RÉPUBLIQUE DU SÉNÉGAL

Un peuple - un but - une foi

¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATON

¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤

L'excellence ma référence

¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤

Mention : Management des systèmes d'information Automatisé

Département : Économie Gestion

UFR : Science Économique Sociale

¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤

THÈME :

Web Scraping et veille concurrentielle : généralité, état de l'art et étude de cas sur les boutiques en ligne.

¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤

Présenté par :

Oumar DIAGNE

Sous la direction de:

Dr Edouard Ngor SARR

Sous la supervision de :

Pr Ousmane SALL

Membres du jury :

Pr Abdou Aziz NIANG (**Président**)

Dr Abel DIATTA (**Examinateur 1**)

Dr Lamine FATY (**Examinateur 2**)

Année universitaire : 2022 – 2023

Soutenu publiquement le 18 Décembre 2023

Dédicaces

Je dédicace ce mémoire à :

- Toute ma famille,
- Tous mes amis.

Remerciements

DIEU MERCI,

A l'heure de conclure ce travail de recherche, je tiens à prendre un moment pour exprimer ma sincère gratitude envers toutes les personnes qui ont contribué à la réalisation de ce mémoire.

Tout d'abord, je souhaite exprimer ma gratitude envers le Pr Ousmane SALL et le DR Edouard Ngor SARR pour leur accompagnement précieux, leur soutien constant et leurs conseils éclairés tout au long de ce travail. Leur expertise et leur dévouement ont été une véritable source d'inspiration, et j'ai acquis une immense richesse de connaissances à leurs côtés

Je souhaite également exprimer ma reconnaissance envers mes enseignants et mentors qui ont partagé leurs connaissances et leur expertise, m'aidant ainsi à développer une compréhension approfondie du sujet.

Mes amis et ma famille méritent un immense remerciement pour leur soutien inébranlable, leurs encouragements et leur compréhension lorsque j'avais besoin de temps pour me consacrer à cette étude.

Enfin, je tiens à remercier l'Université Assane SECK de Ziguinchor (UASZ) qui a rendu ce projet possible et qui continue de promouvoir l'excellence en recherche et en éducation.

Ce mémoire représente une étape importante de mon parcours académique, et il n'aurait pas été possible sans l'aide et le soutien de chacun d'entre vous. Vos contributions ont été précieuses, et je suis profondément reconnaissant de les avoir à mes côtés.

Avec toute ma gratitude

Résumé

Ce mémoire traite la contraignante problématique liée à l'application du Web scraping pour une veille concurrentielle optimale. En effet, l'art d'**extraire des données depuis un site web** a un nom : c'est le **web scraping**, aussi appelé harvesting. Cette technique permet de récupérer des informations d'un site, grâce à un programme ou un logiciel et de les réutiliser ensuite. En automatisant ce processus, nous évitons ainsi de devoir récolter les données manuellement, nous gagnons du temps et nous accédons à un fichier unique et structuré. Le **web scraping** est une technique informatique qui a de nombreux usages. D'autres **applications du web scraping** sont particulièrement utiles dans le cadre de la prospection ou de la veille concurrentielle d'une entreprise. Nous pouvons collectés les données d'un site concurrent pour surveiller ses variations de prix ou bien l'évolution de ses offres. Des données, comme les prix pratiqués par la concurrence, les différentes gammes de produits proposées ou encore celles qui sont le plus mises en avant peuvent par exemple être des indicateurs précieux pour adapter son positionnement. Toutefois, lors du processus de collecte, il est important de bien préparer les données pour les rendre propres. Cela va permettre d'éviter les doublons, les données aberrantes et tout autre risque capable de biaiser les résultats d'analyse et du traitement. Dans ce contexte, le contenu extrait peut subir différentes manipulations à partir desquelles des informations clés peuvent en découler pour aider les décideurs à éclairer leur vision et mieux adapter leurs stratégies. Cette pratique permet de récupérer les données externes pour les confronter avec ses propres données et ainsi dégager des axes d'amélioration ou avoir une meilleure compréhension de son environnement.

Mots clés : Veille concurrentielle, Web scraping, Web crawling, E-commerce, Machine Learning

Abstract

This thesis addresses the challenging issue related to the application of web scraping for optimal competitive intelligence. Indeed, the art of extracting data from a website has a name: it's web scraping, also known as harvesting. This technique allows for the retrieval of information from a site, using a program or software, and then reusing it. By automating this process, we thus avoid having to collect data manually, saving time and accessing a unique and structured file. Web scraping is a computer technique with numerous uses. Other applications of web scraping are particularly useful in the context of prospecting or competitive intelligence for a company. We can collect data from a competitor's site to monitor its price variations or the evolution of its offers. Data such as prices practiced by the competition, the different product ranges offered, or those highlighted the most can be valuable indicators, for example, to adjust positioning. However, during the collection process, it is important to prepare the data well to make it clean. This will help avoid duplicates, aberrant data, and any other risks that could bias the results of analysis and processing. In this context, the extracted content can undergo various manipulations, from which key information can emerge to help decision-makers clarify their vision and better adapt their strategies. This practice allows for the retrieval of external data to confront it with one's own data and thus identify areas for improvement or gain a better understanding of the environment.

Keywords: Competitive monitoring, Web scraping, Web Crawling, E-Commerce, Machine Learning

Sommaire

Liste des figures

Liste des tableaux

Sigles et abréviations

- **HTTP** : HyperText Transfer Protocole ou Protocole de Transfert HyperText
- **3W** : World Wide Web
- **RI** : Recherche d'Information
- **EI** : Extraction d'Information
- **ML** : Machine Learning
- **IA** : Intelligence Artificielle
- **DOM** : Document Object Model
- **HTML** : HyperText Markup Langage
- **CSS** : Cascading Style Sheet ou Feuille de Style en Cascade
- **Regex** : Expression Régulière
- **NLP** : Natural Langage Processing ou Traitement du Langage Naturel
- **URL** : Uniform Ressource Locator
- **JSON** : JavaScript Object Notation
- **CSV** : Comma-Separated Values
- **XML** : eXtensible Markup Langage
- **API** : Application Programming Interface ou interface de programmation applicative
- **SAAS** : Software As A Service ou logiciel en tant que service
- **VC** : Veille Concurrentielle
- **CAPTCHA** : Completely Automated Public Turing test to tell Computers and Humans Apart
- **4P** : Produit, Prix, Place et Promotion
- **BI** : Business Intelligent
- **CFA** : Communauté Financière d'Afrique
- **JMV :** Machine Virtuelle Java

Introduction Générale

La veille concurrentielle est un enjeu crucial pour toutes les entreprises commerciales, notamment pour les boutiques en ligne qui cherchent à se différencier et à s'adapter aux évolutions du marché. Elle se traduit comme étant « *une démarche organisée visant à améliorer la compétitivité de l'entreprise par la collecte, le traitement d'informations et la diffusion de connaissances utiles à la maîtrise de son environnement (menaces et opportunités) et à la prise de décision* [1] ». C'est une approche qui consiste à surveiller régulièrement son environnement afin d'éviter tout effet de surprise venant des concurrents. Cette vigilance permet aux acteurs d'avoir de l'avance sur le marché. Elle favorise un type de management basé sur des statistiques, des tendances, des informations concrètes. La veille concurrentielle vise à informer sur le *Quoi* ; le *Quand,* le *Comment*, le *Pourquoi* et l'*Avec qui* ? Avec quoi ? S'y prendre dans le but d'assurer une réussite totale des actions entreprises.

En effet, Les données jouent un rôle pivot dans cette pratique, offrant un levier stratégique majeur dans une compétition féroce entre acteurs. L'acquisition et l'analyse des renseignements provenant de diverses sources permettent aux entreprises d'adopter une approche éclairée en matière de prise de décision.

Toutefois, pour faire une analyse concurrentielle efficace, le web scraping s'impose comme un moyen très pratique pour accompagner ces dernières à avoir un regard approfondi sur leur environnement et à écouter efficacement le marché. Par web scraping, nous sous-entendons une automatisation des processus de récupération, de structuration et de stockage de données depuis les pages web dans un but précis [2, 3, 4]. Son déroulement automatique garantit l'efficacité et l'efficience. En plus, l'extraction des informations (EI) donne vie au laboratoire d'analyse et de traitement qui à son tour expose aux destinataires des motifs fondés sur lesquels une entreprise peut s'appuyer pour renforcer sa position. Ainsi, dans l'un comme dans l'autre cas, les entreprises de commerce en ligne s'en servent pour optimiser les résultats et pérenniser leur existence. Or ce mécanisme trouve ses repères dans cette dynamique en pleine croissance des sciences de données au sein des organisations. Face à cette ère de la révolution numérique, les cybers boutiques n'ont plus de choix si ce n'est de se donner les ressources nécessaires pour maîtriser les flux d'information [5].

C'est d'ailleurs dans ce sens que nous nous demandons : ***Dans quelle mesure l'utilisation du web scraping sur les boutiques en ligne, apportent-elles une valeur ajoutée pour la veille concurrentielle*** ? Cette question nous amène à poser la problématique de notre thématique comme suit : « comment utiliser les données web pour optimiser l'analyse de la concurrence dans le contexte des cyber boutiques ? »

Cette problématique suggère l'exploration de plusieurs dimensions clés :

- Comment identifier nos sources de données ?
- Comment extraire leurs contenus ?
- Comment les structurer, les stocker ?
- Comment les fusionner, les traiter, les analyser ?
- Comment ensuite s'en servir ?

