

Projet Allociné



Franck Le Fur

Table des matières

1	Présentation	2
1.1	Présentation du projet	2
1.2	Spécifications techniques	2
1.3	Organisation du travail	3
2	Préliminaires	4
2.1	Préparation de l'environnement	4
2.2	Création d'un repository GitHub	4
3	Webscrapping	5
3.1	Introduction	5
3.2	Présentation des outils de web scraping	6
3.3	Scraper les informations des films de 1960 à 2024	6
3.4	Utilisation de Selenium	9
3.5	Scraper les films à l'affiche	9
3.6	Quelques difficultés rencontrées	9
4	Requêtage d'une API publique	11
4.1	Présentation de l'API OMDB	11
4.2	Récupération des informations	11
5	Nettoyage et agrégation des données	12
5.1	Nettoyage des données	12
6	Création de la base de données	13
6.1	Introduction au SGBD	13
6.2	Modèles MCD et MPD	13
6.3	Création de la base et des tables MySQL	16
6.4	Remplissage de la base MySQL	17
6.5	Création et remplissage de la base MongoDB	18
6.6	Exemples de requêtes SQL	18
7	Création d'une API	19
7.1	Généralités sur les APIs	19
7.2	Création de l'API	19
7.3	Règle d'authentification	21
7.4	Exemples d'utilisation de l'API	22
8	Automatisation	23
8.1	Présentation de Crontab	23
8.2	Création d'un environnement virtuel sous WSL	23
8.3	Script à exécuter	23
8.4	Création d'une tâche Crontab	23

1 Présentation

1.1 Présentation du projet

Ce projet présente une application permettant de faire des requêtes sur une base de données concernant le cinéma. L'application offre les options classiques de filtres de films selon des mots clés dans le titre, les noms d'acteurs, le réalisateurs, le compositeur, l'année de production, les catégories de films.

Elle offre également la possibilité de combiner ensemble tous ces filtres et ainsi de répondre à des questions telles que :

- Quels films ont réuni les acteurs Pierre Richard et Gérard Depardieu ?
- Dans combien de films de la franchise "Terminator" ont joué ensemble Linda Hamilton et Schwarzenegger ?
- Combien de films avec Pierre Richard ont vu leur musique composée par Vladimir Cosma ?

Cette application peut être proposée à des professionnels du cinéma, par exemple des critiques de films, ayant besoin de filtres avancés pour faire des recherches sur des associations dans le cinéma.

L'**objectif fonctionnel** est de construire une base de données et de l'exposer via une API

1.2 Spécifications techniques

Ce projet sera mené en **Python**, ce langage offre toutes les librairies, de façon gratuites, nécessaires à la conduite de ce projet.

Les données seront collectées à partir du site allocine.fr ne utilisant les librairies python **Beautifulsoup** et **Selenium**.

Des données supplémentaires seront récupérées par requêtage de l'api publique **OMDB** en utilisant la librairie python **requests**.

Nettoyage et aggrégation des données seront fait à l'aise de librairies python **Numpy** et **Panda**.

Nous utiliserons deux SGBD : **MySQL** et **MongoDB**, la mise en base et manipulation des données seront faites à l'aide des librairies python **mysql.connector** et **pymongo**.

Les tâches de scrapping, de requêtage de l'api public, de nettoyage et aggrégation des données seront automatisées à l'aide de l'outil **Crontab** disponible sur la machine Linux **WSL**.

L'API sera créée avec **FastAPI**, la documentation sera faite selon le modèle **OpenAPI**.

Un repo **github** sera créé pour le versioning des fichiers.

Enfin, le présent rapport sera rédigé en **Latex** via l'utilitaire en ligne **overleaf**.

1.3 Organisation du travail

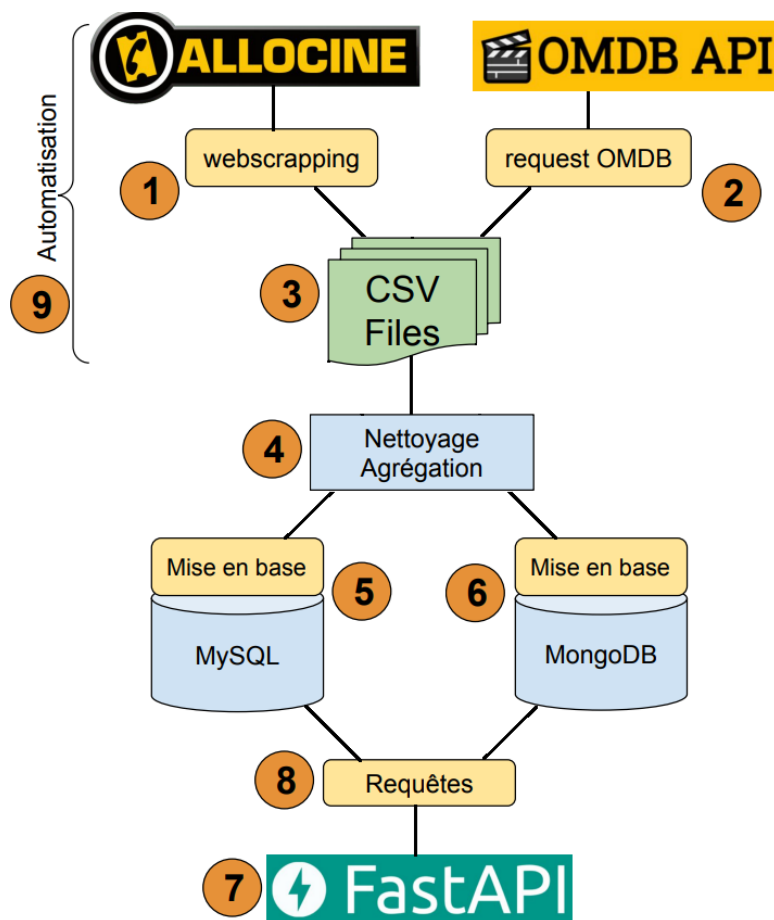


FIGURE 1 – Organigramme

Etapes

- 1 Scrapping du site allociné,
- 2 Requête de l'api publique OMDB,
- 3 Ecriture des données en CSV,
- 4 Nettoyage et agrégation des données,
- 5 Insertion des données dans MySQL,
- 6 Insertion des données dans MongoDB,
- 7 Création de l'API avec Fast,
- 8 Requetes sur les bases pour notre API,
- 9 Automatisation de l'extraction des données.

2 Préliminaires

2.1 Préparation de l'environnement

Dans un premier temps nous créons un nouvel environnement virtuel sous conda avec toutes les librairies python nécessaires.

