

Top 11 Europe FC 25

Scraping Web

13 janvier 2025

Lumière

Porte des Alpes

# Introduction

Le football est une passion commune à travers le monde, et pour nous, c'est bien plus qu'un simple sport. C'est un domaine qui nous fascine par sa richesse en données et ses nombreuses possibilités d'analyse. En tant qu'amateurs de football et de statistiques, nous avons choisi de nous plonger dans l'univers des joueurs européens présentés sur le célèbre jeu EA Sports FC25.

Notre objectif est double : d'une part, recueillir et analyser les statistiques des joueurs des clubs européens afin de mieux comprendre leurs performances et leurs spécificités, et d'autre part, utiliser ces données pour imaginer et concevoir la meilleure équipe possible. Grâce à cette approche, nous espérons identifier des tendances, des points forts et des faiblesses, tout en mettant en lumière les données les plus intéressantes pour les amateurs de football comme nous.

Ce projet, au croisement entre notre passion pour le football et notre intérêt pour l'analyse de données, représente une opportunité idéale pour explorer ces deux mondes et en tirer des enseignements captivants.

# Site Web utilisé :

Pour se faire nous avons choisis le site de FC25 qui référencie les statistiques de tous les joueurs recensé :

<https://www.ea.com/fr/games/ea-sports-fc/ratings>

# Répartition du travail :

La Collecte et le Traitement sera supervisé par Lucas et l’analyse et la visualisation sera supervisé par Ugo.

Table des matières

[Introduction 1](#_Toc187650892)

[Site Web utilisé : 1](#_Toc187650893)

[Répartition du travail : 1](#_Toc187650894)

[1 ère étape, Collecte : 3](#_Toc187650895)

[- Implication : 3](#_Toc187650896)

[- Légalité et éthique du scraping : 3](#_Toc187650897)

[- Choix de la librairie : 3](#_Toc187650898)

[- Docker File : 4](#_Toc187650899)

[- Dossier www : 4](#_Toc187650900)

[- Dossier dump : 4](#_Toc187650901)

[- Architecture du repository : 5](#_Toc187650902)

[Création d’un repository GitHub : 5](#_Toc187650903)

[Déployer et sauvegarder le repository : 7](#_Toc187650904)

[- Paramétrage de Github : 7](#_Toc187650905)

[- Déploiement du repository : 8](#_Toc187650906)

[- Sauvegarde du repository : 9](#_Toc187650907)

[Conclusion : 10](#_Toc187650908)

# 1 ère étape, Collecte :

## Implication :

Ugo s’est occupé de la recherche de la thématique, de la recherche des sites web et de la vérification de la légalité à les scraper.

Lucas s’est occupé de la réalisation du script qui scrap les données du site web avec leurs intégrations dans un fichier CSV.

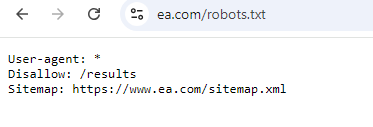
Pourcentage : Ugo -> 30%, Lucas -> 70%

## Légalité et éthique du scraping :

Lorsque l'on effectue du scraping, il est essentiel de respecter les règles et les limitations définies par les sites web que l'on souhaite analyser. Pour cela, le fichier robots.txt, accessible à la racine de nombreux sites, joue un rôle central. Ce fichier indique quelles sections du site sont autorisées ou interdites au scraping pour les robots. Par exemple, dans le cas du site EA.com, on observe que le fichier robots.txt interdit l'accès à la section /results, ce qui signifie qu'il ne faut pas collecter des données provenant de cette partie du site.

De plus, il est crucial de s'assurer que le scraping respecte les conditions d'utilisation du site, souvent précisées dans les "Terms of Service". Toute violation de ces conditions peut entraîner des sanctions légales ou des blocages. L'usage de données collectées doit également rester dans un cadre légal, en évitant tout usage commercial non autorisé ou atteinte à la propriété intellectuelle.

En adoptant une approche éthique et en se limitant à des données accessibles au public sans restrictions explicites, le scraping peut être utilisé comme un outil puissant d'analyse de données, tout en respectant les droits des éditeurs de sites web.



