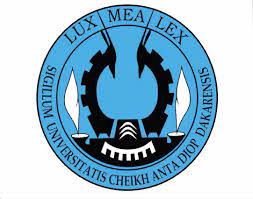


République du Sénégal

Ministère de l’enseignement supérieur et de la recherche

**UNIVERSITE CHEIKH ANTA DIOP**



Faculté des sciences et Techniques

Département de Mathématiques-Informatique

***MASTER 1 SIR JOUR***

**Projet de NOSQL : Couplage d’un web crawler à New SQL**

**Présenté par** : **Sous la direction de** :

Ndéye Khady Niang Pr Idrissa Sarr

Fatou Sow

Serigne Cheikh Ahmet Tidiane Sy Thioune

**Année universitaire 2021-2022**

TABLE DES MATIERES

[Introduction 4](#_Toc125752547)

[I. Le fonctionnement du crawler 4](#_Toc125752548)

[II. Le modèle de données utilisé 5](#_Toc125752549)

[III. L’interface de visualisation et les contraintes rencontrées 7](#_Toc125752550)

[Conclusion 10](#_Toc125752551)

TABLE DES FIGURES

[**Figure 1: Extraction de données du site de Jumia dans un fichier CSV** 5](#_Toc125752563)

[**Figure 2: Stockage des données du ficiher maquillage.csv de jumia dans Mongodb** 6](#_Toc125752564)

[**Figure 3: Stockage des données du fichier ordinateurimprimante.csv de jumia dans MongoDB** 6](#_Toc125752565)

[**Figure 4 : accueil de notre interface** 7](#_Toc125752566)

[**Figure 5: Navigation à travers les onglets** 8](#_Toc125752567)

[**Figure 6: Visualisation des réfrigérateurs** 8](#_Toc125752568)

[**Figure 7: Visualisation des maquillages** 9](#_Toc125752569)

[**Figure 8: Visualisation des tablettes** 9](#_Toc125752570)

[**Figure 9: Visualisation des iPhones et les détails de l’iPhone 6+** 10](#_Toc125752571)

## Introduction

Le web crawler est un programme informatique ou un script automatisé qui parcourt le World Wide Web d'une manière prédéfinie et méthodique pour collecter des données. Ils recherchent sur le web les meilleurs résultats et vous les présentent dans un format indexé.

Le succès du web crawler pour l'indexation des recherches ou l'extraction de données en a fait un outil populaire pour les scientifiques des données, les spécialistes des moteurs de recherche et les grattoirs pour extraire des informations utiles à partir de sources en ligne.

Tout d’abord nous commencerons par vous expliquez le fonctionnement du crawler, ensuite le modèle de données qu’on a utilisé et enfin l’interface de visualisation et les contraintes rencontrées.

## Le fonctionnement du crawler

Le web crawler fonctionne en commençant par une liste d’URL connues, en examinant puis en catégorisant les pages Web. Avant de passer en revue chaque page, le crawler examine le fichier robots.txt de la page Web, qui spécifie les règles pour les robots qui accèdent au site Web. Ces règles définissent quelles pages peuvent être explorées et les liens qui peuvent être suivis.

Pour accéder à la page web suivante, le crawler trouve et suit les liens hypertexte qui apparaissent. Le lien hypertexte suivi par le robot dépend de stratégies définies qui le rendent plus sélectif quant à l’ordre que le robot doit suivre. Par exemple, les stratégies définies peuvent inclure les éléments suivants :

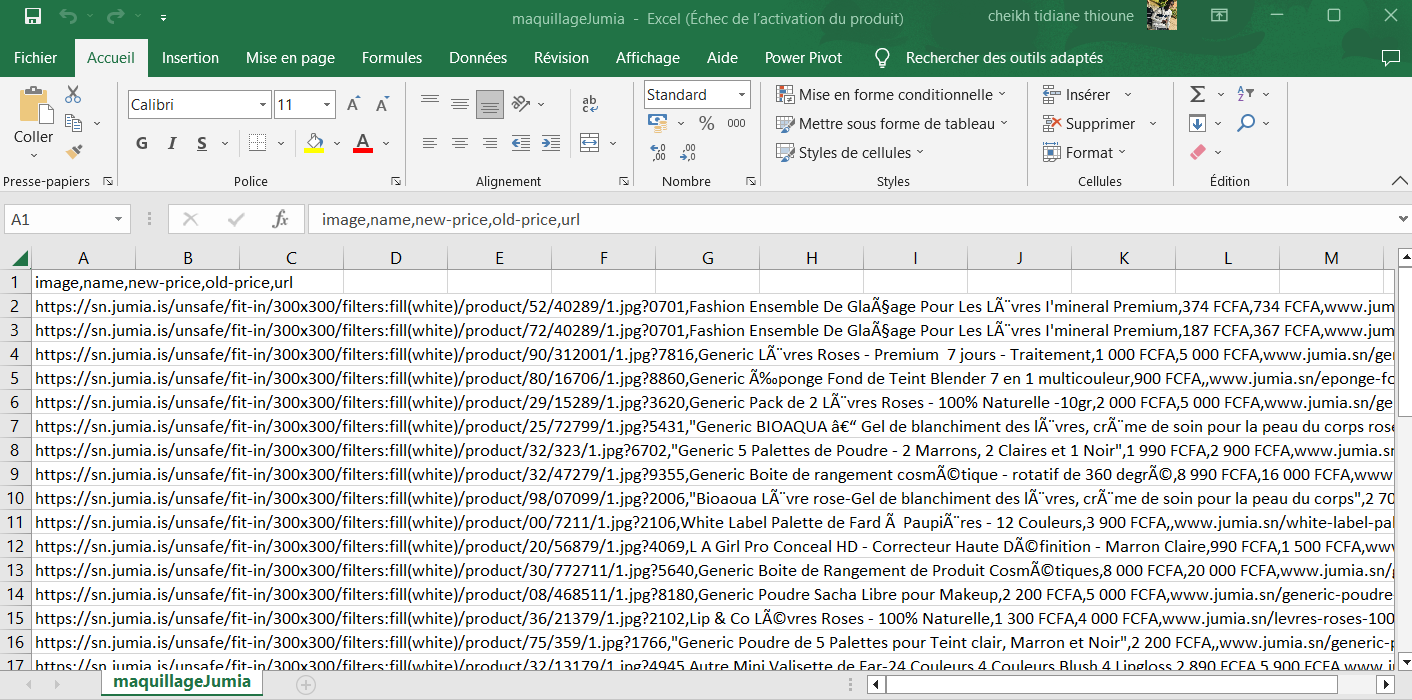
* Combien de pages sont liées à cette page ;
* Le nombre de pages vues ; etc.

