Atelier - Pokémon

Xavier Gendre

Cet atelier reprend différents éléments vus durant cette formation. L'objectif est d'illustrer les relations entre les étapes étudiées et comment elles s'articulent dans le cadre d'un projet data simple. Pour cela, nous utiliserons des données librement accessibles relatives à l'univers des *Pokémon* issues d'un site web à scraper et d'une API à interroger pour stocker les résultat dans une base MongoDB à partir de laquelle nous définirons un pipeline d'agrégation élémentaire. Ainsi, ce projet va de la collecte des données au stockage et à la valorisation.

Les explications de chaque étape ne rentrent pas dans le détail de la mise en pratique afin de vous laisser libre de les implémenter comme vous le pensez et de vous confronter aux problématiques concrètes d'un projet data.

Préparation du projet

La première étape consiste à initialiser un dépôt Git et à créer un service VSCode associé dans Onyxia. Vous devrez faire des *commits* réguliers au fur et à mesure de votre progression. Il n'existe pas de règle absolue pour décider du moment d'un *commit* mais une bonne pratique consiste à s'assurer que le code est fonctionnel à chaque *commit*.

Web scraping

Le site https://scrapeme.live/product-category/pokemon/ se présente comme une boutique en ligne pour acheter des *Pokémon*. Il s'agit d'une plateforme d'entraînement au web scraping qui nous servira de point de départ pour notre projet.

Après une simple exploration du site et de la façon dont sont formées les URL des différentes pages, vous pourrez scraper les données suivantes pour chaque *Pokémon* (nous pourrons nous limiter aux premières pages pour limiter le temps de scraping en pratique) :

- son nom (*e.g.* Bulbasaur),
- son prix (*e.g.* £63.00 à convertir en valeur numérique),
- son poids (*e.g.* 15.2 kg à convertir en valeur numérique).

Le résultat pourra être stocké dans un *DataFrame* intermédiaire.

PokéAPI

Afin de compléter les données obtenues depuis le site web, nous utiliserons l'API pokeapi.co. La page d'accueil donne tous les détails sur son fonctionnement. En particulier, le nom du *Pokémon* en minuscules dans l'URL de requête permet d'obtenir des informations additionnelles au format JSON.

Vous utiliserez cet API pour enrichir votre base de données avec les informations suivantes :

- les capacités de chaque *Pokémon* (`abilities`),
- les statistiques de base (`hp`, `attack`, `defense`, `speed`),
- les types du *Pokémon* (`types`).

Le résultat viendra compléter le *DataFrame* précédent.

MongoDB

L'ensemble des données sera stocké dans MongoDB avec une collection `pokemons`. Il faudra veiller à ce que le type de chaque variable soit cohérent avec l'utilisation que nous pourrons en faire.

Pour finir, ces données seront valorisées par un pipeline d'agrégation qui calculera les valeurs moyennes du prix, du poids et des statistiques de base en fonction de chaque capacité, version ou type (selon votre choix).

L'idée du projet est que l'ensemble des étapes soit autonome et que nous puissions produire un rapport à partir des résultats (ce qui n'est pas attendu ici mais qui représente un prolongement de ce projet).

(Bonus) Vous pourrez produire quelques graphiques à partir des résultats pour les illustrer.