Atelier 2 - Collecte en ligne (scraping)

Afin d'enrichir vos données avec des informations touristiques, vous décidez de collecter des données directement sur le site https://www.bordeaux-tourisme.com (vous supposez que ces informations ne sont pas disponibles en open data).

Etape 1 - Aspect légal

Préciser les limites que vous vous imposerez au niveau du scraping pour restreindre le risque de poursuite juridique de la part du site.

Etape 2 - Structure du projet

Créer un dépôt github pour le projet

Organiser votre projet avec les répertoires regroupant les fichiers suivants (nom au choix) :

- scripts shells ou fichiers de configuration pour l'installation des outils,
- scripts d'ingestion,
- cible du fichier de base de données duckdb

Actualiser votre readme pour rendre compte de cette organisation et la procédure de configuration d'un environnement de développement.

Etape 3 - Récupération d'une liste

Ecrire un script qui récupère un maximum d'informations directement à partir d'une liste proposée sur le site (ex : agenda).

Stocker les données obtenues dans une table de la base duckdb initiée au premier atelier.

Important : Avant d'historiser remplacez les bordeaux-tourisme.com par tourisme.example

Etape 4 - Exploration de lien

Ecrire un script qui suit les liens de la première liste pour obtenir des informations plus détaillées.

Mettre à jour votre/vos insertion-s.

Livrable

Pdf avec la réponse à l'étape 1 et le lien github.