Ces facteurs signifient qu’une page peut contenir des informations plus importantes pour l’indexation.

Sur une page Web, le robot d’exploration stocke la copie et les données descriptives appelées balises meta, puis les indexe pour que le moteur de recherche de trouver les mots-clés. Ce processus décide ensuite si la page apparaîtra dans les résultats de recherche pour une requête et, le cas échéant, renvoie une liste de pages Web indexées par ordre d’importance.

Dans le cas où un propriétaire de site Web ne soumet pas son plan de site pour que les moteurs de recherche explorent le site, crawler peut toujours trouver le site web en suivant les liens des sites indexés qui y sont liés.

Exemple : Extraction de données du site web jumia stockées dans un fichier csv.



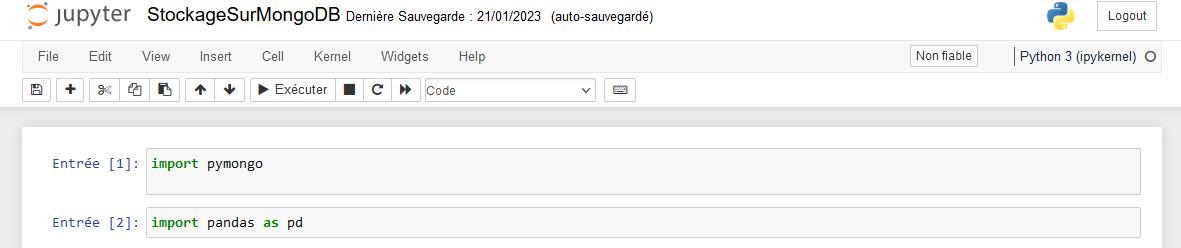
**Figure 1: Extraction de données du site de Jumia dans un fichier CSV**

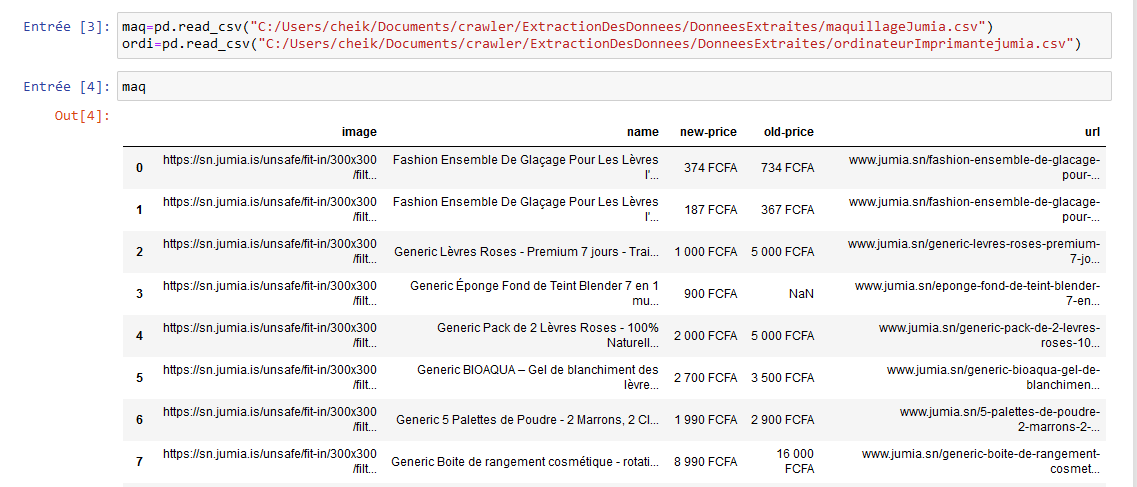
## Le modèle de données utilisé

Nous avons utilisé MongoDB qui est une base de données NOSQL de document, MongoDB nous permet de stocker facilement des données structurées et non structurées. Il utilise un format de type JSON pour stocker des documents. Ce format correspond directement aux objets natifs dans la plupart des langages de programmation modernes, ce qui en fait un choix naturel, car nous n’avons pas besoin de penser à normaliser les données. MongoDB peut également gérer des volumes élevés et peut évoluer verticalement ou horizontalement pour prendre en charge de grandes charges de données.

Voici des exemples de stockage des données que nous avons collecter depuis le site de jumia sur MongoDB.

Ci-dessous le code qu’on a crée avec un Notebook d’anaconda.

