



# R4.C.12 – Initiation au GéoWeb et au Mashup

## Projet

Cartographie des cinémas

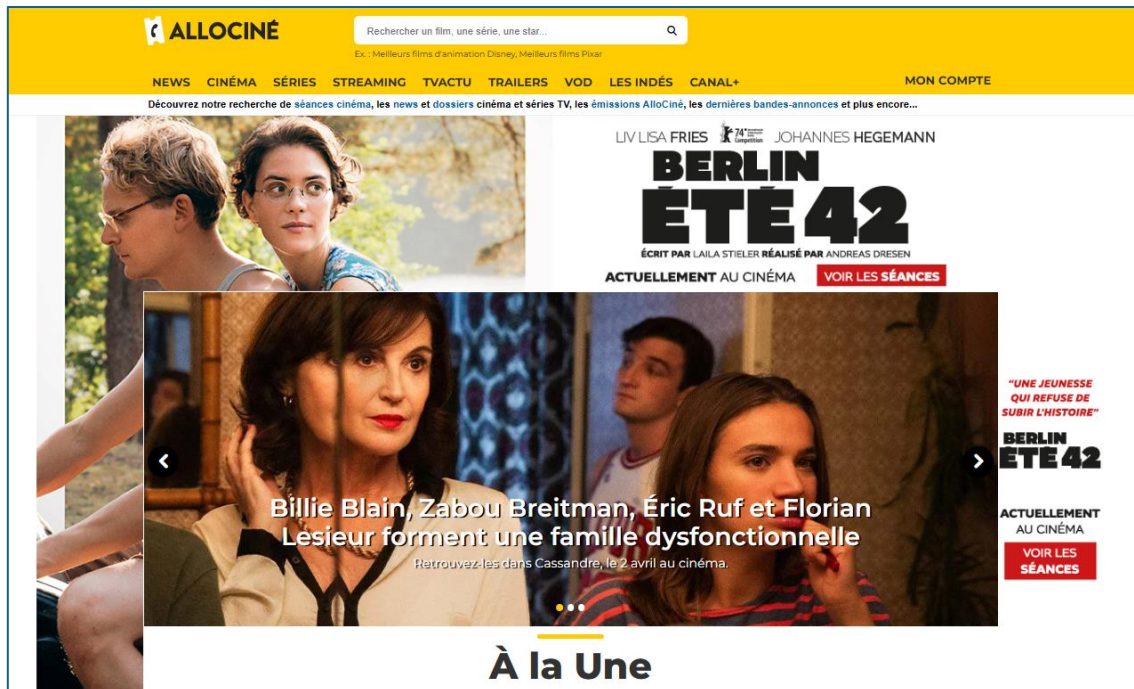
BUT Informatique



Eric Deléchelle

# Projet de Récupération et Visualisation des Informations sur les Séances de Cinéma en France

Ce cahier des charges présente une vue d'ensemble du projet et des attentes. Il peut être ajusté en fonction des besoins spécifiques ou des contraintes supplémentaires.



## Introduction

Le présent projet vise à développer un notebook Python permettant de collecter des informations sur les séances de cinéma en France par le biais du scraping du site web Allociné. Les données récupérées incluront les coordonnées géographiques des cinémas, obtenues via une API externe, ainsi que les films à l'affiche. Ces informations seront ensuite visualisées sur une carte interactive à l'aide de la bibliothèque Folium.

Avertissement : Le scraping est un outil puissant, mais il doit être utilisé de manière responsable et éthique (voir ce que dit la [CNIL](#) à ce sujet).

## Objectifs

- Collecter des données sur les cinémas, incluant les noms, adresses, liens vers les pages des cinémas, et les films à l'affiche.
- Obtenir les coordonnées géographiques des cinémas.
- Visualiser les données sur une carte interactive.
- Sauvegarder les données collectées dans un fichier JSON.

