République Tunisienne
Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique

Université de Monastir
Institut Supérieur d'Informatique
de Mahdia

Projet de fin d'études

N° d'ordre:

MEMOIRE

Présenté à

Institut Supérieur d'Informatique de Mahdia (ISIMa)

En vue de l'obtention

Diplôme National Mastère Professionnel en: Technologies de Sciences des Données (Data Science)

par

AYA GABSI

DEVELOPPEMENT D'UN OUTIL DE WEB SCRAPING

Soutenu le : Décembre 2021, devant la commission d'examen:

Mr.	**Président**
Mr.	**Rapporteur**
Mr. Chaaouri Jaafer	**Encadrant**

Année universitaire 2020/2021

DÉDICACE

*A MON PERE **ABD EL KEFI** ET MA MERE **NAJIA***

A qui je dois ce que je suis, qu'ils trouvent dans ce travail. Le fruit de leurs sacrifices consentis pour mon éducation.

L'expression de mon amour et ma gratitude pour

La bienveillance avec laquelle ils m'ont toujours entouré.

Que dieu leur préserve bonne santé et longue vie.

*A mon mari **ALA** pour l'amour qu'il me porte et pour ses encouragements.*

*A mon frère **SALIM** et sa femme **TESNIME***

Vous étiez toujours présents pour m'aider et m'encourager

Sachiez que vous serez toujours dans mon cœur.

A tous mes amis qui n'ont cessé de m'encourager et de me soutenir

A toute ma famille

Vous occupez une place particulière dans mon cœur. Je vous dédie ce travail en vous souhaitant un avenir radieux, plein de bonheur et de succès.

REMERCIEMENT

Il m'est spécialement agréable, d'exprimer toute ma reconnaissance envers les personnes qui de près ou de loin m'ont apporté leur soutien dans la réalisation de ce projet.

Au premier rang mon encadrant Mr. ***JAAFER CHAAOURI****, son aide, ses conseils précieux, ses critiques constructives, ses explications et suggestions pertinentes qui m'ont donné un coup d'aide pour réaliser mon application convenablement.*

Je remercie également les membres du Jury

Qui ont accepté d'évaluer ce travail et qu'ils y trouvent mes sincères reconnaissances.

Finalement, je remercie tous mes enseignants et le cadre administratif de l'Institut Supérieur D'Informatique « ISIMA » pour leurs aides et leurs efforts à assurer les meilleures conditions possibles pour réaliser ce stage de fin d'études.

TABLE DES MATIERES :

LISTE DES FIGURES :

LISTE DES TABLEAUX :

INTRODUCTION GENERAL :

Le World Wide Web est composé de milliards de documents reliés entre eux et appelés « sites Internet ». Le code source de ces sites Internet est écrit en langage HyperText Mark up Lagunage (HTML). Ce code source HTML est un mélange d'informations lisibles par l'homme et de codes lisibles par les machines, que l'on appelle balises. Le navigateur web par ex. Chrome, Firefox … traite le code source, interprète les balises et met les informations qu'elles contiennent à disposition de l'utilisateur. Des logiciels spécifiques sont utilisés afin d'extraire uniquement du code source les informations intéressantes pour l'être humain. Ces programmes connus sous le nom de « web scrapers », « robots d'indexation », « spiders » ou simplement « bots » parcourent le code source des sites Internet à la recherche de schémas et extraient les informations contenues à ces endroits. Les informations obtenues lors du web scraping sont rassemblées, combinées, analysées ou enregistrées pour une utilisation ultérieure. Si vous avez déjà copié et collé des informations d'un site web, vous avez rempli la même fonction que n'importe quel scraper web, mais à une échelle microscopique et manuelle. Contrairement au processus banal et abrutissant d'extraction manuelle des données, le scraping web utilise une automatisation intelligente pour récupérer des centaines, des millions, voire des milliards de données à partir de la surface illimitée du web. Et cela ne devrait pas être surprenant car le scraping web fournit quelque chose de vraiment précieux que rien d'autre ne peut fournir : il vous donne des données web structurés à partir de n'importe quel site web public. Plus qu'une pratique moderne, la véritable puissance du web scraping réside dans sa capacité à créer et à alimenter certaines des applications commerciales les plus avancées au monde. Le terme " transformation " ne décrit pas la manière dont certaines entreprises utilisent les données collectées sur le web pour améliorer leurs

performances, en éclairant les décisions des dirigeants de l'entreprise jusqu'à l'expérience individuelle de chaque client.

Dans ce cadre s'inscrit notre projet de fin d'études qui consiste à réaliser une application desktop qui permettra aux internautes de scrapper des sites web afin de collecter diverses données de manière très facile. Comme un exemple nous utiliserons, ensuite, l'outil du web scraping dans le domaine e-commerce comme un comparateur de prix. Trois chapitres délimitent le contour de notre projet :

- Le premier chapitre est consacré à la présentation du cadre du projet avec une présentation de l'organisme d'accueil. Ensuite, la définition de la problématique ainsi qu'un critique de l'existant pour mieux planifier les objectifs à atteindre. Enfin, la proposition d'une solution en plus d'une démarche à suivre dans notre projet.

- Le deuxième chapitre présente une étude préalable et état de l'art, dans laquelle, nous expliquons les approches et les techniques de modélisation adoptés au sein de notre réalisation du projet. Ce chapitre est divisé en trois sections. La première section pour l'étude préalable. La deuxième section en vue de définir les approches nécessaires. Et la troisième section présente le concept général de notre projet.

- Le troisième chapitre est réservé à la réalisation qui représente l'environnement de développement du projet puis nous allons donner une explication sur nos interfaces en utilisant des imprimes d'écran.

CHAPITRE 1. PRÉSENTATION GÉNÉRALE DU PROJET :

I. INTRODUCTION:

Avant de commencer notre sujet «Développement d'un outil de web scraping », nous commençons par mettre le projet dans son cadre général. Ensuite, nous discutons de la problématique avec un critique de l'existant afin de conclure une solution proposée. Enfin, nous présentons une étude détaillée de la méthodologie de travail que nous avons poursuit lors de réalisation de notre projet.

II. CADRE DE PROJET:

Le projet « Développement d'un outil de web scraping » est réalisé dans le cadre de projet de fin d'étude, pour l'obtention du diplôme de Mastère professionnel en technologies de sciences des données au sein de l'Institut Supérieur d'Informatique de Mahdia pour l'année universitaire 2020/2021. En fait, le rôle du data scientiste se manifeste dans la manipulation de données afin d'aide à la décision. Dans ce contexte, se base notre projet.

III. PRÉSENTATION DE L'ENTREPRISE :

Mon stage est effectué au sein de l'institut supérieur d'informatique de Mahdia (ISIMA). C'est un établissement universitaire tunisien de l'université de Monastir.

Figure 1: Logo ISIMA

IV. PROBLÉMATIQUE ET CRITIQUE DE L'EXISTANT :

1. Problématique :

Les données peuvent être collectées en plusieurs façons, mais ce qui compte, c'est la précision et la propreté des données. En effet, c'est facile d'extraire des données du web, mais elles ne sont pas forcément très utiles si elles contiennent des erreurs ou si elles sont incomplètes.

En addition, lors de rechercher des données, nous devons garder à l'esprit :

- Nous avons besoin de données propres et prêtes à l'emploi. La qualité des données est donc le critère le plus important dans tous les projets de web scraping.
- Nous voulons que les données soient utilisées pour prendre les bonnes décisions et, pour cela, nous avons besoin de données de qualité en permanence.

2. Critique de l'existant :

Les données sont un facteur important nos jours. Ce sont le moteur de la relation client, de la stratégie commerciale et de tout projet. La valeur d'une donnée dépend avant tout de sa qualité. La qualité de données se retrouve ainsi aux cœurs des problématiques d'entreprises. Le processus de qualification doit commencer dès l'intégration des données au sein de l'organisation. Un enjeu majeur quand nous savons

qu'une fois stockées, croisées et analysées ; Les données qualifiées apportent une forte valeur ajoutée à tous les niveaux de l'entreprise.

La solution manuelle de collecte de données contribue à perdre le temps et ralentir les services, ce qui entraine l'insatisfaction des clients.

3. Objectifs à atteindre :

En tenant compte des critiques et des besoins, la solution est de développer une application qui a pour but :

- Automatiser l'extraction des données.
- Analyser les données.
- Simplifier la visualisation des données.

V. SOLUTION PROPOSÉE ET DÉMARCHE À SUIVRE:

1. Solution proposée :

Le web scraping se présente comme une solution afin de satisfaire nos objectifs.

Il se définit comme un processus de collecte automatisée de données structurées sur le web. Il est également appelé extraction de données web. Parmi les principaux cas d'utilisation du web scraping, nous pouvons citer la veille concurrentielle, la surveillance des tarifs, le suivi de l'actualité et les études de marché, entre autres.

L'extraction de données sur le web est utilisée par les personnes et les entreprises qui veulent utiliser la vaste quantité de données disponibles sur le web pour prendre des décisions plus intelligentes.

2. Etapes de développement d'une application :

Le développement d'une application désigne l'ensemble du processus consistant à bâtir tout type d'applications informatiques fiables et performantes et va de l'étude du besoin, jusqu'à la maintenance de l'application. Ces diverses étapes sont possibles grâce à un langage informatique spécifique.

Cette figure résume ces étapes :

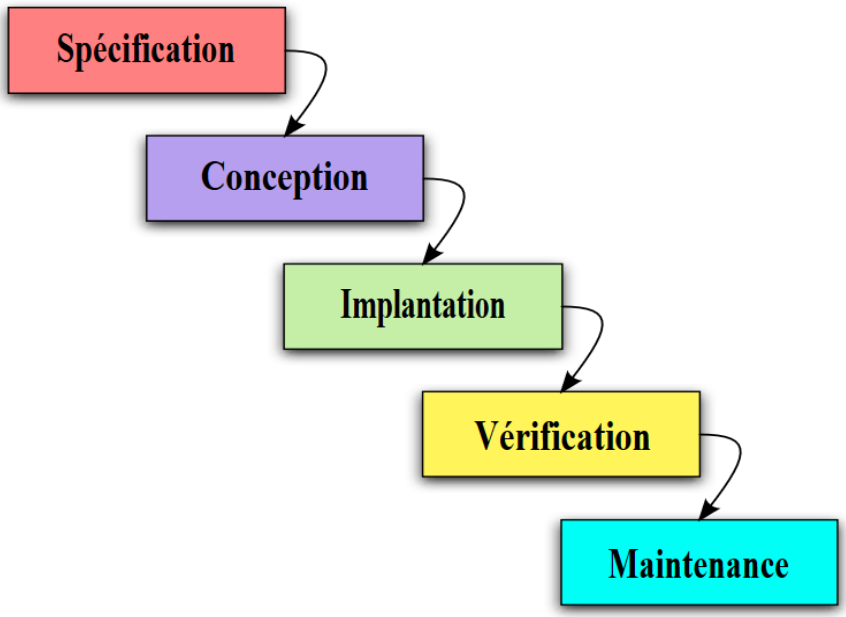

Figure 2: Etapes de la construction d'un programme

- **Spécification** : Décrire un programme en s'appuyant sur les besoins fonctionnels.
- **Conception :** Définir les diagrammes de cas d'utilisation, de classes et de séquences de votre application.

- **Implémentation** : Partie du code.
- **Vérification** : Tester l'application avec un groupe d'utilisateurs.
- **Maintenance** : Publier votre application.

3. Etapes de développement d'un projet data science :

- **Compréhension du besoin métier :**

C'est la faite de discuter avec les responsables métiers pour comprendre la problématique à résoudre et identifier les variables à prédire.

- **Collecte des données :**

Le data scientiste doit avoir une vision claire et exhaustive des données à collecter, identifier les sources où obtenir ces données, savoir y accéder et les stocker.

- **Nettoyage des données :**

Les données provenant de différentes sources peuvent avoir des formats différents (csv, json, XML…) et contenir des anomalies ou des valeurs incorrectes.

Il est nécessaire de nettoyer et restructurer cette donnée. L'objectif de cette manipulation est de mettre à un « meilleur » format pour faciliter l'exploration des données.

- **Formulation des hypothèses :**

L'objectif est de croiser les différentes natures de données et d'établir des liens de corrélation entre ces dernières. Ces liens doivent se matérialiser par la formulation d'hypothèses.

- **Détermination des variables synthétiques :**

Cette étape consiste à concevoir et sélectionner des variables synthétiques : des combinaisons de données brutes sur lesquelles tourneront les algorithmes. Les

variables synthétiques ont pour objectif de mieux représenter le problème à résoudre et donc d'améliorer la performance du modèle.

- **Construction du modèle :**

Cette étape correspond à la phase de machine Learning à proprement parler du projet de Data Science.

Il s'agit de choisir les différents modèles de machine Learning qui permettent de modéliser au mieux la variable cible à expliquer.

- **Présentation & Communication :**

La visualisation des données (Data visualisation / Dataviz) permet de tirer rapidement des informations grâce à des représentations graphiques pertinentes et dynamiques.

Les étapes de développement d'un projet data science sont présentés dans la figure suivante :

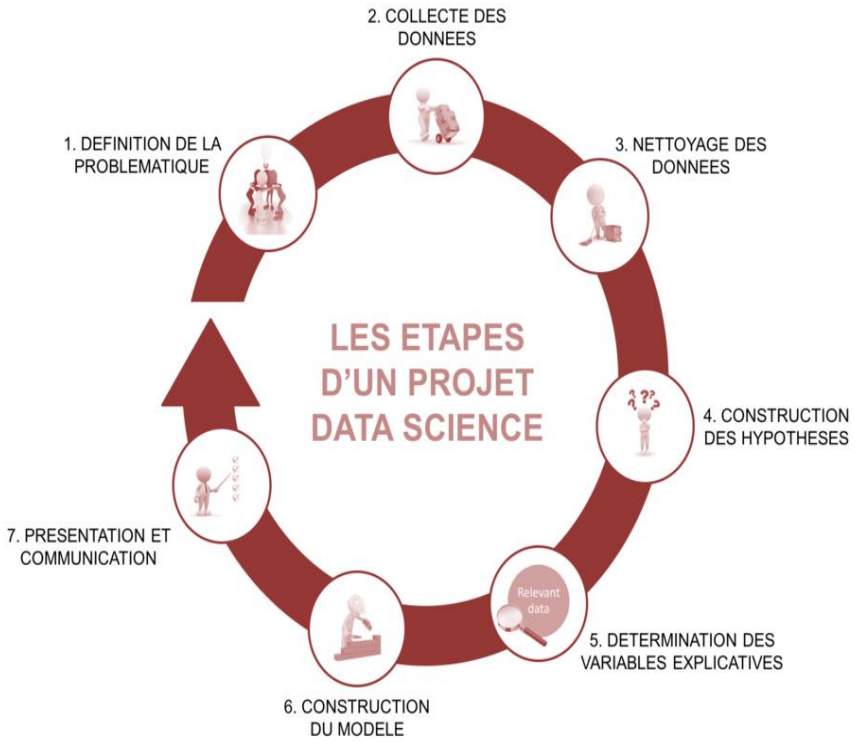

Figure 3: Les étapes de développement d'un projet data science

4. Analyse et spécification des besoins :

L'analyse et la spécification des besoins sont deux étapes nécessaires dans le cycle de vie de développement d'un projet. Cette partie est réservée à analyser les besoins fonctionnels et non fonctionnels ainsi qu'identifier les acteurs.

4.1. Identification des acteurs :

- **Utilisateur :**

Notre application est destinée à un utilisateur qui peut gérer tous ses fonctionnalités.

4.2. Spécification des besoins :

- **Besoins fonctionnels :**

Les besoins fonctionnels sont les besoins spécifiant un comportement d'entrée sortie du système. Ils comprennent une description des fonctions requises, des aperçus, des rapports associées ou des requêtes en ligne, ainsi que des détails sur les données à conserver dans le système.

Dans le cadre de notre projet, on cite les besoins fonctionnels suivants pour notre utilisateur :

- Extraire des données automatiquement.
- Visualiser les données scrapés.
- Consulter les prix d'un tel produit dans les différents sites e-commerce.

- **Besoins non fonctionnels :**

Les besoins non fonctionnels sont les besoins techniques décrivant les contraintes à prendre en considération pour assurer le bon fonctionnement et la qualité de notre système.

- **La fiabilité :** L'application doit avoir une forte probabilité pour fonctionner sans tomber en panne.
- **La Maintenance :** Les différents modules de l'application doivent être faciles à maintenir, par conséquent, le code doit être lisible et bien structuré.
- **La disponibilité :** L'application doit être disponible pour être utilisé par n'importe quel utilisateur.
- **La sécurité :** La confidentialité des données doit être respectée.

VI. CONCLUSION :

Dans ce chapitre, nous avons présentés les grands points du projet, l'organisme d'accueil, le contexte général du projet et la problématique.

Ensuite, nous avons identifié les différents besoins de notre projet, afin de déterminer les spécifications fonctionnelles et non fonctionnelles des solutions.

Dans le chapitre suivant, nous nous concentrerons sur l'étude détaillée des méthodes et techniques utilisées dans ce projet.

CHAPITRE 2. ETUDE PREALABLE ET ETAT DE L'ART :

► SECTION 1 : ETUDE PREALABLE :

I. INTRODUCTION :

Cette section est consacrée à une étude détaillée de l'art des domaines informatiques, en particulier, les sciences de données, que nous utilisons dans la solution proposée.

Ainsi que, une présentation de quelques concepts.

II. L'ÉVOLUTION DU WEB:

1. Le web 1.0 :

Le web 1.0 ou web traditionnel (1991-1999), est un web statique. Il est centré sur la distribution d'informations. Ce web est caractérisé par des sites orientés produits, qui demandent peu l'intervention des utilisateurs. Les premiers sites d'e-commerce datent de cette époque.

2. Le web 2.0 :

Le web 2.0 ou web social (2000-2009), change totalement de perspective. Il privilégie la dimension de partage et d'échange d'informations et de contenus (textes, vidéos, images ou autres). Il voit l'émergence des réseaux sociaux, des smartphones et des blogs. Ce web se démocratise et se dynamise. L'avis du consommateur est sollicité en permanence et il prend goût à cette socialisation virtuelle.

3. Le web 3.0 :

Le web 3.0 ou aussi nommé web sémantique (2010-XX), vise à organiser la masse d'informations disponibles en fonction du contexte et des besoins de chaque utilisateur, en tenant compte de sa localisation, de ses préférences, etc. C'est un web qui tente de donner sens aux données. C'est aussi un web plus portable et qui fait de plus en plus le lien entre monde réel et monde virtuel. Il répond aux besoins d'utilisateurs mobiles, toujours connectés à travers une multitude de supports et d'applications malines ou ludiques.

4. Le web 4.0 :

Le web 4.0 est évoqué par certains comme le web intelligent (2020-XX), il vise à immerger l'individu dans un environnement (web) de plus en plus prégnant. Il pousse à son paroxysme la voie de la personnalisation ouverte par le web 3.0 mais il pose par la même occasion de nombreuses questions quant à la protection de la vie privée, au contrôle des données, etc.

III. LE BIG DATA :

L'explosion quantitative des données numériques a obligé les chercheurs à trouver de nouvelles manières de voir et d'analyser le monde. Il s'agit de découvrir de nouveaux ordres de grandeur concernant la capture, la recherche, le partage, le stockage, l'analyse et la présentation des données. Ainsi est né le « Big Data ». Il s'agit d'un concept permettant de stocker un nombre indicible d'informations sur une base numérique.