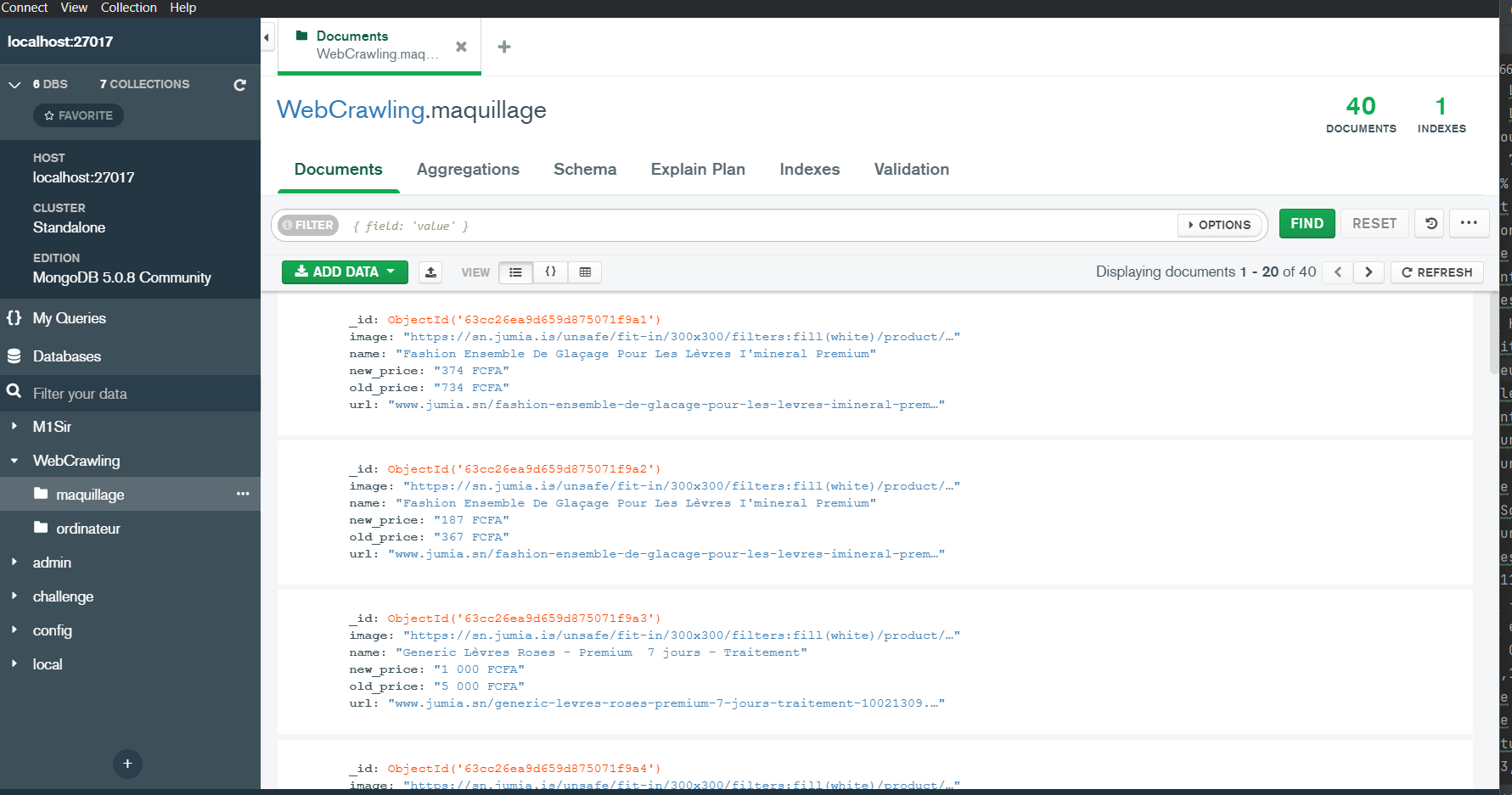






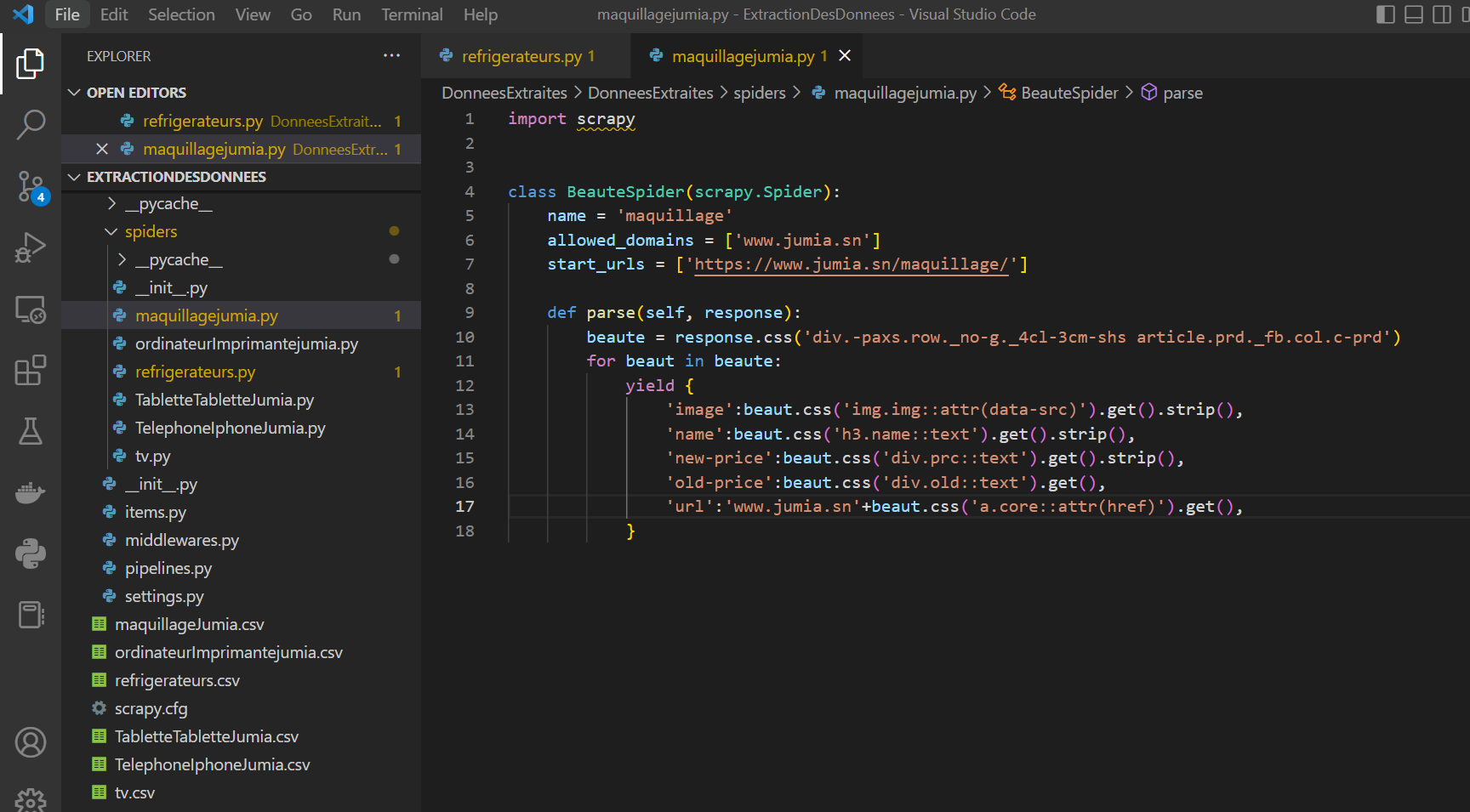


Les données importées sont bien dans MongoDB.

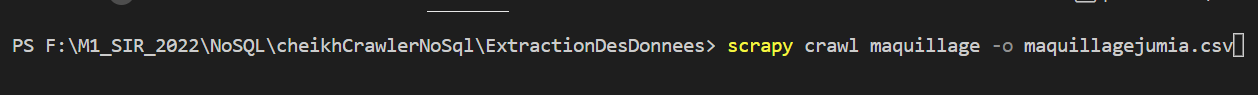


**Figure 2: Stockage des données du ficiher maquillage.csv de jumia dans Mongodb**

Voici un exemple de l’extraction de données avec scrapy en utilisant un code python



Voici un exemple de la commande d’exécution pour générer les fichier csv.