Voilà entre autres limites sous-jacentes à cette équation et dont une démonstration serait nécessaire pour trouver la formule parfaite.

Cependant, au Sénégal, la vente en ligne couvre l'essentiel des transactions commerciales (offres et demandes), soit une valeur estimée à près de 26 700 milliards de francs CFA ; d'où l'importance pour nous de réfléchir sur la manière dont les acteurs du secteur pourront accroître leurs productivités dans un environnement fortement challengé. C'est ainsi que nous nous sommes orientés vers les systèmes autonomes d'acquisition [6, 7] de faits afin de faciliter la prise de décisions éclairées. En d'autres termes, cette étude permettra aux entreprises :

- De développer une meilleure capacité d'écoute du marché ;
- D'automatiser les processus de collecte et traitement de données ;
- De visionner les résultats ;
- De prendre des décisions stratégiques ;
- D'analyser le marché …

Ainsi, nous nous fixons comme objectif général d'explorer l'utilisation du grattage de contenus web pour développer une meilleure forme de vigilance dans un environnement de compétition entre acteurs d'e-commerce. Autrement dit, nous cherchons à :

- Comprendre les contours et types de web scraping ;
- Analyser l'environnement concurrentiel ;
- Promouvoir l'intérêt de pratiquer une veille concurrentielle dans une boutique en ligne ;
- Voir l'impact du Machine Learning pour une meilleure efficacité de la collecte de données ;
- Évaluer les pratiques de web scraping et d'en déduire un mode d'usage optimal…

Afin d'apporter des réponses aux questionnements soulevés par le sujet, nous avons établi une méthodologie de recherche en plusieurs niveaux consistant à :

- Mener une étude approfondie sur le sujet ;

- Identifier les sources de données pertinentes ;
- Sélectionner les techniques de web scraping adaptées ;
- Décliner les pistes de solutions pratiques ;
- Définir les critères de surveillance des concurrents…

En somme, nous mettons en place un système capable de collecter, stocker, traiter, analyser et visualiser les données web pour répondre aux besoins de surveillance des challengers. Sur ce, nous veillerons à utiliser des outils et des techniques efficaces pour automatiser le processus.

Ce travail est organisé en quatre chapitres.

- D'abord, nous allons mener une étude panoramique visant à découvrir les généralités sur la veille concurrentielle et le commerce électronique.
- Ensuite, nous allons traiter les concepts de base liés au web scraping.
- Puis, nous allons établir le lien entre le web scraping et la veille concurrentielle.
- Enfin, nous mettrons en pratique le web scraping pour le compte de l'observation concurrentielle dans le contexte du commerce électronique.

Chapitre I : Généralité sur la veille concurrentielle et le E-commerce

1.1. La veille concurrentielle

C'est une méthode cyclique [8] consistant à collecter, analyser, interpréter puis diffuser les informations en rapport avec une entreprise rivale dans l'espérance de saisir leur positionnement sur le marché, leurs stratégies commerciales, leurs forces et faiblesses et identifier toutes opportunités et menaces potentielles de son environnement.

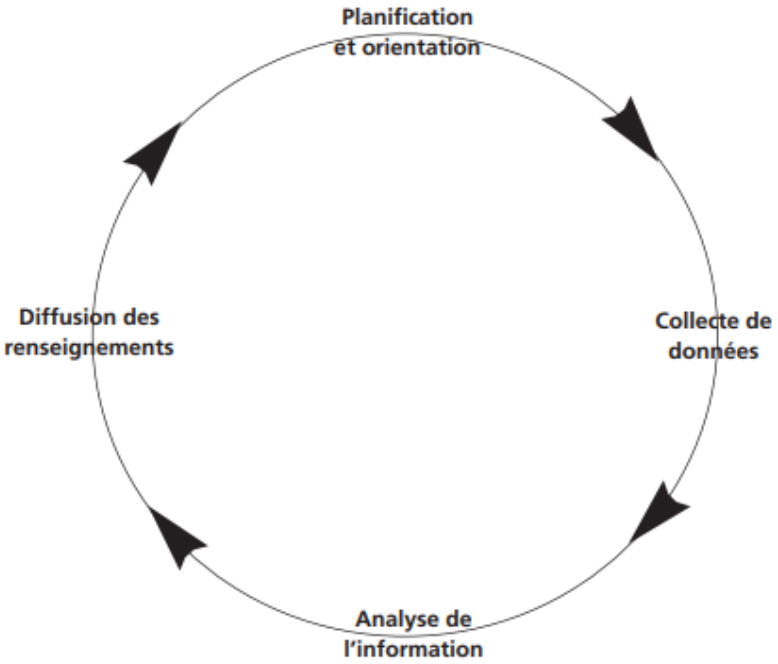


Figure 1 : ***Le cycle de la veille concurrentielle***

Elle permet aux organisations appartenant au même secteur d'activité de se surveiller régulièrement pour trouver de nouvelles perspectives d'évolution en s'appuyant sur les connaissances que cette pratique permettra de produire. Une veille est dite efficace lorsque [9, 10] :

- L'information est acheminée vers une seule destination ;
- L'information est régulièrement transformée en renseignements stratégiques ;
- Ces renseignements sont communiqués à la haute direction ;
- La haute direction les utilise ;
- Les renseignements sont utilisés pour percer des secrets, aussi inviolables soient-ils ;
- Les renseignements sont utilisés pour donner une liberté d'action à tous les employés ;
- La haute direction prend la VC au sérieux.

Autrement, elle consiste à mettre en place un dispositif d'ordre informationnel dynamique et continu capable d'informer avec précision sur l'environnement et de capter l'ensemble des signaux de nature à aider les décideurs à être décisifs.

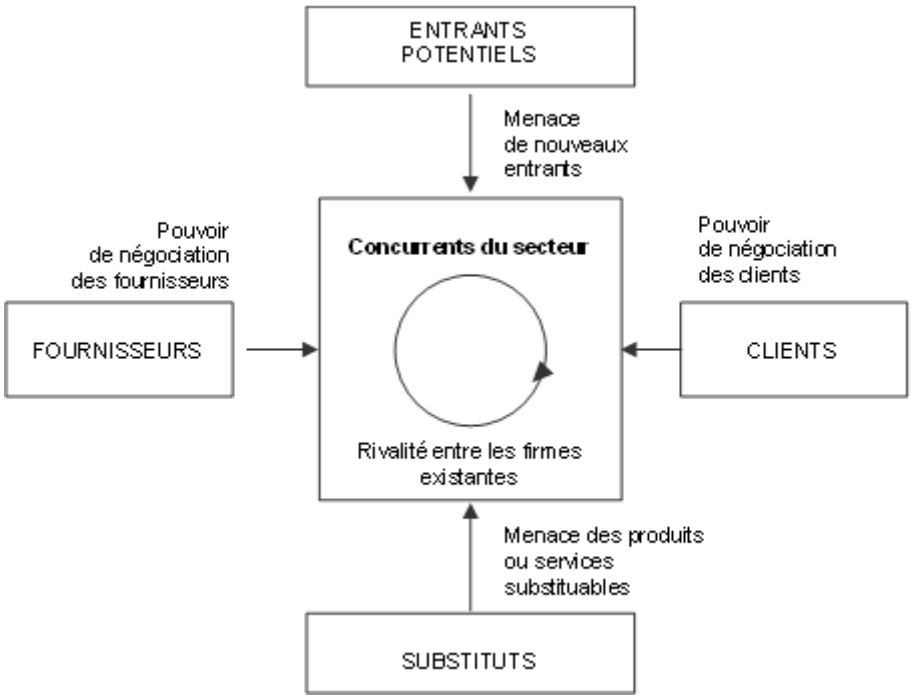


Figure 2 : ***Les types de concurrents chez PORTER1***

Selon PORTER, la veille concurrentielle prend tout son sens lorsque les entreprises qui la pratiquent prenne en compte :

- **Menace des nouveaux entrants** : cette force évalue la facilité avec laquelle de nouvelles entreprises peuvent entrer sur un marché donné. Si les barrières à l'entrée sont élevées (coûts initiaux élevés, réglementations strictes, accès limité aux canaux de distribution, etc.), la menace des nouveaux entrants est faible. En revanche, si les barrières sont faibles, de nouveaux concurrents peuvent plus facilement pénétrer le marché, augmentant ainsi la concurrence.
- **Pouvoir de négociation des fournisseurs :** cette force mesure le pouvoir que les fournisseurs ont sur les entreprises. Si les fournisseurs sont peu nombreux et proposent des produits ou des services essentiels sans substituts faciles, ils ont un fort pouvoir de négociation. Les entreprises peuvent alors être contraintes d'accepter des conditions moins favorables ou des prix plus élevés. Si les fournisseurs sont nombreux et interchangeables, leur pouvoir est moindre.
- **Pouvoir de négociation des acheteurs :** cette force examine le pouvoir que les clients ont sur les entreprises. Si les clients sont peu nombreux mais achètent en grande quantité, ils ont un pouvoir de négociation élevé. Ils peuvent exiger des prix

[1] Michael Eugene PORTER, *Économiste, ingénieur en aéronautique, ingénieur mécanicien, professeur d'université*

plus bas ou des conditions plus favorables. En revanche, si les clients sont nombreux et leurs achats sont dispersés, leur pouvoir est limité.