1. C’est quoi le big data ?

Le big data, méga données en français, est une combinaison de données structurés, semi structurés et non structurés, collectés par des organisations qui peuvent être collectés et traités dans le but d’étirer les informations. Ces informations sont utilisées dans des projets d’apprentissage automatique, de modélisation prédictive et d’autres applications analytiques avancés. Les données brutes collectées sont stockées, ensuite, des outils dotés de l’intelligence artificielle agissent comme des supercalculateurs et analyse avec des algorithmes complexes avant d’être finalement utilisé pour la prise de décision. On cherche alors à donner un sens aux données pour mieux comprendre ses clients.

2. Caractéristiques du big data :

Le Big Data se caractérise par la problématique des 3V :

- **Volume** : Le volume correspond à la masse d’informations produite chaque seconde. Selon des études, pour avoir une idée de l’accroissement exponentiel de la masse de données, on considère que 90 % des données ont été engendrées durant les années où l’usage d’internet et des réseaux sociaux a connu une forte croissance. L’ensemble de toutes les données produites depuis le début des temps jusqu’à la fin de l’année 2008, conviendrait maintenant à la masse de celles qui sont générées chaque minute

- **Variété** : Seulement 20% des données sont structurées puis stockées dans des tables de bases de données relationnelles similaires à celles utilisées en gestion comptabilisée. Les 80% qui restent sont non-structurées. Cela peut être des images, des vidéos, des textes, des voix, et bien d’autres encore… La technologie Big Data, permet de faire l’analyse, la comparaison, la

reconnaissance, le classement des données de différents types comme des conversations ou messages sur les réseaux sociaux, des photos sur différents sites etc. Ce sont les différents éléments qui constituent la variété offerte par le Big Data.

- **Vélocité** : La vélocité équivaut à la rapidité de l'élaboration et du déploiement des nouvelles données. Par exemple, si on diffuse des messages sur les réseaux sociaux, ils peuvent devenir « viraux » et se répandre en un rien de temps. Il s'agit d'analyser les données au décours de leur lignée (appelé parfois analyse en mémoire) sans qu'il soit indispensable que ces informations soient entreposées dans une base de données.

- Plus récemment, plusieurs V ont été ajoutés aux différentes descriptions du big data notamment : Véracité, Valeur, Variabilité.

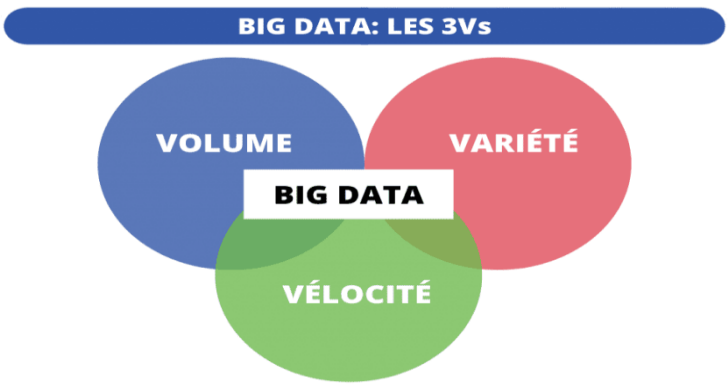

Figure 4: Règles des 3V en Big Data

3. Les types du big data :

- **Données structurées : (format défini)**

Les données qui ont été organisées pour faciliter leur traitement, contribuent une structure et un format fait pour simplifier l'analyse et peuvent s'agir d'éléments comme : le nom, une adresse, l'âge respectent un format défini. C'est ce type de données quand vous remplissez un formulaire.

- **Données non structurées : (format indéfini)**

Les données qui ne suivent pas de structure ou de formats. Ce sont les éléments comme photos, vidéos ou encore de texte brut. Elles sont plus difficiles à analyser.

4. Le marché du big data :

Les entreprises utilisent le big data pour améliorer leurs opérations, fournir un meilleur service à la clientèle, créer des compagnes de marketing personnalisées : en fin de compte, augmenter leur rentabilité.

Ces entreprises qui utilisent le big data détiennent un avantage concurrentiel potentiel sur celles qui ne le font pas, car elles sont en mesure de prendre des décisions commerciales plus rapide et plus éclairés, à condition, d'utiliser les données efficacement. Par exemple, le big data peut fournir des informations précieuses sur leurs clients qui peuvent être utilisés pour affiner les compagnes et les techniques de marketing afin d'augmenter l'engagement des clients et les taux de conversion.

5. Conclusion :

Chaque seconde, d'énormes quantités de données sont générés, rapidement partagés et en permanence mise à jour. Bien que le big data ne possède pas un volume spécifique

de données, l'utilisation implique souvent même des Exaoctets de données capturées au fil du temps.

Les big data ne sont pas seulement un aspect important de l'avenir, elles peuvent être l'avenir lui-même.

IV. L'INTELLIGENCE ARTIFICIELLE :

L'IA est un ensemble de techniques permettant à des machines informatisées d'accomplir des tâches et de résoudre des problèmes algorithmiques ou logiques. Ces machines peuvent prendre leur propre décision lorsqu'elles sont confrontées à de nouvelles situations de la même manière que les humains. Le nouvel essor de l'IA est dû aux nouvelles capacités d'apprentissage des machines. La grande majorité des avancées et des applications de l'IA se réfère à une catégorie d'algorithmes connus sous le nom d'apprentissage automatique (Machine Learning). Ces algorithmes utilisent des statistiques pour trouver des modèles dans le big data. Ils utilisent ensuite ces modèles pour faire des prédictions. Le Machine Learning et ses sous ensemble le Deep Learning sont incroyablement puissants.

La combinaison entre l'intelligence artificielle, la Machine Learning et le Deep Learning se manifeste dans la figure ci-dessous :

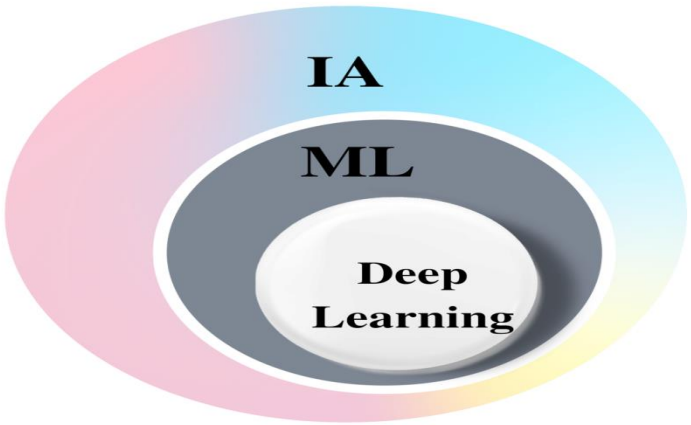

Figure 5: Schéma explicative

1. L'apprentissage automatique (Machine Learning) :

La Machine Learning (ou ML) est un domaine qui étudie comment des algorithmes peuvent apprendre en étudiant des exemples.

Les algorithmes d'apprentissage automatique :

- Trouvent des schémas récurrents dans le big data et apprennent à partir de ces modèles pour améliorer leur performance à résoudre des taches.
- Utilisent des statistiques pour trouver des modèles dans des quantités massives de données, et les données ici englobent beaucoup de choses : des nombres des mots, des images, des cliques…
- Sont en grande partie responsable de la majorité des progrès et des applications utilisant l'IA.

Le ML est le processus qui alimente de nombreux services que nous utilisons aujourd'hui : les systèmes de recommandations comme YouTube, les moteurs de recherches comme Google, les flux de réseaux sociaux comme Facebook. Chaque

plateforme recueille autant de données sur vous que possible ; et utilise l'apprentissage automatique pour faire une estimation très précise de ce que vous pourriez vouloir ensuite.

2. Deep Learning :

Le Deep Learning ou apprentissage profond est une manière particulière de faire du ML. C'est un ML qui utilise une technique qui donne aux machines une capacité accrue à trouver et amplifier les plus petits modèles du Big data et Small data. Cette technique est appelée Deep neural network soit réseau neuronal profond, parce qu'elle comporte de très nombreuses couches de nœuds de calcul qui travaillent ensemble pour extraire des informations à partir d'un ensemble de données historiques et fournir des résultats finaux sous la forme d'une prédiction. Ainsi, c'est un système capable de travailler à partir de données non structurées et en toute autonomie d'où son avantage sur le ML classique.

3. ML vs Deep Learning :

Le ML et le Deep Learning se déclinent en 3 méthodes d'apprentissages différentes :

- **Supervisé :**

Les données sont étiquetées pour dire à la ML exactement quels modèles elle doit rechercher.

- **Non supervisé :**

Les données n'ont pas d'étiquettes, la ML se contente de rechercher les modèles qu'elle peut trouver.

- **Par renforcement :**

Un algorithme de renforcement apprend par essais et erreurs pour atteindre un objectif clair. Ils essayent beaucoup de choses différentes et est récompensé ou pénalisée selon que ces comportements l'aide ou l'empêchent d'atteindre son objectif.

Le tableau suivant déduit une comparaison entre la Machine Learning et le Deep Learning :

	ML	Deep Learning
Organisation des données	Données structurées	Données non structurées
BD	Contrôlable	>1 million de données
Entrainement	Entrainement par l'humain	Autonome
Algorithme	Algorithme modifiable	Réseau neuronal d'algorithme
Champ d'application	Actions simples	Taches complexes

Tableau 1: Comparaison entre ML et Deep Learning

4. Conclusion :

L'intelligence artificielle, ou IA est un concept qui vise à donner aux machines la capacité de raisonner comme un être humain. D'une manière précise, un ensemble des techniques permettant d'intégrer des comportements intelligents dans une machine.

La figure suivante illustre précisément à quoi sert l'intelligence artificielle :

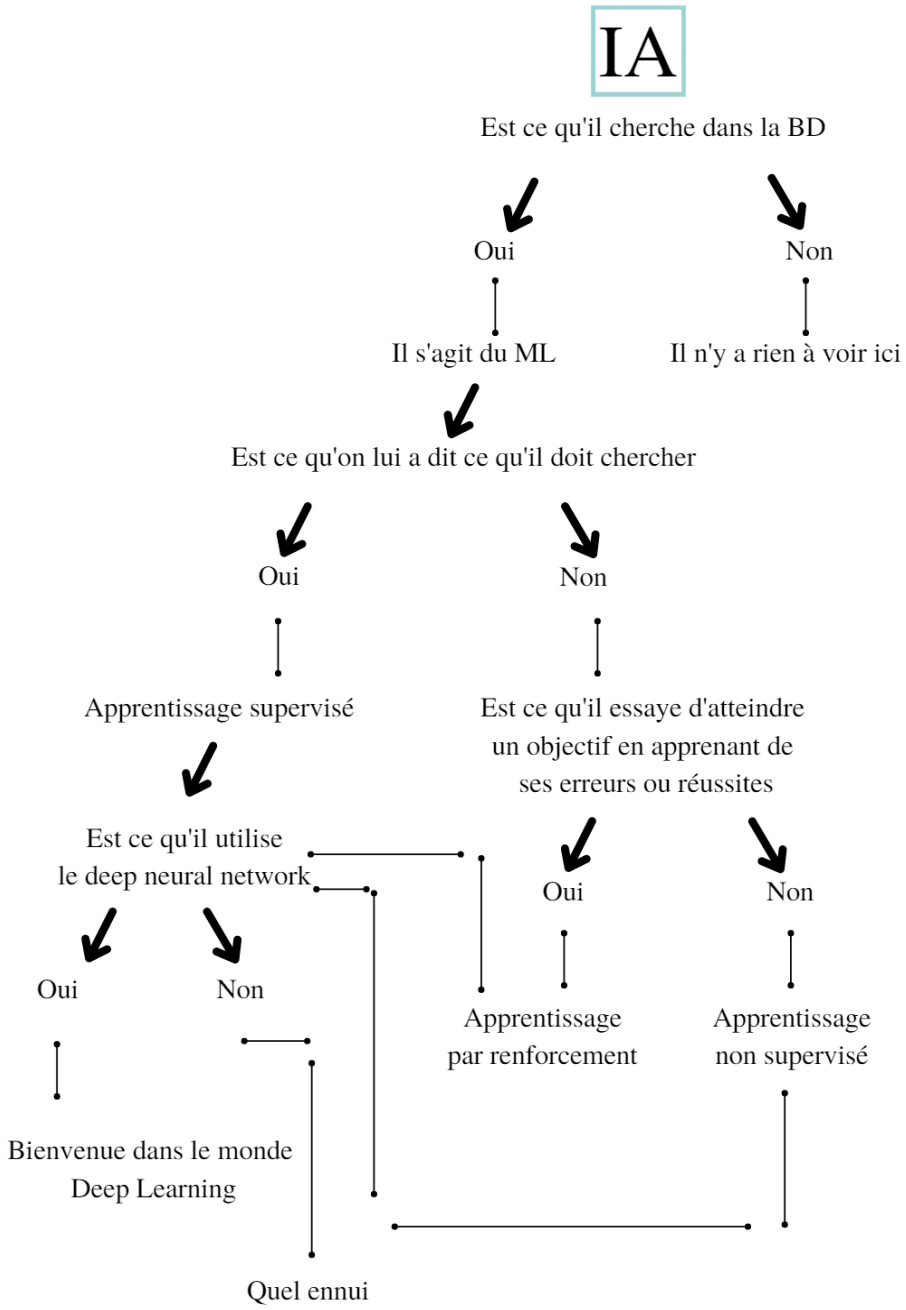

Figure 6: Modèle explicatif du IA

V. VISUALISATION DES DONNÉES :

La visualisation des données est un moyen qui permet aux professionnels de mettre en place des représentations visuelles d'un grand ensemble de données. Cette représentation visuelle aussi connu sous le nom de dataviz, permet d'augmenter le processus de compréhension et d'assimilation. Comme, la visualisation des données prend en compte des centaines, des milliers ou encore des millions de données, cela permet de faciliter les analyses. Elle se situe à l'intersection des domaines des sciences de l'information de la communication et du design. Le principal avantage de la visualisation des données n'est pas qu'elle donne un aperçu d'ensembles de données complexes en communiquant leur clé aspect les plus intuitifs de manière significative. La visualisation des données a été identifiée comme une compétence de recherche clé du 21ème siècle.

Les avantages du dataviz :

- Gagner du temps.
- Accélérer le processus de décision.
- Comprendre et enregistrer des informations.
- Accélérer le processus de réaction vis-à-vis des changements.
- Interagir directement avec les données.

Exemples :

- Analyser et prévoir des tendances.
- Procéder à un suivi de performance d'une compagne et de l'optimiser.
- Créer du contenu visuel qui apporte un certain engagement.
- Comparer des données.

VI. CONCLUSION :

Les sciences des données et le machine Learning (ou apprentissage automatique) sont deux mots très en vogue lorsque l'on parle de la révolution Big Data, de prédiction des comportements ou tout simplement de la transformation numérique des entreprises. Le premier objectif du data scientiste est de produire des méthodes (automatisées, autant que possible) de tri et d'analyse de données, afin d'en extraire des informations utiles.

L'intersection des modules de science des données est définie dans la figure suivante :

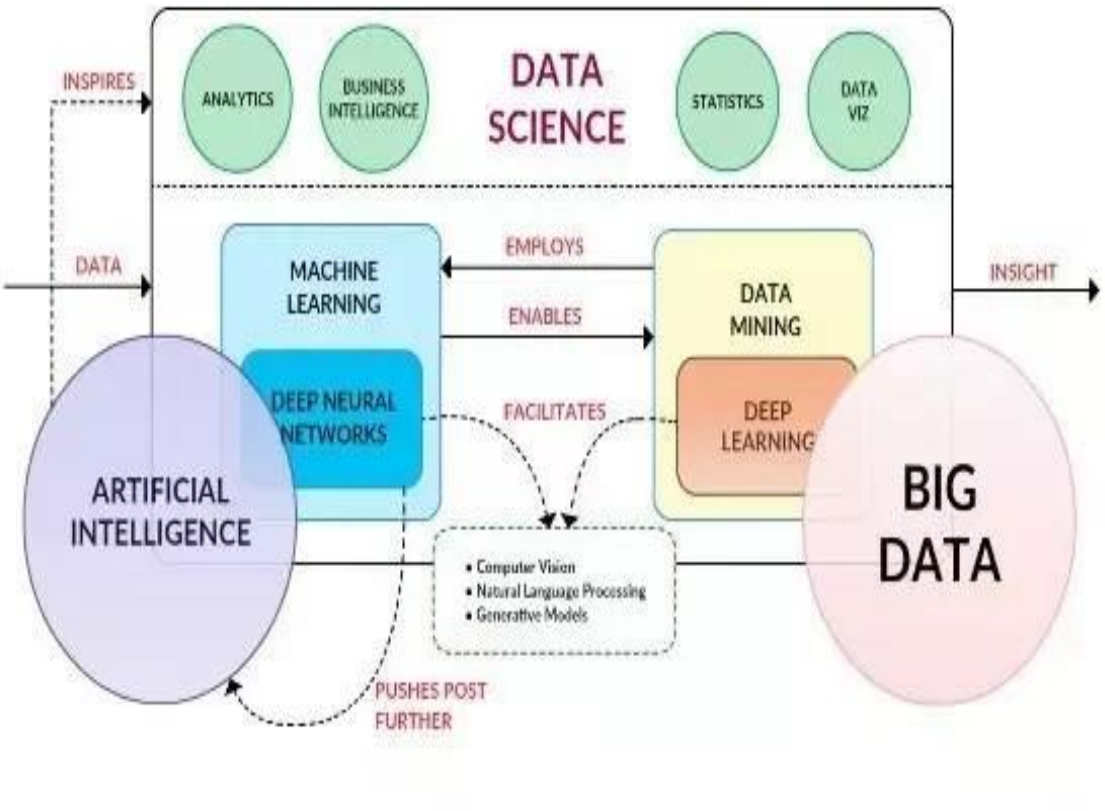

Figure 7: Schéma explicative de l'intersection entre les modules de science de données

► SECTION 2 : Extraction automatique de données :

I. INTRODUCTION :

Après la naissance du World Wide Web en 1989, le nombre de sites web a eu une croissance exponentielle qui dépasse aujourd'hui 5,611 milliards de pages, donc le volume de données est très important. Ces données sont variées (texte, images, vidéo ...) et renouvelable, ce qui fait que l'accès à ces données représente un challenge, qui nécessite des techniques diffèrent des méthodes traditionnelles comme le copier-coller, le capture d'écran ou encore accéder aux bases de données à travers les requêtes. Ces techniques ne permettent pas d'extraire les informations d'une façon optimale. Là intervient une technique intéressante, efficace, très rapide et prometteuse On parle d'une technologie baptisée le « web scraping ». Dans ce chapitre nous allons détailler cette technologie.

II. WEB CRAWLING :

1. Définition :

Un crawler ou également appelé un robot d'indexation est un bot internet qui parcourt systématiquement le web dans le monde entier. C'est un programme automatisé pour naviguer sur le net dans le but d'indexer les pages web et leur contenu. Ils sont utilisés par les moteurs de recherche pour explorer des pages web afin de mettre à jour leur index avec de nouvelles informations.

Le web crawling passera généralement par chaque page du site web plutôt que par un sous-ensemble de pages.

2. Fonctionnement :

Les robots d'indexation commencent à partir d'une liste d'URL connues. Ils exploreront d'abord les pages web de ces URL. En indexant ces pages web, ils trouveront des hyperliens vers d'autres URL et les ajouteront à la liste des pages à indexer ensuite.