Création d'un nouvel environnement

```
1 C:\Users\Utilisateur>conda create --name block1
2 C:\Users\Utilisateur>conda activate block1
```

Installation des packages python

```
1 C:\Users\Utilisateur>conda install numpy pandas tqdm
2 C:\Users\Utilisateur>conda install requests beautifulsoup4 selenium
3 C:\Users\Utilisateur>conda install mysql-connector-python
4 C:\Users\Utilisateur>conda install pymongo
5 C:\Users\Utilisateur>conda install unicodecode
6 C:\Users\Utilisateur>conda install fastapi pyjwt uvicorn
```

2.2 Création d'un repository GitHub

Création d'un repository **Github** via l'interface github puis connexion de notre répertoire de travail au repository **Github**.

```
git init
git add .gitignore
git branch -m master main (pour renommer la branche)
git commit -m "first commit"
git remote add origin https://github.com/Franck-LF/projectBlock1
git push -u origin main
```

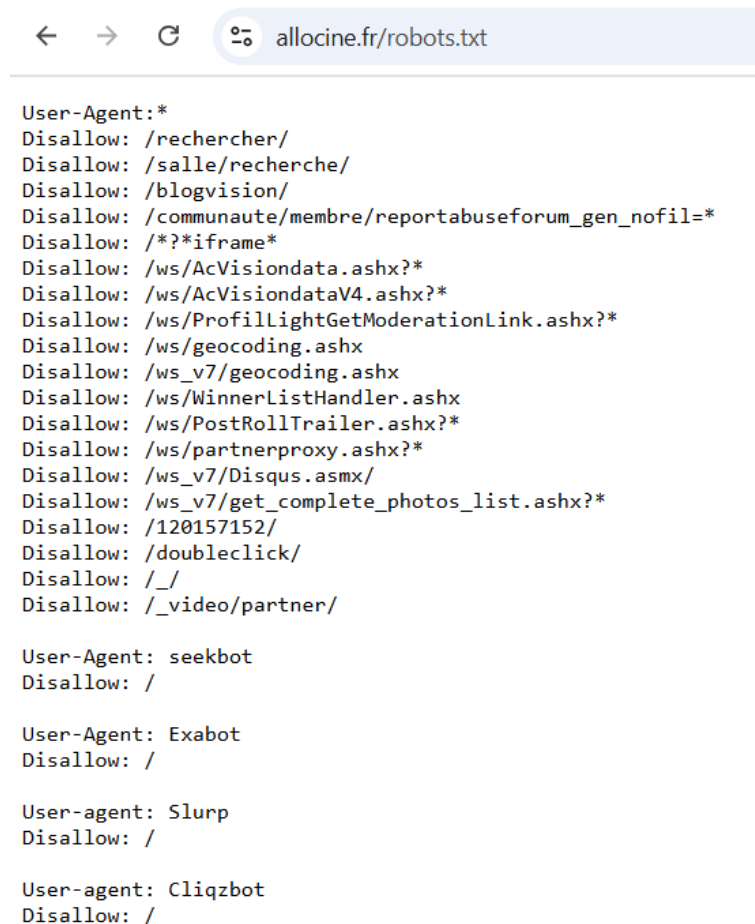
3 Webscrapping

3.1 Introduction

Le **webscrapping** est un processus d'extraction automatique de données à partir d'un site web. Il s'agit de parcourir des pages web et d'y récupérer le contenu souhaité. Ce processus doit être fait en respectant les conditions d'utilisation des sites et les lois sur la protection des données.

Dans ce projet les données seront scrappées à partir du site **allocine.fr**, ces données sont entièrement publiques, en effet il s'agit exclusivement de titres de films, noms d'acteurs, réalisateurs et compositeurs.

Il nous faut vérifier que le site **allocine.fr** autorise ce genre de pratiques, pour cela nous consultons le fichier "robots.txt".



```
User-Agent:*
Disallow: /rechercher/
Disallow: /salle/recherche/
Disallow: /blogvision/
Disallow: /communaute/membre/reportabuseforum_gen_nofil=*
Disallow: /*?*iframe*
Disallow: /ws/AcVissiondata.ashx?*
Disallow: /ws/AcVissiondataV4.ashx?*
Disallow: /ws/ProfillLightGetModerationLink.ashx?*
Disallow: /ws/geocoding.ashx
Disallow: /ws_v7/geocoding.ashx
Disallow: /ws/WinnerListHandler.ashx
Disallow: /ws/PostRollTrailer.ashx?*
Disallow: /ws/partnerproxy.ashx?*
Disallow: /ws_v7/Disqus.asmx/
Disallow: /ws_v7/get_complete_photos_list.ashx?*
Disallow: /120157152/
Disallow: /doubleclick/
Disallow: /_/
Disallow: /_video/partner/

User-Agent: seekbot
Disallow: /

User-Agent: Exabot
Disallow: /

User-agent: Slurp
Disallow: /

User-agent: Cliqzbot
Disallow: /
```

FIGURE 2 – fichier robots.txt

Le fichier robots.txt n'indique aucune restriction sur le chemin `/films/`, point d'entrée exclusif de notre web scraping, nous pouvons donc collecter en toute légalité les données souhaitées.

3.2 Présentation des outils de web scraping

Nous utilisons la librairie python **requests** pour récupérer le contenu de pages html, ensuite le web scraping se fera à l'aide de la librairie **Beautifulsoup**, cette librairie permet d'analyser et de parser le contenu html d'une page web.

(Un usage réduit et très spécifique de la librairie **Selenium** sera détaillée [ici](#).)

Exemple classique d'utilisation de BeautifulSoup

On commence par utiliser l'inspecteur du navigateur pour détecter les balises html qui contiennent les informations souhaitées, ensuite on utilise la méthode **find** sur un objet **soup** de BeautifulSoup pour récupérer le contenu de ces balises, par exemple la commande

```
elt_categories = soup.find('div', class_='filter-entity-section')
```

renvoie toutes les balises html **div** ayant un attribut **class** égal à "filter-entity-section".

Ensuite on peut récupérer le texte d'une balise :

```
elt.a.text
```

ou bien la valeur d'un attribut :

```
elt.get_attribute('title')
```

3.3 Scraper les informations des films de 1960 à 2024

Nous allons scraper les films de 1960 à 2024, nous nous limiterons aux films ayant plus de 30 avis utilisateurs et nous nous fixons un maximum de 250 films par année.

Finalement 8800 films seront collectés.

Nous commençons par scraper les liens des années disponibles à partir du menu **années**.

Remarque :

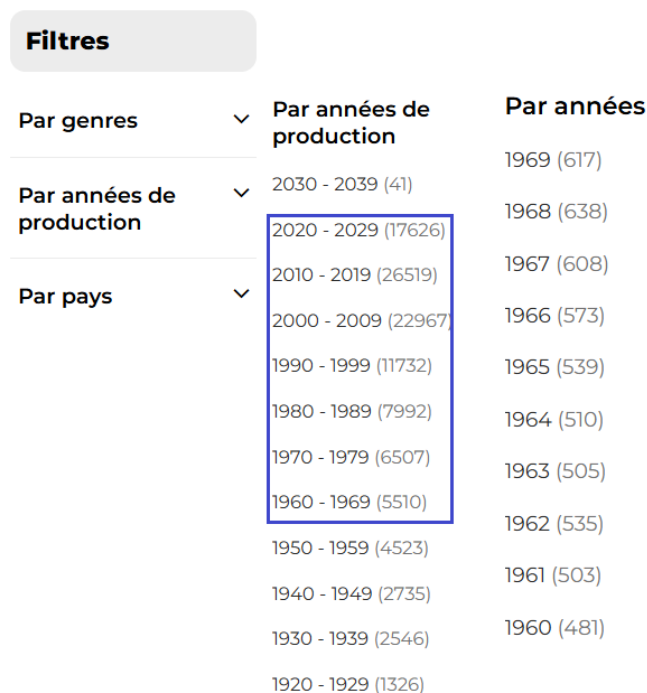


FIGURE 3 – filtre par années

Ensuite pour chaque année nous parcourons la liste de films puis à partir de la page du film, de

la section **infos techniques** ainsi que de la section **Casting complet et équipe technique** nous collectons les informations suivantes :

- Le titre du film
- Le titre original du film
- La date de sortie
- La durée
- La liste des catégories associées
- Les pays de production
- La liste des acteurs
- La liste des réalisateurs
- La liste des compositeurs
- La note associée au film
- Le nombre de notes spectateurs
- Le nombre d'avis spectateurs

Lors du scraping ces données seront stockées dans des **DataFrame Pandas**. Notons que toutes ces informations sont structurées et seront destinés à aller en base SQL.