- **Menace des produits de substitution :** cette force considère la disponibilité de produits ou services alternatifs qui pourraient satisfaire les mêmes besoins que les produits actuels. Si des produits de substitution sont facilement accessibles et offrent une meilleure valeur, ils peuvent constituer une menace pour les entreprises existantes.
- **Rivalité entre les concurrents :** cette force évalue l'intensité de la concurrence entre les entreprises existantes sur le marché. Si la concurrence est intense (nombre élevé de concurrents, peu de différenciation entre les produits, guerre des prix), cela peut réduire les marges bénéficiaires des entreprises. Une concurrence moins intense peut permettre des marges plus élevées.

1.2. E-commerce / boutique en ligne

1.2.1. Définitions

Encore appelé boutique en ligne, cyberboutique, vente en ligne, e-commerce etc., le commerce électronique traduit un déroulement virtuel des boutiques traditionnelles. Il permet aux parties prenantes (offreur et demandeur) de réaliser des transactions de vente, d'achat, de règlement jusqu'au suivi après-vente sans pour autant marquer un contact physique entre consommateur et entreprise. En d'autres termes, c'est un commerce qui se repose essentiellement sur internet, utilisant les sites, plateformes, applications web et mobile pour réaliser des opérations de vente entre acteurs [11]. Le schéma ci-après décrit les caractéristiques des sites de vente en ligne :

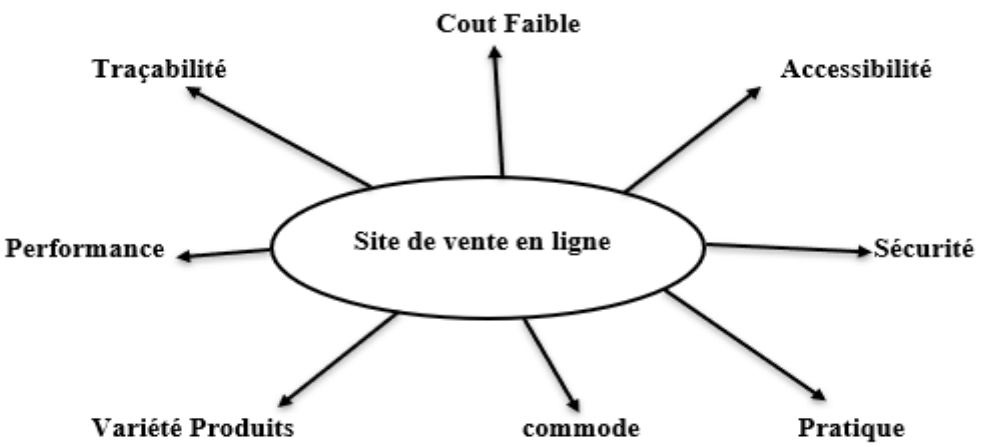


Figure 3 : ***Les Caractéristiques de base d'un E-Commerce***

À noter aussi que ces types d'entreprises sont aussi soumis à des règlements spécifiques notamment en matière de protection des données, la sécurité des paiements en ligne, le respect des conditions utilisateurs.

Figure 4 : ***Schéma du fonctionnement de la vente ligne***

1.2.2. Les acteurs du E-commerce

Plus connus sous l'appellation de marchand en ligne, les acteurs du e-commerce regroupent les entreprises et individus qui opèrent dans la vente de produits et services sur internet. Ils utilisent des outils technologiques pour commercialiser et distribuer leur marchandise à un grand public. Ils facilitent aux consommateurs un accès à une large gamme de produits et services en plus d'une expérience d'achat pratique et personnalisée. L'e-commerce est un domaine qui regroupe des acteurs de différents secteurs d'activités parmi lesquels [12] :

- **Les marketplaces en ligne ou place de marché :** ils sont réputés dans la mise en relation entre les offreurs et demandeurs. C'est des plateformes qui font intervenir 3 principaux acteurs à savoir les vendeurs, les clients et les responsables de la place de marché en question. Son rôle consiste à aider les e-marchands à faire héberger leurs marchandises et les clients à satisfaire leurs besoins en consommation de produits et services. Les marketplaces sont des intermédiaires commerciaux. Exemple : Alibaba, Amazon …
- **Les Entreprises de vente en ligne :** c'est un type de business entièrement mené sur internet ce qui lui vaut le nom de commerce électronique. Ces sociétés se donnent le moyen de vendre en détail les produits et services légalement reconnus à travers leur dispositif. Exemple : Auchan.sn, Jumia…
- **Les services de Livraison :** c'est des types d'entreprises qui s'activent dans la livraison de produits. C'est d'ailleurs ce qui les rend importants. Sachant que les achats se font par l'intermédiaire du digital, ils s'assurent que les produits parviennent aux acheteurs.
- **Les agrégateurs de comparaison de prix :** ils donnent aux acteurs la possibilité de

comparer les prix et caractéristiques de produits proposés par différents vendeurs afin de mieux choisir. Exemple : Google Shopping …

- **Les Fournisseurs de solution de commerce électronique :** c'est des entreprises ou des particuliers évoluant dans la conception et développement de solution, de système informatique de ce genre avec toutes les fonctionnalités que ça nécessite.
- **Les fournisseurs de services de paiement en ligne :** ils sont aussi des acteurs clé qui permettent de dématérialiser tout ce qui est règlement des transactions par le moyen de paiement en ligne. Exemple : Wave, Orange Money, Banque …

1.2.3. Les forme de E-commerce

Il existe différentes formes d'e-commerces qui varient en fonction des transactions entre parties prenantes. Nous en avons noté 6 [13] :

- **B2B (Business-to-Business) :** est une forme de commerce virtuel impliquant des transactions commerciales entre entreprises. C'est lorsqu'une organisation de vente en ligne commande ses produits chez une autre plateforme e-commerce (grossiste) pour mener ces activités qu'on parle la forme B2B [14].
- **B2C (Business-to-Client) :** c'est la forme de transaction la plus récurrente des ventes en ligne ou, les sociétés vendent directement leurs produits et services aux consommateurs finaux.
- **C2C (Customer-to-Customer) :** elles sont aussi des formes de commerce électronique qui impliquent le fait que les consommateurs vendent directement des produits et services à d'autres consommateurs à travers des canaux de distribution digitale.
- **C2B (Consumer -to-Business) :** dans ces cas de figure, les transactions de commerce sur internet vont dans un sens consommateur vers les entreprises. Par exemple, une entreprise X achète des produits chez d'autres consommateurs puis les revendent par le moyen d'outils d'e-commerce.
- **B2A (Business-To-Administration) :** ce modèle fait réfère aux relations commerciales entre les entreprises (entreprises privées ou organisations) et les administrations publiques [15]. C'est le cas de figure ou un appel d'offre émis par un administrateur est gagné par une entreprise donnée.
- **C2A (Consumer-to-Administration) :** représente les interactions directes entre les consommateurs (utilisateurs ou citoyens) et les administrations publiques. Dans ce cas, les consommateurs interagissent avec les administrations pour accéder à des services publics, effectuer des transactions gouvernementales, soumettre des demandes administratives, ou obtenir des informations officielles.

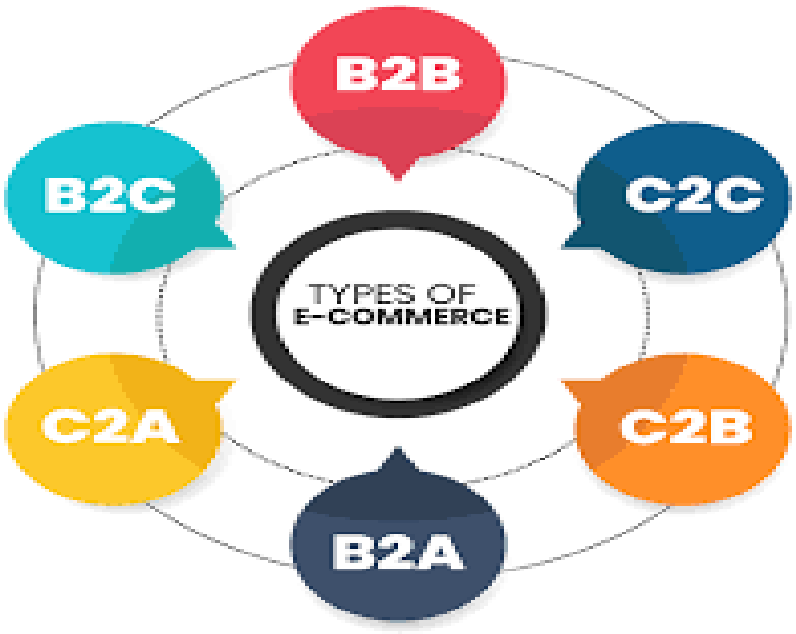


Figure 5 : ***Types de commerce électronique***

1.3. La veille concurrentielle et le commerce en ligne

L'observation régulière entre acteurs d'e-commerce porte : « *essentiellement sur les clients, et sur les fournisseurs de l'entreprise, mais aussi sur ses sous-traitants et ses partenaires dans l'élaboration des produits et services. Elle s'intéresse aux produits ou aux composantes du mix produit (Distribution - prix - produits - Publicité). Tout ce qui compose le marché d'un produit* [16] ». En fait, une surveillance efficace des challengers permet aux e-marchands de rester compétitif dans un environnement en constante évolution. Le dire de cette manière sous-entend dans ce cas que surveiller des vis-à-vis dans ce domaine est un moyen de subsistance. En réalité, cette pratique s'offre comme étant un facteur de succès pour toutes entreprises, à plus forte raison celles qui évoluent dans le digital. Leurs performances dépendent d'un certain niveau de maîtrise des tendances de marché qui se traduit par leur capacité à se servir de la veille pour attirer et fidéliser les clients d'autres entreprises [17, 18]. C'est ce qui explique, le fait que plus une veille est pointue, plus la rivalité est dominée avec un part de marché nettement supérieur à ceux qui ne la pratique pas ou du moins qui l'exerce de manière inefficace. Grâce à une bonne veille, un cyber boutique obtient une compréhension approfondie de l'environnement concurrentiel pour ensuite s'en servir pour s'orienter vers des services de qualité conforme aux attentes et exigences des utilisateurs finaux. En outre, la veille e-commerce permet aux vendeurs en ligne de :