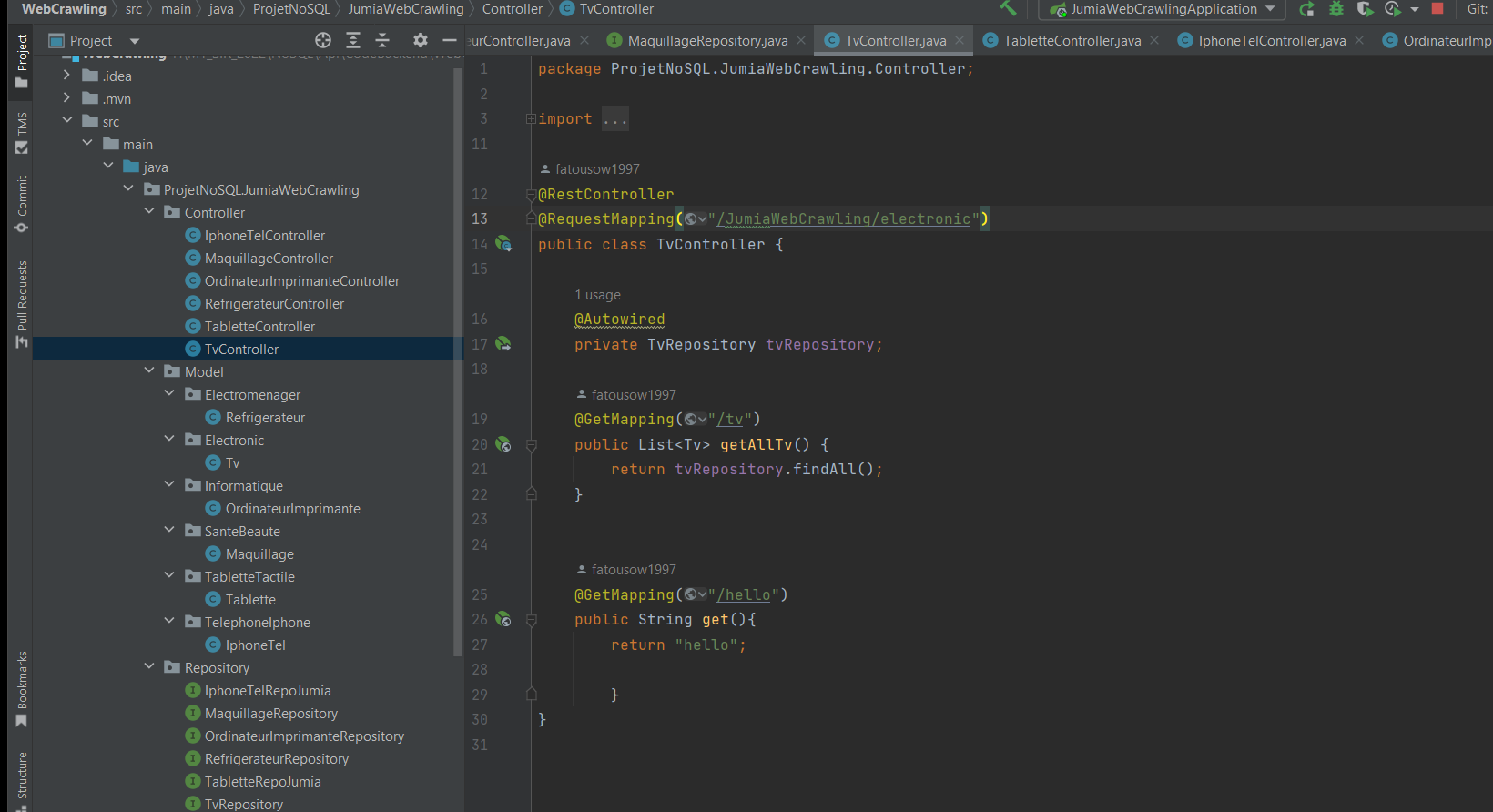


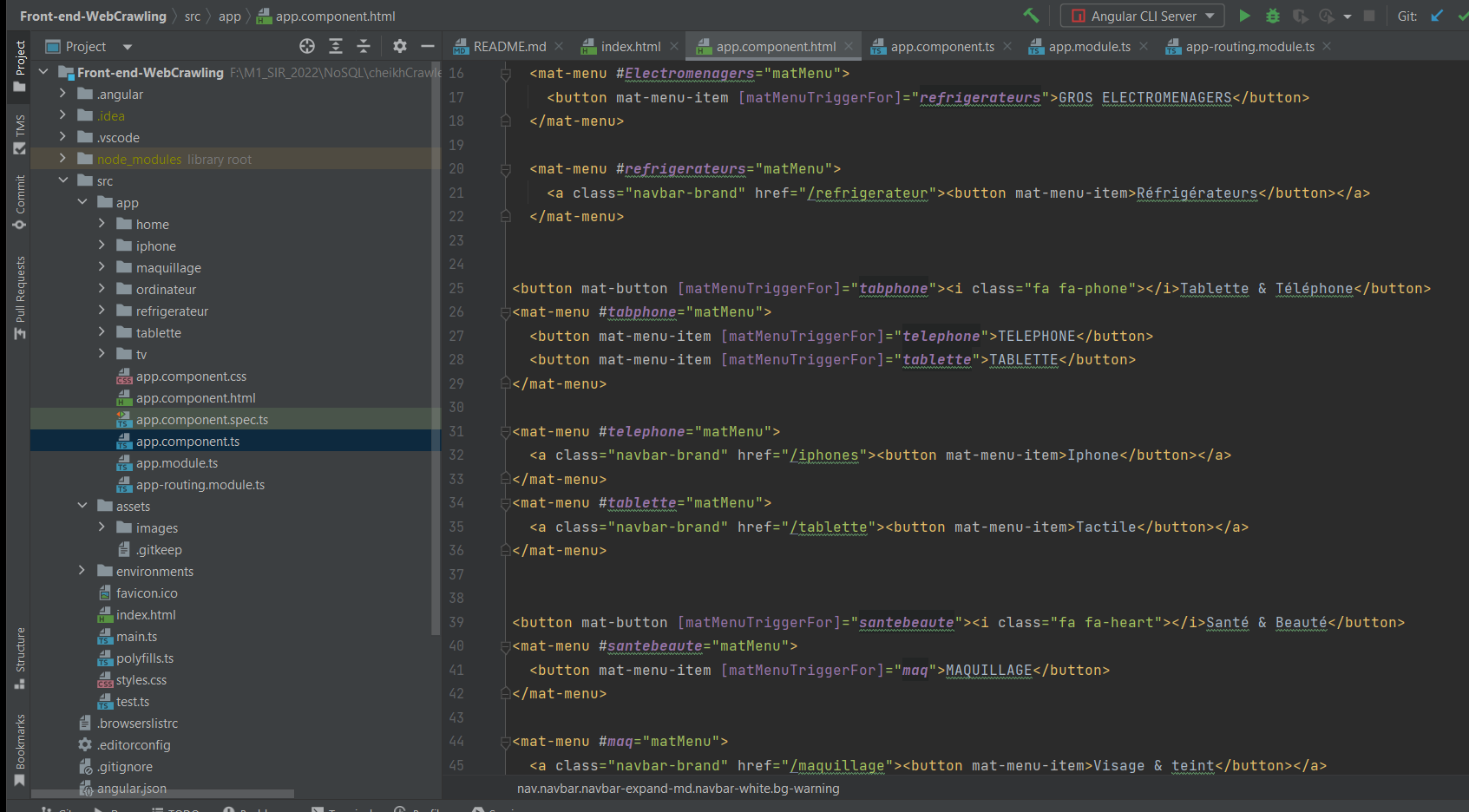
## L’interface de visualisation et les contraintes rencontrées

