$compteur_tableau-$compteur</a></td>
150 <td align=\"center\">$encodage<br/><br/>avec
charset<br/>inconnu de iconv</td>
151 <td align=\"center\"><a
href=\"../DUMP-TEXT/$compteur_tableau-$compteu
r.txt\">DUMP n°
$compteur_tableau-$compteur</a></td>
152 <td align=\"center\">&empty;</td>
153 <td align=\"center\">&empty;</td>
154 <td align=\"center\">&empty;</td>
155 <td align=\"center\">&empty;</td>
156 </tr>" >> $fichier_tableau ;
157 else
158 lynx -dump -nolist -assume_charset=$encodage
-display_charset=$encodage $line > ./
DUMP-TEXT/$compteur_tableau-$compteur.txt ;
159 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
160 egrep -i $motif ./DUMP-TEXT/$compteur_tableau
-$compteur-utf8.txt > ./CONTEXTES/
$compteur_tableau-$compteur.txt ;
161 nombre_motif=$(egrep -coi $motif
./DUMP-TEXT/$compteur_tableau-$compteur-utf8.t
xt) ;
162 perl ./PROGRAMMES/
minigrepmultilingue-v2.2-regexp/
minigrepmultilingue.pl "utf-8" ./DUMP-TEXT/
$compteur_tableau-$compteur-utf8.txt ./
PROGRAMMES/minigrepmultilingue-v2.2-regexp/
motif.txt ;
163 mv resultat-extraction.html ./CONTEXTES/
$compteur_tableau-$compteur.html ;
164 echo "ECRITURE RESULTAT dans le tableau" ;
165 echo "<tr>
166 <td align=\"center\">$compteur</td><td
align=\"center\"><a href=\"$line\">lien
n°$compteur</a></td>
167 <td
align=\"center\">$status2</td><td>$status1</td>
>
168 <td align=\"center\"><a
href=\"../PAGES-ASPIREES/$compteur_tableau-$co

```

```

169     mpteur.html\">PA n°
170     $compteur_tableau-$compteur</a></td>
171     <td align=\"center\">$encodage<br/>avec
172     charset</td>
173     <td align=\"center\"><a
174     href=\"../DUMP-TEXT/$compteur_tableau-$compteu
175     r.txt\">DUMP n°
176     $compteur_tableau-$compteur</a></td>
177     <td align=\"center\"><a
178     href=\"../DUMP-TEXT/$compteur_tableau-$compteu
179     r-utf8.txt\">DUMP n°
180     $compteur_tableau-$compteur</a></td>
181     <td><a
182     href=\"../CONTEXTES/$compteur_tableau-$compteu
183     r.txt\">CONTEXTE n°
184     $compteur_tableau-$compteur</a></td>
185     <td align=\"center\"><a
186     href=\"../CONTEXTES/$compteur_tableau-$compteu
187     r.html\">CONTEXTE n°
188     $compteur_tableau-$compteur</a></td>
189     <td align=\"center\">$nombre_motif</td>
190 </tr>" >> $fichier_tableau ;
191
192     fi
193
194     fi
195     else #sinon
196     echo "<tr>
197     <td align=\"center\">$compteur</td>
198     <td align=\"center\"><a href=\"$line\">lien n°$compteur</a></td>
199     <td align=\"center\">$status2</td>
200     <td align=\"center\">$status1</td>
201     <td align=\"center\"><a
202     href=\"../PAGES-ASPIREES/$compteur_tableau-$compteur.html\">PA
203     n° $compteur_tableau-$compteur</a></td>
204     <td align=\"center\">Aucun encodage extrait...</td>
205     <td align=\"center\">&empty;</td>
206     <td align=\"center\">&empty;</td>
207     <td>&empty;</td>
208     <td>&empty;</td>
209     <td>&empty;</td>
210     </tr>" >> $fichier_tableau ;
211
212     fi
213
214     fi
215     let "compteur+=1"; #on rajoute 1 à notre compteur
216 }
217 echo "</table>" >> $fichier_tableau ; #on ferme notre table html
218 let "compteur_tableau=compteur_tableau+1";
219 }
220 echo "</body></html>" >> $fichier_tableau ; #on ferme nos balises html ouvertes plus
haut, lorsque nous avons crée la page

```