- Développer des stratégies de vente plus rentable
- Identifier ses potentiels clients

- Surveiller les prix et promotions pratiqués par ses concurrents
- Évaluer les avis des consommateurs sur leurs produits
- Suivre les activités des concurrents sur les médias sociaux ainsi que les appréciations des utilisateurs
- Analyser les tendances de marché ainsi que les comportements d'achat des clients…

Elle peut cibler chez les concurrents :

- Un nouveau produit développé
- Une action sur la communication marketing
- Un événement de lancement
- L'évolution du chiffre d'affaires
- Une promotion
- La politique des prix …

Bref, le fait de savoir ce que font nos semblables, comment ils le font et comment ils s'en sortent permet de bien se positionner dans le marché, de proposer les produits attendus par les consommateurs, d'être à jour sur les tendances en vogue. Pour un e-commerçant, l'observation continue des concurrents du secteur permet d'identifier de nouveaux prospects [19]. La veille est dans ce sens un processus informationnel par lequel l'organisation se met à l'écoute anticipative des signaux précoces de son environnement socio-économique

Chapitre 2 : Généralité sur le web Scraping

1.1. Acquisition de données / Recherche d'information (RI)

Voici des notions redoutables qui forment le tissu principal de l'informatique en général et celles des sciences de données en particulier. La recherche d'information et l'acquisition de données sont étroitement liées et complémentaires. Nous irons même jusqu'à dire qu'elles sont trop souvent confondues. Cette subtilité réside dans un objectif commun qui tourne autour de l'information. Néanmoins, la plupart des utilisateurs ne font pas la part des choses les concernant. La RI est le processus de recherche, de filtrage, et d'indexation des données jugées intéressantes. En principe, aucun type de restriction n'est mise sur le type d'objet manipulés dans une RI. Par essence, elle vise à donner accès à l'information recherchée. Pour se faire, la RI favorise l'utilisation des mots-clés, des références etc. Nous en déduisons alors :

RI = Rechercher + Filtrer + Indexer / Données

Elle peut mêler la recherche sur le web, dans les bases de données, dans des documents textuels, des articles scientifiques, livres, sur des appareils connectés … Et peut également avoir plusieurs types [20] :

- Factuelle (structurée, numérique, alphanumérique …) ;
- Documentaire (référence document physique) ;
- Bibliographique (référence document numérique) ;
- Contextuelle (mot clé, groupe de mot).

Sa mission consiste à faciliter l'acquisition des données. Cette dernière permet à son tour de collecter, d'organiser et de stocker les données brutes Comme ceci :

Acquisition de données = RI + collecter + Structurer + Stoker

Cette phase de collecte reçoit sa matière de la RI. Son mode opératoire consiste à acquérir des renseignements utiles et capables de répondre à une requête, un besoin spécifique.

Toutefois, les données acquises peuvent être caractérisées soit :

⇨ **Par leur nature**

- **Quantitative :** elles couvrent l'ensemble des données quantifiable, mesurable, exprimable numériquement. Ces données facilitent les analyses statistiques, les calculs mathématiques et fournissent des mesures précises ce qui les rendent objectives. Exemple : le nombre d'employés d'une entreprise.
- **Qualitative :** alors que les données qualitatives sont subjectives. Elles sont plus

descriptives, contextuelles et ne peuvent surtout pas être mesurées numériquement. Exemple : les avis clients sur un produit.

⇨ **Par leur accessibilité**

- **Données noires :** est appelé données noires, des renseignements confidentiels, privés ou qui ne nous sont pas accessibles légalement. L'accès à ces types d'informations n'est autorisé qu'à une tierce personne ce qui rend leur utilisation illégale et moralement répréhensible. Exemple : le piratage, l'espionnage, le vol de documents entre autres qui sont considérés comme non éthique.
- **Données grises** : il s'agit des informations semi confidentielles. C'est-à-dire que leurs accessibilités nécessitent un certain niveau d'accréditation. Leurs sources sont très souvent non officielles. Par exemple : les fuites d'information, les discussions informelles …
- **Données blanches** : quant aux données blanches, elles indexent toutes informations officielles, publiques et facilement accessibles. C'est donc toutes les données transparentes à savoir les sites web officiel, les données publiques, les publications médiatiques…

⇨ **Par leur volume :** qui fait allusion à la quantité d'un ensemble de données mesurable en (bit, octet, mégaoctet, gigaoctet, téraoctet, pétaoctet etc.)

NB : Quelques soit leurs caractéristiques, les informations trouvées doivent être pertinentes. Ceci découle de leurs capacités à mener une recherche vers des résultats plus spécifiques, plus pointus, plus satisfaisants concernant un problème X que nous visons à résoudre ou bien un phénomène Y que nous cherchons à connaître. Une data est dite pertinente quand elle est : fiable, de qualité, précise, à jour, compréhensible et claire. C'est des types de données utiles et apportent une valeur ajoutée. La pertinence d'une information est mesurée sur deux niveaux [21] :

- Niveau utilisateur (Capacité de recherche, précision des requêtes, Bon référence) ;
- Niveau système (qualité de l'information retournée).

1.2. Acquisition de données dans le web

Les données sont partout mais c'est lorsqu'elles nous parviennent du 3W que l'on parle d'acquisition de données web. Le constat est que de plus en plus les données produites s'augmentent de manière exponentielle et peuvent venir des médias sociaux, des sites web, des forums, des blogs, des bibliothèques numériques etc. Dans un contexte de web scraping, ces informations doivent être obtenues dans des formats compréhensibles, bien organisés pour faciliter leur réutilisabilité [22]. Le processus d'extraction d'information peut être manuel,

automatisé ou semi-automatique et peut aussi viser des éléments structurés ou non structurés pour diverses raisons.

Figure 6 : ***Schéma d'acquisition de données web***

1.2.1. Les systèmes d'acquisition de données web

C'est un ensemble de ressources humaines, matérielles, technologiques interagissant ensemble de façon cohérente pour :

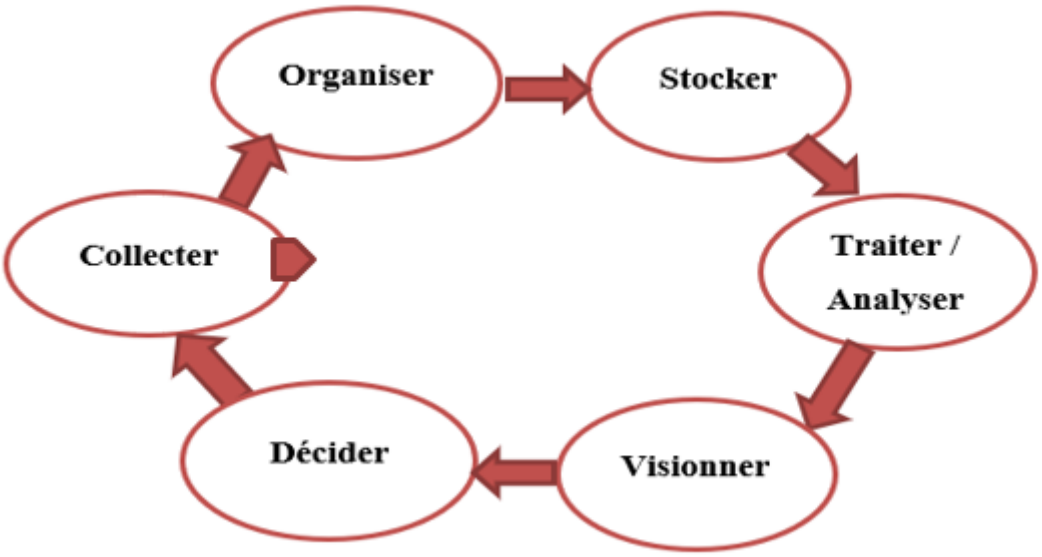


Figure 7 : ***Objectifs d'un système d'acquisition de données***

La structure du Web et son contenu évoluent continuellement. Des informations apparaissent et disparaissent et de nouveaux concepts et outils sont créés et changent très rapidement. Les systèmes de collecte d'information doivent pouvoir suivre cette évolution. Ils sont appelés à être extensibles, adaptatifs et flexibles pour exploiter les informations issues du Web [23].