FIGURE 4 – Vignette du film

Acteurs et actrices



Anthony Hopkins
Rôle : Dr. Frederick Treves (chirurgien)



John Hurt
Rôle : John Merrick, Elephant Man



Anne Bancroft
Rôle : Mrs. Madge Kendal (actrice)



John Gielgud
Rôle : Carr Gomm (directeur de l'hôpital)



Wendy Hiller
Rôle : L'infirmière en chef



Freddie Jones
Rôle : Bytes (propriétaire d'Elephant Man)



Hannah Gordon
Rôle : Mrs. Treves



Michael Elphick
Rôle : Le gardien de nuit

Lesley Dunlop infirmière Nora
Helen Ryan La princesse Alex
Kenny Baker Nain à plumes
John Standing Fox
Dexter Fletcher Le garçon de Bytes
Phoebe Nicholls La mère de Merrick
Pat Gorman Bobby au champ de foire
Claire Davenport Femme obèse
Orla Pederson Homme squelettique
Patsy Smart Femme désespérée
Frederick Treves Alderman
Richard Hunter Hodges
James Cormack Pierce
Alfie Curtis Le livreur de lait
Robert Lewis Bush Le messenger
Roy Evans Le chauffeur de taxi
Joan Rhodes Le cuisinier
Nula Conwell L'infirmière Kathleen
Tony London Porter jeune

FIGURE 5 – Casting complet et équipe technique

Infos techniques

Nationalité	U.S.A.
Distributeur	Carlotta Films
Récompenses	4 prix et 16 nominations
Année de production	1980
Date de sortie DVD	11/12/2001
Date de sortie Blu-ray	03/11/2009
Date de sortie VOD	08/02/2007
Type de film	Long métrage
Secrets de tournage	23 anecdotes
Budget	5 000 000 USD
Date de reprise	22/06/2020
Langues	Anglais
Format production	-
Couleur	N&B
Format audio	-
Format de projection	-
N° de Visa	54114

FIGURE 6 – Infos techniques

3.4 Utilisation de Selenium

Lors de l'utilisation de BeautifulSoup, certains liens apparaissent **décorés** et sont donc inexploitable, l'utilisation de la librairie **Selenium** permet de résoudre ce problème.

```
<li class="filter-entity-item"> == $0
  <a class="xXx item-content" title="2030 - 2039" href="/films/decennie-2030/">2030 - 2039</a>
  <span class="light">(47)</span>
</li>
▶ <li class="filter-entity-item">...</li>
▶ <li class="filter-entity-item">...</li>
```

FIGURE 7 – Inspecteur Chrome

```
<span class="ACrL2ZACrpbG1zL2RlY2VubmllLTlWmZAv item-content" title="2030 - 2039">
  2030 - 2039
</span>
décoration
```

FIGURE 8 – Résultat BeautifulSoup

Nous n'enregistrons pas les affiches de films dans la base de données, ce n'est pas très pertinent d'un point de vue gestion de la mémoire, nous préférons garder en base les urls des affiches, les affiches étant déjà stockées sur internet.

3.5 Scraper les films à l'affiche

Ici nous ne scrapons pas les films de toute une année mais uniquement les films de la semaine à partir de la page allociné dédiée. Cette tâche sera automatisée de façon hebdomadaire pour récupérer les informations de tous les nouveaux films à l'affiche (voir section Crontab).

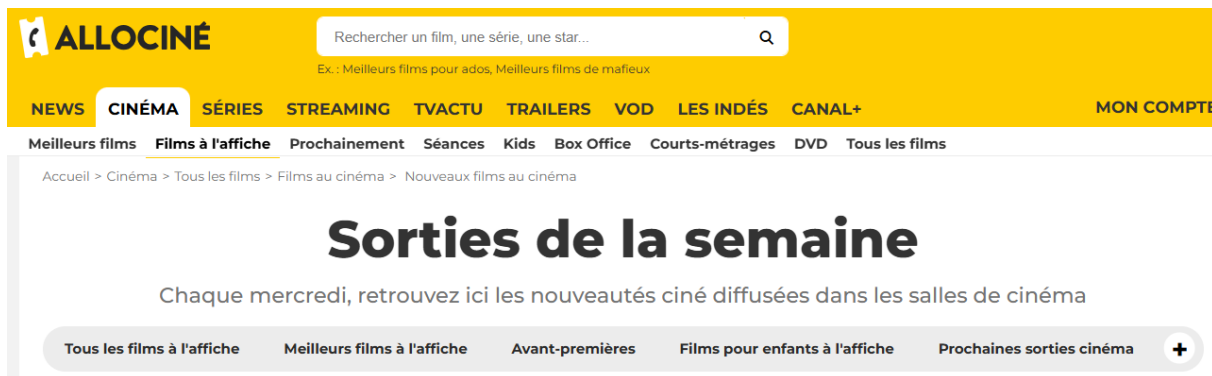


FIGURE 9 – Section "Sorties de la semaine"

3.6 Quelques difficultés rencontrées

- 1 Nous avons commencé par scraper les catégories et les pays à partir des listes du site mais il apparaît que la liste des pays à cet endroit du site n'est pas complète, en effet cette liste ne contient pas le **Bostwana** mais il existe bien des films dont le pays de production est le **Bostwana**.

Finalement, nous construisons la liste des pays à partir des informations des films et non pas à partir de cette liste, ensuite, pour chaque nouveau scrapin nous ajoutons dans la table 'pays' les pays ne s'y trouvant pas déjà pour éviter les doublons. Cette méthode est plus robuste.

- 2 Les films récents n'ont pas de section "ratings" ou bien ont des sections "ratings" vides, des tests supplémentaires ont été ajoutés dans la phase de web scraping.
- 3 Les films récents ont parfois une section "rating" mais avec zéro avis utilisateur alors que jusqu'à présent nous ne prenions que les films avec au moins 30 avis, des options supplémentaires ont été créées et sont transmises pour
- 4 Lors du web scraping des films à l'affiche début 2025, certains films avaient en fait une année de production "2024" et avaient déjà été scrapés. Un nettoyage des doublons a été fait à partir de la base MySQL, ce phénomène ne peut plus arriver car désormais nous scrapons uniquement les films à l'affiche (et non pas des années entières).

4 Requête d'une API publique

4.1 Présentation de l'API OMDB

Nous avons choisi l'api publique [OMDB](#) pour collecter des informations de résumé de films et d'url d'affiche de films, à noter que ces informations de texte et d'url sont non-structurés et seront destinés à aller en base NoSQL.

Pour accéder à l'API d'OMDB nous demandons une clé dans la section dédiée, cela nous offre un accès gratuit à 1000 requêtes par jour.

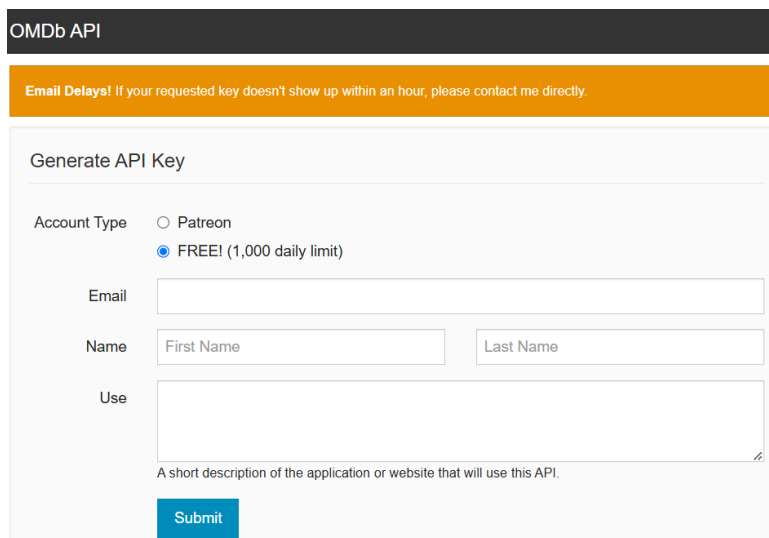
The image shows the 'OMDb API' website interface for generating an API key. At the top, there's a dark header with 'OMDb API' in white. Below it is an orange banner with the text 'Email Delays! If your requested key doesn't show up within an hour, please contact me directly.' The main section is titled 'Generate API Key' and contains several input fields: 'Account Type' with radio buttons for 'Patreon' and 'FREE! (1,000 daily limit)' (which is selected), 'Email' (a single line), 'Name' (split into 'First Name' and 'Last Name'), and 'Use' (a larger text area). Below these is a small text prompt: 'A short description of the application or website that will use this API.' At the bottom right of this section is a blue 'Submit' button.

