FACULTE DES SCIENCES BEN M'SIK
UNIVERSITE HASSAN II DE CASABLANCA

Université Hassan II de Casablanca
Faculté des Sciences Ben M'Sick
Département de Mathématiques et Informatique

Mémoire de fin d'étude pour l'obtention de la
Licence Sciences Mathématiques et Informatique
Option : Base de données

Sujet :

SCRAPING ET ANALYSE DE TEXTES EN DIALECT MAROCAIN

Encadré par :

Pr BENLAHMAR EL HABIB

Réalisé par :

EL MESKINI Rania

ELAOUDI Oussama

Soutenu le 17/ 06/ 2022 , devant le jury :

BENLAHMAR EL HABIB	Président
ELFILALI SANNA	Examinateur
ZAHOUR OMAR	Examinateur
OUAHABI SARA	Examinateur

Année Universitaire : 2021-2022

Remerciement :

Arrivée au terme de notre formation, nous tenons à remercier en premier temps toute l'équipe pédagogique de la faculté des sciences Ben Msik pour avoir assuré notre formation.

Nous remercions également monsieur ELHABIB BENLAHMAR qui n'a pas cessé de nous encourager pendant la durée du projet, ainsi pour sa générosité en matière de formation et d'encadrement qu'il nous a apporté lors des différents suivis et la confiance qu'il nous a témoigné.

Nous remerciements s'adressent aussi à l'ensemble de membres de jury PR ZAHOUR OMAR et PR ELFILALI SANNA qui nous ont fait l'honneur de bien vouloir étudier notre travail avec attention.

Dédicaces :

Nous dédions ce modeste travail et notre profonde gratitude et amour :

À nos chers parents :

Aucune dédicace suffira à exprimer notre grand amour, immense respect et reconnaissance envers les personnes qui ont tous sacrifier afin de nous offrir une meilleure éducation et vie, nous ne pourrions jamais oublier leurs tendresse et amour.

À ma grand-mère :

La femme qui m'a comblé d'amour et d'affection, qui a œuvré toujours pour ma réussite.

À nos frères et sœur :

Qui n'ont pas cessé de nous conseiller, encourager et soutenir tout au long de ce parcours.

À nos amis :

Notre seconde famille pour leur appuis, soutien moral et leurs aides durant toute cette année.

Résumé :

Le web scraping est une technique permettant l'extraction de donnés d'un site web via un programme ou un logiciel automatique. L'objectif est donc d'extraire les données d'une façon structurée afin de le pouvoir les réutiliser et les analyser en classifiant les textes extrait soit par sentiments, par sujet ou par intentions.

Dans ce travail, nous appliquons le web scraping sur le site web JUMIA afin d'extraire l'ensemble de commentaires sur des articles de différentes catégorie et ensuite les analyser pour savoir la polarité d'avis des clients du site web (positive, négative, ou neutre).

Nous présentons tout au long de ce rapport, les étapes suivies ainsi que les outils et les techniques utilisées afin de réaliser ce projet.

Abstract :

Web scraping is a technique that allows the extraction of data from a website via a program or an automatic software. The objective is to extract the data in a structured way in order to be able to reuse and analyze them by classifying the extracted texts either by feelings, by subject or by intentions.

In this project, we apply web scraping on the JUMIA website in order to extract the set of comments on articles of different categories and then analyze them to know the polarity of customers opinions on the website (positive, negative, or neutral).

Throughout this report, we present the steps followed as well as the tools and techniques used to used in order to realize this project.

Table des matières

Liste des figures :

Introduction générale :

Le web scraping désigne le processus d'extraction de données, c'est une méthode automatique qui permet d'obtenir de grande quantité de données à partir de sites web. Ces données sont des données non structurées au format HTML qui seront ensuite transformé ou converties en format plus utile pour l'utilisateur dans un tableur ou une base de données dans le but de pouvoir être ré-utilisées par la suite dans plusieurs applications.

Le text Mining ou analyse de texte est une pratique qui permet aux machines de comprendre et de traiter le langage humain automatiquement en classifiant les textes par sentiment, par sujet ou par intention. L'analyse de texte est souvent utilisée par des entreprises sur leurs sites web pour savoir le feedback des clients et leur besoin.

Dans ce projet, nous allons extraire un ensemble de commentaires d'un site web et les structurer d'une manière plus simple et efficaces pour pouvoir ensuite les analyser et les classifier afin de savoir s'ils sont positifs, négatifs ou neutres.

Ce projet est proposé par le professeurs Mr BENLAHMAR EL HABIB pour faire l'objet de notre projet de fin d'étude. Ce projet nous permettons d'avoir un sujet à traiter sous la lumière des connaissances acquises tout au long de nos études.

Ce rapport présente les différentes étapes de réalisation de notre projet. Il comporte deux parties dont la composition est comme suit : La première partie concerne le scraping du site web JUMIA afin de collecter le maximum possible de commentaires sur des produits de différentes catégorie. La deuxième partie est dédié à l'analyse de ces commentaires en leurs classification par sentiments. Nous conclurons par la suite par un bilan personnel et les problèmes rencontrer.

Chapitre 1 : contexte général du projet

Introduction :

Dans ce chapitre, nous passerons en revue les connaissances de base sur notre projet. Nous présenterons aussi une problématique.

1.1 Problématique :

Pour extraire les données ou bien précisément les commentaires d'un site web il faut en premier temps savoir leur emplacement et parfois il existe des articles sans aucun commentaires du coup la collection des informations pose quand même un problème.

En outre, pour avoir une vraie analyse des commentaires et avoir les feedbacks valables, il faut travailler sur un nombre massif de commentaires.