Ces systèmes obtiennent leurs données dans différentes sources et peuvent être expérimenté dans divers domaines tels que la recherche scientifique, la veille concurrentielle, renseignement

et sécurité, finance et investissement, santé, tourisme, marketing et dans bien d'autres secteurs. Leur particularité est qu'ils sont pragmatiques car assurent :

- Automatisme ;
- Sécurité ;
- Fiabilité ;
- Efficacités et efficience ;
- Rapidité ;
- Réactivité …

Aussi, ils font intervenir plusieurs composantes pour se constituer :

- Agent scraping ;
- Bases de données ;
- Mécanisme de traitement et de transformation ;
- Outils d'analyse et de visualisation …

De pareils dispositifs sont capables également de constituer des corpus hétérogènes d'information à grande échelle en un temps record.

1.2.2. Crawling Web

Est appelé crawler ou bot d'indexation un programme informatique permettant de survoler le web, d'analyser sa structure puis de stocker son contenu de manière organisée dans un index. Les crawleurs sont des programmes parcourant le W3 au travers des hyperliens entre documents. Il est considéré comme un agent logiciel chargé d'indexer des contenus à la recherche d'éventuelles mises à jour. Le crawling est très utilisé par les moteurs de recherche qui s'en servent pour mémoriser l'ensemble des informations véhiculées via un protocole de communication hypertexte en vue de faciliter l'ordonnancement et l'affichage des résultats de recherche à l'exemple de Google Bot[2]. Les actions de crawl peuvent être de deux sortes [24] :

- Les actions de navigation pointent vers des URL à ajouter dans la file de crawl ;
- Les actions d'extraction pointent vers des objets sémantiques individuels à extraire de la page web.

Toutefois, le champ d'application de cette science ne s'arrête pas là. En parlant de collecte, il s'impose aussi comme un acteur incontournable. Grâce aux crawleurs, un programme de

[2] Bot d'indexation de Google

scraping maîtrise mieux ses cibles. En effet, le spider lui permet de parcourir les données telle une araignée sur sa toile, de les fixer avant de les soumettre à l'extracteur. C'est comme dire, dans ce cas, si le grattoir veut parvenir aux informations, il fait appel à l'index pour les atteindre et les saisir.

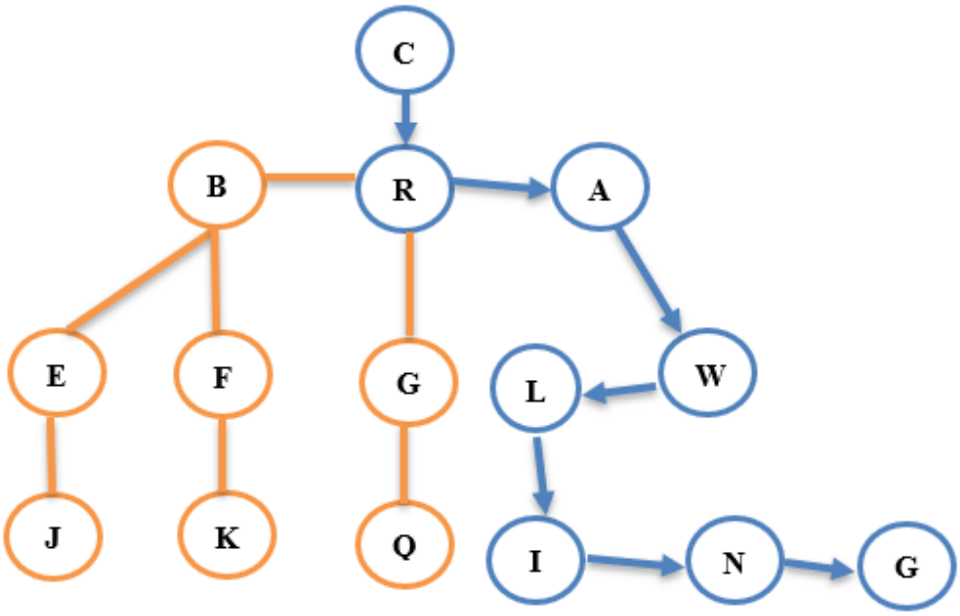


Figure 8 : ***Exemple d'indexation de contenu web***

A supposer que cette schématisation constitue le DOM du site web cible et que les flèches indiquent à notre script le chemin où se trouve les données indexées. Cela signifie que si le programme est lancé, nous aurons accès au tableau de données [C - R - A -W - L- I - N - G]. Le bot d'indexation après avoir dans le cas présent mémorisé tout le contenu avec les itinéraires, facilite ainsi la recherche et l'exploitation de ces données. Ici elles sont obtenues grâce à :

Index [Position]

Exemple : C = index [0], R = index [1], L = index [4], G = index [7] …

1.2.3. Web Scraping

1.2.3.1. Définitions

Selon Wikipédia[3], le web scraping (parfois appelé harvesting ou en français moissonnage) est une technique d'extraction du contenu de sites Web, via un script ou un programme, dans le

[3] https://fr.wikipedia.org/wiki/Web_scraping

but de le transformer pour permettre son utilisation dans un autre contexte comme l'enrichissement de base de données, le référencement ou l'exploration de données. » Le terme nous vient du verbe « To scrap » en anglais qui signifie littéralement « gratter ». Il répond favorablement aux attributions d'acquisition de données de la partie 1.2 de ce chapitre dans le sens de la collecte, du formatage et du stockage des informations issues du web. La particularité avec le web scraping est qu'il acquière exclusivement ces données à l'aide de programmes informatiques. Ainsi, nous serons tentés de dire que le web scraping désigne le processus d'extraction de données web. Ces informations sont collectées et ensuite exportées dans un format plus utile pour l'utilisateur [25]. C'est une automatisation des processus qui stimule la conduite d'un humain cherchant à constituer un corpus de données à travers des copier/coller manuels sur le web avec :

- Plus de productivité ;
- Moins d'erreurs humaines.

C'est une manière optimale de saisir uniquement les données dont nous avons besoin.

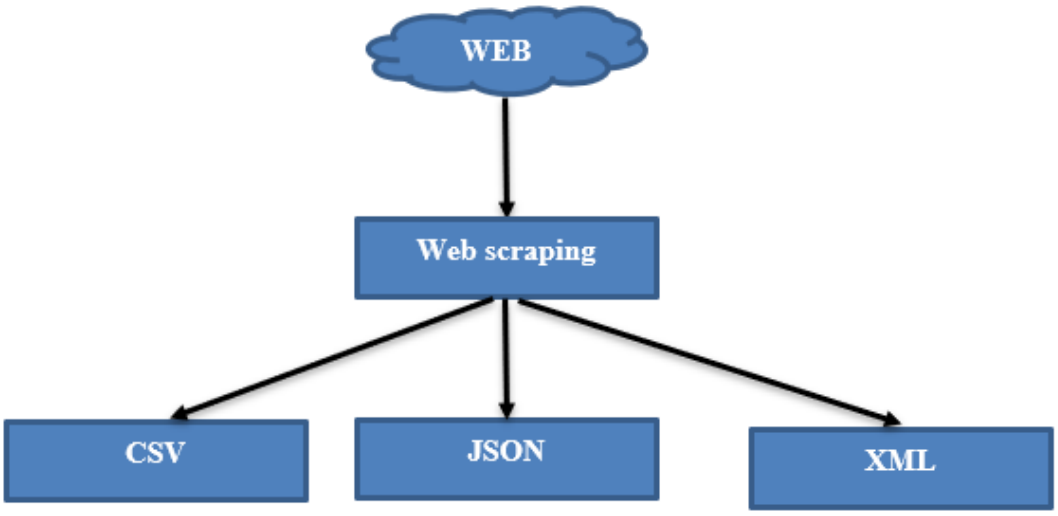


Figure 9 : ***Schéma web scraping***

Ci-dessous, nous avons un exemple d'algorithme de scraping.

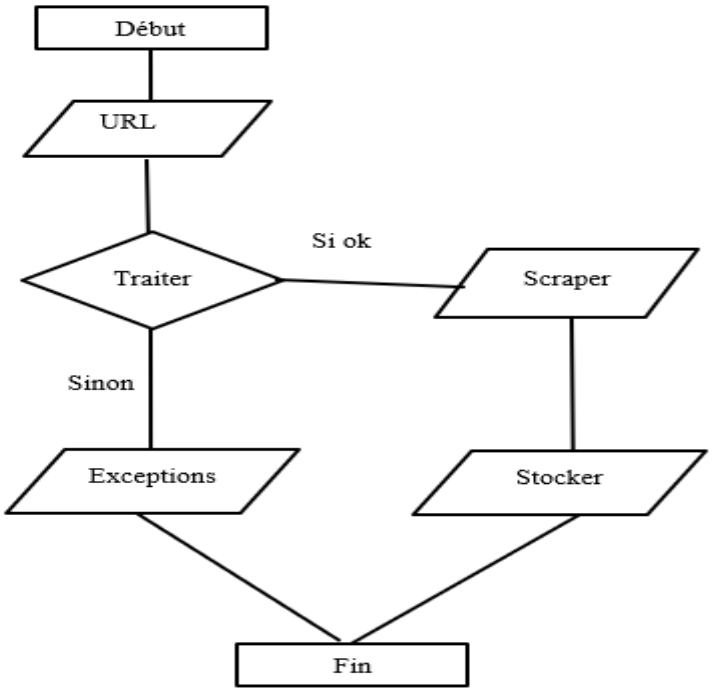


Figure 10 : ***Exemple d'algorithme de web scraping***

L'algorithme prend en entrée une url qu'il envoie comme appât à son site cible à travers une requête HTTP Get[4] en attendant sa réaction. Ensuite, il analyse le document HTML. Lors de ce travail, si l'algorithme rencontre des erreurs (requête, DOM …), il arrête de s'exécuter. Sinon le script va rechercher un modèle particulier de données, les extraire, les convertir dans le format de son programme (CVS, JSON, XML, XLSML, XLS …) puis les stocker avant de s'interrompre.