## Fonctionnalités principales

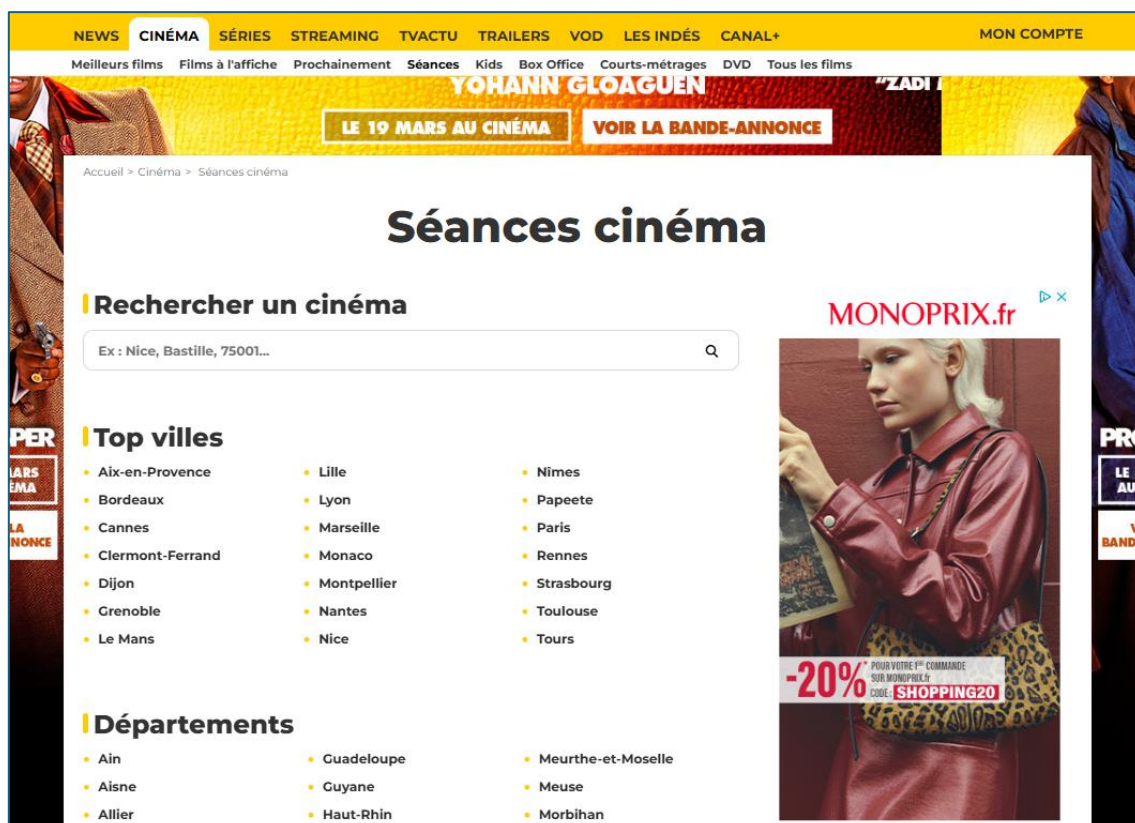
- Scraping des pages web : Utilisation de BeautifulSoup pour extraire les informations des pages web.
- Appels API : Utilisation d'API externes pour obtenir les coordonnées géographiques des adresses.
- Gestion des erreurs : Mise en place d'une gestion robuste des erreurs pour les requêtes HTTP et les appels API.
- Visualisation des données : Utilisation de Folium pour créer une carte interactive affichant les cinémas et les films à l'affiche.
- Sauvegarde des données : Stockage des données dans un fichier JSON.

## Technologies utilisées

- Python : Langage de programmation principal.
- BeautifulSoup : Bibliothèque pour le scraping des pages web.
- Requests : Bibliothèque pour effectuer des requêtes HTTP.
- Pandas : Bibliothèque pour la manipulation des données.
- Folium : Bibliothèque pour la création de cartes interactives.
- JSON : Format de stockage des données.

## Détails techniques

- Base URL : <https://www.allocine.fr>
- API de Géolocalisation : Utilisation de l'API d'Etalab (<https://api-adresse.data.gouv.fr>) ou Nominatim (<https://nominatim.openstreetmap.org>).
- Structure des données : Les données seront structurées en dictionnaires imbriqués pour les départements, villes, cinémas, et films à l'affiche.



## Étapes du Projet

### Initialisation

- Configuration de l'environnement de développement.
- Installation des bibliothèques nécessaires (requests, beautifulsoup4, pandas, folium).