Figure 1 - un article sans commentaire

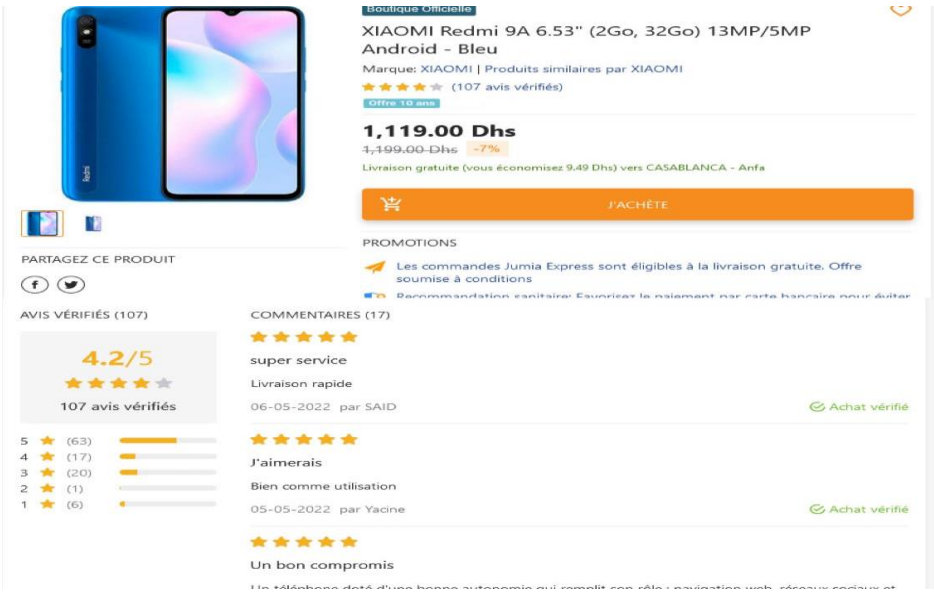

Figure 2 - un article avec des commentaires

1.2 Objectifs :

Afin de réussir notre projet, nous nous sommes fixées des objectifs à atteindre à la fin de ce projet :

- Collecter le maximum de commentaires d'un site web
- Classifier ces commentaires en 3 catégories : positive, négative ou neutre.

1.3 Travail connexe :

Le web scraping est utilisé pour plusieurs taches. Il permet par exemple, de récolter rapidement des données de contact ou des informations spécifiques. Dans le domaine professionnel, le web scraping est souvent utilisé pour accéder à des avantages par rapport à des concurrents. Le scraping de données permet à une société de consulter tous les produits d'un concurrent er les comparer avec ses propres produits. Il apporte également une valeur ajoutée pour les données financières : les informations sont lues depuis un site Internet externe, transférées dans un tableau et peuvent ensuite être analysées et traitées.

Google constitue un bon exemple de Web scraping. Le moteur de recherche utilise cette technologie afin d'afficher des informations météorologiques ou des comparatifs de prix pour des hôtels et des vols. Parmi les comparateurs de prix, nombreux sont ceux à également utiliser le Web scraping afin d'afficher des informations de nombreux sites Internet et prestataires.

Conclusion :

Ce chapitre a permis de situer le contexte général du projet et d'expliquer la vision stratégique à laquelle il adhère. A ce niveau, il y a lieu de s'interroger sur les besoins que ce projet doit rencontrer, c'est ce que détaillera le prochain chapitre.

Chapitre 2 : analyse et conception

Introduction :

Dans ce chapitre, nous allons présenter toutes les méthodes et techniques utilisées pour obtenir les meilleurs résultats possibles.

2.1 Le site web :

Durant ce projet nous avons travaillé sur le site web JUMIA.

JUMIA est un site d'achats en ligne fondé en 2012 et ayant son siège à Casablanca, il fait parti du top 10 site web au Maroc et il est incontestablement la première destination du shoping en ligne. Le site web contient +300 000 produits de différentes catégories, et 8,5 millions du visite mensuelles. Nous avons posé notre choix sur le site web JUMIA parce qu'il

propose une grande quantité de commentaires et de données pour scraper et analyser par la suite.