FIGURE 10 – Générer une clé API

La documentation de l'API OMDB nous indique la forme des requêtes, par exemple pour le film **In The Lost Lands** :

`https://www.omdbapi.com/?apikey=xxxx&t=in+the+lost+lands`

A l'aide de l'utilitaire **requests** nous obtenons le résultat sous forme d'un json.

```
1 { 'Title': 'In the Lost Lands',
2   'Year': '2025',
3   'Genre': 'Action, Adventure, Fantasy',
4   'Director': 'Paul W.S. Anderson',
5   'Writer': 'Constantin Werner, Paul W.S. Anderson, George R.R. Martin',
6   'Actors': 'Milla Jovovich, Dave Bautista, Arly Jover',
7   'Plot': 'A queen sends the powerful and feared sorceress Gray Alys to the
8           ghostly wilderness of the Lost Lands in search of a magical power, where
           the sorceress and her guide, the drifter Boyce, must outwit and outfight
           man and demon.',
9   'Poster': 'https://m.media-amazon.com/images/M/
              MV5BOWYxYjEyYTUtY2FkZC00jMtMTAyMTAyMzQ3MDZiXkEyXkFqcGc@._V1_SX300.jpg',
```

4.2 Récupération des informations

Nous disposons d'un Dataframe Pandas avec la liste des films ainsi que les informations déjà scrapées, pour requêter l'api OMDB nous créons une fonction qui créer une requête OMDB au bon format à partir du titre du film puis une fonction qui requêtes l'api OMDB et insère les données collectées dans de nouvelles colonnes du DataFrame.

A l'issue de cette étape les données sont enregistrées dans des fichiers CSV.

5 Nettoyage et agrégation des données

5.1 Nettoyage des données

Avant d'insérer les données en bases, nous devons nous assurer qu'elle sont au bon format.

Exemples de nettoyage

- 1 Les informations scrapées concernant les catégories de films, les acteurs, réalisateurs et compositeurs ont été mis sous la forme d'une chaîne de caractères où les entités sont séparées par une virgule.

```
"Alain Delon,Olga Georges-Picot,Charles Bronson,Brigitte Fossey,Bernard Fresson,
```

```
Jean-Paul Tribout,Ellen Bahl,Stéphane Bouy"
```

 Nous transformons cette chaîne de caractères en liste python de chaînes de caractères

```
["Alain Delon", "Olga Georges-Picot" ...]
```

 en nous assurant qu'il n'y a pas de doublon.

- 2 Certaines informations sont manquantes sur le site allociné (durée du film, nom du réalisateur, date de sortie et même la liste des acteurs pour les films d'animation etc), les champs dans nos Dataframe Pandas sont alors représentés par des `NaN` en python, nous les remplaçons par des chaînes de caractères vides ou bien par des valeurs nulles selon le type de champs souhaités.
- 3 Les durées sont stockées sous forme de chaînes de caractères de la forme `1h 35min`, nous les transformons en durée en minutes (95) de type **entier**.
- 4 Les notes sont stockées sous la forme d'une chaîne de caractères "3,5" où parties entière et décimale sont séparées par une virgule, nous remplaçons les virgules par des points.
- 5 Les dates de sortie de film sont stockées au format **5 mars 2025** sont converties au type Pandas **datetime**, pour se faire nous devons au préalable convertir les mois du français vers l'anglais.

A noter qu'un nettoyage a déjà été fait lors du scrapping.
Quel est le type des données dans mongodb ? Mixte ?

6 Création de la base de données

6.1 Introduction au SGBD

Un SGBD (Système de Gestion de Base de Données) est un logiciel prévu pour stocker, manipuler et extraire des données.

Pour ce projet, nous avons choisi les SGBD :

- **MySQL** pour toutes les **données structurées** c'est-à-dire les données numériques ou catégorielles, les entités concernées (films, acteurs, réalisateurs, compositeurs, catégories) ont des relations étroites et des requêtes de jointures ou de filtres complexes seront nécessaires, **MySQL** semble adaptée à ce besoin,
- **MongoDB** pour toutes les **données non-structurées** comme du texte ou des urls, dans notre projet il s'agira des synopsis des films et des urls vers les affiches de films, données qui sont uniquement liées avec l'entité film et ne nécessite donc pas de requête complexe.

6.2 Modèles MCD et MPD

MCD : Modèle conceptuel de données, il est indépendant du SGBD choisi, c'est "une carte mentale des données offrant une compréhension partagée entre développeurs et décideurs").

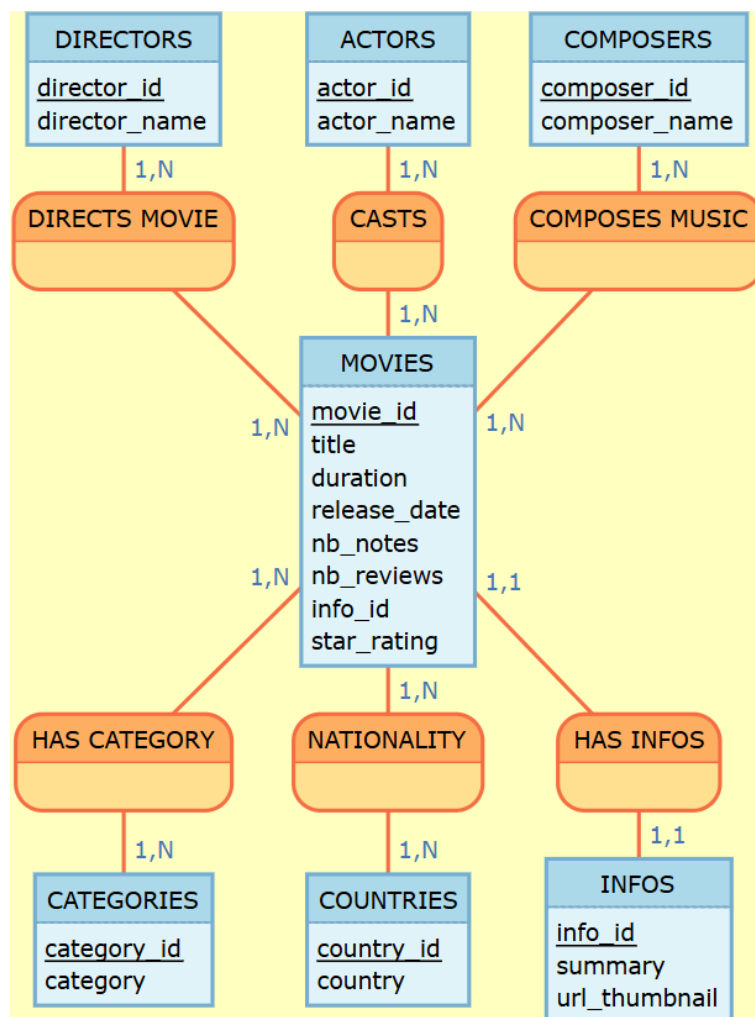


FIGURE 11 – MCD avec l'utilitaire Mocodo

Pour concevoir un **MCD** nous devons comprendre "le métier", c'est-à-dire :

- connaître les entités mises en jeu (ici les films, acteurs, catégories ...),
- identifier leurs attributs (titre, nom etc ...),
- savoir si ces attributs sont de type quantitatif ou catégoriel,
- identifier les relations entre les entités,
- définir les cardinalités dans ces relations (c'est-à-dire le nombres d'entités qui peuvent reliés à une autre entité).