### Scraping des départements

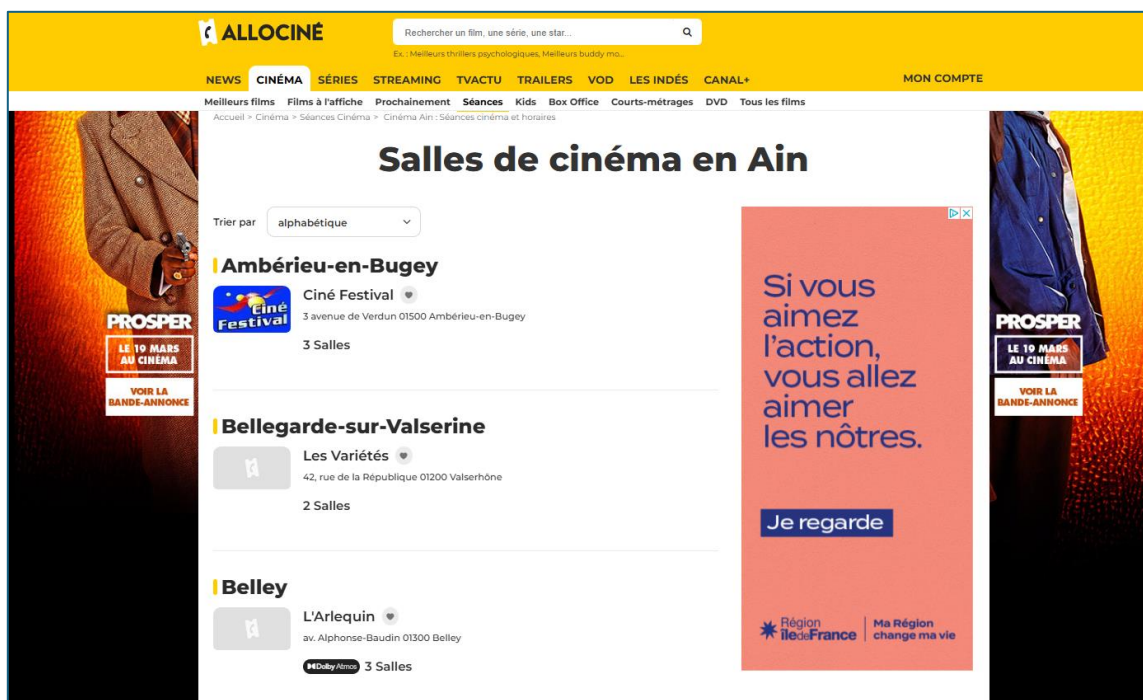
Récupération de la liste des départements et de leurs liens.

### Scraping des villes

Pour chaque département, récupération de la liste des villes et de leurs liens.

### Scraping des cinémas

Pour chaque ville, récupération de la liste des cinémas, de leurs adresses, et de leurs liens.



**Attention :** il est possible que la liste des communes soit soumise à une pagination (voir bas de page). Il faut donc en tenir compte afin d'obtenir la liste complète des communes.

Par exemple, pour l'Ain, la liste des commune occupe deux pages. La page #2 de la liste est accessible via l'URL paramétré <https://www.allocine.fr/salle/cinema/departement-83191/?page=2>

### Récupération des films à l'affiche

Pour chaque cinéma, récupération de la liste des films à l'affiche.

### Géolocalisation

Utilisation d'une API externe pour obtenir les coordonnées géographiques des adresses des cinémas.

## Visualisation des données

Utilisation de Folium pour créer une carte interactive affichant les cinémas et les films à l'affiche.

## Sauvegarde des données

- Sauvegarde des données récupérées dans un fichier JSON.
- Sauvegarder la carte au format HTML.

## Contraintes et critères de réussite

### Les contraintes

- Respect des règles de scraping : Ne pas surcharger le serveur avec des requêtes trop fréquentes.
- Gestion des erreurs : Assurer une gestion robuste des erreurs pour éviter les interruptions du script.
- Confidentialité des données : Ne pas stocker ou divulguer des informations sensibles.

### Les critères

- Complétude des données : Le script doit récupérer les informations pour tous les départements, villes, cinémas, et films à l'affiche.
- Précision des coordonnées : Les coordonnées géographiques doivent être précises.
- Qualité de la visualisation : La carte interactive doit être claire et informative.
- Fiabilité du script : Le script doit fonctionner sans interruption et gérer les erreurs de manière appropriée.

## Livrables

- Notebook Python : Un notebook fonctionnel permettant de collecter, visualiser, et sauvegarder les données.
- Fichier JSON : Un fichier JSON contenant les données des cinémas et des films à l'affiche.
- Carte Interactive : Une carte Folium affichant les cinémas et les films à l'affiche dans une page HTML.
- Documentation : Un guide d'utilisation du notebook et une description des données collectées.