Programmes perl

ExtractModif.pl

lebon.pl / lebon2.pl

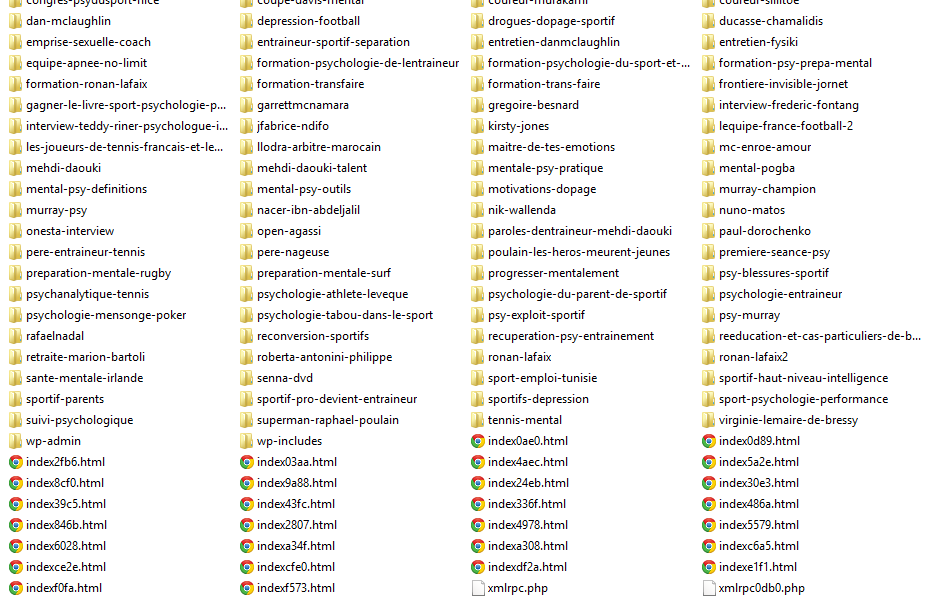
ligne.pl

url.pl

bd.pl

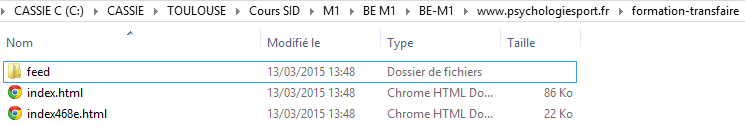
Utilisation de HTTRACK //TODO

Après avoir récupéré tous les articles du web grâce au logiciel HTTRACK, nous devions trier ces résultats pour en retirer les articles utiles.



En étudiant les résultats retournés, sous forme de dossiers (un dossier par article), nous avons remarqué que seuls les fichiers HTML nommés index.html étaient utiles.

Les fichiers en questions n’étaient pas stockés dans un dossier spécifique. Par exemple, le fichier index.html pouvaient être présent directement à l’ouverture du dossier correspondant à l’article



Dossier correspondant à l’article

Fichier HTML à ne pas récupérer

Fichier HTML à récupérer