Étant donné le grand nombre de pages web qui pourraient être indexées pour la recherche, ce processus pourrait se poursuivre presque indéfiniment. Toutefois, un robot d'indexation suivra certaines politiques qui le rendent plus sélectif sur les pages à indexer, dans quel ordre les indexer et à quelle fréquence ils doivent les indexer à nouveau pour vérifier les mises à jour de contenu.

- ✓ **Importance relative de chaque page web :** la plupart des robots d'indexation n'indexent pas l'intégralité d'Internet accessible au public et ne sont pas prévus pour effectuer cette tâche. En fait, ils décident quelles pages indexer en premier en fonction du nombre des autres pages liées à cette page, du nombre de visiteurs que cette page reçoit ainsi que d'autres facteurs qui indiquent la probabilité que la page contienne des informations importantes.
  L'idée est qu'une page web qui est citée par de nombreuses autres pages et qui compte de nombreux visiteurs est susceptible de contenir des informations de grande qualité qui font autorité. Il est donc particulièrement important qu'un moteur de recherche l'index, de la même façon qu'une bibliothèque veillera à avoir en rayons plusieurs exemplaires d'un livre qui est emprunté par de nombreuses personnes.

- ✓ **Nouvelle visite des pages web :** le contenu du Web est continuellement mis à jour, supprimé ou déplacé vers de nouveaux endroits. Les robots d'indexation

doivent régulièrement revisiter les pages pour s'assurer que la dernière version du contenu est indexée.

3. Architecture :

Le schéma suivant présente l'architecture du web crawling :

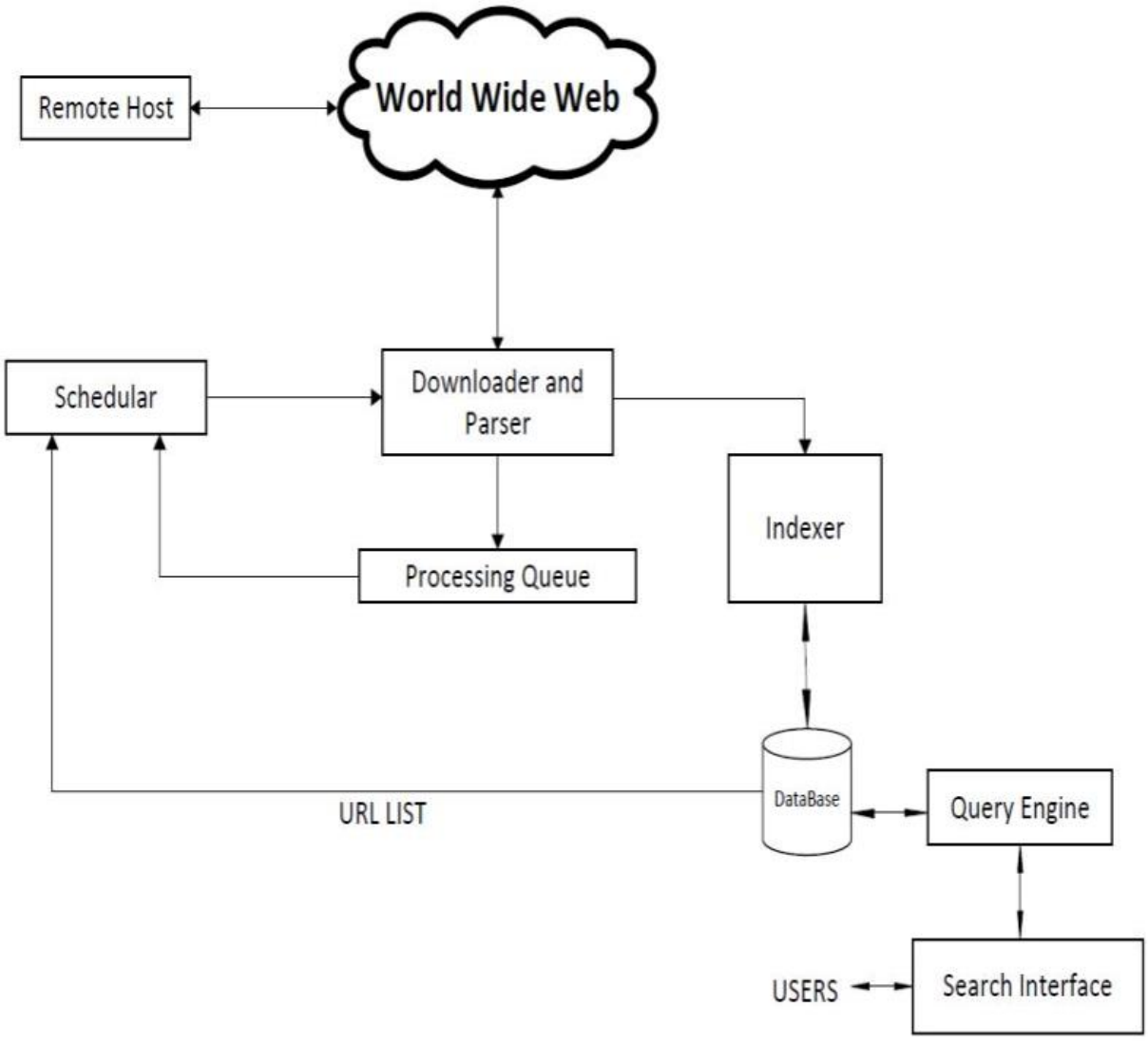

Figure 8: Architecture général du web crawler

III. WEB SCRAPING :

1. Définition :

Le web scraping est une technique qui permet de récupérer de manière automatisée des données provenant du web. Il permet d'extraire des données et des informations présentent sur des sites web et de les transformer en d'autres formats plus exploitables, comme Excel ou csv.

2. Fonctionnement :

- **Demande-réponse :**

La première étape simple de tout programme de grattage Web (également appelé « grattoir ») consiste à demander au site Web cible le contenu d'une URL spécifique.

En retour, le grattoir obtient les informations demandées au format HTML. N'oubliez pas que HTML est le type de fichier utilisé pour afficher toutes les informations textuelles sur une page Web.

- **Analyser et extraire :**

HTML est un langage de balisage, ayant une structure simple et claire. L'analyse s'applique à n'importe quel langage informatique, prenant le code sous forme de groupes de texte. Il produit une structure en mémoire, que l'ordinateur peut comprendre et utiliser. Pour faire simple, nous pouvons dire que l'analyse HTML prend le code HTML, l'attend et extrait les informations pertinentes - titre, paragraphes, en-têtes. Liens et mise en forme comme du texte en gras. Donc, tout ce dont vous avez besoin est une expression régulière, définissant le langage régulier, afin

qu'un moteur d'expression régulière puisse générer un analyseur pour ce langage spécifique. Ainsi, la correspondance de motifs devient possible, ainsi que l'extraction de texte.

- **Télécharger les données :**

La dernière étape - télécharger et enregistrer les données dans le format de votre choix (CSV, JSON ou dans une base de données). Une fois accessible, il peut être récupéré, implémenté dans d'autres programmes.

En d'autres termes, le scraping vous permet non seulement d'extraire des données, mais de les stocker dans une base de données ou une feuille de calcul local central et de les utiliser plus tard lorsque vous en avez besoin.

3. Architecture :

L'architecture du web scraping est illustrée dans la figure ci-dessous :

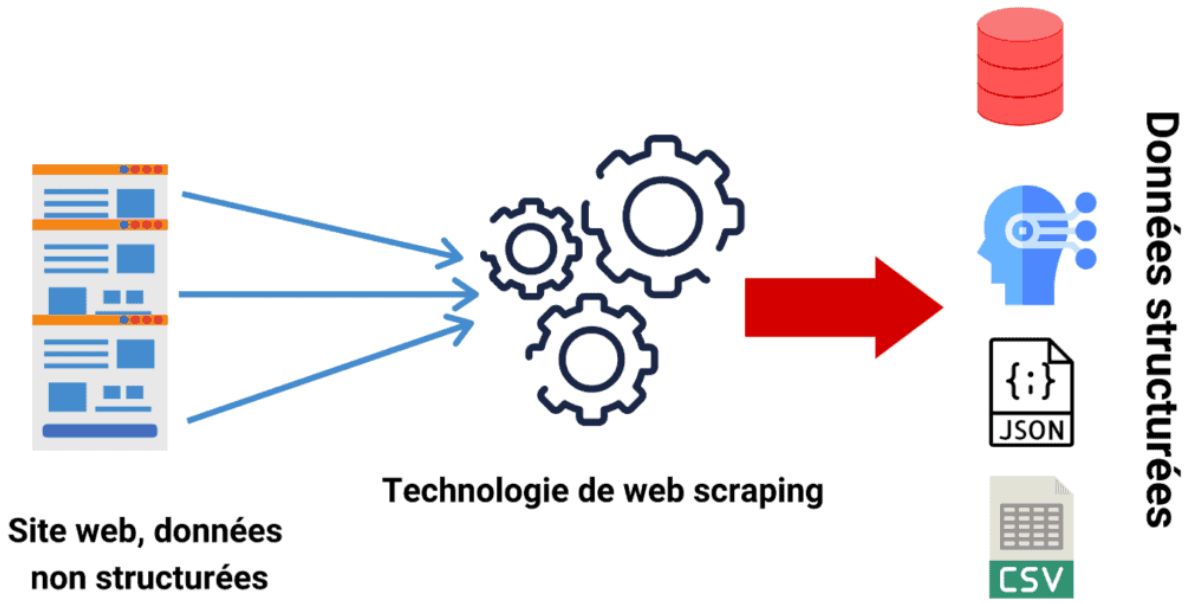

Figure 9: Architecture du Web Scraping

IV. WEB SCRAPING VS WEB CRAWLER :

Les termes web scraping et web crawling sont souvent utilisés de manière interchangeable, mais bien que ces termes partagent de nombreuses similitudes, il existe des différences clés qui les distinguent.

Le web scraping a une approche et un objectif beaucoup plus ciblés, tandis que le web crawling qui numérisera et extraira toutes les données d'un site web. En raison des différences, des objectifs et des applications pour le web scraping et le web crawling est également radicalement différence.

- **Web crawler** : Guide le web scraper à travers les adresses web.
- **Web scraper** : Collecte et extrait les données sur les URLs spécifiées par le web crawler.

Une comparaison simple entre le web scraping et le web crawling est présentée dans la figure suivante :

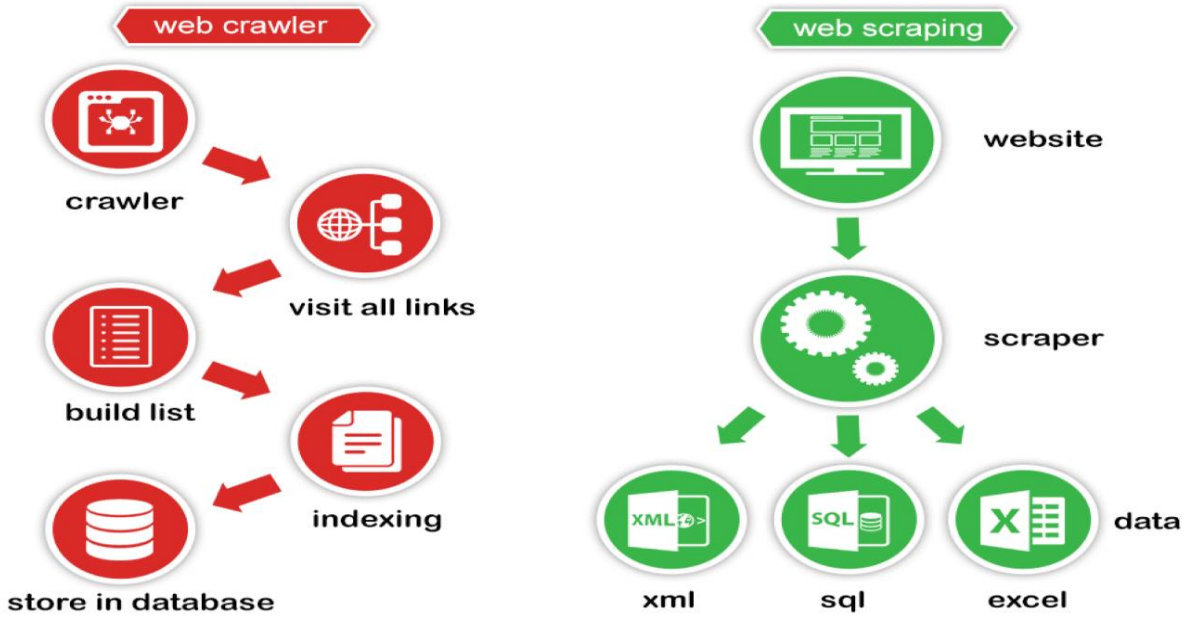

Figure 10: Web scraping VS Web crawling

V. LES OUTILS DU WEB SCRAPING :

- **Extraction de données :**

- BeautifulSoup :

BeautifulSoup est une bibliothèque très populaire pour le crawling sur le web et le scraping de données. Cette bibliothèque vous aide à collecter des données disponibles sur un site web afin dès les scrapers et les organiser dans le format dont vous avez besoin.

- **Traitement et modélisation des données :**

- NumPy :

NumPy (pour Numerical Python) est un outil parfait pour le calcul scientifique et la réalisation d'opérations de base et avancées avec des tableaux.

La bibliothèque offre de nombreuses fonctionnalités pratiques permettant d'effectuer des opérations sur des tableaux (n-arrays) et des matrices en Python. Elle permet de traiter des tableaux qui stockent des valeurs du même type de données et facilite l'exécution d'opérations mathématiques sur les tableaux (et leur vectorisation).

- Pandas :

Pandas est une bibliothèque créée pour aider les développeurs à travailler intuitivement avec des données « étiquetées » et « relationnelles ». Elle est basée sur deux structures de données principales : « Série » (unidimensionnelle, comme une liste Python) et « DataFrame » (bidimensionnelle, comme un tableau à plusieurs colonnes). Pandas permet de convertir des structures de données en objets DataFrame, de gérer les données manquantes et d'ajouter/supprimer des colonnes de DataFrame, d'imputer

les fichiers manquants et de tracer les données avec un histogramme ou une boîte à moustache.

- Requests :

La bibliothèque Requests sert à envoyer des requêtes HTTP en utilisant Python. Elle est simple et facile à utiliser avec de nombreuses fonctionnalités allant de la transmission de paramètres dans les URL à l'envoi d'en-têtes personnalisés et à la vérification SSL.

- **Visualisation de données :**

- Matplotlib :

Matplotlib est une bibliothèque scientifique de données standard qui aide à générer des visualisations de données telles que des diagrammes et des graphiques bidimensionnels (histogrammes, diagrammes de dispersion, graphiques de coordonnées non cartésiennes). Matplotlib est l'une de ces bibliothèques de tracés qui sont vraiment utiles dans les projets de science des données, elle fournit une API orientée objet pour intégrer des tracés dans les applications.

C'est grâce à cette bibliothèque que Python peut rivaliser avec des outils scientifiques comme MatLab ou Mathematica.

VI. LES DOMAINES D'APPLICATION DU WEB SCRAPING :

Le web scraping peut être appliqué dans plusieurs activités au sein des entreprises.

Parmi ces domaines on peut citer :

- **Etude du marché :**

Les entreprises peuvent utiliser le web scraping pour leurs études de marché. Les données de haute qualité obtenues en grands volumes peuvent être très utiles aux entreprises pour analyser les tendances de consommation et comprendre dans quelle direction l'entreprise doit se diriger à l'avenir.

- **Apprentissage automatique :**

Les modèles d'apprentissage automatique ont besoin de données brutes pour évoluer et s'améliorer en précision. L'apprentissage automatique alimente les merveilles technologiques d'aujourd'hui, comme les voitures sans conducteur, les vols spatiaux, la reconnaissance d'images et de la parole. Cependant, ces modèles ont besoin d'énormément de données variées pour améliorer leur précision et leur fiabilité. Les outils du web scraping peuvent permettre de scraper une grande variété de données, de textes et d'images en un temps relativement court, pour alimenter automatiquement ces modèles.

Un bon projet de web scraping garantit que vous obtenez les données que vous recherchez tout en ne perturbant pas les sources de données.

- **Analyse des sentiments :**

Si les entreprises veulent comprendre le sentiment général des consommateurs à l'égard de leurs produits, l'analyse des sentiments est indispensable. Les entreprises

peuvent utiliser le web scraping pour collecter des données à partir des réseaux sociaux tels que Facebook et Twitter afin de connaître le sentiment général sur leurs produits. Cela les aidera à créer des produits que les gens désirent et à prendre de l'avance sur leurs concurrents.

- **Marketing par e-mail :**

Les entreprises peuvent également utiliser le Web scraping pour le marketing par e-mail. Elles peuvent collecter des identifiants d'email à partir de divers sites en utilisant le web scraping et ensuite envoyer des emails promotionnels et marketing en masse à toutes les personnes possédant ces identifiants d'email. Il existe encore plein d'autres applications du web scraping mais nous avons vus les principales. Dans la section suivante nous parlerons des outils les plus utilisés pour le web scraping.

- **Surveillance des actualités :**

Le web scraping des sites d'actualités peut fournir des rapports détaillés sur l'actualité à une entreprise. C'est d'autant plus essentiel pour les entreprises du secteur du journalisme ou qui dépendent de l'actualité quotidienne pour leur fonctionnement. Après tout, les rapports d'actualité peuvent faire ou défaire une entreprise en une seule journée.

- **Contrôle des prix :**

Le web scraping peut être utilisé par les entreprises pour récupérer les données des produits concurrents et ensuite de les utiliser pour fixer le prix optimal de leurs produits afin d'obtenir un revenu maximal.