* L’interface d’utilisation

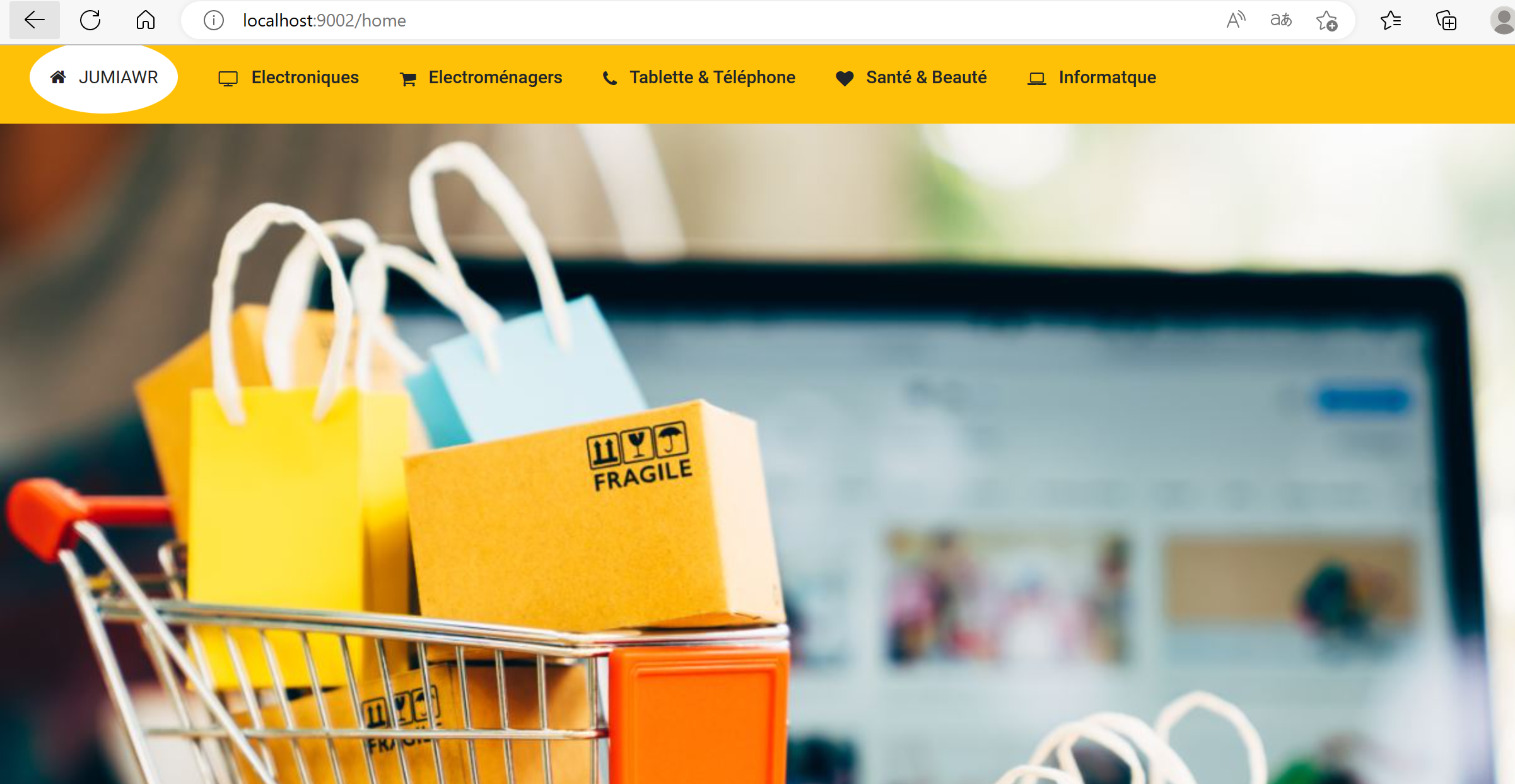
Pour pouvoir visualiser les stocker dans MongoDB, nous avons créer une API et un client angular.