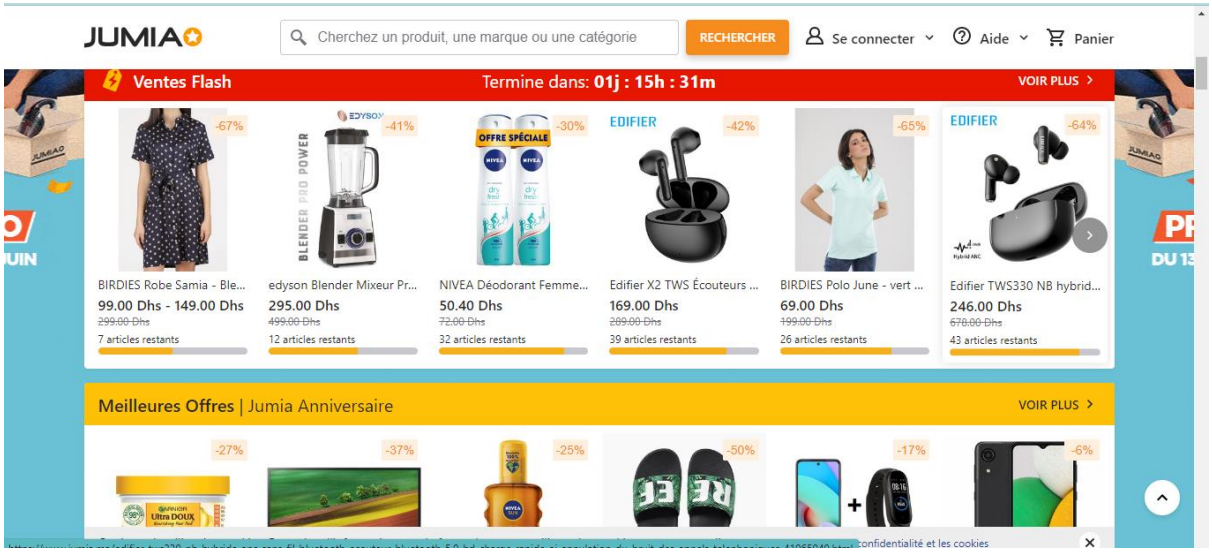

Figure 3 - site web JUMIA

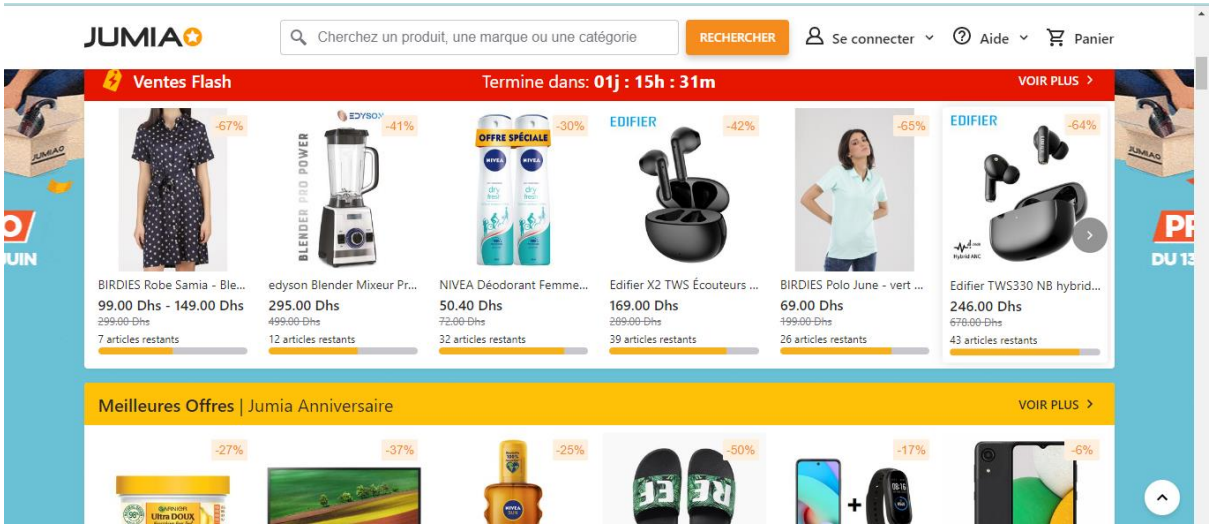

Figure 4 - site web JUMIA

2.2 Outils de développements :

- Python :

Figure 5 - logo python

Python est un langage de programmation puissant et facile à apprendre. Il dispose de structures de données de haut niveau et permet une approche simple mais efficace de la programmation orientée objet. Parce que sa syntaxe est élégante, que son typage est dynamique et qu'il est interprété, Python est un langage idéal pour l'écriture de scripts et le développement rapide d'applications dans de nombreux domaines et sur la plupart des plateformes. Ainsi, Python contient plusieurs bibliothèques du machine Learning, ce qu'il le rend plus adapté à notre projet.

- Natural Language Processing (NLP) :

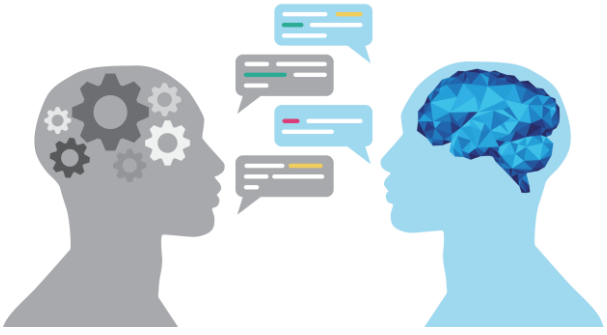

Figure 6 - natural language processing

Le NLP pour Natural Language Processing ou Traitement du Langage Naturel est une discipline qui porte essentiellement sur la compréhension, la manipulation et la génération du langage naturel par les machines. Ainsi, le NLP est réellement à l'interface entre la science informatique et la linguistique. Il porte donc sur la capacité de la machine à interagir directement avec l'humain.

- Visual studio :

Figure 7 - logo visual studio

Microsoft Visual Studio est une suite de logiciels de développement pour Windows conçu par Microsoft.

- Selenium :

Figure 8 - logo selenium

Selenium est un Framework développé en java, qui offre des passerelles pour s'exécuter avec différents langages comme Python et PHP. Il s'agit d'un outil puissant pour contrôler les navigateurs web grâce à des programmes et effectuer l'automatisation du navigateur. Il prend en compte tous les navigateurs, tous les principaux systèmes d'exploitation et ses scripts sont écrits dans différents langages comme Python, java, C# …

Selenium permet une automatisation efficace des tests de l'interface graphique des applications Web. Il est composé principalement de 4 composants à savoir : Selenium IDE, Selenium RC, Selenium Webdriver et Selenium GRID.

Avantages :

- Est un framework open source et portable.
- Peut être utilisé avec de nombreux navigateurs et plateformes différents.
- Peut explorer un site Web à l'aide d'un navigateur spécifique : bien que de nombreux logiciels de scraping de sites Web utilise un véritable navigateur Web pour l'extraction de données, dans la plupart des cas, le navigateur qu'ils utilisent est WebBrowser Control, c'est-à-dire Internet Explorer. Selenuim, cependant, fonctionne non seulement avec Internet Explorer, mais aussi avec une variété de navigateurs tels que Google Chrome, Firefox….

- Selenium WebDriver :

Figure 9 - selenium WebDriver

Selenium Webdriver accepte les commandes (envoyées en Selenese, ou via une API cliente) et les envoient à un navigateur. Ceci est implémenté via un pilote de navigateur spécifique au navigateur, qui envoie des commandes à un navigateur et récupère les résultats. Il n'a pas besoin d'un serveur spécial pour exécuter les tests.