Dans le schéma suivant, illustre quelques domaines applicatifs du web scraping :

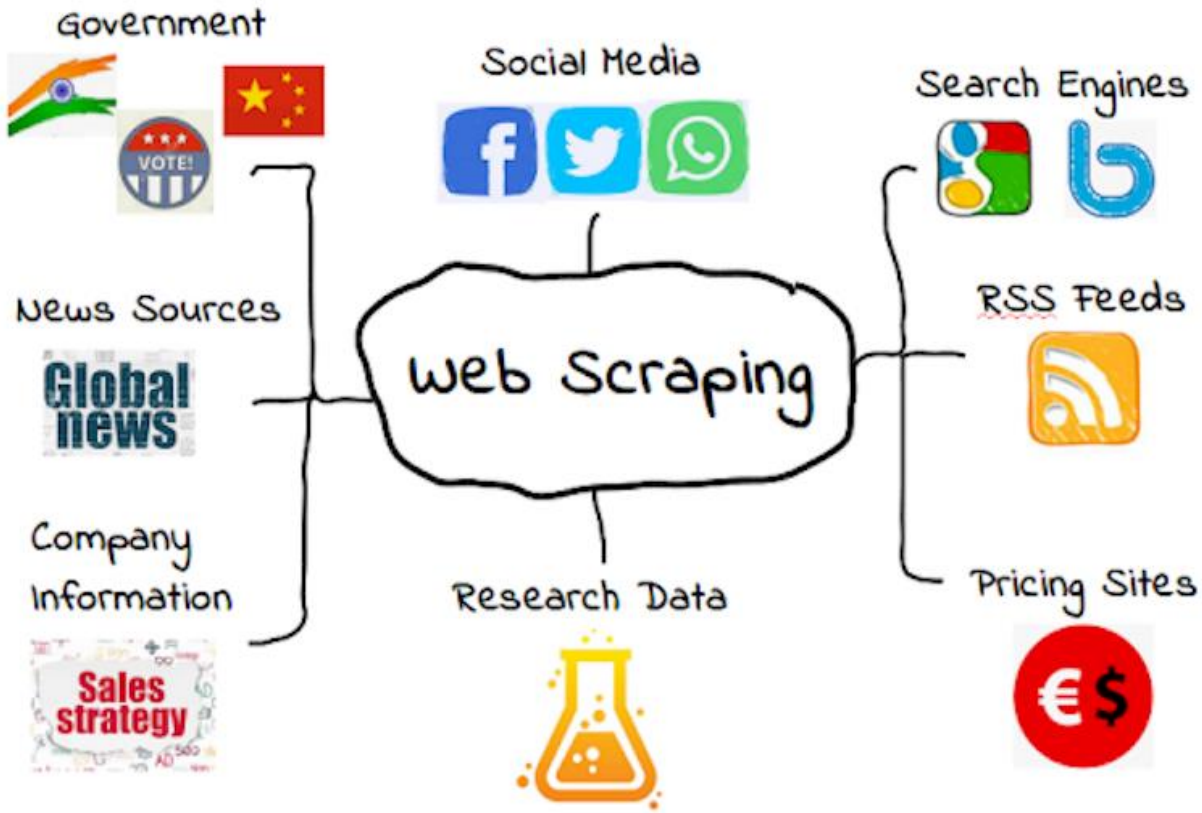

Figure 11: Les domaines d'application du web scraping

VII. CONCLUSION :

Le tableau ci-dessous illustre des bibliothèques en python nécessaires dans le développement d'un outil web scraping, ainsi que, leurs avantages et inconvénients :

Librairie python	Avantages	Inconvénients
Scrapy	+ Un outil de référence pour les amateurs de Python + Framework très efficace & bien documenté	- Limite concernant les pages générées en JavaScript
Requests	+ Simplifie l'écriture de code en python + Rendre les requêtes HTTP plus simples et plus conviviales	
BeautifulSoup	+ Composé de différents outils d'analyse syntaxique tels que html, parser, lxml et HTML5lib + Facilité d'utilisation + Large documentation	
Selenium	+ Puissant, permet de passer sur à peu près tous les sites + Coûteux à mettre en place (en temps ou en argent)	- Connaissances préalables requises

Tableau 2: Tableau comparatif de 4 outils du web scraping

Pour plus de détail, la figure suivante présente une large comparaison entre les bibliothèques en python nécessaires pour le web scraping :

	Scrapy	Requests	Beautiful Soup	Selenium
c'est quoi?	framework web scraping	bibliothèque	bibliothèque	bibliothèque
but?	complète les solutions web scraping	simplifie les requetes HTTP	analyse de données	navigateur web scriptable pour rendre javascript
cas d'utilisation idéal	développement de récurrents ou à grande échelle les projets web scraping	taches simples, non récurrentes du web scraping	taches simples, non récurrentes du web scraping	web scraping à petite échelle de sites web lourds en javascript
stockage de données	json, xml, csv	développer votre propre type de stockage	développer votre propre type de stockage	personnalisable
sélecteurs disponibles	Jcss & Xpath	N/A	css	Jcss & Xpath
asynchrone	oui	non	non	non
support javascript	oui	N/A	non	oui
documentation	excellente	excellente	excellente	bonne
courbe d'apprentissage	facile	trés facile	trés facile	facile
écosystème	large écosystème de développeurs contribuant aux projets ey supports sur github et stackoverflow	quelques projets	quelques projets	quelques projets

Figure 12: Comparaison des librairies python pour le web scraping

► Section 3 : le web scraping pour le e-commerce :

I. INTRODUCTION :

Bien que le web scraping a était implémenté dans plusieurs domaines, l'un des domaines les plus mis en œuvre le e-commerce. Dans ce chapitre, on va mettre l'accent sur l'extraction des prix des sites d'achat et de vente, les plus populaires, afin de réaliser une comparaison des prix pour faciliter le choix du client.

II. QU'EST-CE QUE LE E-COMMERCE ?

Le commerce électronique ou vente en ligne, désigne l'échange de biens et de services entre deux entités sur les réseaux informatiques, notamment Internet. Le magasinage en ligne gagne de plus en plus en popularité en raison de la rapidité et de la facilité d'utilisation qu'il offre aux clients.

Selon les modèles d'affaires, le e-commerce se manifeste en quatre formes :

- **B2C** (Business-to-Consumer)

Le e-commerce B2C englobe les transactions effectuées entre une entreprise et un consommateur. C'est l'un des modèles de vente les plus utilisés dans le contexte du commerce électronique. Lorsque vous achetez des chaussures chez un détaillant de chaussures en ligne, il s'agit d'une transaction d'entreprise à consommateur.

- **B2B** (Business-to-Business)

Le commerce électronique inter-entreprises concerne les ventes effectuées entre des entreprises, comme un fabricant et un grossiste ou un détaillant. Ce type de e-commerce n'est pas orienté vers le consommateur et n'existe qu'entre les entreprises.

Le plus souvent, les ventes inter-entreprises se concentrent sur les matières premières ou les produits qui sont reconditionnés ou combinés avant d'être vendus aux clients.

- **C2C** (Consumer-to-Consumer)

L'une des premières formes de e-commerce est le modèle C2C. Il se rapporte à la vente de produits ou de services entre, vous l'avez deviné : les clients. Il s'agit notamment des relations de vente de consommateur à consommateur comme celles observées sur eBay ou Amazon, par exemple.

- **C2B** (Consumer-to-Business)

Le modèle du C2B renverse le modèle traditionnel du commerce électronique (et c'est ce que l'on voit couramment dans les projets de crowdfunding). Le C2B signifie que les consommateurs individuels mettent leurs produits ou services à la disposition des acheteurs commerciaux. Un exemple de cela serait un modèle commercial comme iStockPhoto. Les photos d'archives sont disponibles en ligne pour achat directement auprès de différents photographes.

III. LES AVANTAGES ET INCONVÉNIENTS DU E-COMMERCE :

1. Les avantages :

- **Plus de clients :**

Ni une boutique locale ni une entreprise implantée dans plusieurs villes ne peut atteindre autant de personnes qu'un e-commerce. Pouvoir acheter et vendre depuis n'importe quel endroit du globe élargit considérablement le public cible et permet d'obtenir davantage de clients.

- **Pas d'horaire :**

A l'inverse des boutiques traditionnelles, qui sont rarement ouvertes 24/24h, le e-commerce n'a pas d'horaires. Le site web reste ouvert et accessible au public toute la journée et le client peut donc faire ses achats à n'importe quelle heure.

- **Moindre cout :**

Pouvoir se passer d'un établissement physique permet de réduire les coûts par rapport au fonctionnement d'un commerce traditionnel. Et si le e-commerce fonctionne en mettant en contact des fournisseurs avec des acheteurs, il n'y aura même pas de frais de production (cas du drop shipping, dont nous vous parlions plus haut).

- **Davantage de marge :**

La réduction des coûts et l'augmentation du nombre de clients permettent d'atteindre une plus grande marge qu'avec un commerce traditionnel, même en baissant les prix. On vend davantage et on gagne plus d'argent.

- **Scalabilité :**

Dans un e-commerce, vous pouvez vendre à une ou mille personnes en même temps. Dans une entreprise physique, il y a toujours une limite au nombre de clients que vous pouvez servir à la fois ; dans le commerce électronique, la limite est votre capacité d'attirer des visiteurs. Et bien sûr, celle de votre serveur informatique.

2. Les inconvénients :

- **Manque de confiance :**

Bien que les passerelles et les moyens de paiement aient fait d'énormes progrès et soient aujourd'hui aussi sûrs que dans les boutiques physiques, beaucoup de personnes continuent de ne pas faire entièrement confiance aux transactions en ligne.

- **Produits et services non touchable :**

En tant que client, on aime avoir la sensation de faire un bon achat. On aime voir le produit et le toucher pour se rendre compte de sa qualité et cela ne peut pas se faire dans un e-commerce.

- **Connexion internet indispensable :**

C'est évident, mais afin de vendre et d'acheter sur internet, un dispositif connecté à internet est nécessaire. Cela ne concerne pas la majorité des activités en ligne, mais peut représenter un problème pour certains secteurs où le public cible est plus âgé ou moins familiarisé avec les nouvelles technologies.

- **Difficultés techniques :**

Faire face à des thématiques inconnues est le quotidien des entrepreneurs, que ce soit hors ligne ou en ligne. Dans le cas d'un e-commerce, la partie technologique requiert un minimum de connaissances technologiques, dont tout le monde ne dispose pas.

- **Concurrence :**

La barrière d'entrée économique pour créer un e-commerce n'est pas aussi élevée que pour un commerce physique. La concurrence est donc plus importante, et il faut se montrer plus compétent que les autres.

- **Temps pour obtenir des résultats :**

Quand un commerce physique ouvre ses portes, les clients qui passent devant le voient. Obtenir de la visibilité pour un commerce en ligne est plus difficile qu'il n'y paraît.

IV. EXEMPLE DU WEB SCRAPING POUR L'EXTRACTION DES PRIX :

Bien que le web scraping a été émergé dans plusieurs domaines. Cette figure représente une statistique d'utilisation du web scraping dans les différents domaines :

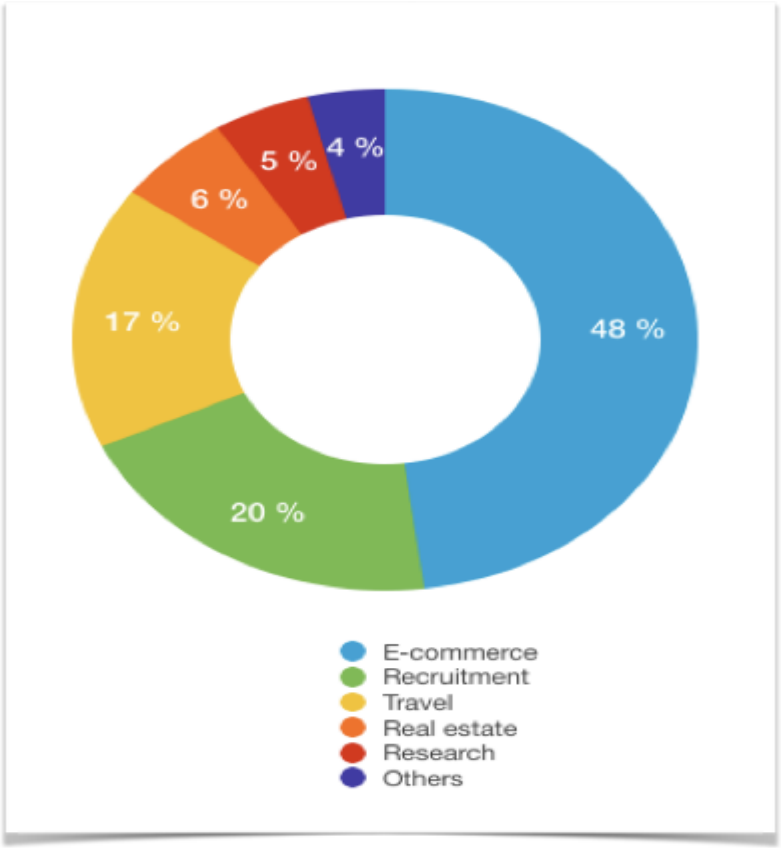

Figure 13: Statistiques d'utilisation du web scraping

1. Price scraping :

Les transactions d'achat et de vente s'effectuent chaque jour. Plusieurs internautes visitent souvent les sites du e-commerce afin de chercher un tel produit. Sans doute le facteur principal est le prix d'article. Pour un produit cible, le prix varie entre un site web et un autre. Le client doit trouver, alors, le prix optimal.

Au lieu de mettre en œuvre la recherche manuellement, on a besoin d'un outil du web scraping pour montrer les listes des prix d'un produit particulier automatiquement.

Cet outil a pour mission la navigation des sites web, en fonction d'une technologie, afin d'extraire les différents prix d'un article donné.

2. Les avantages d'un Price scraping :

- **En tant que vendeur :**

Tout d'abord, les sites de comparaison de prix pour les achats en ligne offrent aux entreprises la possibilité d'élargir leurs canaux de vente. En outre, les fournisseurs ont la possibilité d'utiliser les données de l'agrégateur pour fixer des prix compétitifs, ainsi que de recevoir du trafic supplémentaire du public cible. Ainsi, les principaux avantages comprennent:

- L'augmentation rapide du trafic sur le site en ligne.
- Plus de ventes réalisées.
- Un canal supplémentaire d'interaction avec le public.
- L'opportunité de recueillir des avis d'utilisateurs et de travailler sur le développement de la ressource.
- Fonctionnalité simple et abordable qui permet de travailler avec un site Web de comparaison de prix sans aide tierce.

➢ **En tant que client :**

Un bon site de comparaison de prix est particulièrement nécessaire pour ceux qui font constamment des achats en ligne et ne veulent pas passer trop de temps à essayer de trouver «où est le moins cher». Ainsi, les avantages comprennent:

- La possibilité de voir un grand nombre de boutiques en ligne sur une seule plateforme.
- Une réponse rapide à la question de savoir où est le produit le moins cher;
- Beaucoup d'avis de vrais clients.
- Sélection des produits les plus populaires (ils sont généralement placés en haut des listes).
- Différents tris d'offres de produits.

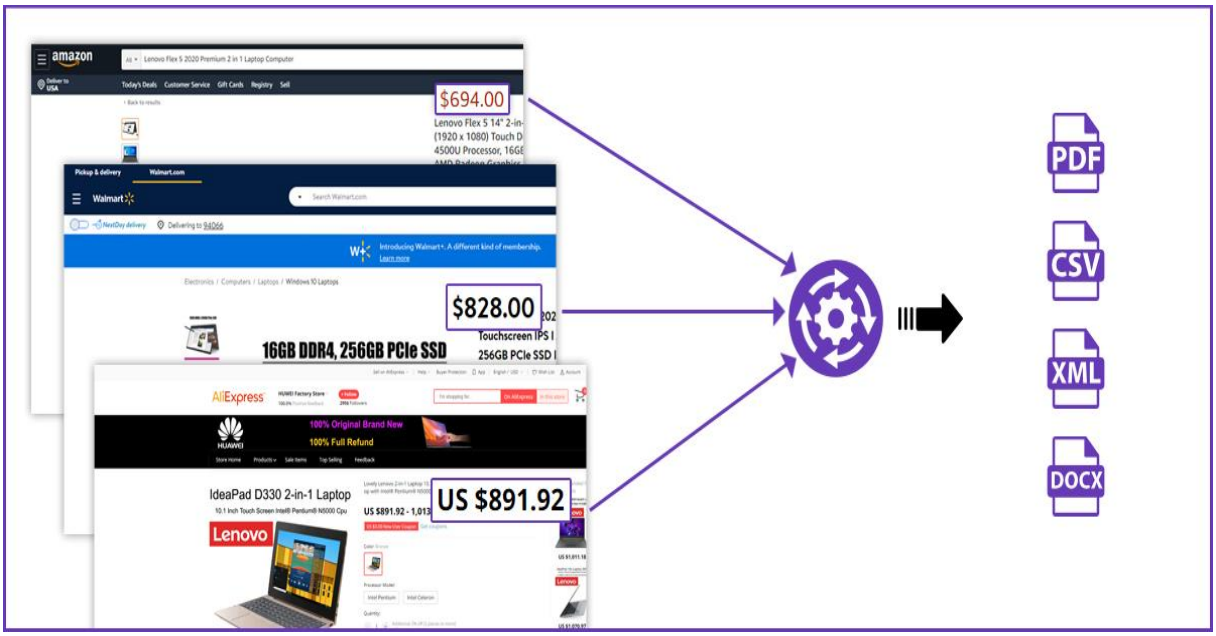

Figure 14: Exemple de Price scraping

3. Etude du marché :

Parmi les outils du Price scraping sur le marché, nous allons présenter quelques exemples :

Pro WebScraper

Parsehub

Octoparse

Apify

Import.io

La figure ci-dessous représente une liste des outils du Price scraping en ordre :

Figure 15: Trie de 5 meilleurs outils du web scraping

La figure suivante explique une comparaison entre des outils du Price scraping selon des différents facteurs :

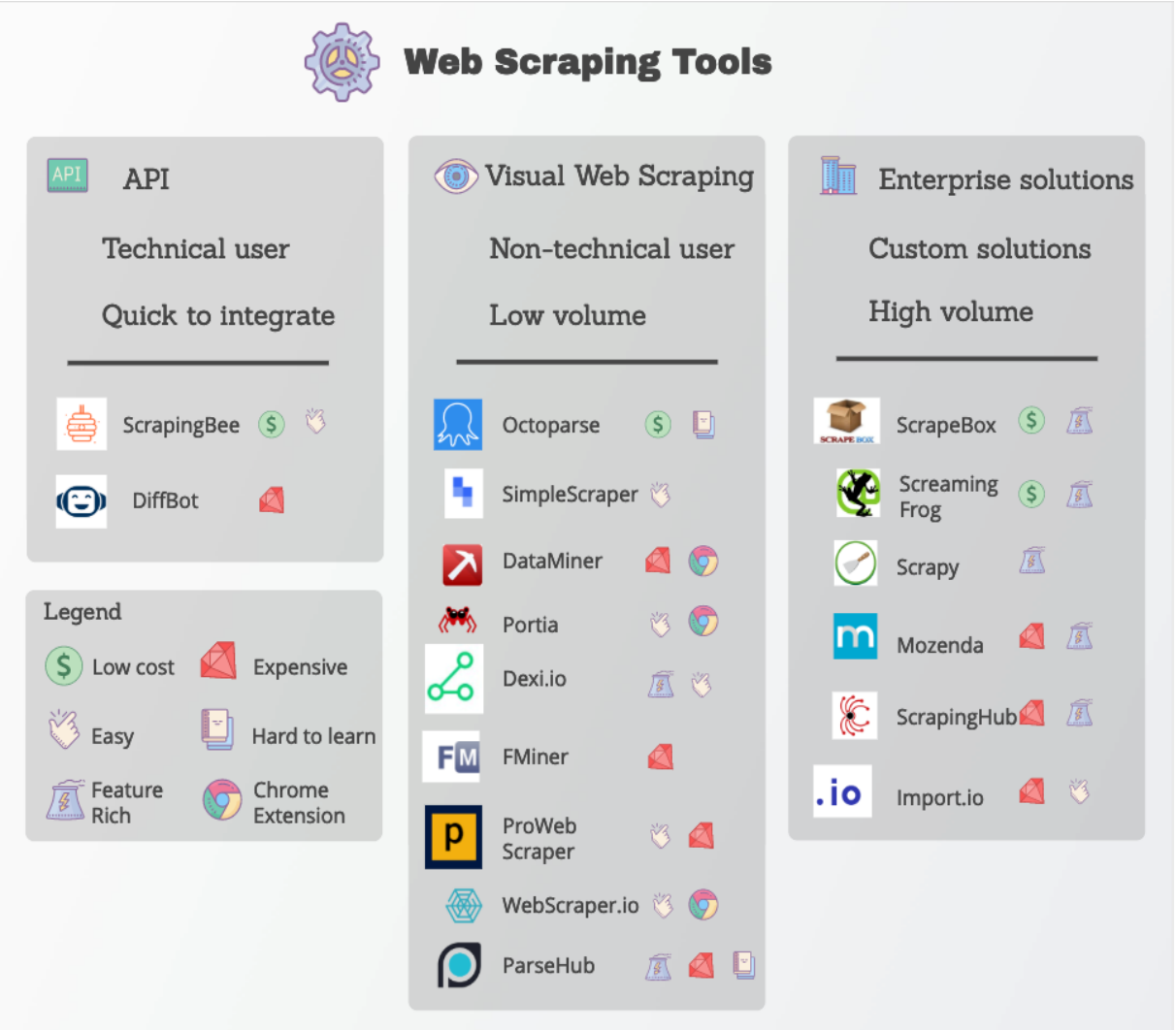

Figure 16: Différents outils du web scraping

VII. CONCLUSION :

Dans ce chapitre nous avons essayé, en premier lieu, d'élaborer une étude détaillée sur le domaine de l'intelligence artificielle et ses composants. En second lieu, nous avons présenté le domaine du web scraping. Enfin, nous avons mis l'accent sur le domaine de Price scraping.