Voici quelques extraits de codes backend et front qu’on a créé

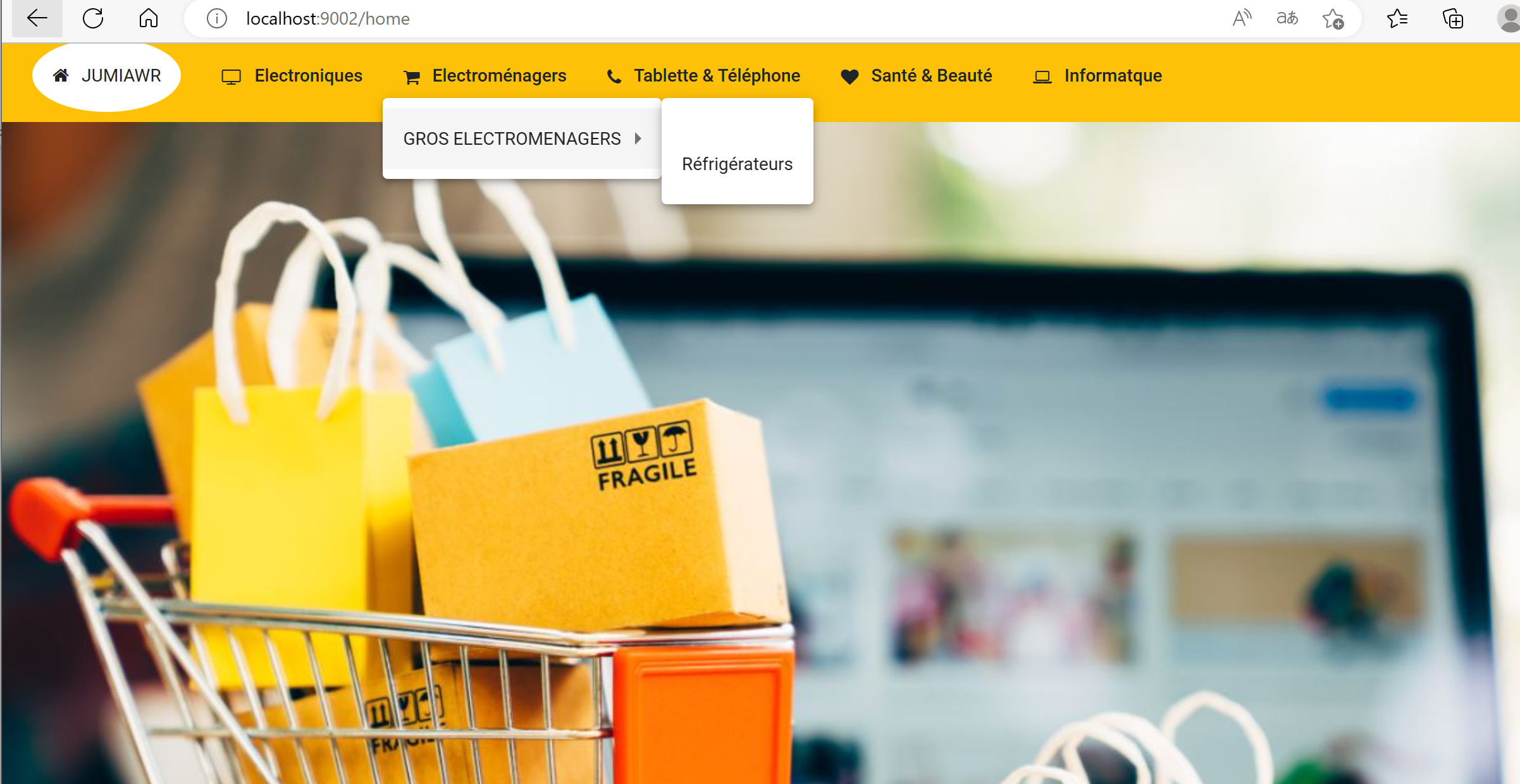




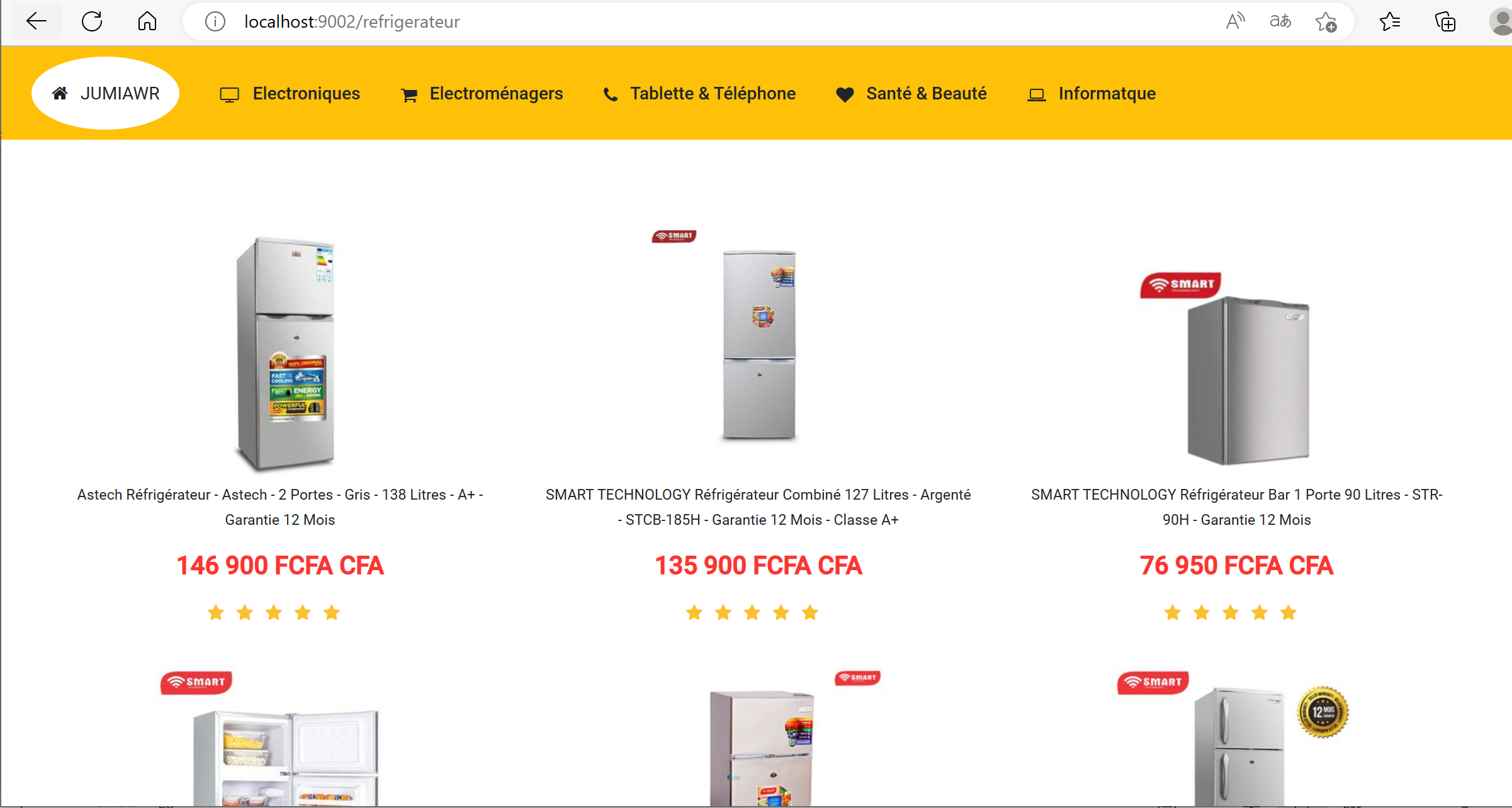
Ci-dessous la page d’accueil de notre interface angular avec les onglets des informations récupérées depuis le site de jumia à savoir Electronique, Electroménagers, Tablettes & Téléphone, Santé & Beauté et Informatique.