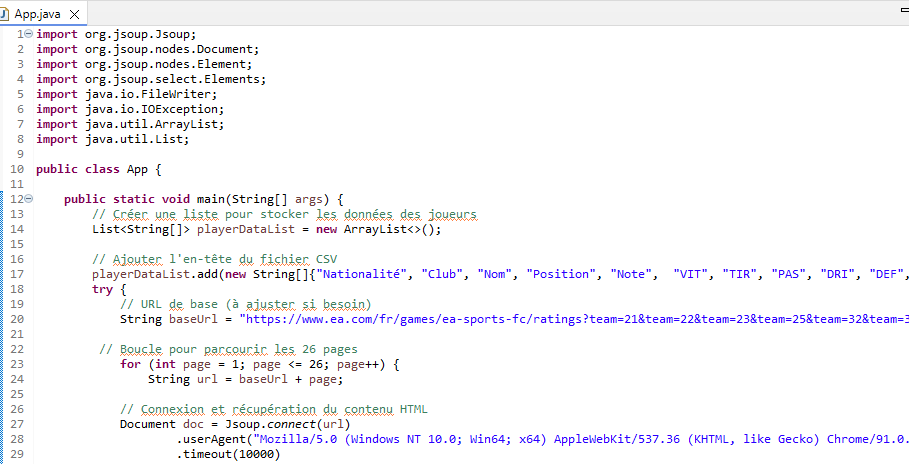
## Choix de la librairie :

Pour effectuer le scraping, nous avons choisi d'utiliser la librairie **JSOUP** en raison de sa simplicité d'utilisation et de sa performance. JSOUP est une librairie Java puissante qui permet de manipuler et d'extraire facilement des données HTML. Elle offre des fonctionnalités complètes pour se connecter à des pages web, parcourir leur structure DOM, et extraire les éléments souhaités à l'aide de sélecteurs CSS, ce qui est particulièrement pratique pour analyser des sites organisés.

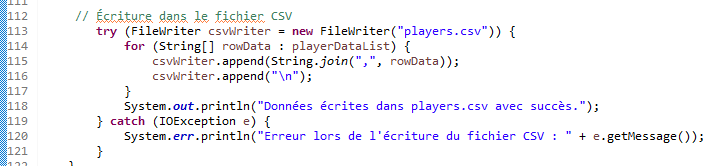
De plus, JSOUP intègre des mécanismes robustes pour gérer des cas complexes, comme les pages mal formées ou des structures HTML non standard, ce qui améliore la fiabilité du scraping. Sa documentation claire et son intégration fluide avec des projets Java la rendent idéale pour des projets nécessitant une analyse approfondie et rapide des données.

## Script et compilation :

Réalisation du code :



Compilation avec insertion dans un document CSV :



# 2 ème étape, Traitement :

## Implication :

Lucas s’est occupé de rendre conforme le contenu du fichier CSV pour l’intégration dans le logiciel d’analyse.

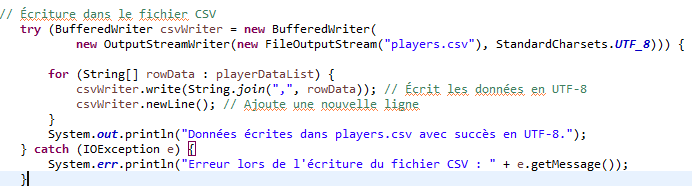
Ugo s’est occupé d’injecter correctement le fichier dans le logiciel d’analyse et de vérifier qu’il n’y a pas d’erreurs.

Pourcentage : Ugo -> 50%, Lucas -> 50%

## Conformité :

Soucis d’encodage dans le fichier CSV exemple d’une ligne : NationalitÃ©,Club,Nom,Position,Note,VIT,TIR,PAS,DRI,DEF,PHY

La prise en compte des accents est mauvaise donc je vais modifier ça :



L’ajout du StandardCharsets.UTF\_8 permet l’encodage correcte et d’afficher les accents.