1.2.3.2. Mode de fonctionnement générique

Pour se faire, le programme parcourt et indexe d'abord les informations du site cible grâce au crawling, les récupère et les stocke dans une base de données dans un format compréhensible et structuré.

Figure 11 : ***Fonctionnement générique du scraping***

[4] Méthode informatique permettant de formuler une demande via une requête

1.2.4. Web Scraping Vs Web Crawling

Pour un novice, ces deux approches ne forment qu'une seule. Il en fait partie des raisons de cette confusion entre scraping et crawling, le fait est qu'ils sont généralement effectués ensemble. Cela relève aussi de leur caractère complémentaire dans cette mission permettant aux utilisateurs (humains ou logiciels) d'exploiter pleinement les potentialités d'internet. En Voici quelques missions que poursuivent spécifiquement chacun d'entre eux :

Tableau : Comparaison entre scraping et crawling

Web scraping	**Web crawling**
Extrait des informations web à partir des pages	Mémorise et indexe les données web
Se sert du crawling pour atteindre ces cibles	Sert de guide au scraping
N'indexe pas	N'extrait pas non plus
Cible des données spécifique au but du scraping	Parcourt les toutes données d'un site pour trouver davantage de mise à jour

En résumé, le web scraping est utilisé pour extraire des données spécifiques à partir de pages web ciblées, tandis que le web crawling est utilisé pour explorer, indexer des informations globales de plusieurs sites web.

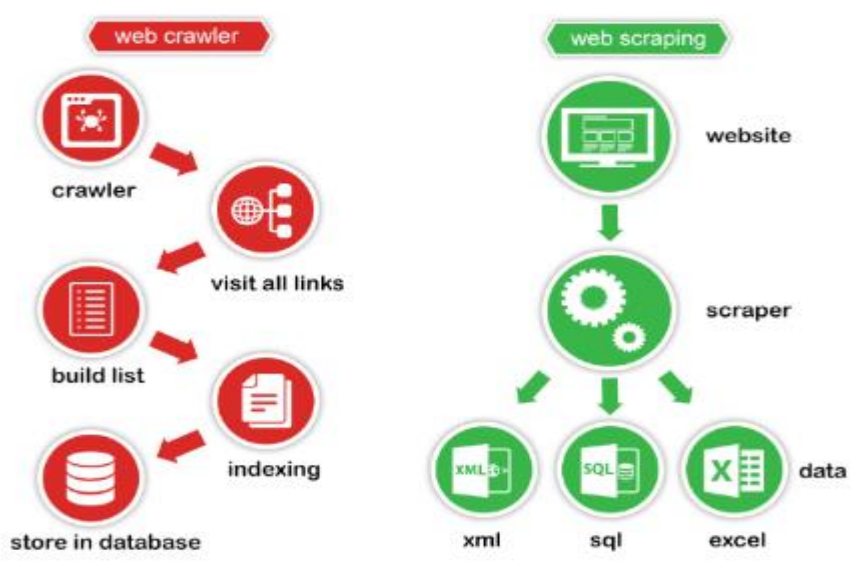


Figure 12 : ***Web scraping contre web crawling***

1.2.5. Web Scraping intelligent

1.2.5.1. Définitions

Par définition**,** « *le Machine Learning ou apprentissage automatique est la capacité d'un ordinateur à apprendre de manière autonome à prendre des décisions. En d'autres termes, il imite la capacité des humains à acquérir de nouvelles connaissances sans avoir été programmés pour cela. Tout comme les humains qui s'améliorent dans une tâche à chaque nouvelle expérience, un modèle de machine Learning devient de plus en plus performant à mesure qu'il s'exécute.* » [26]. Il nous fait passer d'une étape d'analyse descriptive à une autre étape d'analyse prédictive, prescriptive [27, 28, 29]. Il repose sur l'utilisation d'algorithmes et de modèles statistiques qui identifient des motifs dans les données, permettant aux systèmes de prendre des décisions ou de faire des prédictions. Les trois types de Machine Learning les plus courants sont l'apprentissage supervisé, non supervisé et par renforcement [30].

- Dans l'apprentissage supervisé, le modèle est formé sur un ensemble de données étiqueté, apprenant à faire des prédictions précises.
- L'apprentissage non supervisé implique l'utilisation de données non étiquetées pour découvrir des structures et des relations intrinsèques.
- L'apprentissage par renforcement consiste à entraîner un modèle à prendre des décisions en fonction de récompenses ou de punitions.

Les domaines d'application de l'apprentissage automatique sont vastes, englobant la reconnaissance vocale, la vision par ordinateur, la recommandation de produits, la prévision météorologique, le scraping et bien d'autres. Pour le mettre en œuvre, il est essentiel de comprendre les concepts tels que les ensembles d'entraînement, les algorithmes d'apprentissage, l'évaluation du modèle et l'optimisation des hyper paramètres. Son intégration avec le web scraping permet d'assurer une faible marge d'erreur dans cette mission de collecte [31].

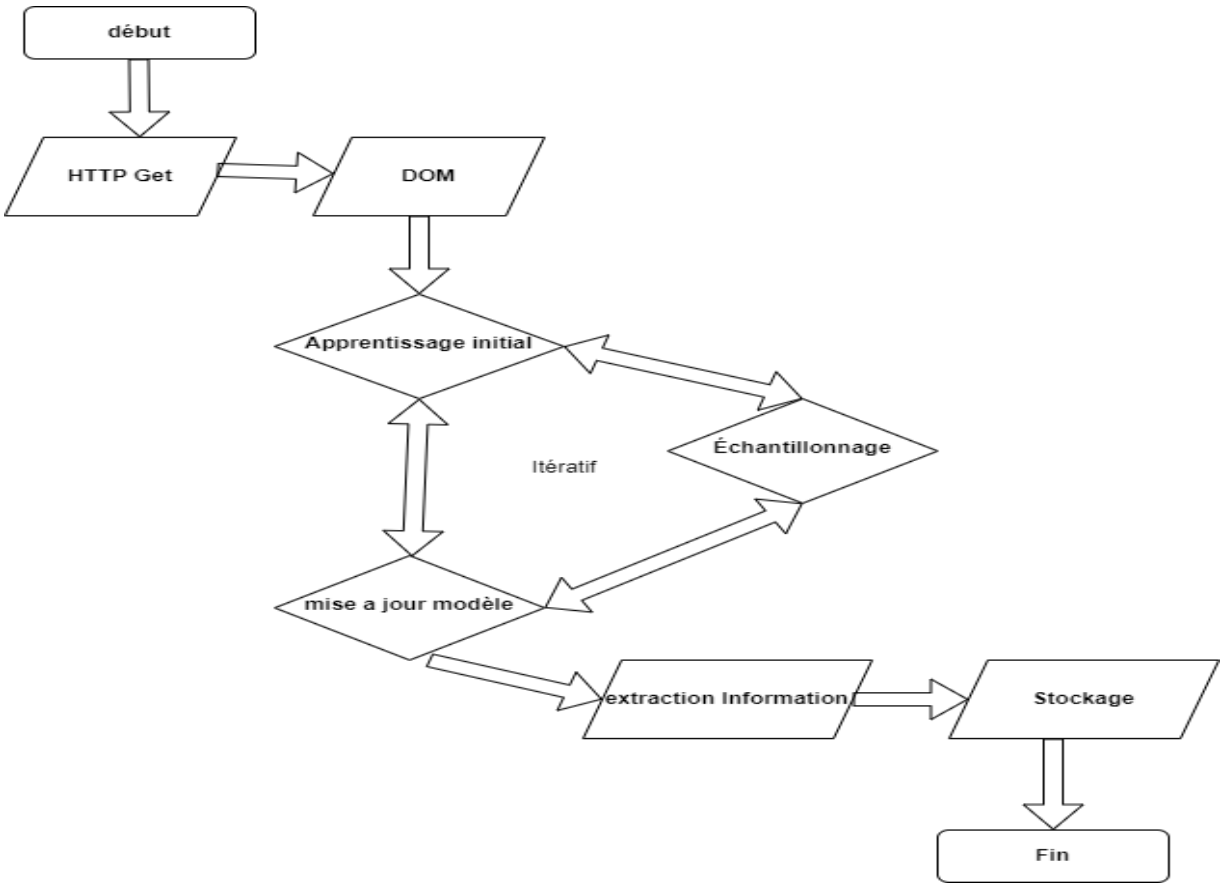


Figure 13 : ***Système de Web scraping intelligent***

1.2.5.2. Approche de base

C'est le lieu précisément d'aborder des approches de scraping avancées. Ils utilisent des techniques d'apprentissage automatique et de traitements sur langage naturel pour s'adonner à l'acquisition programmée d'éléments sur internet. Ces systèmes sont en mesure d'analyser les sites web, d'interpréter leur structure, de reconnaître les modèles utilisés puis de s'adapter aux changements pour ne pas rater leur cible en cas d'action. Aussi, ils sont en mesure de tromper les programmes anti-scraping à l'image du CAPTCHA et autres programmes cherchant à les bloquer.