Le chapitre suivant représente une implémentation de notre application.

CHAPITRE 3 : REALISATION :

I. INTRODUCTION :

Ce chapitre est consacré à l'étude de l'environnement matériel et logiciels employé pour l'implémentation de cette application. Ensuite, on va présenter les interfaces nécessaires à partir des figures de captures d'écran.

II. ENVIRONNEMENT DE TRAVAIL :

1. Environnements matériels :

L'application est réalisée sur mon ordinateur portable Lenovo ayant les caractéristiques suivantes :

- Processeur : Inter® core™ i3-6006 CPU @ 2.00 GHZ 2.00GHZ
- RAM : 4 GO
- Système d'exploitation : SE 64 bits

2. Environnements logiciels :

- **Python :**

Python est le langage de programmation open source le plus employé par les informaticiens. Ce langage s'est propulsé en tête de la gestion d'infrastructure, d'analyse de données ou dans le domaine du développement de logiciels. En effet, parmi ses qualités, Python permet notamment aux développeurs de se concentrer sur ce qu'ils font plutôt que sur la manière dont ils le font. Il a libéré les développeurs des

contraintes de formes qui occupaient leur temps avec les langages plus anciens. Ainsi, développer du code avec Python est plus rapide qu'avec d'autres langages.

- **Vs Code :**

Visual Studio Code est un éditeur de code extensible développé par Microsoft pour Windows, Linux et macOS2. Les fonctionnalités incluent la prise en charge du débogage, la mise en évidence de la syntaxe, la complétion intelligente du code, les snippets, la refactorisation du code et Git intégrer. Les utilisateurs peuvent modifier le thème, les raccourcis clavier, les préférences et installer des extensions qui ajoutent des fonctionnalités supplémentaires.

- **UI design :**

Le terme UI fait référence à l'interface par le biais de laquelle l'utilisateur interagit : être un site web, une application mobile ou un logiciel. Le travail de l'UI designer est important pour toute entreprise ou organisation souhaitant marquer sa présence sur le web. UI est l'abréviation d'user interface ou interface utilisateur. L'UI design se rapporte donc à l'environnement graphique dans lequel évolue l'utilisateur d'un logiciel, d'un site web ou d'une application. La mission de l'UI designer consiste à créer une interface agréable et pratique, facile à prendre en main.

Ainsi, l'UI design fait partie de l'UX design, en cela qu'il travaille à donner la meilleure expérience possible à l'utilisateur, mais il s'attache plus particulièrement aux éléments perceptibles : éléments graphiques, boutons, navigation, typographie…

3. Réalisation :

La partie réalisation du projet est dédiée à mettre en œuvre les caractéristiques du notre projet, en se référant à des captures écrans avec des explications.

Notre travail est divisé en trois parties :

- **Quick Scrap** : Un simple scraping ; Scraper les liens, les classes, les paragraphes, les titres d'un site web. Ainsi que, télécharger automatiquement la page html du site, la partie CSS, la partie JavaScript, les PDF et les images. Cette partie est nécessaire afin de comprendre la structure générale d'un site web.

- **Advanced Scrap** : Un scraping plus avancé ; Collecter des données d'un site web. Ces données peuvent être des noms des produits avec leurs prix, des contacts des médecins, des tableaux des statistiques… Après avoir scrapés les informations nécessaires, nous pouvons les enregistrer sous la forme des fichiers CSV. Cette partie du web scraping nous permet aussi de généraliser un graphe correspond à des données sélectionnées et de le former dans un PDF.

- **Commercial Scrap** : Un comparateur des prix; Exemple d'un outil commercial. Il suffit de passer en entrée un nom de produit, le programme affiche les différents prix de ce produit dans des sites e-commerce. Ce résultat aide l'utilisateur à prendre une décision d'achat.

3.1. Quick Scrap :

En premier lieu, nous avons commencé par un simple scraping. Cette partie est consacrée à scraper les Links, Classes, ID, Titles, Paragraphes. En plus, nous pouvons télécharger la page html du site, Css, JavaScript, Images et PDF.

Après avoir scraper les données nécessaires, nous pouvons générer un graph correspond à la partie statistique.

Le processus relatif à la première interface « Quick Scrap » est illustré à partir de la figure suivante :

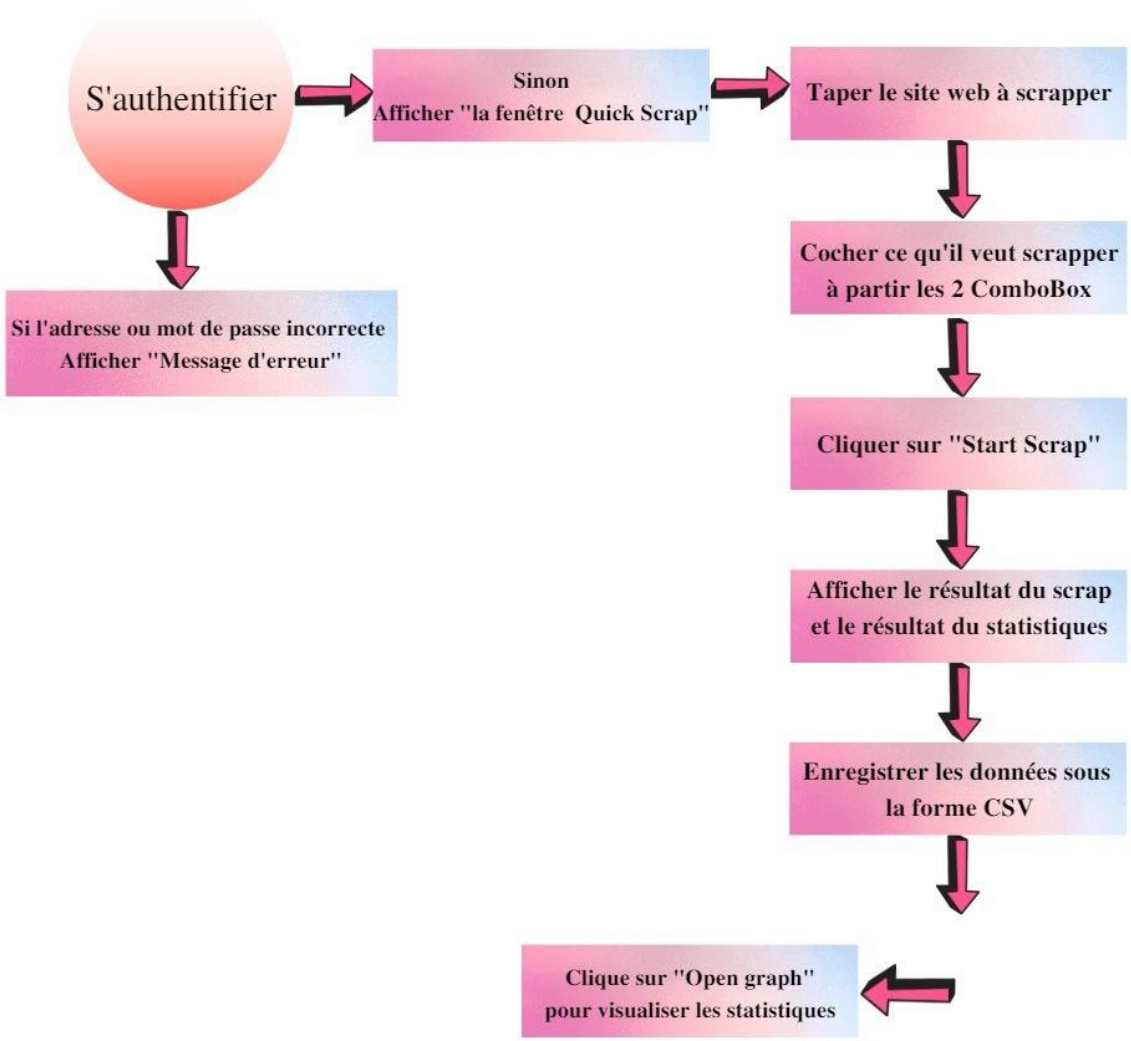

Figure 17: Le processus de l'interface Quick Scrap

❖ Interface « Login » :

En premier lieu, l'utilisateur doit s'authentifier. Une interface de login se présente comme la figure ci-dessous.

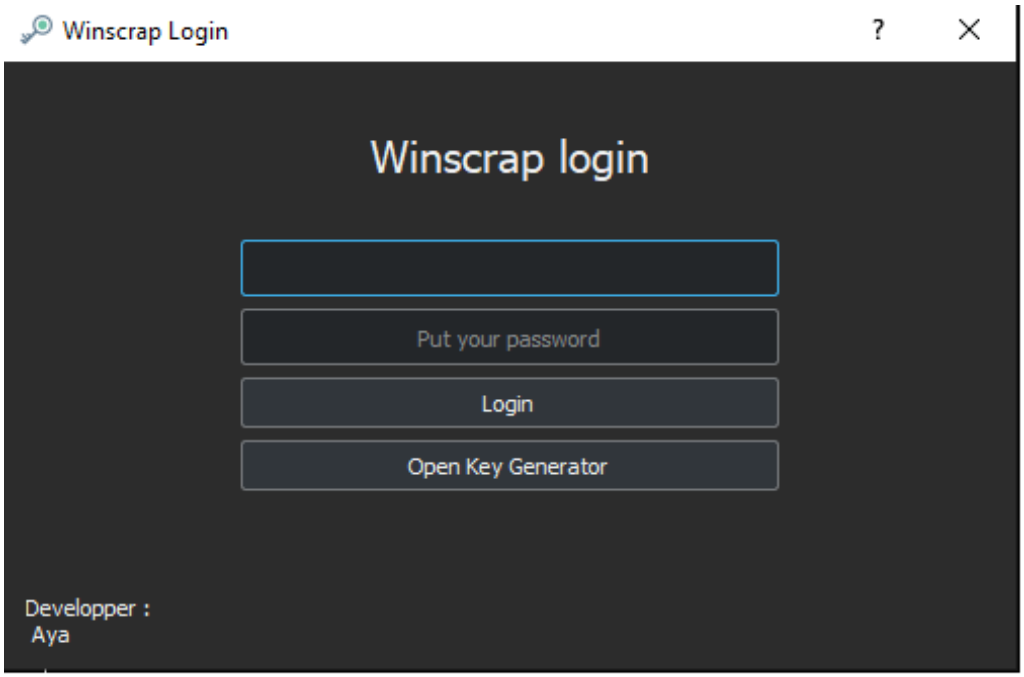

Figure 18: Interface de login

❖ Interface Quick Scrap :

Après avoir entré les paramètres corrects de login, la deuxième fenêtre se présente. L'utilisateur doit passer en entrée le lien du site web à scrapper. Ensuite, il faut sélectionner le bon choix depuis les deux comboBox. Si l'URL est valide le résultat s'affiche au-dessus avec une simple statistique. Comptant les lignes scrapés, l'utilisateur peut généraliser un graphe à l'aide des statistiques apparut.

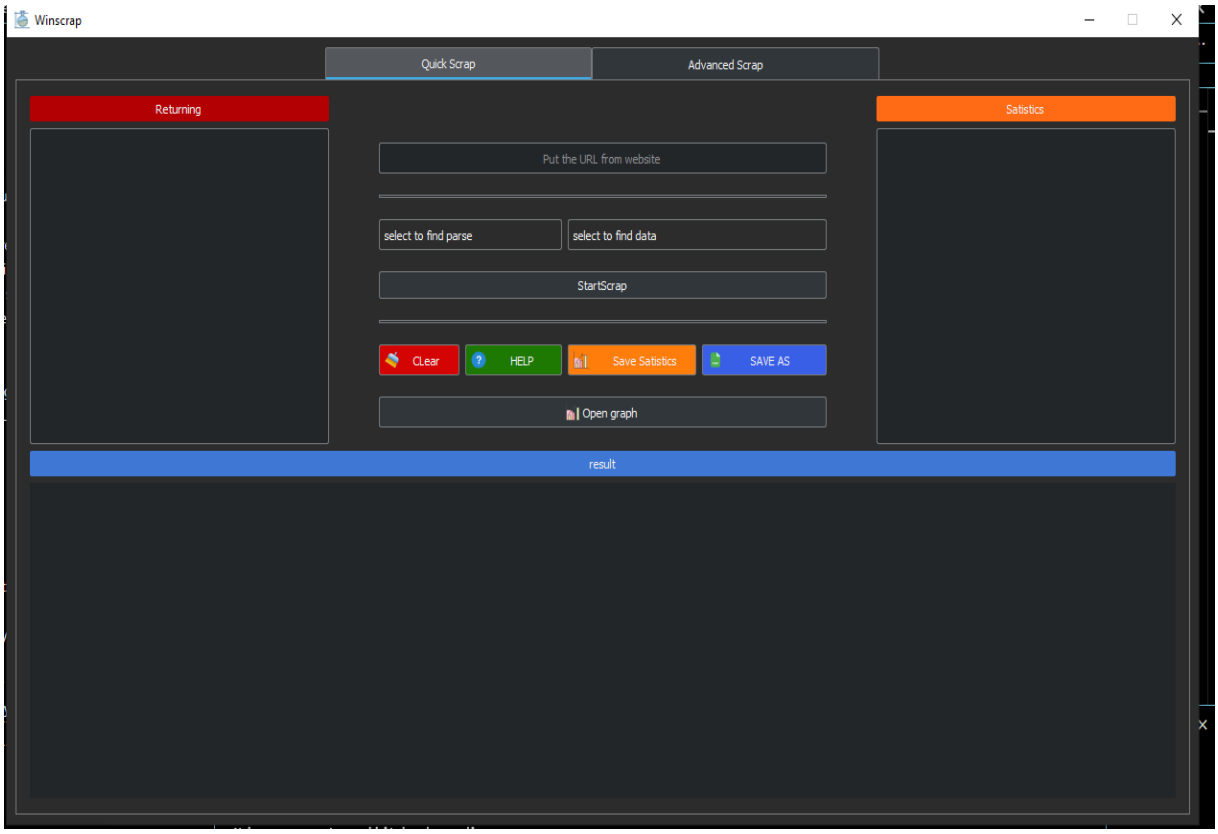

Figure 19: Interface Quick Scrap

❖ Exemple sur le site web www.jumia.com.tn:

➢ Scrapper les liens :

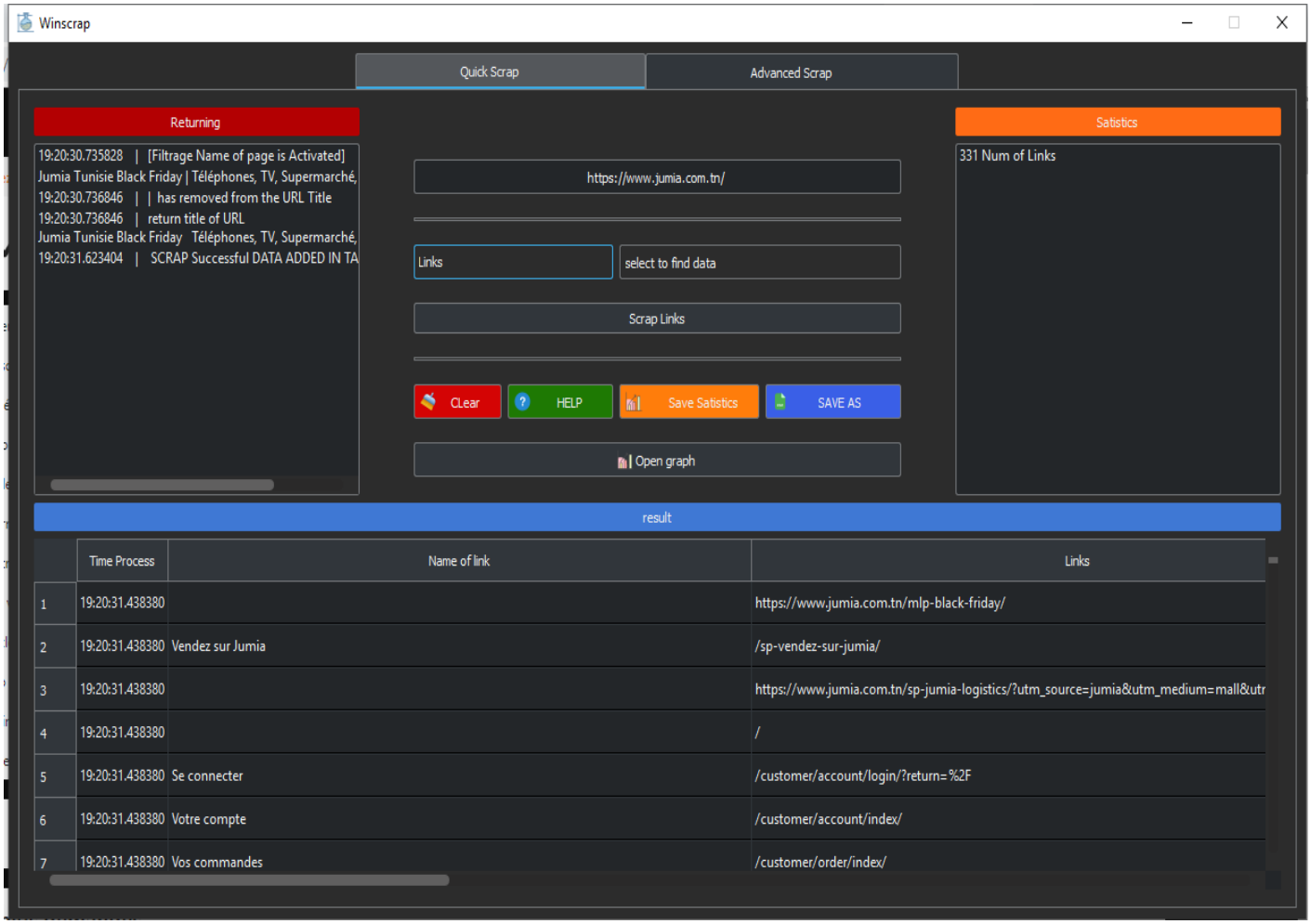

Figure 20: Exemple de Scrape des liens

➢ Scrapper les classes :