- Pandas :

Figure 10 - logo pandas

Pandas est une bibliothèque open source conçue principalement pour travailler avec des données relationnelles ou étiquetées à la fois facilement et intuitivement. Il fournit diverses structures de données et opérations pour

manipuler des données numériques et des séries chronologiques.
Cette bibliothèque est construite sur le dessus de la bibliothèque NumPy.
Pandas est rapide et offre des performances et une productivité élevée pour les utilisateurs.

Conclusion :

Ce chapitre, nous a permis d'analyser et étudier l'ensemble de techniques et outils utilisées lors de la réalisation de notre projet.

Chapitre 3 : Réalisation et résultat

Introduction :

Nous présentons dans ce chapitre les détails de la mise en œuvre, ensuite les résultats de chaque étape du projet.

3.1 Choix de l'URL de la page :

La première étape du web scraping est de choisir l'URL. Dans notre cas nous avons choisi 3 URL de de vente de 3 catégories de produits (qui contiennent le plus de commentaires possibles) :

- Catégorie supermarché :

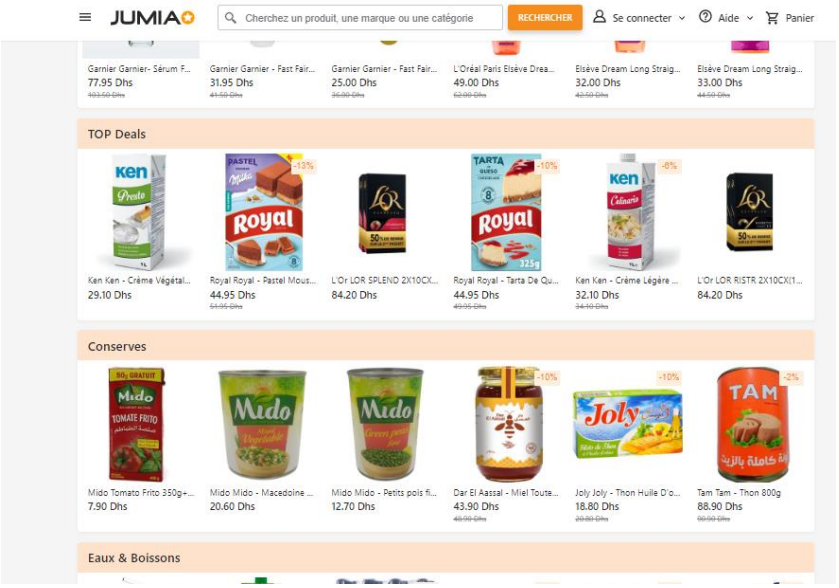

Figure 11 - catégorie supermarché

- Catégorie vêtements et chaussures :

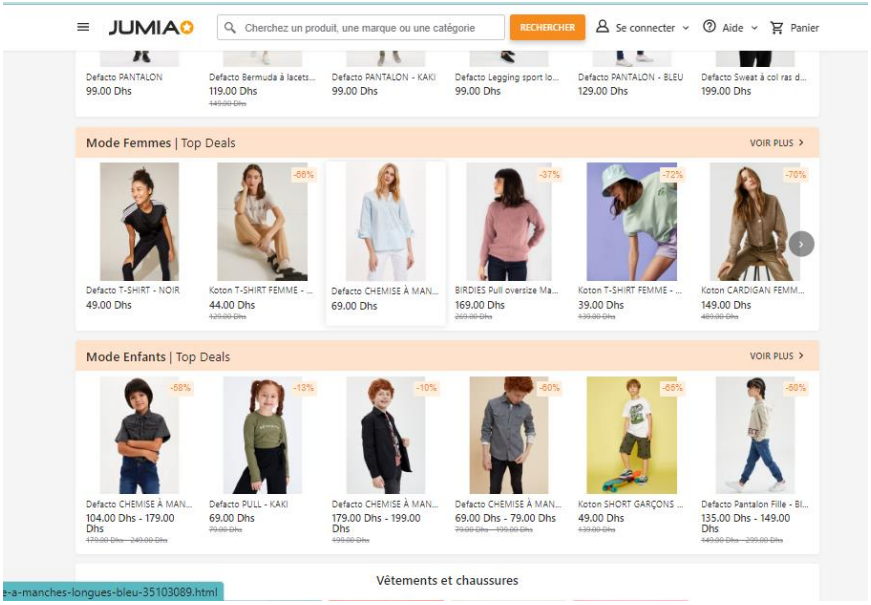

Figure 12 - catégorie vêtements et chaussures

- Catégorie téléphone et tablette :