Remarque sur les cardinalités

- 1 La cardinalité dans la relation **movies - directors** est de type **many-to-many** avec cardinalité minimal 1, l'interprétation est qu'un film est réalisé par au minimum un réalisateur et peut avoir été réalisé par plusieurs réalisateurs et réciproquement un réalisateur a réalisé au moins un film (sinon il ne serait pas dans nos données puisque nous sommes partis des films pour récupérer les informations) et peut aussi avoir réalisé plusieurs films.
- 2 La cardinalité dans la relation **movies - infos** est de type **1 - 1** car un film possède exactement une fois les informations qui le concernent (synopsis et url vers l'affiche de film) et que réciproquement une entité **infos** correspond exactement à un film.

Ensuite nous pouvons passer à la conception de notre **MPD** (Modèle physique de données), il dépend directement du SGBD choisi, ici **MySQL**, il représente les tables, les colonnes, les clés primaires et étrangères, on y précise les types de données, les contraintes etc...

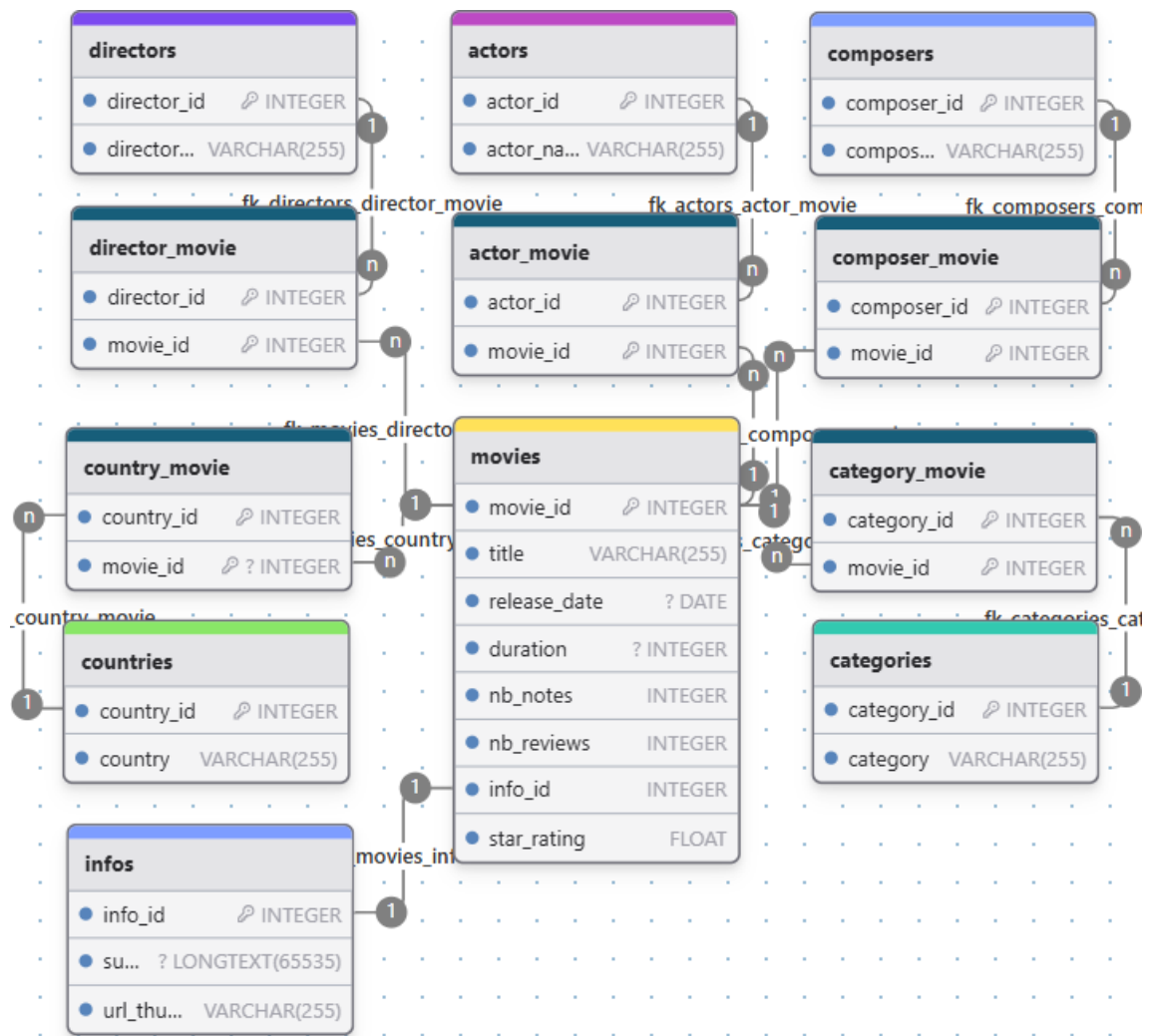


FIGURE 12 – MPD avec l'utilitaire DrawDB

Pour concevoir un **MPD** nous devons :

- convertir chaque entité du MCD en table du MPD,
- convertir les attributs des entités du MCD en colonne,
- définir le type de données pour chaque colonne,
- définir les clés primaires (identifiant unique de chaque enregistrement),
- convertir les relations en clés étrangères.

Pour exprimer une relation **many-to-many** nous avons besoin de créer une nouvelle table, appelée **table de jonction**, elle sert uniquement à faire le lien entre deux tables. Sa clé primaire est le couple composée des 2 clés étrangères des tables qu'elle relie. Dans le cadre de notre projet nous avons ajouté 5 tables de jonctions.

ToDo **Remarque** : Contrairement au **MCD** il n'y a pas de table **infos**, les informations de cette table seront stockées dans une collection MongoDB.

6.3 Création de la base et des tables MySQL

A partir du diagramme de tables conçu avec l'utilitaire **DrawDB** nous générons un fichier .sql contenant la structure de nos tables SQL, voici un extrait de ce fichier légèrement adapté.

```
1 CREATE TABLE actors (  
2     actor_id  VARCHAR(255)      NOT NULL,  
3     actor_name VARCHAR(255)      NOT NULL,  
4     PRIMARY KEY (actor_id),  
5     UNIQUE KEY (actor_name)  
6 );  
7  
8 CREATE TABLE actor_movie (  
9     actor_id  VARCHAR(255) NOT NULL,  
10    movie_id   VARCHAR(255) NOT NULL,  
11    PRIMARY KEY (actor_id, movie_id),  
12    FOREIGN KEY (actor_id) REFERENCES actors (actor_id) ON DELETE CASCADE,  
13    FOREIGN KEY (movie_id) REFERENCES movies (movie_id) ON DELETE CASCADE  
14 );
```

Remarque : La clause **ON DELETE CASCADE** sur une clé étrangère spécifie le comportement lors de la suppression d'un enregistrement dans la table parente. Concrètement si un acteur est supprimé de la base alors seront supprimés toutes les références à cet acteur dans la table de jonction **actor_movie** afin de ne pas laisser d'enregistrements "orphelins".