1.3. Domaine d'application du web scraping

Pour toutes les entreprises qui souhaitent se maintenir dans cet écosystème en pleine croissance, la pratique du web scraping est une nécessité. Ainsi, grâce aux données collectées sur des semblables, les dirigeants peuvent les assujettir à des traitements variés pour diverses applications :

- **Le Journalisme** : Le web scraping est un outil essentiel pour le journalisme en permettant la collecte rapide d'informations sur le web. Il aide les pratiquants de ce métier à accéder à des données volumineuses et à découvrir des tendances. Ils l'utilisent pour surveiller les événements, analyser les réseaux sociaux et extraire des données pertinentes [32]. Cela permet une recherche approfondie, la création d'histoires basées sur des données et la vérification des faits.
- **La recherche :** Le grattage web facilite la collecte rapide d'informations à grande échelle. Il contribue à l'analyse des tendances, à la comparaison des données et à la création de bases de connaissances. Il est utilisé pour rassembler des informations de recherche dispersées sur le web dans le but de favoriser la synthèse de différentes sources pour des analyses approfondies [33].
- **L'intelligence Artificiel :** Dans ce domaine, la constitution des bases de données à très grange échelle est essentielle pour entraîner des modèles IA. Le harvesting alimente l'apprentissage supervisé et non supervisé, et actualise les ensembles de données. Les données textuelles issues du scraping alimentent le traitement du langage naturel (NLP) [34, 35]. Il permet également la formation de modèles généraux capables de traiter diverses informations. Toutefois, le scraping doit être effectué de manière éthique et légale.
- **Les agrégateurs de données** : Exploitent le web scraping pour collecter des informations variées sur diverses sources en ligne. Ils automatisent le processus de récupération d'informations, permettant une compilation rapide de données

provenant de multiples sites. Cette approche facilite la création de bases de données complètes en extrayant des données pertinentes à partir de différentes plates-formes et sites web. Les agrégateurs de données sont couramment utilisés dans des domaines tels que la recherche de produits, la veille concurrentielle et la comparaison de prix.

- **Veille concurrentielle** : le web scraping est largement utilisé pour surveiller les activités et les stratégies des concurrents sur le marché, en collectant des informations sur les prix, les produits, les promotions, etc.

C'est d'ailleurs ce dernier point (application du web scraping dans la veille concurrentielle) qui fait l'objet de notre étude.

Chapitre 3 : Etat de l'art

3.1. Etat de l'art du web Scraping

3.1.1. Les approche de web scraping

Il existe différentes formes de grattage, chacune adaptée à des besoins spécifiques.

3.1.1.1. Scraping basé sur les expressions régulières

En informatique, une expression régulière ou encore regex fait allusion à une séquence de caractère permettant de définir un motif de recherche. Il s'agit d'un pattern pour conditionner la formation de mot ou groupe de mot de telle sorte qu'il soit fidèle à l'expression qui la régule. Elle est utilisée pour manipuler, valider, rechercher, recueillir des renseignements qui sont conformes au format sur lequel sa composition est développée. En effet, associer les regex au scraping signifierait une collecte de données à partir d'une expression régulière. En d'autres termes, c'est une approche flexible pour traiter des chaînes de caractères en spécifiant les règles qui décrivent ses comportements. Elle est composée de lettres, de chiffres, de caractères spéciaux à l'image de ce regex pour formater un email : \b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,}\

3.1.1.2. Scraping basé sur l'arborescence du DOM

Le DOM représente la structure hiérarchique d'une page web. C'est une API qui permet au programme de lire le squelette du document, de le manipuler. Il est assimilé à l'agencement du code source HTML qui supporte le contenu web (texte, image, vidéo, photo …). C'est la base du web scraping. Il faut noter qu'avant de faire quoi que ce soit, les programmeurs inspectent, analysent et scannent l'arborescence du DOM avant de se donner les moyens techniques d'escalader le cadre pour parvenir aux données. Nous pouvons dire sans risque de nous tromper que toutes les formes de web scraping cités dans cette partie de notre travail passent par cette méthode. Cette forme permet d'atteindre les contenus même des sites dynamiques générés à l'aide JavaScript car stimule dans leur fonctionnement un navigateur qui manœuvre pour contourner les agissements asynchrones de JavaScript quand il retourne les résultats. Voici une copie de DOM avec une architecture simple.

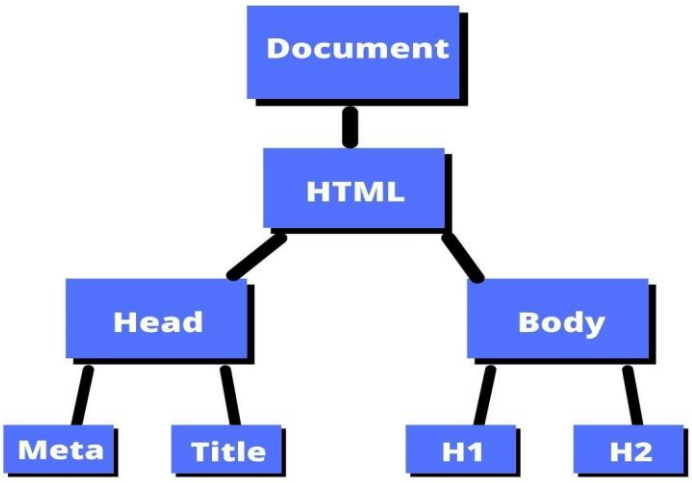


Figure 14 : ***Le DOM HTML***

3.1.1.3. Scraping basé sur le CSS

Un grattoir est dit basé sur le CSS lorsque pour effectuer sa mission d'acquisition d'éléments sur internet, il passe par des sélecteurs (class et id). C'est un mécanisme très courant dans ce domaine surtout lorsque notre besoin se porte sur des données partageant la même classe. C'est après avoir analysé DOM d'un document web pour qui nous manifestons de l'intérêt que l'on s'en sert pour pointer notre script vers les données souhaitées. Néanmoins, il serait bien de savoir que cette méthode est trop fragile aux changements d'architecture car il suffit de changer le nom du sélecteur ciblé pour que le programme rate sa piste.

3.1.1.4. Scraping basé sur les API

Nous voici face à une forme de toile de raclage très légal parce que les moyens techniques permettant d'accéder aux informations nous proviennent des propriétaires que nous visons à exploiter. Certains sites web fournissent une clé dont une API correspondante pourrait se servir pour constituer une collection très riche capable de fournir les informations que nous cherchons.

Cette astuce ne fonctionne que si et seulement si la clé en disposition est conforme avec l'API conçue pour gratter les données de l'outil ciblé.

3.1.1.5. Collecte basée sur l'apprentissage automatique :

Encore appelé collecte active, ce modèle utilise des techniques comme l'apprentissage automatique pour entraîner les algorithmes à reconnaître les pièges anti-scraping, à s'adapter aux changements d'architecture, à recueillir les informations quoi qu'il arrive. Ce procédé s'appuie sur des modelés presque entraînés à tous ces cas de figures ce qui leur donne la facilité de déjouer toutes tentatives visant à les empêcher d'acquérir le contenu des sites concurrents.

Son objectif consiste à développer des systèmes intelligents capables de sélectionner un échantillon de données à collecter en fonction du niveau de confiance du modèle qu'ils utilisent [36, 37]. Pour fonctionner correctement, ce type d'acquisition d'informations suit généralement ces étapes :

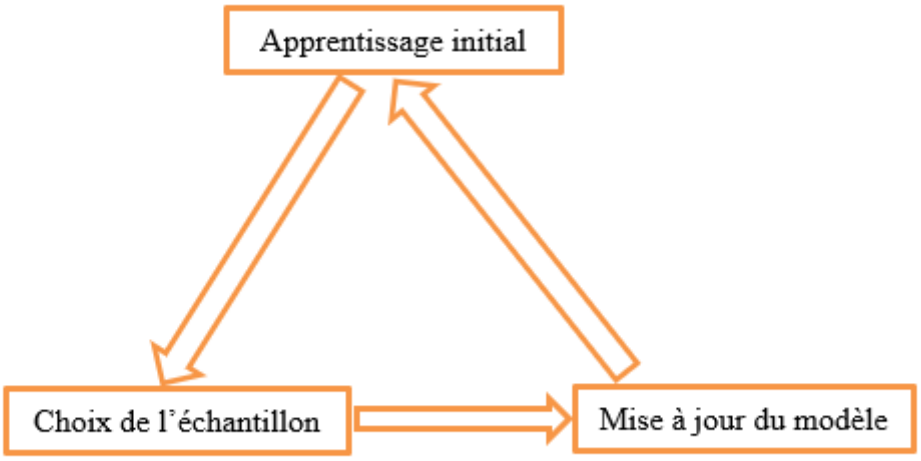


Figure 15 : ***Les étapes de la collecte basée sur l'auto-apprentissage machine***

Ce processus itératif permet au système de s'améliorer au fur et à mesure qu'il s'exécute en adaptant son modèle en fonction de ses expériences.