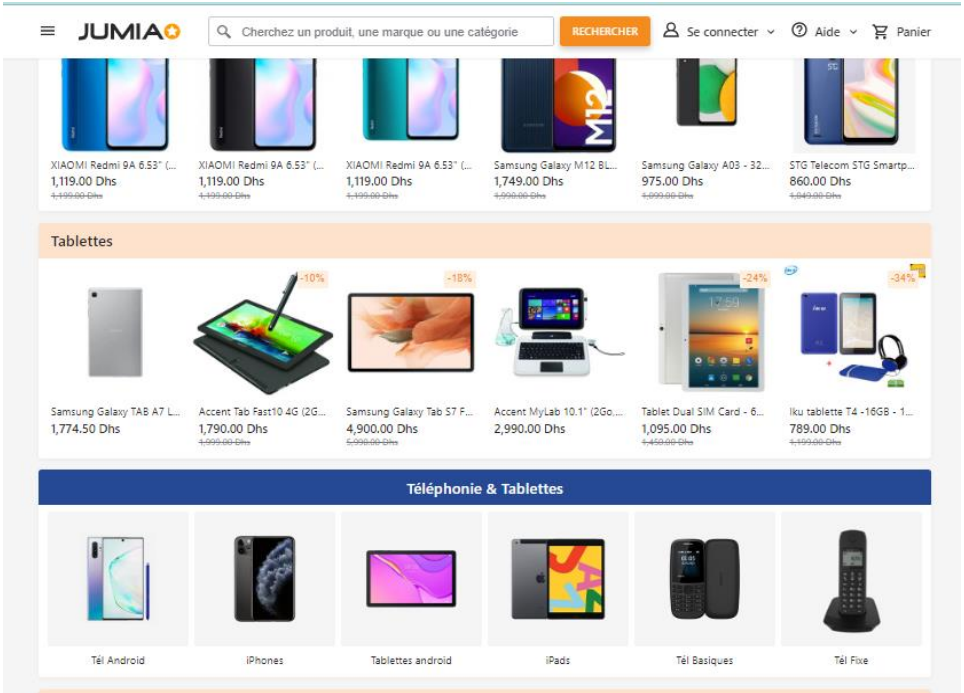

Figure 13 - catégorie téléphone et tablette

3.2 Inspecter la page web :

La deuxième étape du processus consiste à inspecter la page web. Généralement les données sont imbriquées dans des structures imbriquées qu'on appelle tags. Ce qui consiste à inspecter les pages et chercher l'emplacement des données que nous voulons extraire dans notre cas les commentaires des produits.

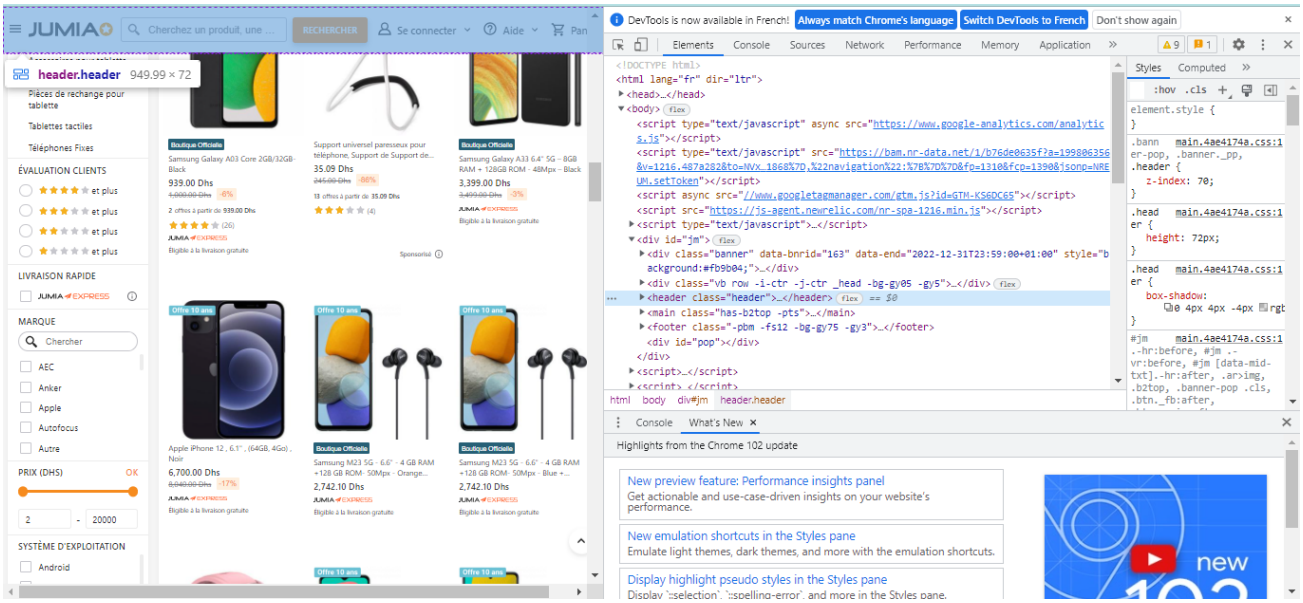

Figure 14 - page web inspecté

3.3 Classe des commentaires :

Nous cherchons la classe des commentaires des produits.