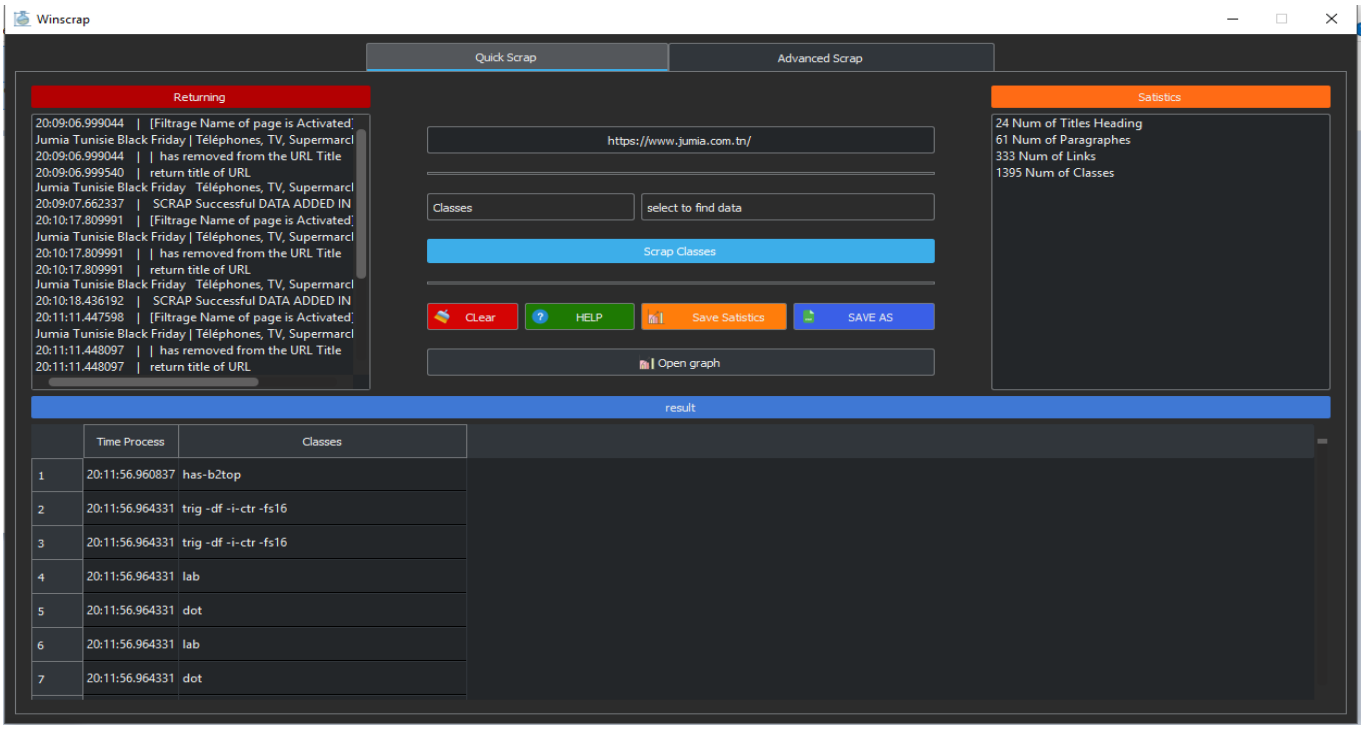

Figure 21:Exemple de Scrape des classes

➢ Scrapper les titres :

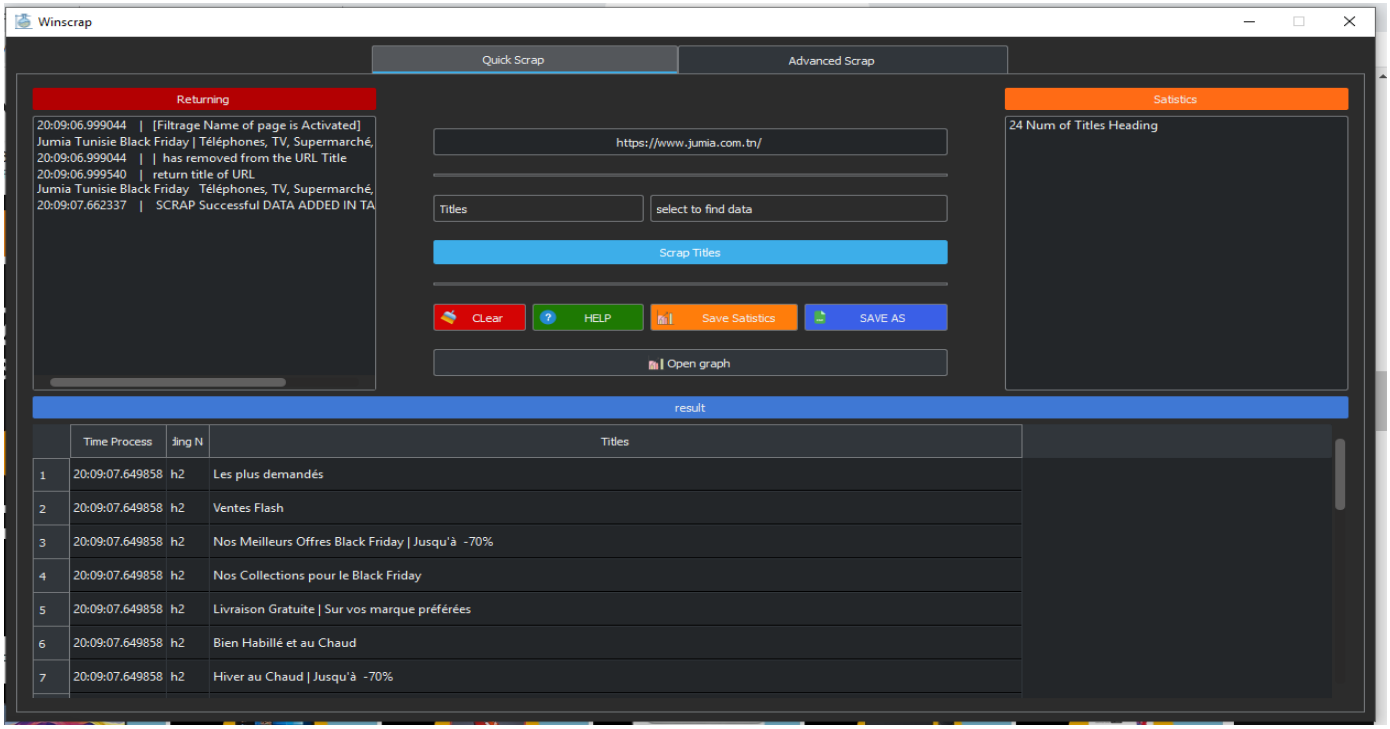

Figure 22:Exemple de Scrape des titres

- **Visualisation des statistiques :**

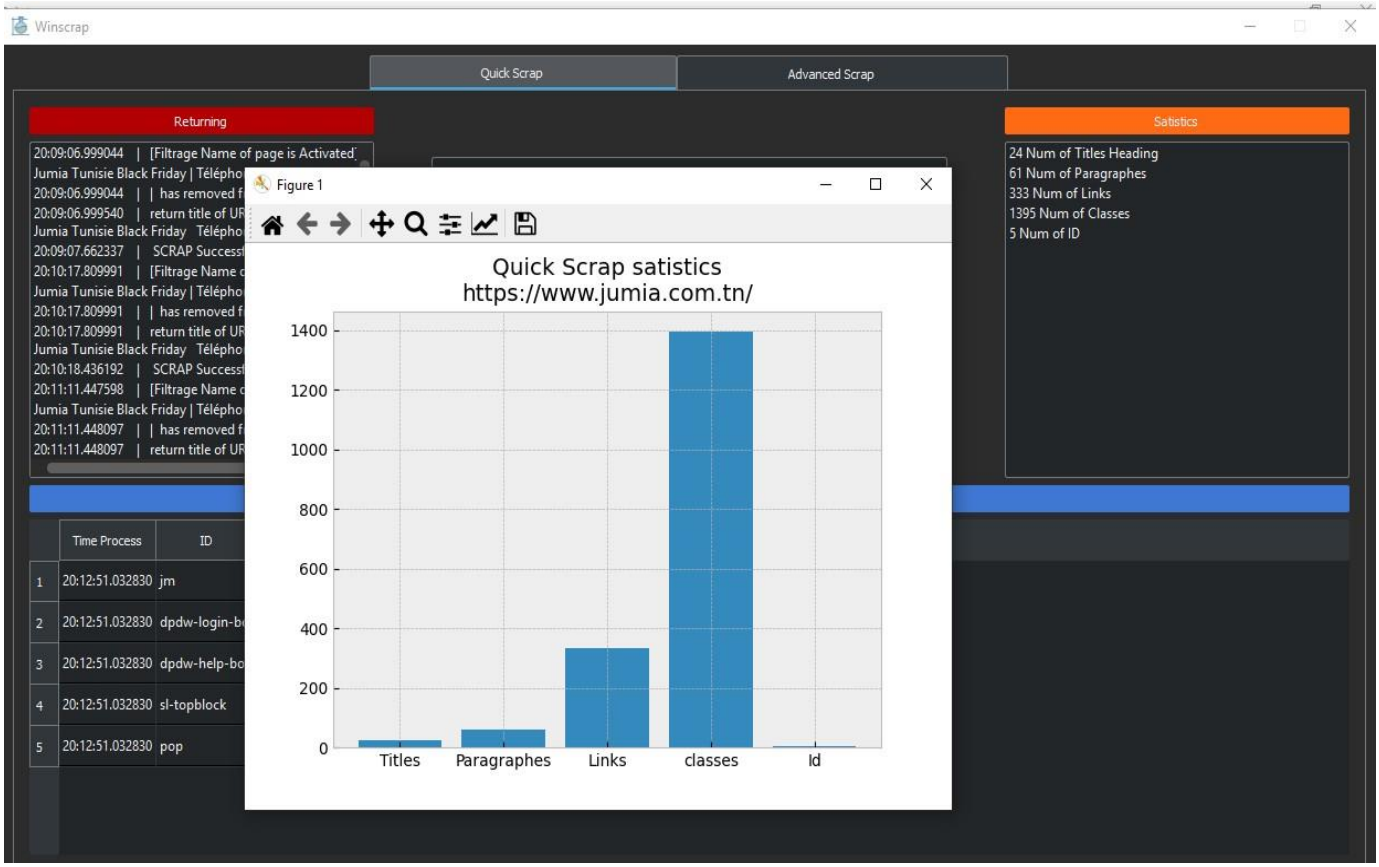

Figure 23:Exemple des statistiques

❖ Enregistrement des données : (Exemple d'enregistrement des liens scrapés) :

	A	B	C	D	E
1		Time Process	Name of link	Links	
2	0	19:20:31.438380		https://www.jumia.com.tn/mlp-black-friday/	
3	1	19:20:31.438380	Vendez sur Jumia	/sp-vendez-sur-jumia/	
4	2	19:20:31.438380		https://www.jumia.com.tn/sp-jumia-logistics/?utm_source=jumia&utm_medium=mall&utm_campaign=venturebar	
5	3	19:20:31.438380		/	
6	4	19:20:31.438380	Se connecter	/customer/account/login/?return=%2F	
7	5	19:20:31.438380	Votre compte	/customer/account/index/	
8	6	19:20:31.438380	Vos commandes	/customer/order/index/	
9	7	19:20:31.438380	Votre liste d'envies	/customer/wishlist/index/	
10	8	19:20:31.438380	Centre d'assistance	/faq/	
11	9	19:20:31.438380	Passer et suivre ma commande	/sp-Aide-avec-commande/	
12	10	19:20:31.438380	Annuler ma commande	/sp-annulation-de-commande/	
13	11	19:20:31.438380	Retour & Remboursement	/sp-retour-remboursement/	
14	12	19:20:31.438380	Paiement et compte Jumia	/sp-paiement/	
15	13	19:20:31.438380	Panier	/cart/	
16	14	19:20:31.438380	Notification sur la confidentialité et les cookies	/sp-confidentialite/	
17	15	19:20:31.438380	Superette	/epicerie/	
18	16	19:20:31.438380	Maison & Bureau	/maison-cuisine-jardin/	
19	17	19:20:31.438380	Santé & Beauté	/beaute-hygiene-sante/	
20	18	19:20:31.438380	Téléphone & Tablette	/telephone-tablette/	
21	19	19:20:31.438380	Mode	/fashion-mode/	
22	20	19:20:31.438380	Informatique	/ordinateurs-accessoires-informatique/	
23	21	19:20:31.438380	Électroniques	/electronique/	
24	22	19:20:31.438380	Jeux vidéos & Consoles	/jeux-videos-consoles/	
25	23	19:20:31.438380	Articles de sport	/sports-loisirs/	
26	24	19:20:31.438380	Auto & Moto	/automobile-outils/	
27	25	19:20:31.438380	Jardin & Plein air	/terrasse-jardin-exterieur/	
28	26	19:20:31.438380	Autres catégories		
29	27	19:20:31.438380		https://www.jumia.com.tn/mlp-black-friday/	

Figure 24: Exemple des liens scrapés

3.2. Advanced Scrap :

En second lieu, nous avons développé une interface d'un scraping plus avancé. Cette partie est réservée à collecter des données via des différents sites web.

La figure ci-dessous illustre le processus général concernant l'interface « Advanced Scrap ».

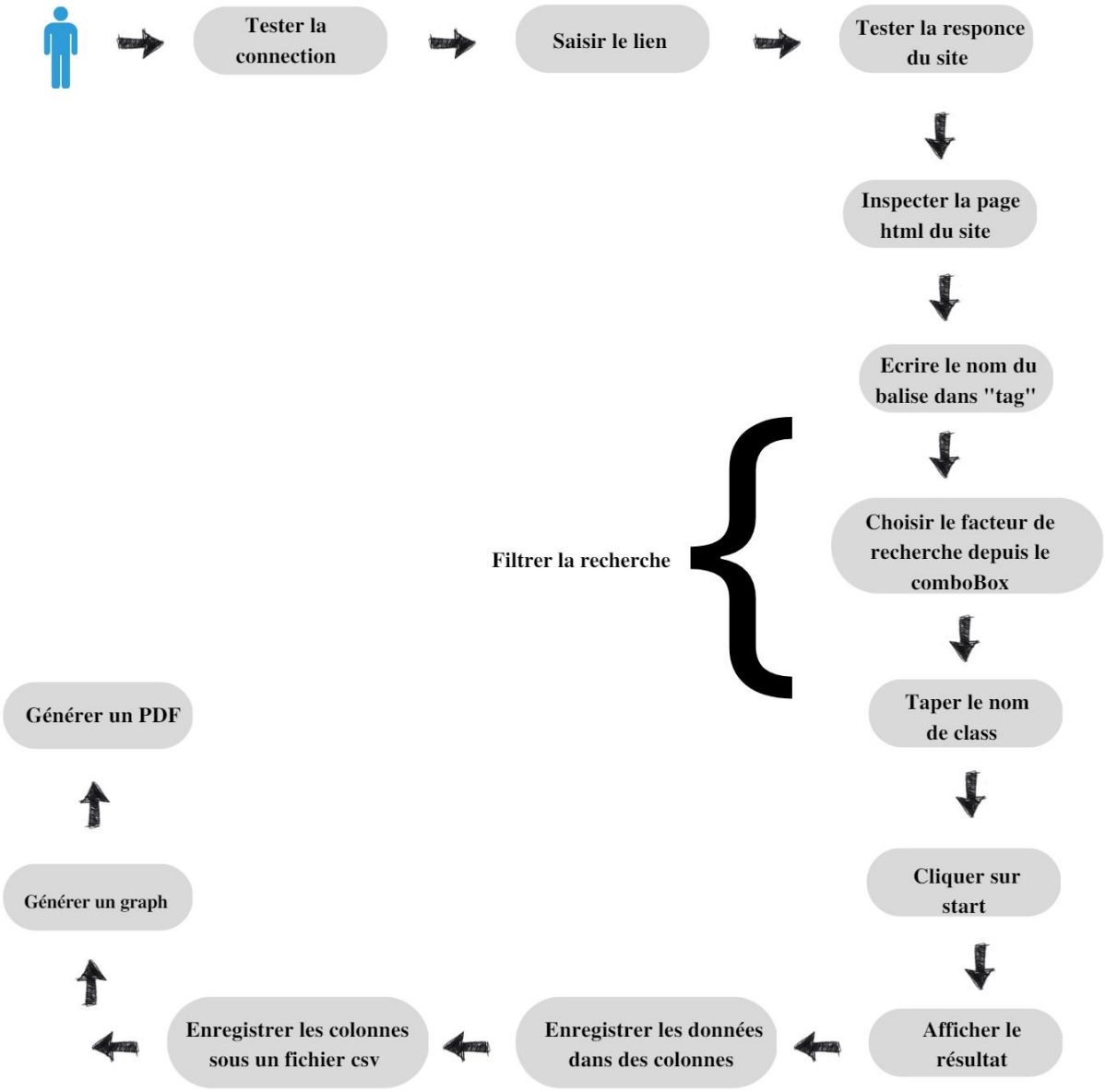

Figure 25: Processus de l'interface Advanced Scrap

❖ Interface « Advanced Scrap » :

A partir de cette interface, l'utilisateur peut scraper n'importe quelles données. Il doit premièrement inspecter la page web nécessaire afin de mieux comprendre l'intégralité des données à partir les balises.

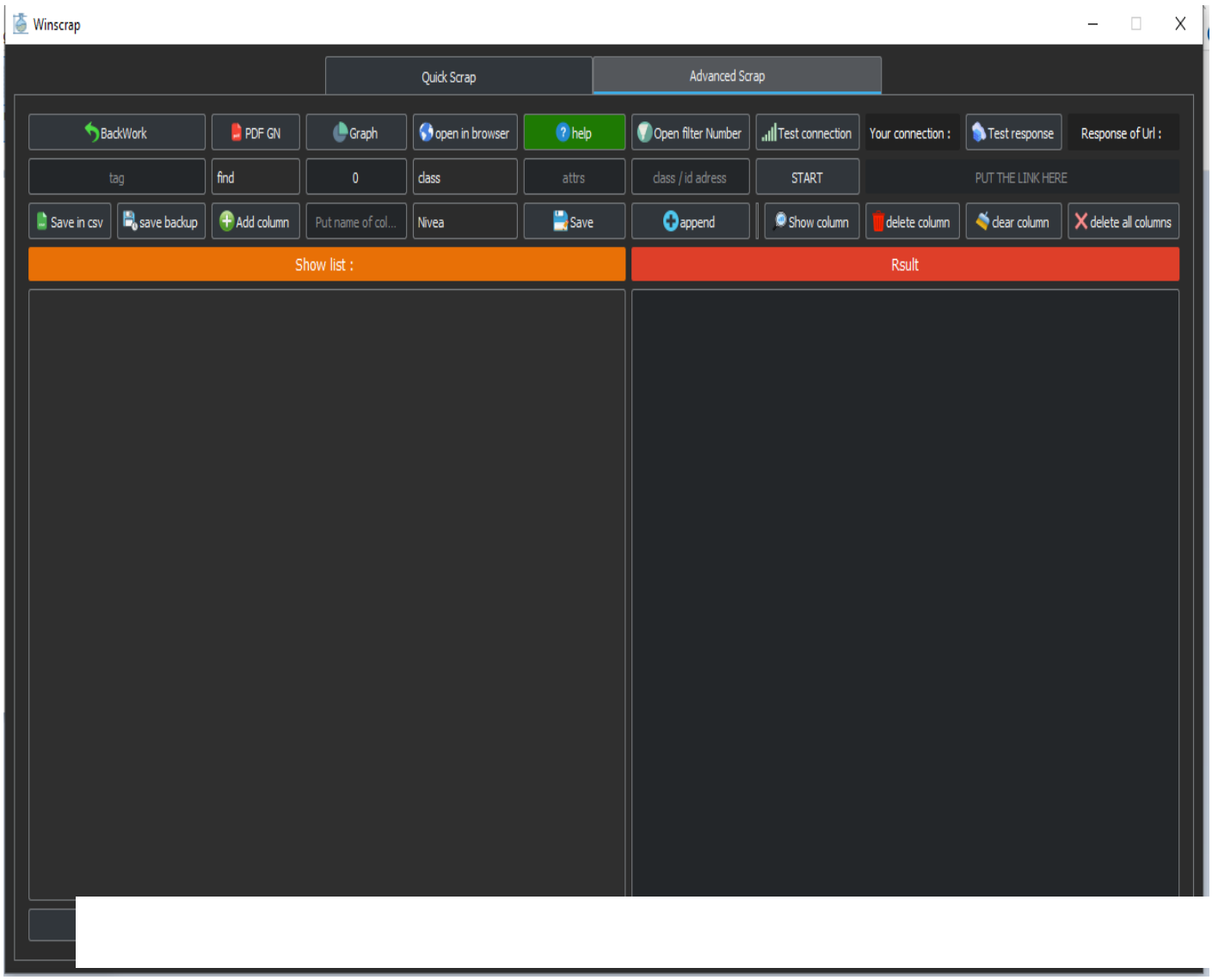

Figure 26: Interface "Advanced Scrap"

❖ Exemple sur le site web www.jumia.com.tn :

Dans cet exemple nous avons collectés les données des meilleures marques sous la catégorie « Santé et beauté ». Notre but est de déduire quelle est la marque la plus populaire.