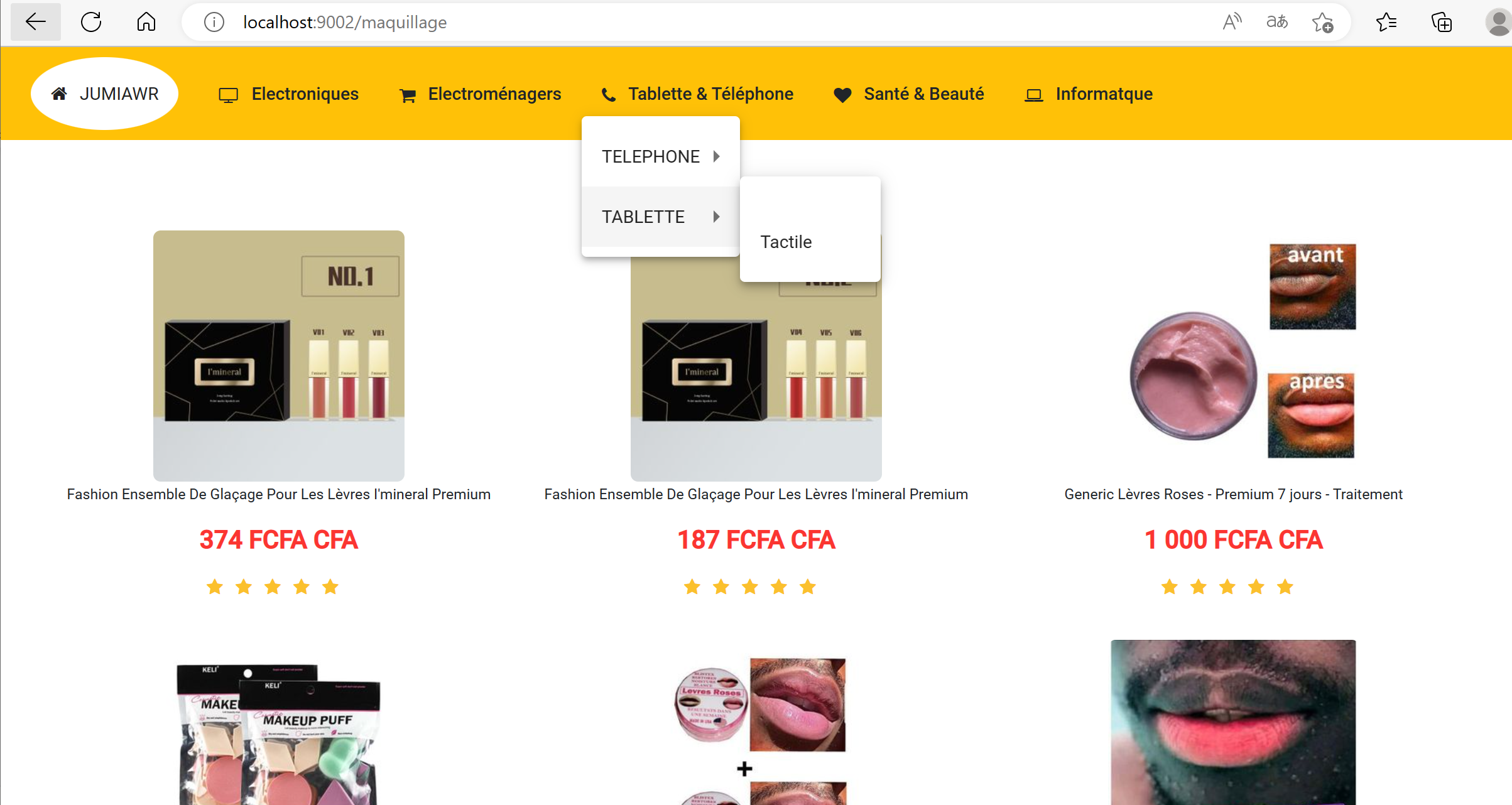
**Figure 4 : accueil de notre interface**



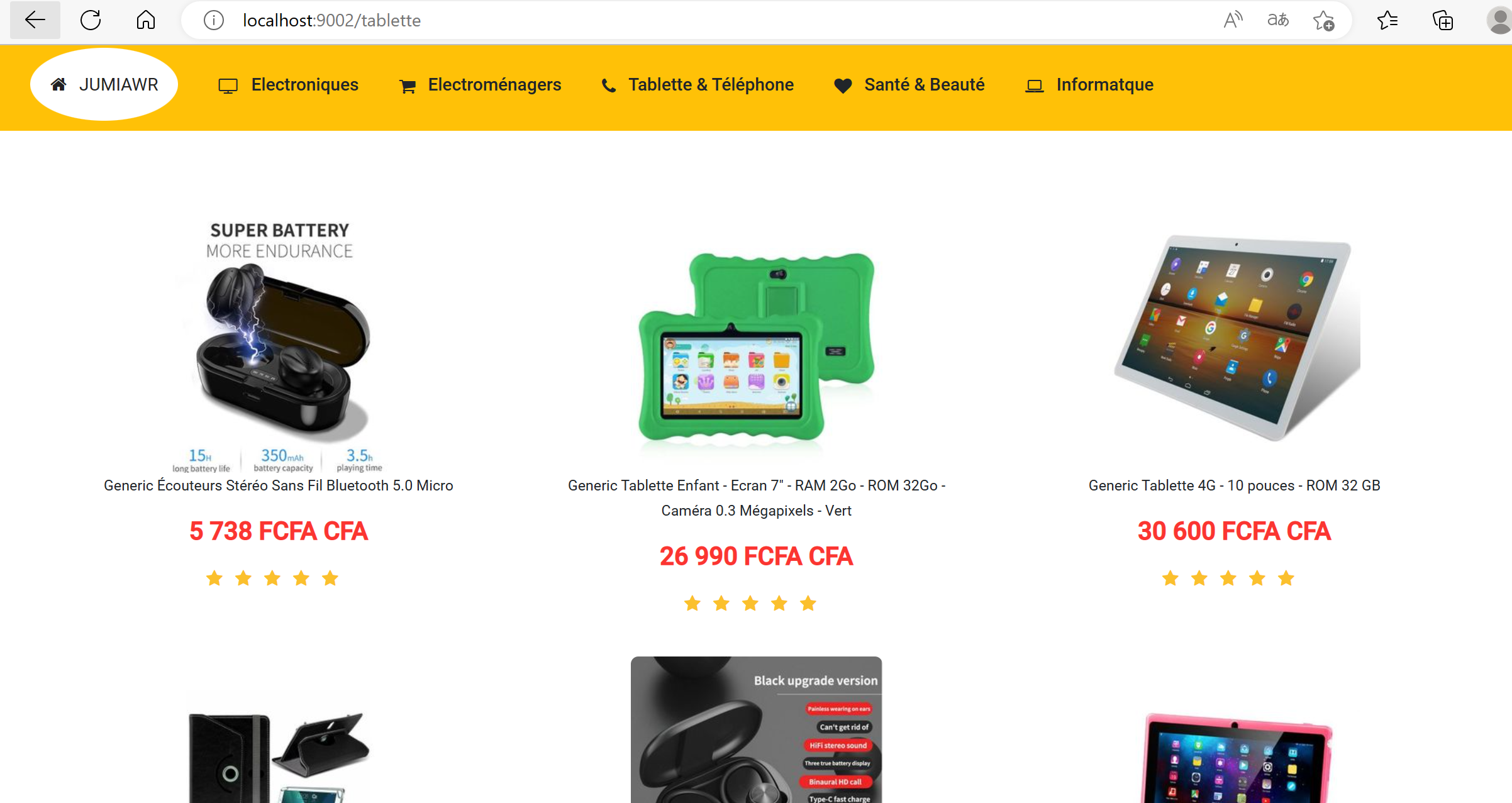
**Figure 5: Navigation à travers les onglets**