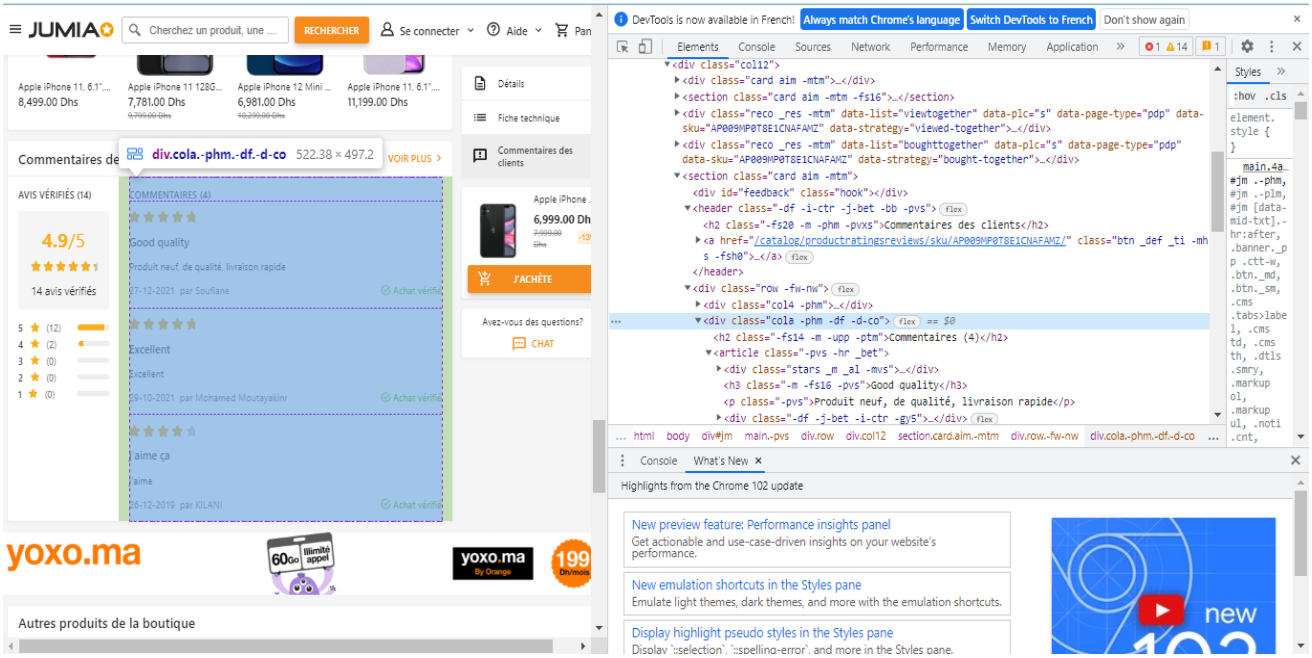

Figure 15 - section des commentaires

Nous pouvons constater que la classe de la division qui contient la section les commentaires des produits est : .

Cependant cette classe ne donne que les quatre premiers commentaires. Pour avoir le reste il faut récupérer la classe qui contient l'URL des restes du commentaires.

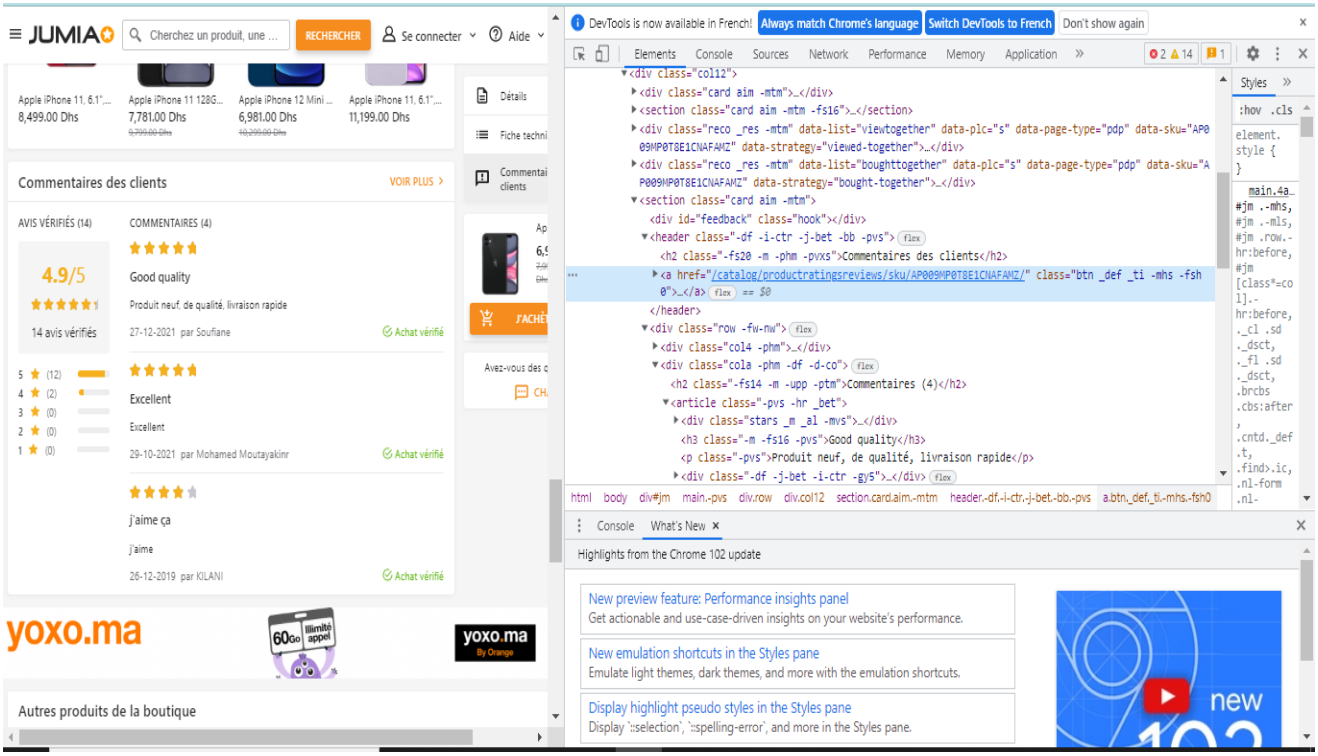

Figure 16 - reste des commentaires

Alors la classe qui contient l'ensemble des commentaires est : <a href="/catalog/productratingsreviews/sku/AP009MP0T8E1CNAFAMZ/ " class= "btn _def _ti -mhs -fsh0".

Ensuite, nous inspectons la classe du paragraphe qui contient le contenu de commentaire.