Nous pouvons lancer l'exécutable **mysql.exe** en lui donnant le fichier .sql qui contient l'architecture complète de notre base MySQL.

```
1 C:\Users\Utilisateur\Documents\Block1>"C:\Program Files\MySQL\MySQL Server  
2 8.0\bin\mysql.exe" < movies.sql -u root -p  
3 Enter password: *****  
4 CREATING DATABASE STRUCTURE  
5 storage engine: InnoDB  
6 EVERYTHING IS OK
```

Nous pouvons visualiser nos tables à partir de la console **MySQL Shell**,

```
MySQL localhost:33060+ ssl SQL > SHOW DATABASES;
```

Database
employees
games
gamesfromdumps
information_schema
list_books
movies
mysql
onetoone
performance_schema
sakila
sys
world

```
MySQL localhost:33060+ ssl movies SQL > SHOW TABLES;
```

Tables_in_movies
actor_movie
actors
categories
category_movie
composer_movie
composers
countries
country_movie
director_movie
directors
infos
movies

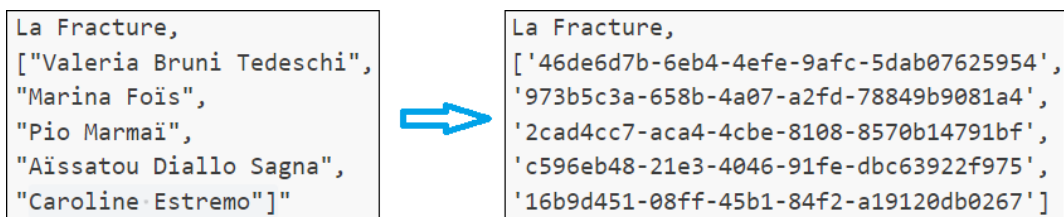
6.4 Remplissage de la base MySQL

Nous commençons par remplir les tables **categories**, **actors**, **directors**, **composers**, **countries**, ces tables n'ont pas de clé étrangère, elles ne dépendent pas directement d'autres tables et doivent être créées en premier.

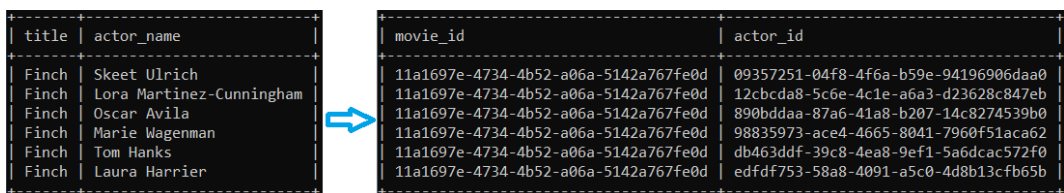
Chaque enregistrement aura un id unique créé avec la librairie **uuid**.

```
1 id = str(uuid.uuid4())
2 f"INSERT INTO actors (actor_id, actor_name) VALUES (%s, %s)"
3 val = (id, "Cameron Diaz")
4 cursor.execute(query, val)
```

Ensuite nous parcourons la liste de nos films, pour chaque film nous remplaçons la liste d'acteurs par la liste des ids d'acteurs.



Ensuite pour chaque film nous générons un id de film, nous disposons alors des couples (movie_id, actor_id) pour tous les acteurs ayant joué dans le film, nous remplissons donc la table de jonction **actor_movie**, idem avec les autres tables de jonction.



6.5 Création et remplissage de la base MongoDB

Nous créons nos bases et collections directement en python à l'aide de la librairie **pymongo** et nous insérons les documents à partir des dataframe Pandas.

```
1 # Connect to MongoDB
2 client = pymongo.MongoClient("mongodb://localhost:27017/")
3
4 # Create database "allocine" (or selects it if already exists)
5 mydb = client["allocine"]
6
7 # Create a collection "movies" (or select it if already exists)
8 col_movies = mydb["movies"]
9
10 # Insertion of movie plots in MongoDB database
11 col_movies.insert_many(df_movies.to_dict(orient='records'))
```

6.6 Exemples de requêtes SQL

Requête : Lister les films avec **Pierre Richard** et dont la musique a été composée par **Vladimir Cosma**.

```
SELECT m.title AS 'title',
       m.release_date AS 'date',
       d.director_name AS 'director'
FROM movies AS m
JOIN actor_movie AS am ON am.movie_id = m.movie_id
JOIN actors AS a ON a.actor_id = am.actor_id
JOIN composer_movie AS cm ON cm.movie_id = m.movie_id
JOIN composers AS c ON c.composer_id = cm.composer_id
JOIN director_movie AS dm ON dm.movie_id = m.movie_id
JOIN directors AS d ON d.director_id = dm.director_id
WHERE a.actor_name = 'Pierre Richard'
AND c.composer_name = 'Vladimir Cosma'
ORDER BY date;
```

	title	date	director
	Alexandre le Bienheureux	1968-02-09	Yves Robert
	Le Distrait	1970-12-09	Pierre Richard
	Le Grand Blond avec une chaussure noire	1972-12-05	Yves Robert
	La Moutarde me monte au nez	1974-10-09	Claude Zidi
	Le Retour du grand blond	1974-12-18	Yves Robert
	Le Jouet	1976-12-08	Francis Veber
	Je suis timide, mais je me soigne	1978-08-23	Pierre Richard
	C'est pas moi, c'est lui	1980-01-23	Pierre Richard
	Le coup du parapluie	1980-10-08	Gérard Oury
	La Chèvre	1981-12-09	Francis Veber
	Les compères	1983-11-23	Francis Veber
	Le Jumeau	1984-10-10	Yves Robert
	Les Rois du gag	1985-03-06	Claude Zidi
	Les Fugitifs	1986-12-17	Francis Veber
	Les Malheurs d'Alfred	2011-04-01	Pierre Richard
	La course à l'échalote	2019-12-11	Claude Zidi

Requête : Afficher les acteurs ayant jouer dans plusieurs films de la franchise **Terminator**.

```
SELECT DISTINCT a.actor_name, COUNT(*)
FROM actors AS a
JOIN actor_movie AS am ON am.actor_id = a.actor_id
JOIN movies AS m ON m.movie_id = am.movie_id
WHERE m.title LIKE '%Terminator%'
GROUP BY a.actor_name
HAVING COUNT(*) > 1
ORDER BY a.actor_name;
```

name	nb
Arnold Schwarzenegger	5
Earl Boen	3
Edward Furlong	2
Linda Hamilton	3
Michael Papajohn	2

7 Création d'une API

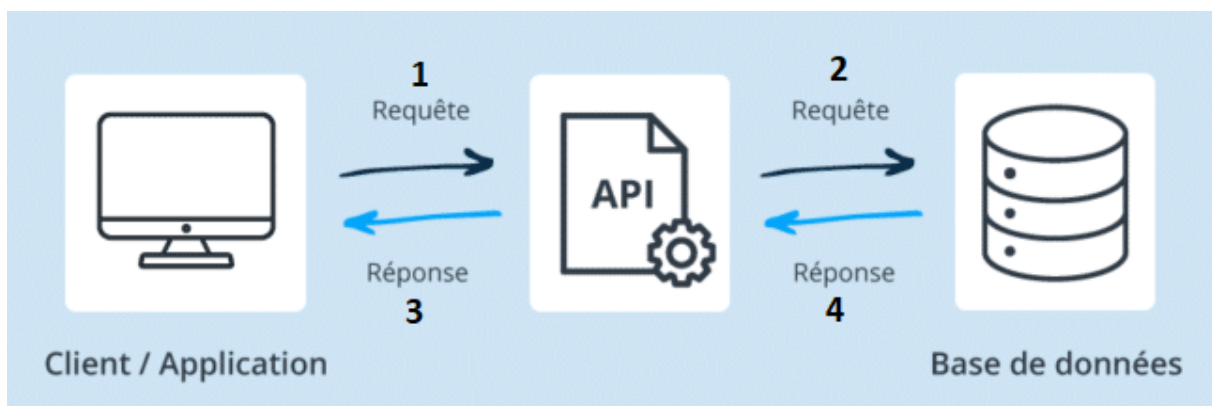
7.1 Généralités sur les APIs

une **API** est un programme permettant à deux logiciels (ou programmes) de communiquer entre eux. Dans notre cas, l'objectif de l'API est d'exposer notre base de données à l'utilisateur sans que celui-ci ne connaisse la base (nom des tables, des colonnes etc ...), l'API fait donc la communication entre la base de données et l'interface utilisateur.