➢ Inspecter le code html d'une page du site :

La figure suivante correspond le code html de la page web https://www.jumia.com.tn/mlp-boutique-nivea/. D'après, ce code nous avons déduit que les marques de beauté sont intégrées dans des classes sous le nom « Name » avec la balise « h3 ».

Figure 27: Code html d'une page web du site Jumia

➢ **Commencer la recherche :**

Nous pouvons maintenent déclenché la recherche des données nécessaires :

- Copier le lien de la page web.
- Passer en entrée le nom du balise « h3 » et le nom du classe « Name ».
- Filtrer la recherche : soit « find all » pour collecter tout les lignes, soit « find » pour afficher juste la première ligne.
- Cliquer sur le bouton « start ».
- Quand le résultat s'affiche, il faut enregistrer les données sous formes des colonnes CSV. Chaque marque est affecté à un colonne.

Figure 28: Exemple de collecte de données.

- **Visualiser le résultat :**

Après avoir enregistré tous les données scrapés ; nous pouvons généraliser un graphe des statistiques comme indique la figure ci-dessous :

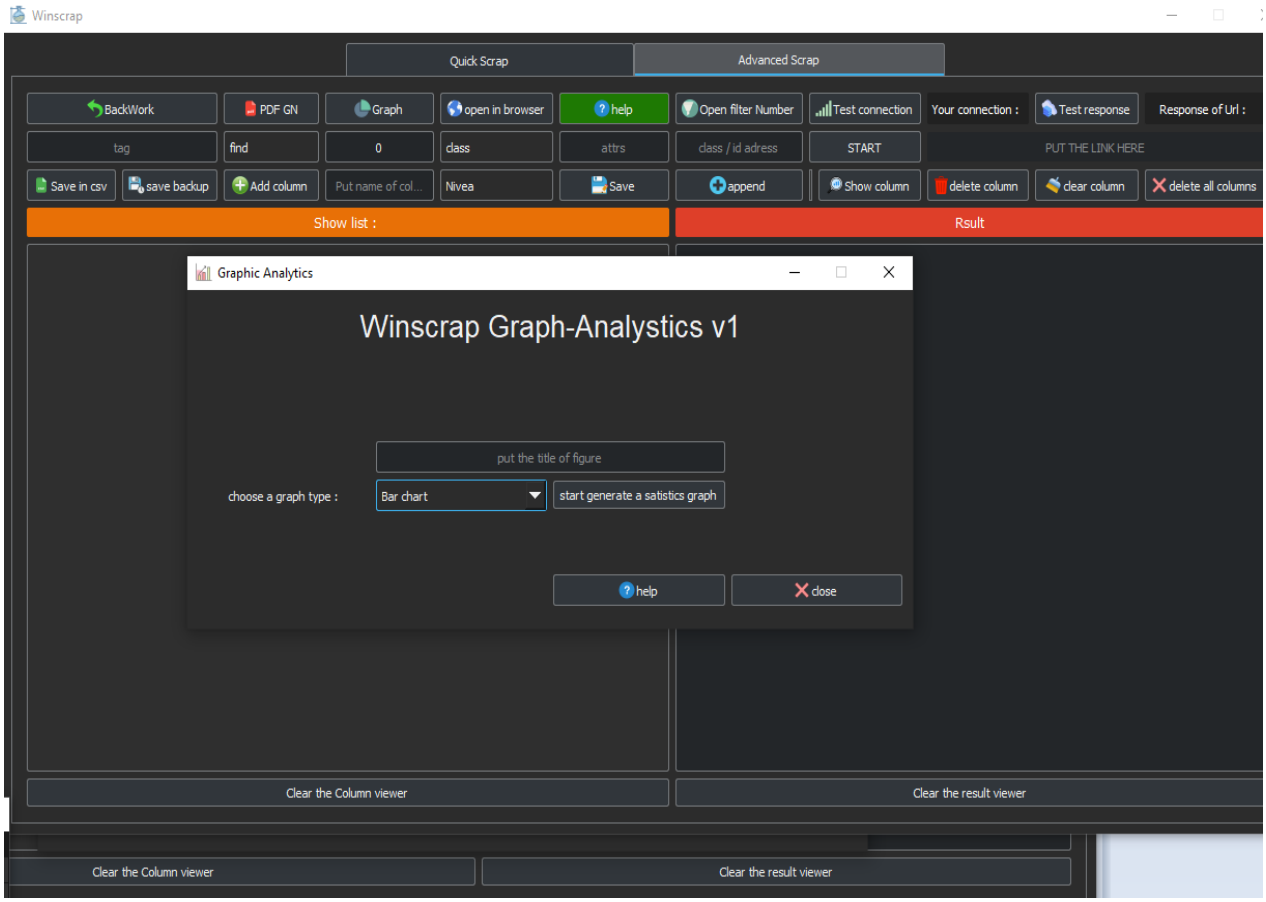

Figure 29: Exemple d'un graphe à paramétrer

Le graphe est affiché comme suit :

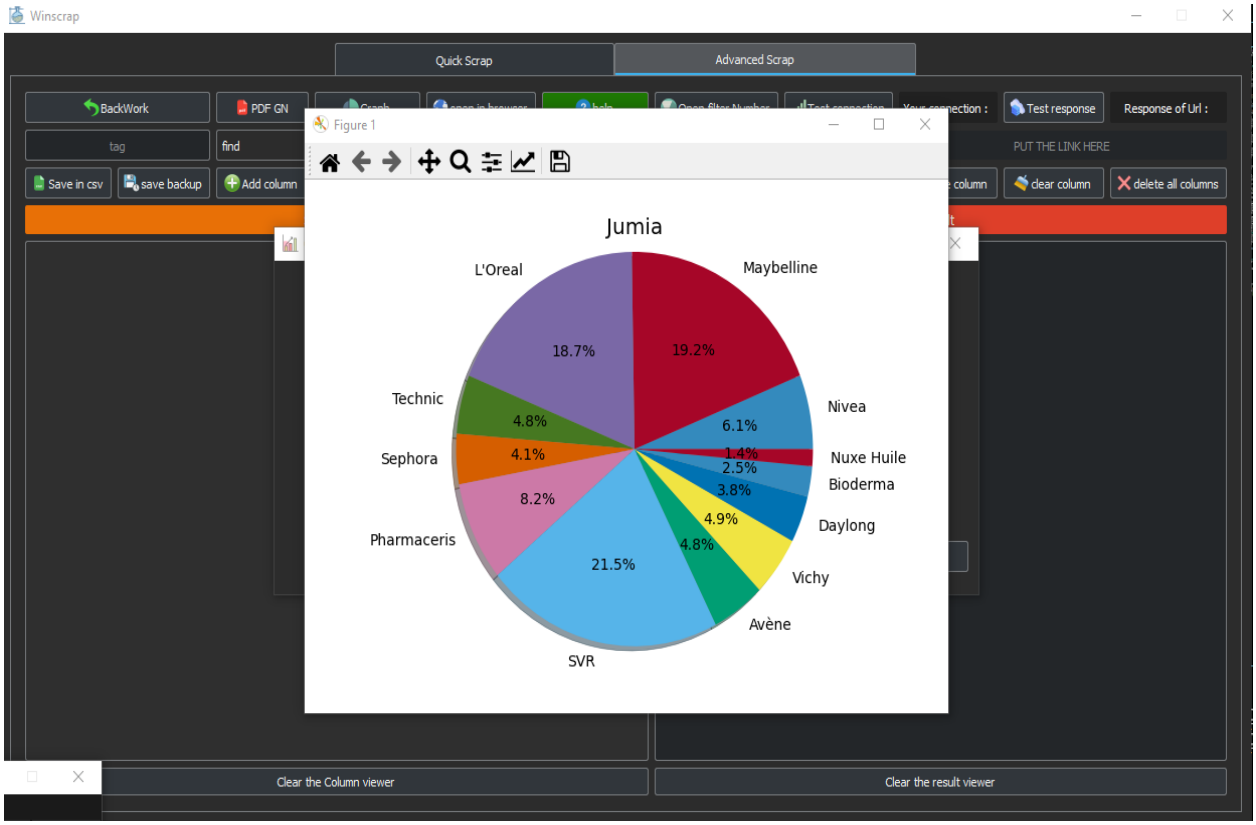

Figure 30: Visualisation des statistiques

D'après ce graphe, nous pouvons conclure que la marque de beauté « SVR » est la marque la plus populaire.

- **Généraliser un PDF :**

Le graphe obtenu peut être manipulé dans un PDF.

La figure suivante présente la fenêtre de généralisation d'un PDF.

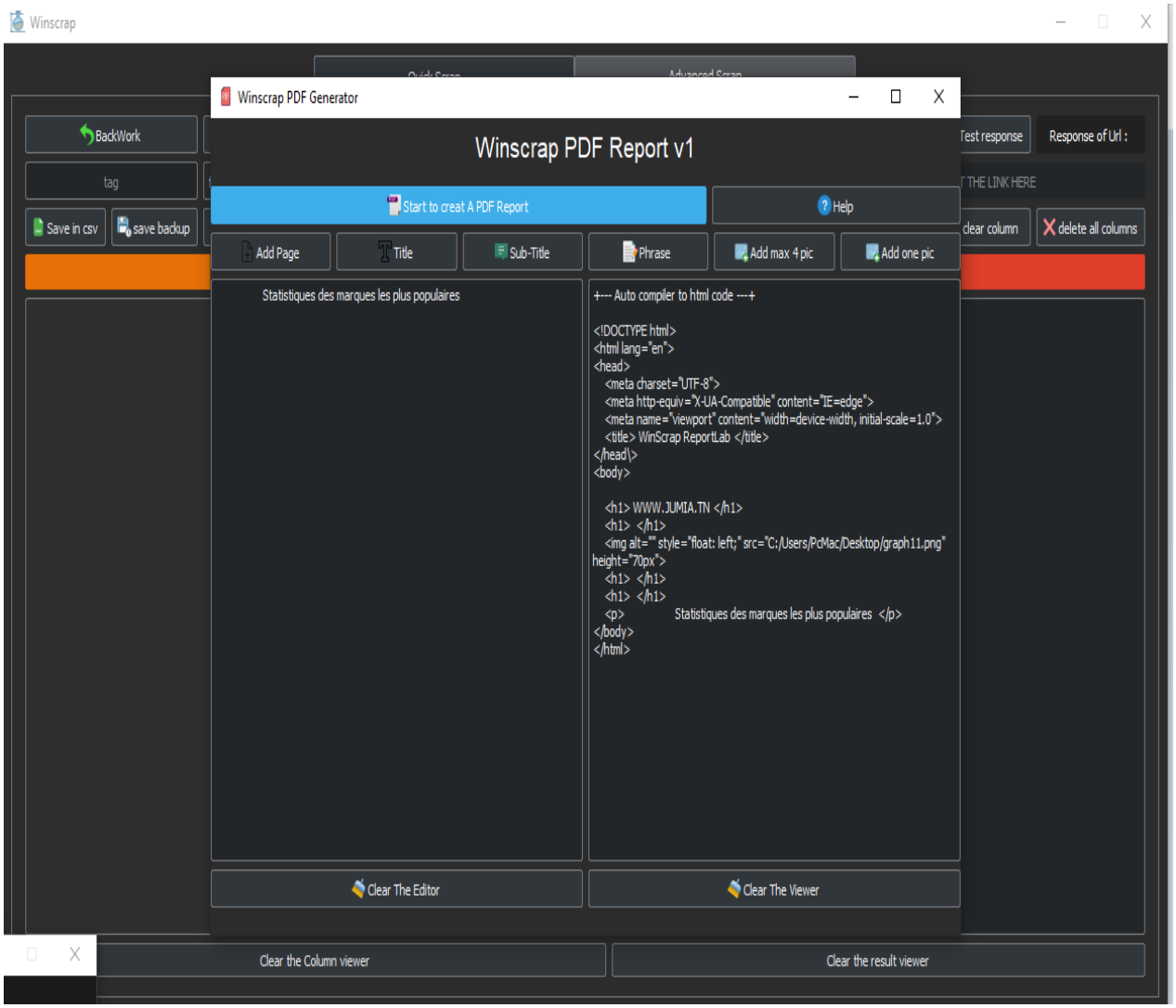

Figure 31: Exemple d'un rapport PDF généré par l'application

Le rapport PDF est sous la forme présenté par la figure suivante :

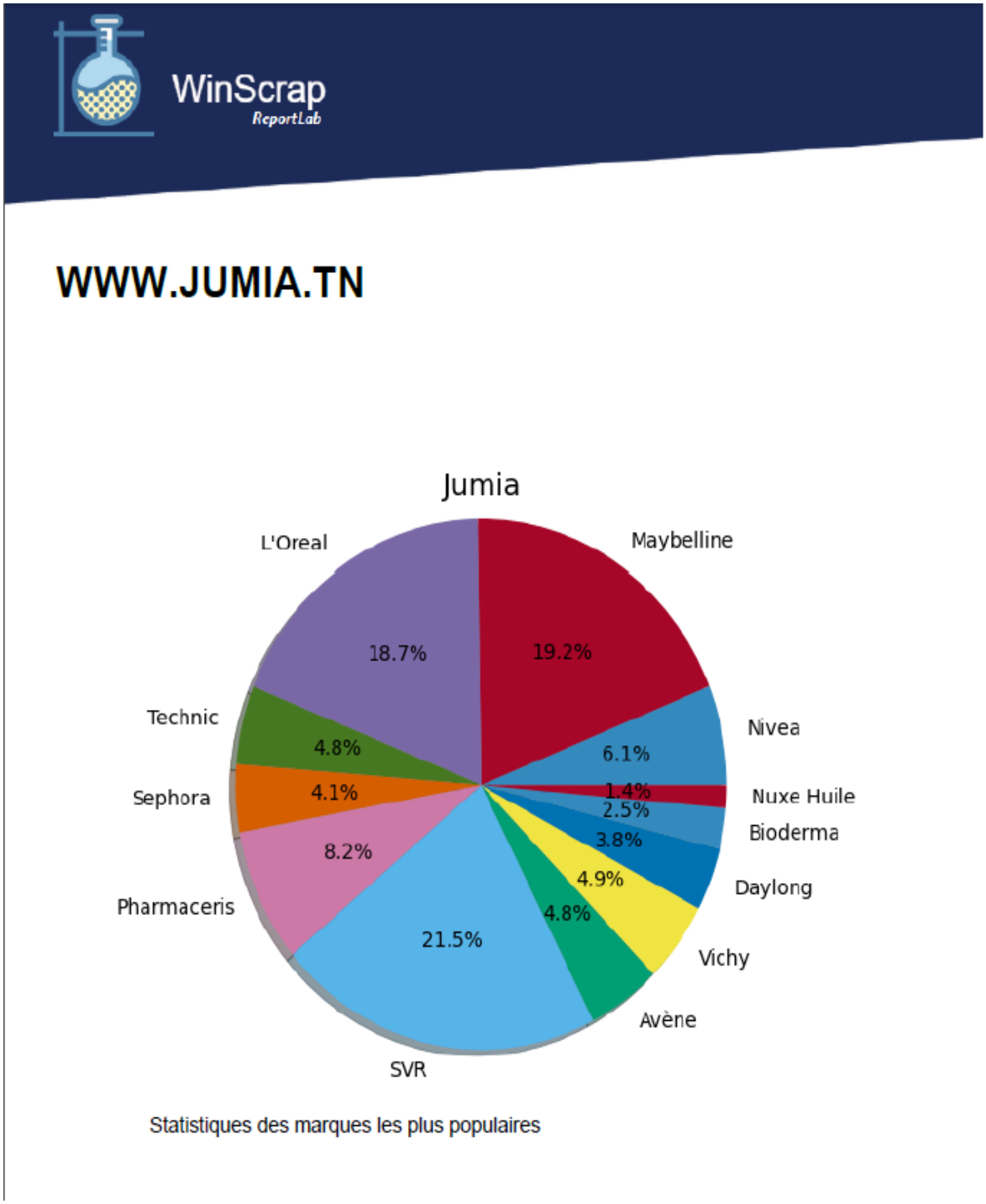

Figure 32: Exemple du rapport PDF

❖ Exemple des données scrapés :

➢ Collecter les noms des PC avec leurs prix :

	A	B	C
3	1	Hp Pc HP15 - N4020 4M - 4G - 1T - Win10 - Argent - Garantie 1an	38490.0
4	2	Asus Pc X543MA - N4020 - 4G- 1Tb - Gris - W10 - Garantie 1an	47490.0
5	3	Lenovo IdeaPad3 - Intel I3 10émé Gén - 8G - 1Tb - Garantie 1an	38490.0
6	4	Lenovo Portable - V15 - Intel N4020 - 4Go - 1To - Garantie 1 an	12390.0
7	5	Hp Pc HP15 - I3 11éme Gén - 4 G - 256 SSD - W10 - Garantie 1an	9990.0
8	6	Lenovo IdeaPad3 - I3 10éme - 8G - 1T + Sac T210 - Garantie 1an	9490.0
9	7	Hp Portable i3-1005G1 - 8Gb - 1Tb + Sac HP Essential Top Load	13690.0
10	8	Hp Portable - i3-1005G1 - 4Gb - 1Tb + Sac HP Essential Top Load	9480.0
11	9	Asus Pc X543MA - N4020 - 4G- 1Tb + Souris ASUS - Gris - W10	15490.0
12	10	Lenovo Pack pc portable - Sacoche - Flash disque - Imprimante - Garantie 1 an	14090.0
13	11	Alessandro DellAcqua Pc Portabel 3583 Intel Celeron 4205U 4 GO -Garantie 1 an	13990.0
14	12	Apple MacBook Air 13 pouces M1 - 256 Go - Clavier Azerty - Gris Sidéral - Garantie 1 an	13290.0
15	13	Apple MacBook Pro 16 pouces - Gris sidéral - Clavier Qwerty -Garantie 1an	9690.0
16	14	Asus PC Portable X509FA I3 10È 1TO 4 GO Blue / Garantie 1an	10990.0
17	15	Apple MacBook Pro 13 pouces Puce Apple M1 - 512 Go - Clavier Azerty - Gris Sidéral - Garantie 1 an	9900.0
18	16	Apple MacBook Air 13 pouces - Ecran retina - 256 Go - Space grey	38490.0
19	17	Hp Pc Portable I5-11éme-8GO-1TB-Windows10 -Garantie 1An	87600.0
20	18	Asus Portable X543MA - Intel N4000 - 1Tb - Win10 - Garantie 1an	13290.0
21	19	Dell Publishing Company Pc portable inspirons 3501 I3 10è Gén 8Go - Garantie 1an	54490.0
22	20	Lenovo PC Portable -v15 4020 - DualCore - 4Go - 1To - Silver - Garantie 1An	45690.0
23	21	Apple MacBook Air 13 pouces M1 - 512 Go - Gris Sidéral - Garantie 1 an	18990.0
24	22	Hp Pc - 15.6" - i5 10è Gén - 4Go - 1To - Gris - Garantie 1an	9990.0
25	23	Hp Portable - I3 11éme Gén - 4 G - 1Tb - Win10 -Garantie 1an	14490.0
26	24	Hp Pc HP15 - N4020 4M - 8G - 1T - Win10 - Argent - Garantie 1an	9490.0
27	25	Lenovo Pc Portable V15 N4020 Dualcore 4G 1T - Garantie 1an	47490.0
28	26	Apple MacBook Pro 13 pouces Puce Apple M1 - 256 Go - Clavier Azerty - Gris Sidéral - Garantie 1 an	18390.0
29	27	Lenovo Pc Portable I3 10è Gén 4Go 512Go SSD -Gris Garantie 1ans .	15490.0
30	28	Asus Pc portable X543M - N4020 - 4 G - 1Tb - Win10 -Garantie 1an	10990.0
31	29	Hp Pc portable 15-dw3014nk - i5-1135G7 -8GB 1TB -Garantie 1an	9790.0
32	30	Apple MacBook Pro 13.3" - Puce Apple M1 - Clavier Qwerty - Gris Sidéral - Garantie 1an	47490.0

Figure 33: Exemple des données scrapés

- Remarque : la colonne des prix de ces données présentés est filtrée avec notre application pour extraire seulement les chiffres sans la devise « TND ».