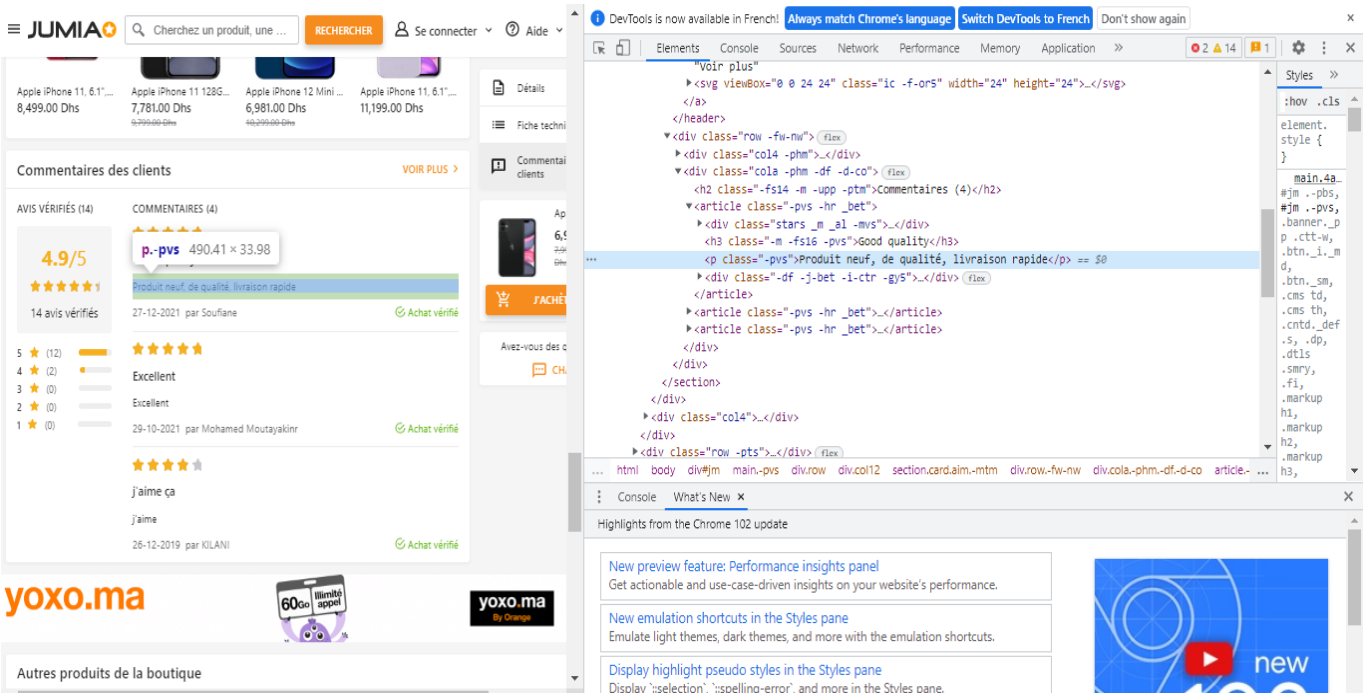

Figure 17 - contenu de commentaire

Nous remarquons que le contenu des commentaires se trouve dans la classe :

<p class="-pvs">.

3.4 Le code :

Pour la partie code, nous commençons par l'importation des librairies : pandas, selenium et selenium WebDriver

```python
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import undetected_chromedriver as uc
import time
import sys
import pandas as pd
```

Figure 18 - importation des librairies

Nous avons utilisé le driver chrome.

```python
def scrap(url):
    options = uc.ChromeOptions()

    options.add_argument("--headless")

    driver = uc.Chrome(options=options)
    wait = WebDriverWait(driver, 5)
    commentaires_list = []
    driver.get("https://www.jumia.ma/telephone-tablette/")

    #content = driver.find_element(By.CLASS_NAME, 'prd _box _hvr')

    content = driver.find_elements(By.CSS_SELECTOR, '.itm.col')
    urls = []
    for element in content:
        urls.append(element.find_element(By.CSS_SELECTOR, '.core').get_attribute('href'))

    for url in urls:
        driver.get(url)
        try:
            content = driver.find_elements(By.CSS_SELECTOR, 'header.-df.-i-ctr.-j-bet.-bb.-pvs')
            link = driver.find_element(By.CSS_SELECTOR, 'a.btn._def._ti.-mhs.-fsh0')
            driver.get(link.get_attribute('href'))

            comments_section = driver.find_elements(By.CSS_SELECTOR, '.cola.-phm.-df.-d-co')
            try:
                comments = comments_section[0].find_elements(By.CSS_SELECTOR, 'p.-pvs')
            except IndexError:
                continue
            for comm in comments:
                commentaires_list.append(comm.text)
```

Figure 19 - le code

Ensuite, après l'importations des librairies, nous avons commencé par créer l'instance du WebDriver chrome. Par la suite la méthode driver.get navigue la page web donnée par l'URL.

Le WebDriver offre plusieurs manières pour trouver les éléments, parmi ces manières la méthode find_element et By.CSS_SELECTOR comme premier paramètre, et en seconde paramètre les classes des données qu'on voulait récuppérer et inspecté dans la première étape.

```
                    continue
                for comm in comments:
                    commentaires_list.append(comm.text)
            except :
                continue

        data = pd.DataFrame(commentaires_list,columns=['Commentaire'])
        return data

if __name__ == '__main__':
    sys.stdin.reconfigure(encoding='utf-8')
    sys.stdout.reconfigure(encoding='utf-8')



    url = "https://www.jumia.ma/telephone-tablette/"

    data = scrap(url)
    data.to_csv('data.csv',index = False)
```

Figure 20 - code (suite)

A l'aide de la librairie Pandas, on stocke les données extraites dans un fichier csv.

3.5 La partie d'analyse :

Dans cette partie, nous travaillons sur la classification des données récupérés, dans notre cas les commentaires des clients sur l'ensemble des produits, dans le but de savoir s'ils sont positifs, négatifs ou neutres.