Nous utilisons le framework python **Fast API** pour la créer une API de type **REST**. Le framework **Fast API** est connu pour sa rapidité et offre une génération automatique d'interface utilisateur ainsi que de la documentation. L'utilisateur n'aura qu'à consulter la documentation puis à renseigner les champs pour requêter la base de données via l'API.

Fonctionnement de l'API

- 1 Requête : L'utilisateur remplit des champs (acteurs, réalisateurs ...) à partir de l'interface créée par FastAPI, cela génère automatiquement une requête,
- 2 Traitement : L'API reçoit la requête, la traite en interrogeant la base de données,
- 3 Réponse de la requête : La base renvoie les données demandées à l'API,
- 4 Réponse de l'API : L'API renvoie le résultat au format JSON au client.



7.2 Création de l'API

Nous créons un script **api.py** puis des **fonctions routes** pour les différentes requêtes que nous souhaitons mettre à disposition de l'utilisateur.

```
1 @app.get("/movies")
2 async def get_movies(
3     actor1: Optional[str] = Query(None, alias="actor1", description="filtrer
4     les films avec 'actor1'"),
5     actor2: Optional[str] = Query(None, alias="actor2", description="filtrer
6     les films avec 'actor2'"),
7     producer: Optional[str] = Query(None, alias="producer", description="
8     filtrer les films realises par 'producer'"),
9 )
```

Une fois l'API lancée avec la commande `uvicorn api:app -reload`, la documentation de l'api est disponible à l'adresse `http://127.0.0.1:8000/docs`.

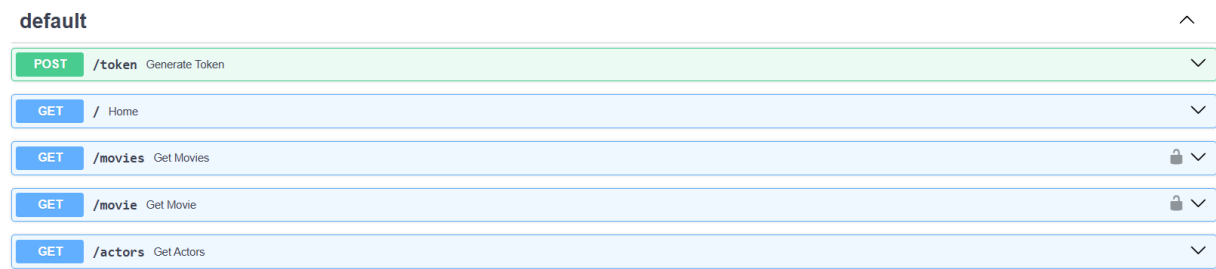


FIGURE 13 – Routes

La documentation des routes est automatiquement générée par **Fast API** dans le standard des documentations **Open API**. L'utilisateur n'a plus qu'à consulter la documentation et à remplir les champs pour obtenir les données.

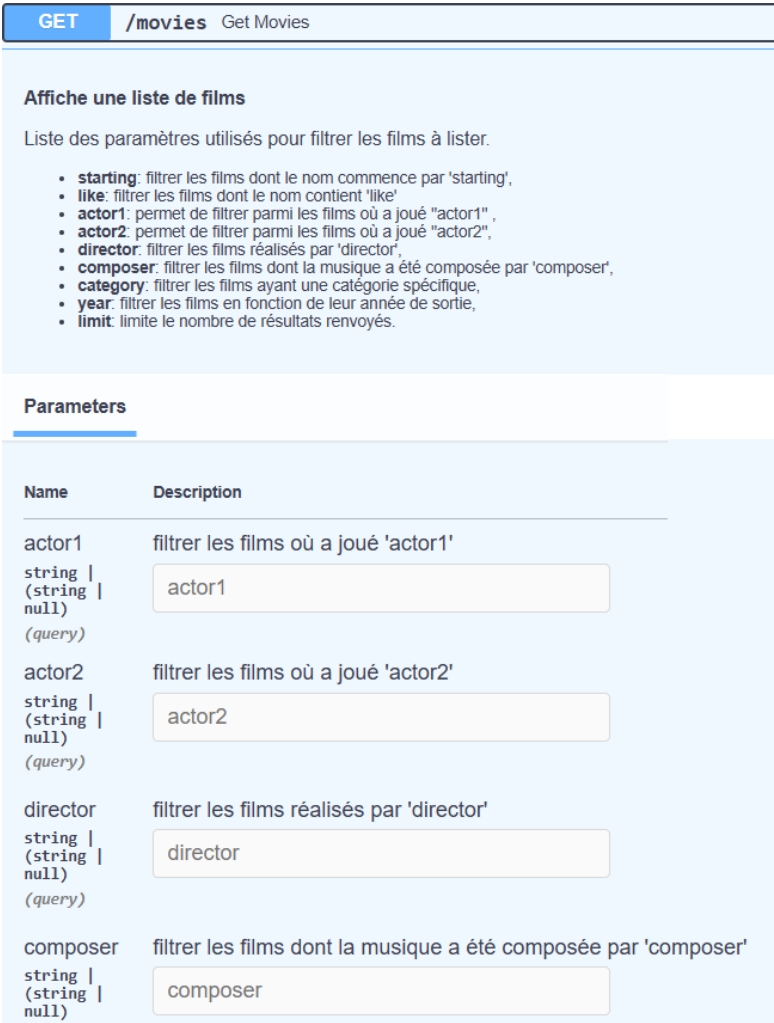


FIGURE 14 – Documentation API

7.3 Règle d'authentification

Nous ajoutons un système d'authentification pour que notre API ne soit disponible qu'aux utilisateurs possédant un mot de passe.

Nous utilisons le système de Tokenisation JWT (JSON Web Token), c'est un standard pour communiquer des données sécurisées à travers des objets JSON (signée numériquement).

Fonctionnement de JWT :

- 1 Le client (navigateur) envoie au serveur les informations d'authentification (password),
- 2 Si l'authentification est acceptée, le serveur génère un jeton JWT (signé avec une clé secrète) et le transmet au client,
- 3 le client peut envoyer des requêtes pour accéder à des ressources, il y joint son token,
- 4 Si le token est bon, le serveur envoie au client les ressources demandées.

