**Projet pour le scraping des données**

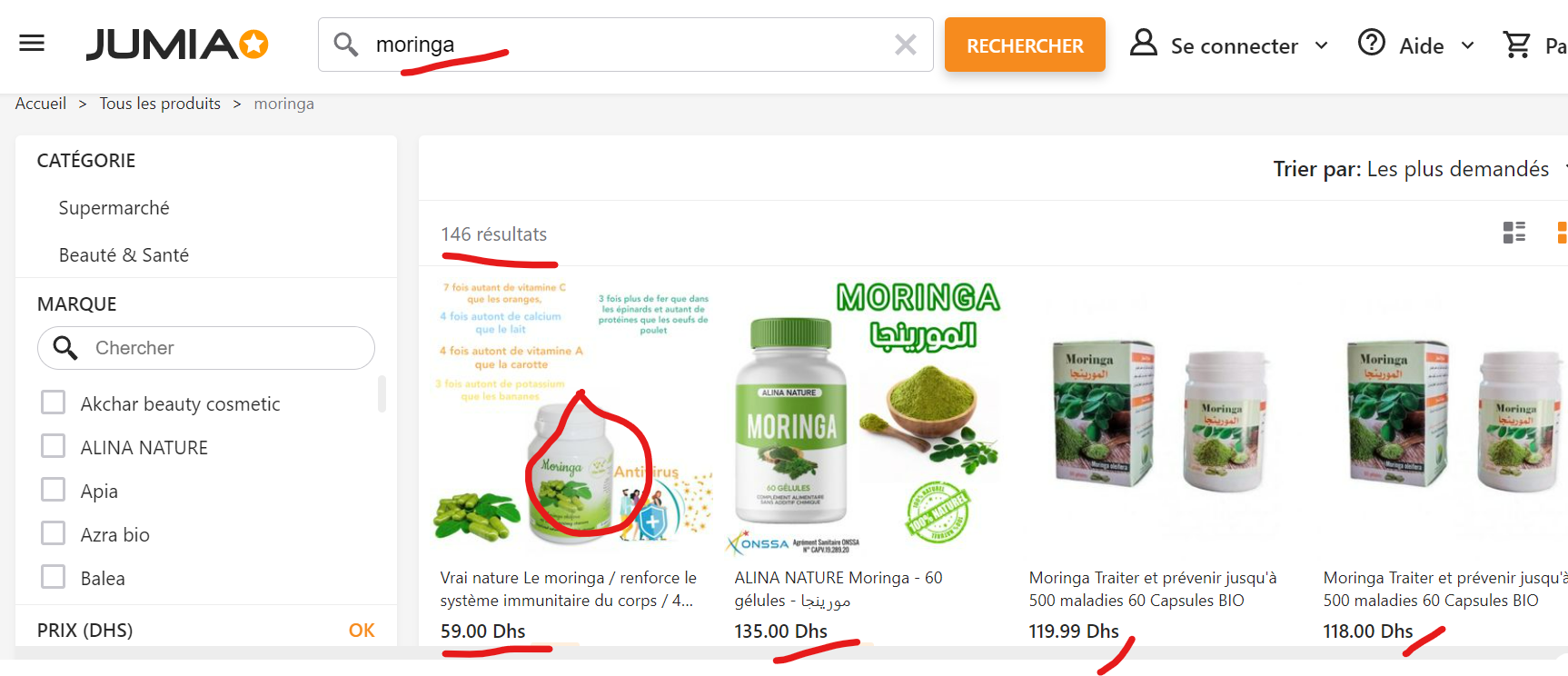
Par équipe de 4 vous devez scraper les données pour le projet qui vous est assigné.

1. **Catalogue de produits Amazon (1 seul type de produit)**

Extraire les données du site web Amazon.com relatives à un produit donné. Nous avons besoin du label du produit, du prix, la description et l’évaluation (review).

1. **Extraire les données de Jumia pour une gamme de produits en particulier**

On souhaite avoir toutes les données relatives au produit « moringa » sur le site jumia. La figure suivante illustre le résultat attendu.



1. **Les données concernant le Covid 19 du site web de l’OMS**

Les données relatives à l’évolution de l’infection covid par pays et par date pour les années 2021 et 2022

1. **Map.ma (**[**https://www.map.ma/**](https://www.map.ma/)**)**

Les nouvelles sur le site de presse map.ma

1. **The newyorktimes**

Les nouvelles sur le site de presse map.ma

Business: <https://www.nytimes.com/section/business>

Et

Health : <https://www.nytimes.com/section/health>

1. **Catalogue des cartes graphiques à partir du site newegg :** <https://www.newegg.ca/gigabyte-geforce-rtx-3080-ti-gv-n308tgaming-oc-12gd/p/N82E16814932436?Description=3080&cm_re=3080->

Liées aux cartes graphiques

1. **Avito Immo**

Extraction des données de l’immobilier sur Avito en vue de faire une analyse des prix et des tendances du marché. Les données seront classées par ville et éventuellement par d’autres critères. On souhaite analyser l’évolution des prix à travers le temps pour certaines villes.

Chaque équipe doit préparer un plan d’extraction de données à faire valider avant de procéder à l’extraction de masse. Il est très recommandé de tester les possibilités d’extraction avec un prototype avant de demander la validation pour s’assurer que vous êtes capable de réaliser l’extraction. La validation peut être demandé auprès de votre professeur, Mme Elasri en tant que consultante.

Les fichiers de données doivent être au format csv avec le séparateur « | »

Les données extraites vous serviront pour le projet ElasticSearch de ce cours et aussi pour le projet BigData. Par conséquent, un investissement dans la qualité des données extraites est requis. Dans ce sens, vous devez aussi présenter dans votre rapport la démarche que vous avez utilisée pour tester la fiabilité de votre extracteur.

La remise des données, du code et du rapport est prévue pour le mardi 18 octobre avant 23h55.