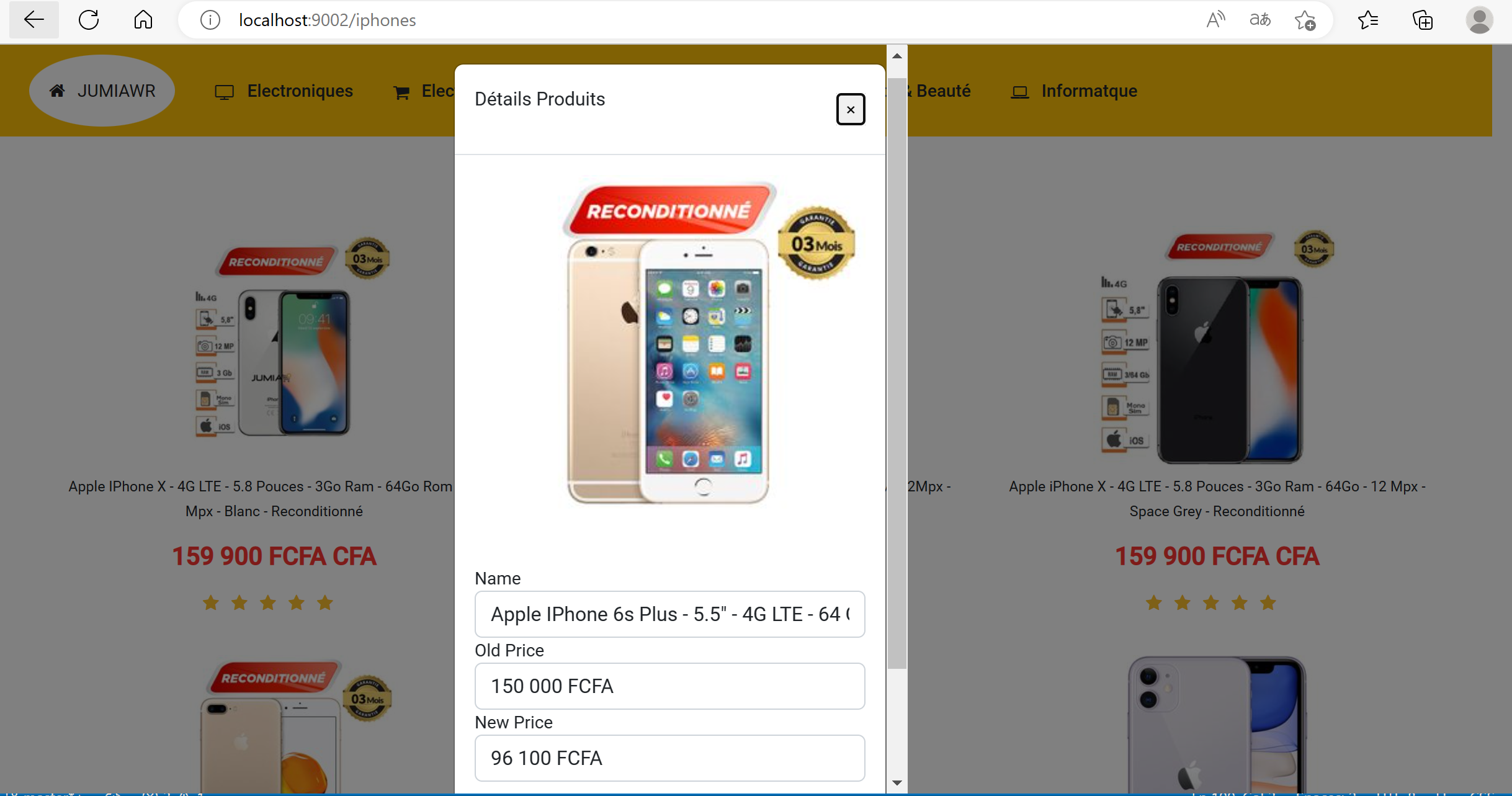
**Figure 6: Visualisation des réfrigérateurs**



**Figure 7: Visualisation des maquillages**



**Figure 8: Visualisation des tablettes**



**Figure 9: Visualisation des iPhones et les détails de l’iPhone 6+**

* Les contraintes rencontrées

Nous avions eu quelques difficultés pour l’extraction de données depuis le site de jumia, ça prenait des heures mais finalement on l’a résolu en indexant les pages.

## Conclusion

En résumé, les crawler parcourent systématiquement les pages web pour savoir de quoi parle chaque page du site web, afin que ces informations puissent être indexées, mises à jour et récupérées lorsqu’un utilisateur effectue une requête de recherche. Nous avons mis en évidence des aspects importants de ce qu’est un web crawler, comment cela fonctionne, le modèle de données que nous avons et l’interface angular que nous avons créé pour mettre en œuvre notre travail.