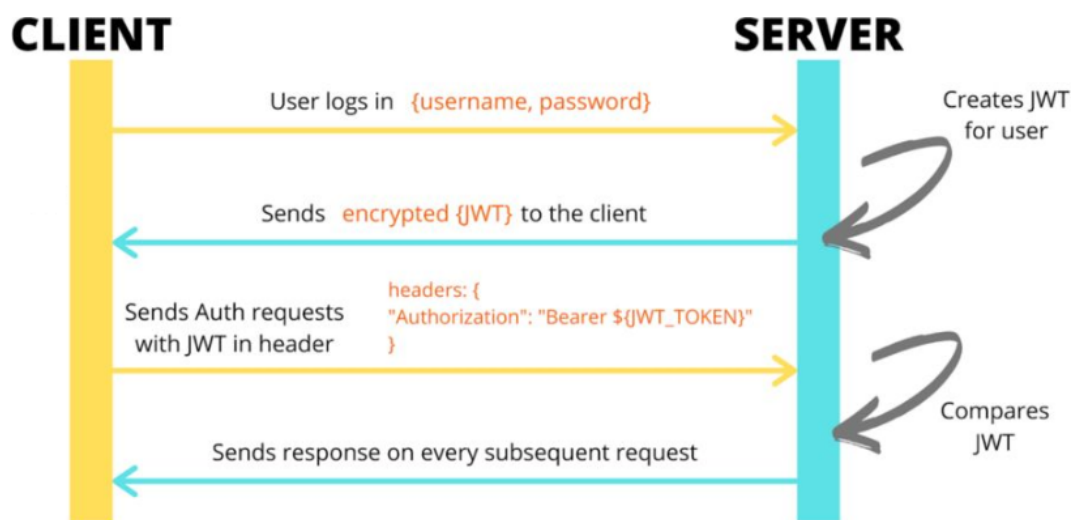


FIGURE 15 – Authentification JWT

7.4 Exemples d'utilisation de l'API

Un système de filtres a été mis en places pour pouvoir filtrer les films à partir de plusieurs noms d'acteurs et/ou du réalisateur et/ou du compositeur. Des filtres SQL pour faire des jointures entre les tables de films, acteurs, réalisateurs et compositeurs et un filtres pymongo pour extraire des informations de la base MongoDB.

GET	/movies	Get Movies
actor1	filtrer les films où a joué 'actor1'	
string (string null) (query)	<input type="text" value="danny glover"/>	
actor2	filtrer les films où a joué 'actor2'	
string (string null) (query)	<input type="text" value="mel gibson"/>	

Code	Details
200	<p>Response body</p> <pre>[{ "movie_id": "5ea69def-2d58-4268-a381-ad982c6ec98a", "title": "L'Arme fatale", "release_date": "1987-08-05" }, { "movie_id": "889bb96b-5791-4ddb-a8db-f3a57975b7e2", "title": "L'Arme fatale 3", "release_date": "1992-08-12" }, { "movie_id": "9ced673a-95bb-4869-81bd-a56b1adbfaa9", "title": "L'Arme fatale 4", "release_date": "1998-07-22" }, { "movie_id": "c247d44b-6bb9-4ba9-a1fc-94f8ce96675e", "title": "Maverick", "release_date": "1994-08-03" }, { "movie_id": "fa6dcca0-239f-428d-a357-7a8bf71e9a30", "title": "L'Arme fatale 2", "release_date": "1989-08-02" }]</pre>

FIGURE 16 – Requête intersection SQL

GET	/movie	Get Movie
Affiche les informations d'un film		
Paramètre : id: identifiant du film recherché.		
Parameters		
Name	Description	
id	identifiant du film recherché	
string (string null)	<input type="text" value="5ea69def-2d58-4268-a381-ad982c6ec98a"/>	
Response body		
<pre>{ "movie": [{ "title": "L'Arme fatale", "release date": "1987-08-05", "url_thumbnail": "https://m.media-amazon.com/images/M/MV5BMWVlNmZlODktMzhhNS00YTdhLWZlZWYtYzU0NDZlMzU0MGVlXkE5XkFqcGc@._V1_SX300.jpg", "plot": "Two newly paired cops who are complete opposites must put aside their differences in order to catch a gang of drug smugglers." }], "actors": [{ "actor_id": "113ddf76-6373-4bd0-b2e6-64282afeb087", "actor_name": "Steve Kahan" }, { "actor_id": "11c5e3d0-d1bb-4539-ab7e-133b871d1e85", "actor_name": "Mel Gibson" }, { "actor_id": "16a8bc/d-b96b-4cad-a3f5-c8046772c9c5", "actor_name": "Lycia Naff" }] }</pre>		

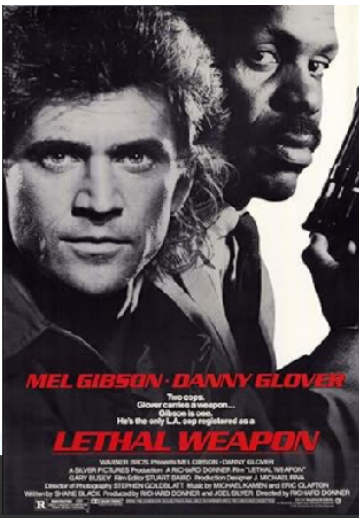


FIGURE 17 – Requête MongoDB

8 Automatisation

8.1 Présentation de Crontab

Crontab est une application d'automatisation de tâches sur système Unix et Linux, plus précisément elle permet l'exécution de scripts en arrière-plan à des moments précis. Nous utiliserons Crontab pour collecter les informations des nouveaux films de la semaine. pour cela nous créerons un environnement sur une machine virtuelle Linux WSL puis une tâche qui exécutera notre script tous les mercredis matins à 8h (date de sortie des films).

8.2 Création d'un environnement virtuel sous WSL

```
python -m venv env_allocine
```

Activation de l'environnement

```
source env_allocine/bin/activate
```

Installation des packages nécessaires à l'exécution du script

```
pip install pandas requests beautifulsoup4 mysql-connector-python
```

8.3 Script à exécuter

Notre script comportera les étapes suivantes :

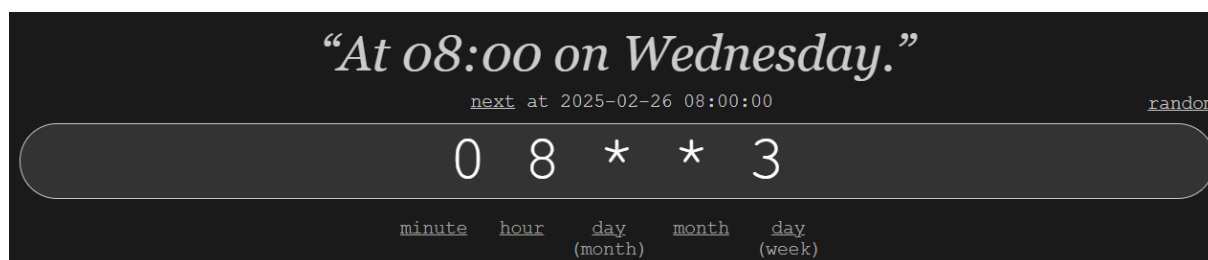
- Scrapping des films de la semaine à partir de l'url dédiée sur allocine.fr,
- Requête de OMDB pour récupérer des informations supplémentaires,
- Enregistrement des données en fichiers CSV.

8.4 Création d'une tâche Crontab

Préparation de l'environnement et création de la tâche Crontab

- 1 Création sous WSL d'un environnement virtuel venv : `python -m venv env`
- 2 Activer l'environnement : `source /env/bin/activate`
- 3 Installer les librairies nécessaires à l'utilisation du script "numpy", "pandas", "requests", "beautifulsoup4", "httpx", "selenium", "mysql-connector-python"
- 4 Copie du script.py dans WSL
- 5 Création de la tâche crontab qui doit se lancer chaque mercredi à 8 heures du matin, ajout d'information dans un fichier log.

```
0 8 3 * * cd /home/franck/testCron && . ~/testCron/env_allocine/bin/activate
&& cd /home/franck/testCron && python script.py »
~/testCron/allocine.log
```



Minute (0-59), Heure (0-23), Jour du mois (1-31), Mois (1-12), Jour de la semaine (0-7, où 0 et 7 représentent dimanche)

Problème lors de l'exécution de la tâche : il ne trouve pas mes modules python