Procédure à suivre pour l'extraction de l'HTML de Lamartine et son nettoyage

Lucie Slavik, stage mai-juillet 2020

Le taux de reconnaissance estimé pour les volumes de Lamartine est de 100 %

Pour l'instant, cette procédure a été suivie pour les deux volumes de correspondance.

1) Étapes préliminaires : extraction + rendre le fichier HTML valide
- Extraire le volume sur Gallica (voir le lien dans le cdc https://catalogue.bnf.fr/ark:/12148/cb30725428p)
- Si la version d'oXygen n'est pas la plus récente, fermer les balises <meta>,
 et <hr>

2) Nettoyage avec regex
- Faire des Regex sur oXygen pour supprimer les titres qui se trouvent dans l'HTML à chaque changement de page.

[0-9]+ CORRESPONDANCE DE LAMARTINE\.[^<\/p>]+[<\/p>](/p>) => 135 matches(vol2)
[0-9]{3} CORRESPONDANCE DE LAMARTINE => 2 matches (vol2)
[0-9]+ CORRESPONDANCE DE LAMARTINE\. => 1 match (vol2)
[0-9]{3} CORRESPONDANCE DE LAMARTINE\. => 9 matches
[0-9]{3} CORRESPONDANCE DE LAMARTINE\. => 25 matches
[0-9]{3} CORRESPONDANCE DE LAMARTINE => 6 matches
[0-9]{3} CORRESPONDANCE DE LAMARTINE\. => 2 matches

S'il en reste, le faire manuellement ou trouver d'autres regex.

- Regex pour enlever les titres des têtes de page mentionnant les années et les pages :
ANNÉE [0-9]{4}. [0-9]+ <\/p> =>153 matches (vol2)
ANNÉE [0-9]+. [0-9]+ =>39 matches(vol2)
ANNÉE [0-9]{4}\. => 2 matches (vol2)
ANNÉE [0-9]{4}\. [0-9]{2} => 3 matches (vol2)
ANNÉE [0-9]{4}\. [0-9]{3} => 6 matches

Attention à cette regex, j'ai dû ensuite supprimer des balises , il faudrait trouver une solution meilleure.
Auparavant, il vaut mieux supprimer les balises inutiles mentionnant simplement un changement de page par un rechercher/remplacer
<\/p><hr \/>

- Certaines années sont difficiles à matcher, comme ici, pour le vol2, ligne 251
dé- ANNÉE
1813. 23
Je choisis de faire les cas difficiles manuellement.

- J’enlève les qui coupent une phrase avec cette regex
https://regex101.com/r/pVJI9T/1
(<\/p>)([a-z])
Je matche les deux balises p et je mets des parenthèses. Je matche celles qui se trouvent avant des minuscules (donc en milieu de phrase).
Pour la substitution, j’écris $2 comme cela les p sont supprimés et le $2 qui fait référence à la deuxième parenthèse garde le contenu et n’enlève pas la première lettre du mot en minuscules.
Tout rechercher : (<\/p>)([a-z])
Tout remplacer : $2
257 matches (vol2)

3) Nettoyage, corrections de fautes
Fautes d’orthographe :
quille au lieu de quitte
l’embrasse au lieu de t’embrasse
loi au lieu de toi env. 10 matches

4) Les destinataires de la correspondance de Lamartine
Les destinataires du tome 1 (au nombre de 2) :
- Aymon de Virieu
- Prosper Guichard de Bienassis

Les destinataires du tome 2 (environ 13)
- A monsieur Aymon de Virieu /monsieur le comte de Virieu (voir si c’est le même)
- A monsieur le marquis de Virieu
- A monsieur Laurent de Jussieu
- A monsieur de Lamartine
- A monsieur Fortuné de Vaugelas
- A mademoiselle Éléonore de Canonge
- A madame la marquise de Raigecourt
- A monsieur le baron de Vignet
- A monsieur le comte de Saint-Mauris
- Au duc de Rohan
- A monsieur l'abbé Dumont
- A monsieur de Genoude / A monsieur Eugène de Genoude
- A monsieur Rocher