La figure suivante indique l'étape de filtrage :

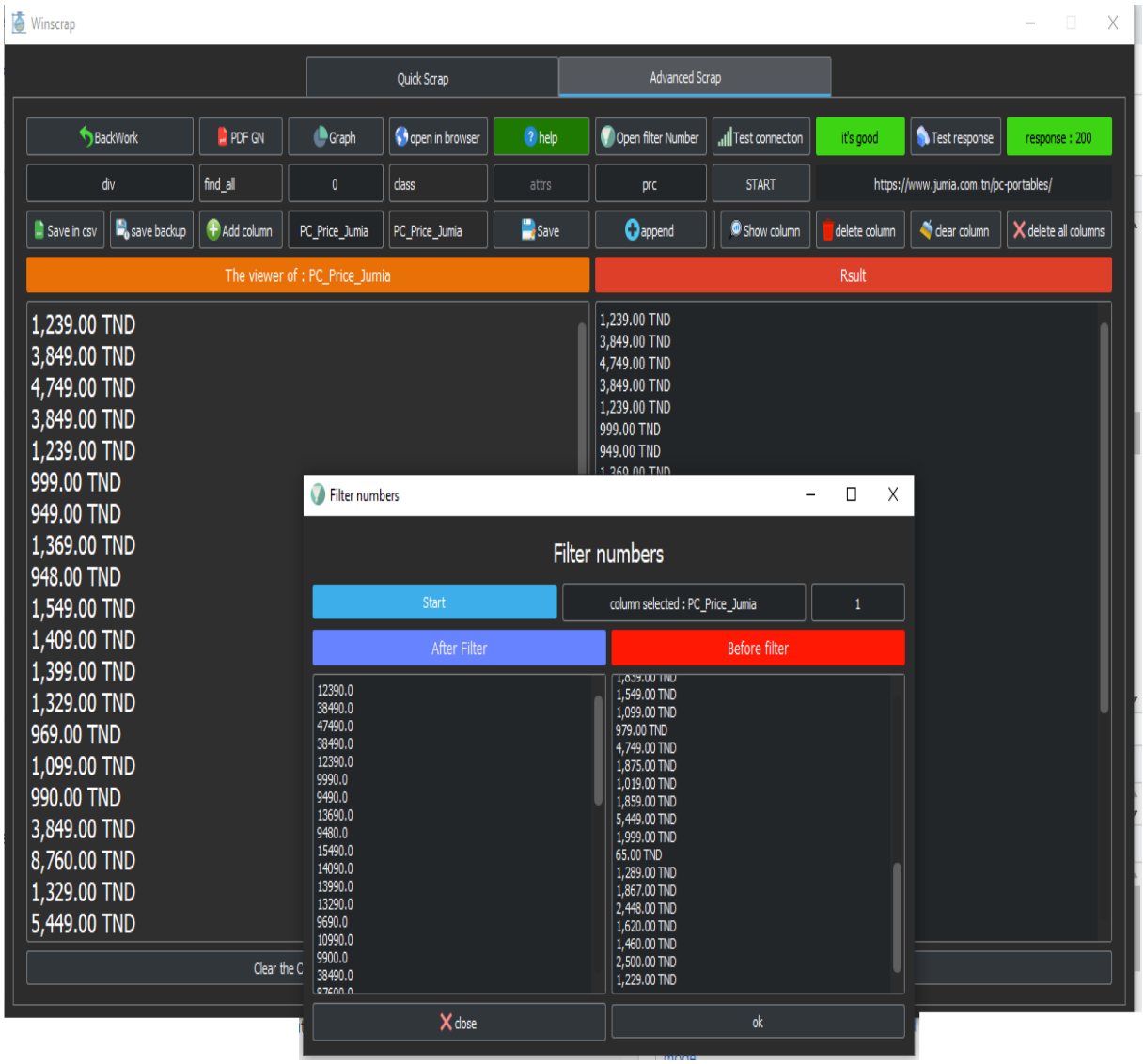

Figure 34: Exemple de filtrage des prix des PC

- **Collecter les tableaux statistiques de Covid19 :**

Le site web correspond à ces statistiques : https://www.coronavirus-statistiques.com/stats-globale/covid-19-par-pays-nombre-de-cas/

	A	B	C	D	E	F
1		DATE	TOTAL CAS CUMULÉS(VARIATION J-1)	TOTAL DÉCÉS CUMULÉS	NBRE DÉCÈS DU JOUR(EN % VARIATION J-1)	GUÉRISONS CUMULÉES
2	0	21/11	7 414 971 (+19749)	118 443	15 (-35%)	
3	1	20/11	7 395 222 (+22678)	118 428	23 (-54%)	
4	2	19/11	7 372 544 (+21220)	118 405	50 (-4%)	
5	3	18/11	7 351 324 (+20366)	118 355	52 (-7%)	
6	4	17/11	7 330 958 (+20294)	118 303	56 (+19%)	
7	5	16/11	7 310 664 (+19778)	118 247	47 (-37%)	
8	6	15/11	7 290 886 (+3241)	118 200	75 (+341%)	
9	7	14/11	7 287 645 (+12496)	118 125	17 (+6%)	
10	8	13/11	7 275 149 (+14646)	118 108	16 (-67%)	
11	9	12/11	7 260 503 (+3860)	118 092	48 (+182%)	
12	10	11/11	7 256 643 (+12603)	118 044	17 (-48%)	
13	11	10/11	7 244 040 (+11883)	118 027	33 (-25%)	
14	12	09/11	7 232 157 (+12476)	117 994	44 (-23%)	
15	13	08/11	7 219 681 (+2197)	117 950	57 (+470%)	
16	14	07/11	7 217 484 (+8547)	117 893	10 (+25%)	
17	15	06/11	7 208 937 (+9605)	117 883	8 (-81%)	
18	16	05/11	7 199 332 (+8998)	117 875	43 (-12%)	
19	17	04/11	7 190 334 (+9502)	117 832	49 (+40%)	
20	18	03/11	7 180 832 (+10050)	117 783	35 (-60%)	
21	19	02/11	7 170 782 (+2039)	117 748	87 (+1350%)	
22	20	01/11	7 168 743 (+1866)	117 661	6 (-50%)	
23	21	31/10	7 166 877 (+6329)	117 655	12 (-45%)	
24	22	30/10	7 160 548 (+7360)	117 643	22 (-21%)	
25	23	29/10	7 153 188 (+6433)	117 621	28 (-18%)	
26	24	28/10	7 146 755 (+6461)	117 593	34 (+750%)	
27	25	27/10	7 140 294 (+6528)	117 559	4 (-94%)	
28	26	26/10	7 133 766 (+6603)	117 555	67 (+219%)	
29	27	25/10	7 127 163 (+1295)	117 488	21 (-19%)	
30	28	24/10	7 125 868 (+5005)	117 467	26 (+13%)	
31	29	23/10	7 120 863 (+6291)	117 441	23 (-21%)	
32	30	22/10	7 114 572 (+6366)	117 418	29 (-22%)	

Figure 35: Exemple de tableau des statistique scrapé

3.3. Commercial Scrap :

Enfin, nous avons développé l'interface commercial Scrap. Cette partie est dédié à collecter les données nécessaires des produits de différents sites e-commerce avec leur prix tel que : Jumia, MyTek, TunisiaNet et MegaPC sous la catégorie « Informatique ». Le collecte de données aide l'utilisateur à consulter les différents prix disponibles concernant un tel produit. Le programme affichera le meilleur prix afin que l'utilisateur prenne une décision d'achat.

❖ Interface Commercial Scrap :

La figure ci-dessous présente l'interface « Commercial Scrap ». L'utilisateur doit entrer le nom d'un produit sous la catégorie informatique. Le programme déclenche une recherche dans les quatre sites de e-commerce « Jumia, MyTek, TunisiaNet et MegaPC » et affiche le résultat selon la disponibilité des articles.

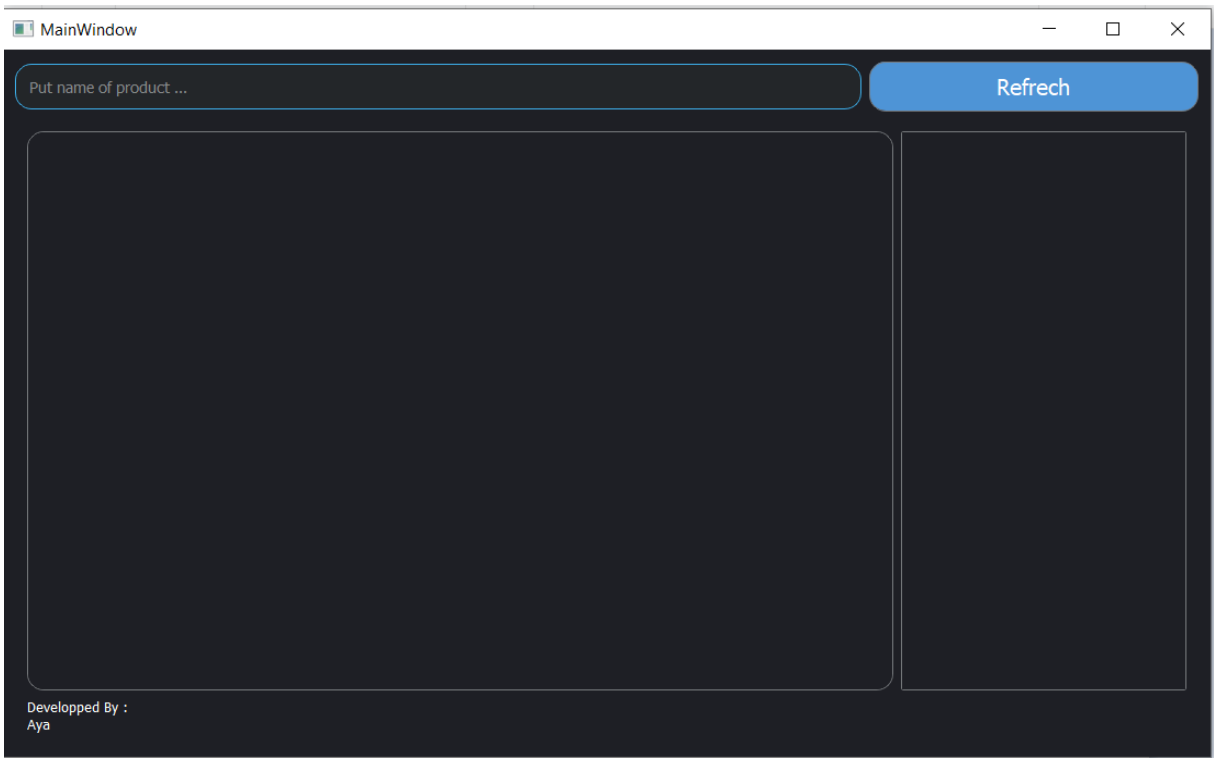

Figure 36: « Interface Commercial Scrap »

❖ Exemple :

Dans cet exemple nous avons cherché un pc portable avec un processeur I3. Le mot clé que nous avons passé en entrer est « I3 » ; Le programme affiche tous données contenant le mot i3 dans les différents sites web. Après avoir terminé l'affichage, une comparaison des prix se présente afin d'aider l'utilisateur à prendre une décision d'achat. Il peut effectuer l'achat en cliquant sur le bouton « Acheter ».

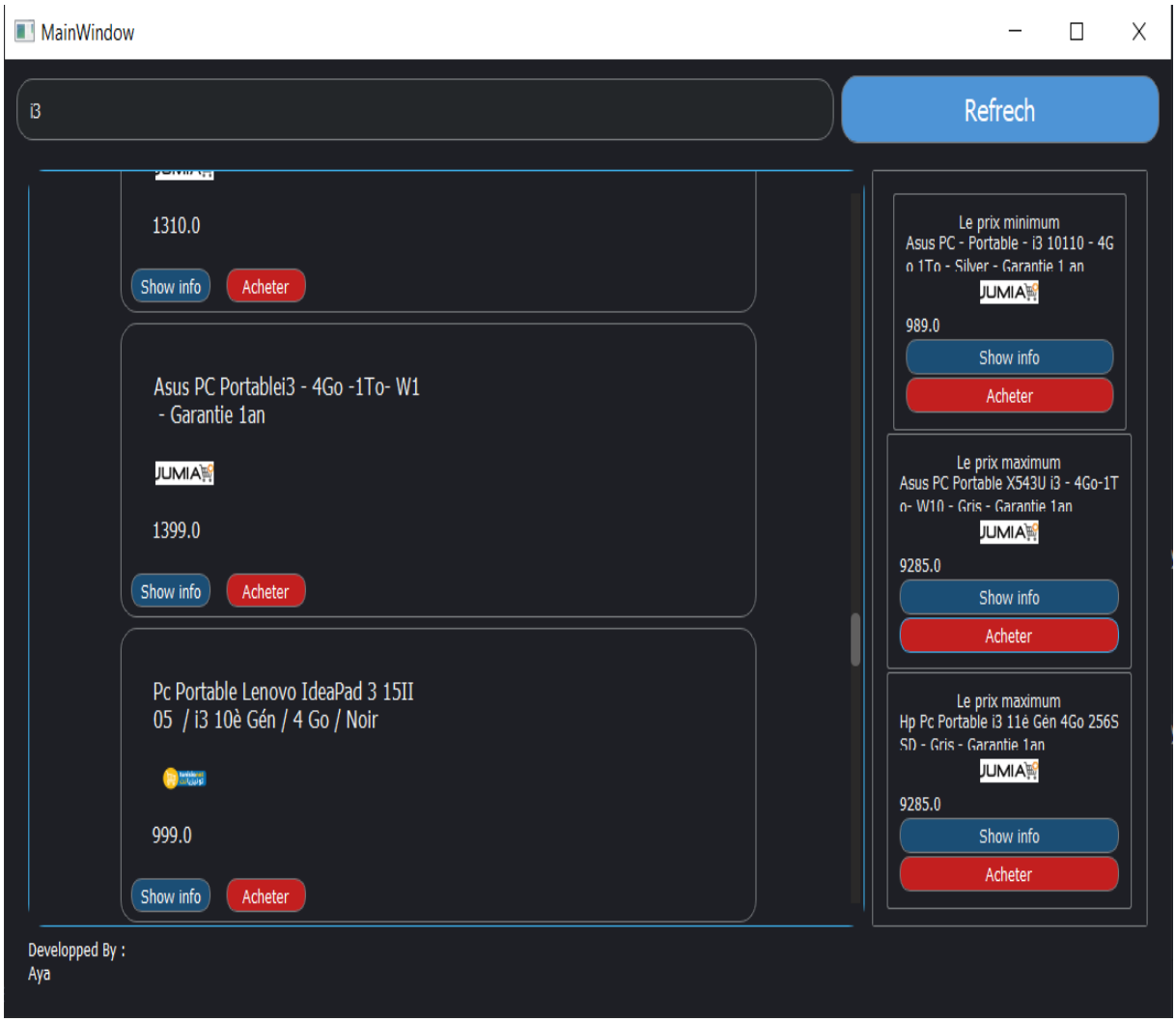

Figure 37: Exemple de recherche d'un produit

Conclusion générale :

Dans le cadre de ce projet de fin d'études, nous sommes appelées à développer en premier lieu, un outil du web Scraping permettant les utilisateurs de collecter des diverses données. En second lieu, nous avons implémenté cet outil à un exemple de comparateur des prix.

Pour mettre en œuvre ce projet, nous étions amenés, premièrement, à établir une

étude conceptuelle du sujet afin de dégager les différents modules et fonctionnalités de cette application ainsi qu'une étude des outils et technologies susceptibles de convenir à sa réalisation.

Deuxièmement, nous avons fait une brève étude de l'art des domaines informatiques, particulièrement les sciences de données, que nous avions abordé dans la solution que nous proposons et qui sont susceptibles de résoudre la problématique.

Finalement, nous avons implémenté les différents modules de notre projet. Le résultat de cette dernière phase répond aux exigences et aux besoins cités dans ce rapport en utilisant du data science (Machine Learning, Deep Learning, Data Visualization).

Pour finir, nous avons fait notre mieux pour bien laisser une bonne impression sur nos

disciplines et nos compétences techniques et présenter un projet à la hauteur de la formation qui nous avons eu au sein de l'ISIMA.

Comme perspectives enrichissantes pour notre projet, nous avons pensé à améliorer le programme « Commercial Scrap ». Par exemple : Travailler sur d'autres sites web de e-commerce sous plusieurs catégories non seulement la catégorie informatique.

Bibliographie

[1]the programming books Practical Web Scraping for Data Science.pdf

[2]Web-Scraping-with-Python-2nd-Edition.pdf

[3]https://www.ionos.fr/digitalguide/sites-internet/developpement-web/web-scraping-avec-python/

[4]https://fr.linkedin.com/pulse/les-7-%C3%A9tapes-dun-projet-data-science-j%C3%A9r%C3%A9my-bouzidi

[5]http://www.mun-balzac.com/medias/files/rapport-oms-comment-le-big-data-et-l-ia-peuvent-ils-transformer-les-soins-en-sante-parlement-balzac-paris-.pdf

[6]https://superdatacamp.com/big-data/definition-et-exemples/

[7]https://rawnote.dinhanhthi.com/files/dataquest/big-data.pdf

[8]https://www.lebigdata.fr/definition-big-data

[9]https://c-marketing.eu/du-web-1-0-au-web-4-0//

[10]https://www.lebigdata.fr/definition-big-data

[11]https://www.cloudflare.com/fr-fr/learning/bots/what-is-a-web-crawler/

[12]https://www.data-transitionnumerique.com/web-scraping-python//

[13]https://moncoachdata.com/blog/python-pour-la-science-des-donnees/

[14]https://actu-ecommerce.fr/quels-sont-les-differents-types-de-e-commerce