3.5 Résultat :

Nous exécutons le code et observons que l'ensemble de commentaires serons stocké de la façon suivante :

```
Commentaire
Économisez et Savourez la Coca-Cola
نحب كوكا كولا ماشي بيبسي بيبسي لا
zwina khasha hitabrad hhhh
Comme la description 8 pièces
Zwin zwin
J'aime
Nice !
Top comme toujours
Les articles reçu il reste deux jours sont perime
On m a livré du coca qui allait expirer Quelques jours après livraison Et il avait mauvais goût !
انا افضل هذا الحجم على الحجم الكبير
J aime bien
Le rendre sans sucre et dans aspartame.
It's is coca cola What wrong gonna happen anyway
شكرا جوميا
Merci coca
????د????د????د????د????د????د
Bon produit
"Livraison rapide, mais mal emballé (pas de carton rien uniquement du scotch partout) et erreur de livraison par
تم توصيل في الوقت المناسب
zwina zero surce
واعرا بزاف
Date hors perimi la honte
Top comme toujours
Bonne offre
Bvvvefvt
Merci jumia
Top livraison rapide
زوينة ولكن لي كيعاني من غازات العصران بلما يشربها فيها بزاف غازات
Super gazeuse wa3ra
J'aime bien
```

Figure 21 - résultat 1

data4.csv

```
Superbe
تقيلة و زوينة و اقتصادية
رجعو البرومو ليّ،كانت، غاليه دبا ا
Très Bon produit
مذاقها رائع جداً
Je préfere la 11
العشق الابدي
9 تاريخ الانتهاء شهر. التاريخ قريب من الانتهاء لديه سطوك قديم
"Bon café, prix et goût"
مروغير مقبول
Bonne qualité et goût
رجعو لبرومو ليّ كانت دبا غليتوها بزاف
Frais et date de peremption loin
Bien !!!!
الحب والعشق، رجعو لبرومو ليّ كانت وشكرا
Livraison rapide
""
LUNGO PROFONDO C LA MEILLEUR
I recommend this product
Bon produit
مزيان.....
très bon prix
Merci jumia
Bien !!!
Qualité top
Ta7ia Ljumia
parfait
```

Figure 22 - résultat 2

data4.csv

```
309  Super expérience sur le shop Nivea
310  رائحة رائعة أحببته
311  Très bien
312  J'ai aimé le gel douche
313  رائع جدا
314  Zwiin w ri7a fih zwina ms l7ajm sghiir
315  ""
316  رائع و أصلي
317  Livraison Rapide et bien emballé
318  Excellent
319  Rapidement
320  كيرطب الجسم و رائحة روعة
321  Très bon service
322  J'aime
323  Bon odeur
324  Jaime normally
325  Bon produit
326  جيد جدا
327  Déodorant de qualité
328  "Très bonne qualité , reddo lbal st3mloha 3la l7em mashi fou9 7waij 7it fiha poudre"
329  il est fort et il laisse des tache blanche
330  Bon pdt
331  Le moitié de la bouteille est vide
332  مكتبقاش 48 ساعة غير شي 5 السوايع
333  ممتازة
334  Ri7tha zwina. Jrebtha 3ajbat hta nas asdi9a dyali
335  Produit normal rien de spécial
```

Figure 23 - résultat 3

- Analyse des commentaires :

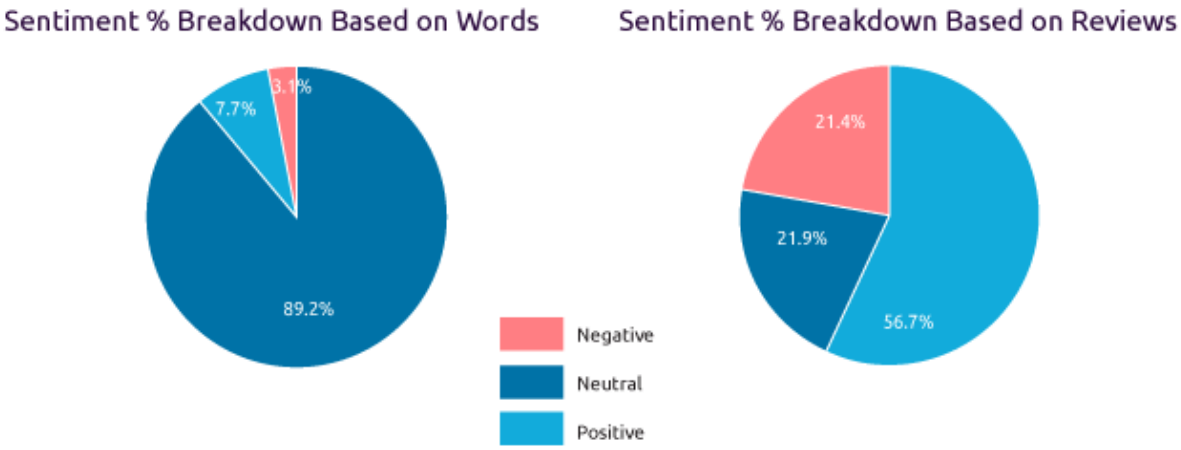

Figure 24 - résultat analyse

En se basant sur ces résultats, nous constatons que la plupart des commentaires sont positives.

Conclusion :

Dans ce chapitre, nous avons présenté les résultats du web scraping, ainsi d'analyse de texte.

Conclusion :

Sur le fond de notre projet, nous avons vu tout au long du rapport les étapes qui commençaient par un contexte général, le traitement des données, la mise en œuvre du projet et enfin les résultats obtenus.

Ce projet a ses racines dans le domaine du e-commerce, et son application serait utile pour les sociétés afin de savoir l'avis des clients et par la suite améliorer leurs produits, ainsi les comparer par rapport aux produits des concurrents. Ce projet a également été l'occasion d'améliorer et de développer nos compétences dans le domaine de programmation avec python.

Au long de ce projet, nous avons rencontré des nouvelles notions tels sentiments analysis et machine learning. Nos perspectives sont maintenant est de bien étudier des deux notions et s'approfondir afin de pouvoir travailler d'autre projets dans le futur.

FSBM
FACULTE DES SCIENCES BEN M'SICK
UNIVERSITÉ HASSAN II DE CASABLANCA

MATHÉMATIQUES & INFORMATIQUE

Réalisé par :

EL MESKINI Rania

Rania.elmeskini37@gmail.com

ELAOUDI Oussama

Oussamaelaoudi49@gmail.co