17/01/2025

Nicolas Morane
M1 SDH Panthéon Sorbonne

Cours Extraction Web avec Python,

Exercice :
Sur le site des Archives Départementales de Savoie, j'ai effectué une recherche libre sur l'ensemble des fonds disponibles avec comme critères :

- Lieux : Beaufort (Savoie)
- Période : 1000 – 1790

https://recherche-archives.savoie.fr/?id=recherche_guidee_inventaire_web

L'idée est de faire l'inventaire de toutes les archives disponibles pour le moyen âge et l'époque moderne sur le territoire de Beaufort en Savoie.
La requête me retourne 371 rubriques avec un affichage de 20 par page.

L'objectif du programme python est de lire les informations (date, cote, description…) résumant chaque document. Ces informations se trouvent dans le code source de chaque page .
Le programme doit donc :

- Ouvrir la première page
- Récolter pour chaque document les informations et les stocker dans une liste
- Faire une boucle sur les pages suivantes, récolter puis ajouter les informations à la suite dans la liste
- Lorsqu'il n'y a plus d'information , arrêter la boucle
- Stocker la liste dans un dataframe puis le dataframe dans un fichier csv

Remarques sur les difficultés rencontrées :

- Le site des archives est très lent : j'ai dû placer un timer de 40sec entre chaque page (entre les get et les response ?)
- Je n'ai pas eu de mal pour la première page, mais j'en ai eu beaucoup par accéder à la seconde ; en effet j'ai mis du temps à comprendre que les paramètres de la requête de recherche étaient stockés sur le serveur des Archives. C'est avec une série de questions / réponses ChatGpt que j'ai compris la notion de « session » permettant enfin de faire comprendre au serveur que mes requêtes html étaient corrélées les unes aux autres avec le contexte des critères de ma recherche.

Pour des facilités de tests, j'ai inclus une boucle avec un nombre limité de pages (ce n'était pas indispensable, car le programme sort de la boucle dès qu'il n'y a plus de données). Mais à cause du timer à 40 sec, l'exécution du programme prend 20 minutes pour les 371 rubriques à récolter.
Pour tester le programme (cf. Archives_Savoie_Beaufort.py joint) la boucle est limitée à 3 pages
Si on veut exécuter le programme pour la totalité des pages il suffit de remplacer 3 par un chiffre élevé (au minimum 19 ici, car 371 rubriques avec 20 rubriques par page))

Le fichier CSV en sortie : Extract_Archives_Beaufort.csv