3.1.1.5.1. Collecte séquentielle :

La collecte séquentielle, également appelée collecte par lot, est un mécanisme dans lequel les données sont divisées, regroupées et importées par série à des intervalles réguliers ou moments spécifiques.

Elle utilise des modèles d'analyses séquentielles pour récolter des informations sur des contenus structurés. Par opposition à l'extraction d'informations en temps réel, la collecte séquentielle en agissant ainsi échappe à tout contrôle vis-à-vis des programmes détecteurs de robot.

Par exemple, dans le cas du web scraping, la collecte par lot peut capturer les données de manière différée en fonction de leur catégorie ou autres.

3.1.1.5.2. Collecte basée sur le langage naturel

Elle fonde tout son art dans la logique dont les informations sont véhiculées du point de vue du langage utilisé. Cette astuce se concentre sur le langage naturel tels que les documentations textuelles, les articles scientifiques, les articles de presse, les transcriptions audio … Voici une approche qui implique l'utilisation des NLPs (Natural Language Processing) ou encore technique de traitement du langage naturel (TLN). Ceci est un domaine interdisciplinaire de la

linguistique, de l'informatique et de l'IA concernant les interactions entre ordinateur [38]. Cette pseudoscience est utilisée pour analyser, extraire des renseignements en juxtaposant le langage des machines (informatique) et celui des hommes (naturel). Le traitement du langage naturel aide les programmes à comprendre le sens et la structure du texte présent sur les pages. Elle sert à extraire des contenus complexes à partir des descriptions. C'est à partir de là que cette méthode valide et saisit les données souhaitées.

Par exemple, un NLP spécialisé dans le traitement des sentiments à tendance à reconnaître le vocabulaire lié à ce domaine car est capable de toucher votre sensibilité en se basant uniquement du langage naturel : c'est le cas des chatbots Ainsi, appliquée au scraping, la collecte basée sur le langage naturel offre au programme collecteur toute la chance de dissocier les informations pertinentes à celles intitulées pour la constitution d'une compilation de données immensément riche.

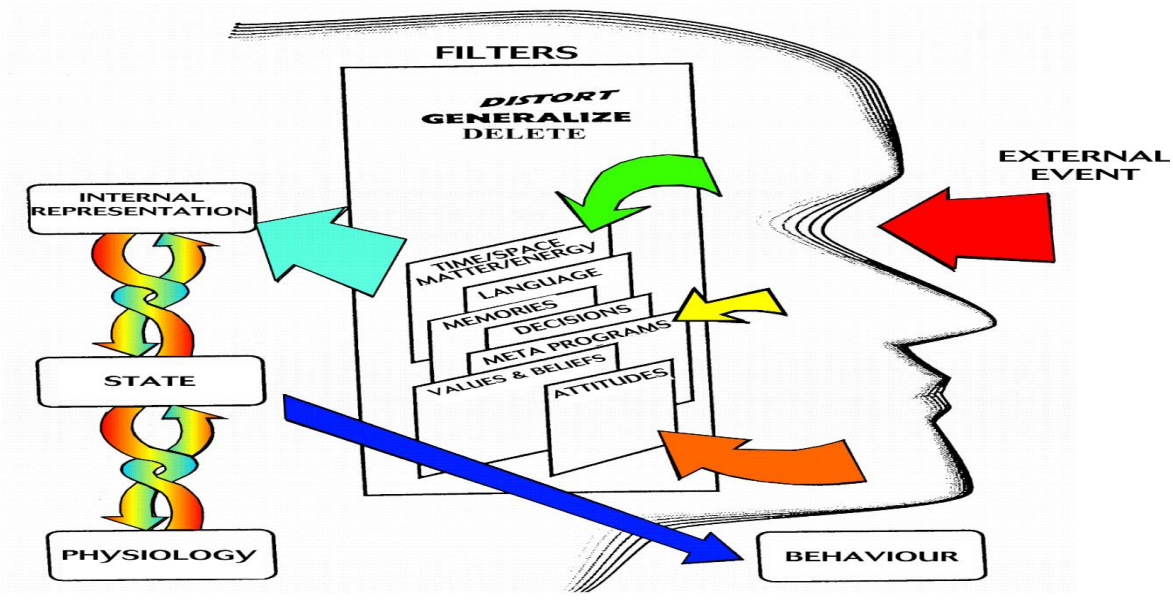


Figure 16 : ***Programmation neurolinguistique***[5]

3.1.2. Web Scraping et langage de programmation

Les langages de programmation sont très nombreux de nos jours. Ils interviennent dans le cadre de la mise en place des solutions informatiques de toute sorte. Par conséquent, les développeurs portent leurs préférences sur les uns par rapport aux autres selon des critères liés à leur environnement de travail, à leurs compétences. En effet, chaque langage a ses spécificités, ses caractéristiques, ses avantages, ses inconvénients. Plusieurs d'entre eux ont développé des outils permettant de capturer des données web [39]. Ceci étant, à travers une étude comparative,

[5] https://bit.ly/3O28K6L

nous allons tenter de décliner les rapports entre les langages informatiques les plus célèbres et le grattage de données.

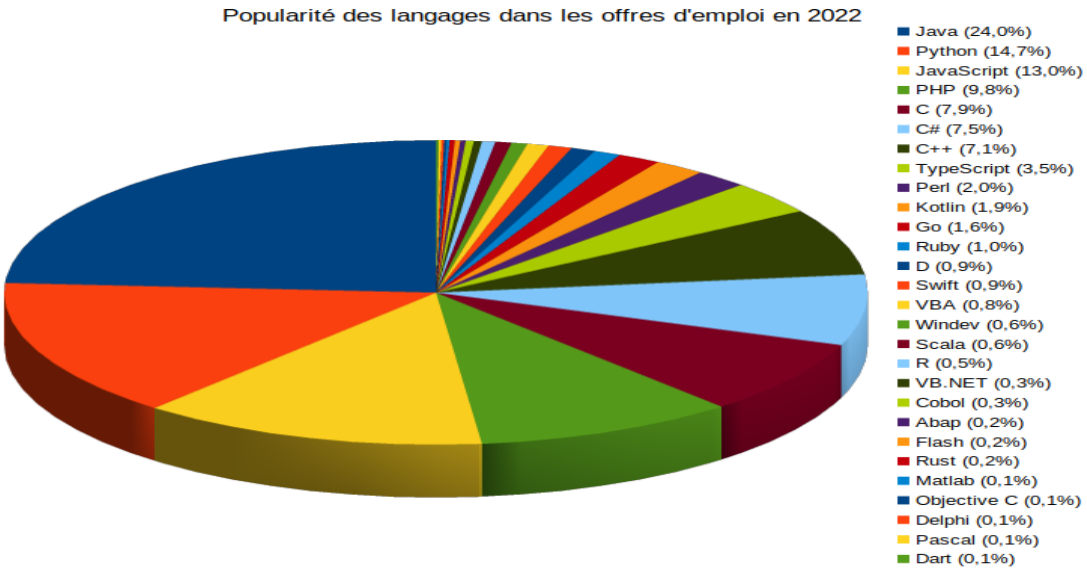


Figure 17 : ***Les langages informatiques les plus utilisés en 2022***[6]

3.1.2.1 Web Scraping et PHP

PHP est un langage de programmation largement utilisé pour le développement web. Bien qu'il soit capable de faire du web scraping, cet outil n'est pas couramment utilisé pour ça mais plutôt pour le développement back office. C'est lorsque le bot est intégré à des projets PHP existants qu'il est souvent convoqué pour collecter des données. En revanche, PHP dispose certaines bibliothèques conçues pour explorer des contenus web :

- **Simple HTML DOM Parser : se** veut un outil simple et léger pour formater et recueillir des renseignements à partir des architectures web. Cet outil est conçu pour faciliter le processus de parsing[7] de pages HTML. Elle fournit une approche simple et légère pour parcourir et extraire des informations à partir de pages web en utilisant des méthodes de sélection similaires à celles de jQuery.
- **Goute** : se trouve être une bibliothèque PHP pour effectuer des requêtes HTTP afin de soutirer des informations venant du web. Elle est basée sur Guzzle, une autre bibliothèque PHP pour effectuer des requêtes HTTP et utilise des tâches de collecte d'informations, de surveillance de sites web, d'automatisation de tests, et bien d'autres cas d'utilisation.

6 *https://www.developpez.net/forums/d2149017/emploi-etudes-informatique/emploi/emploi-informatique-2022-langages-programmation-plus-demandes-mieux-payes/*

7 Analyse syntaxique

- **WebScraping.io** : est un service API d'extraction de données basé sur PHP avec comme vocation la constitution de collections d'éléments à partir de la toile. Comme ScrapingBee, webScraping.io offre une solution pour collecter des informations à partir de pages web à grande échelle grâce à une automatisation du processus de web scraping.
- **ScrapingBee :** il consiste aussi à un service API PHP qui permet à son tour de faire du web scraping à grande échelle. Cette API offre une protection contre les blocages. Il prend en charge plusieurs langages, PHP y compris. Concrètement, c'est une plateforme qui permet aux développeurs de collecter des informations à partir de sites web en automatisant le processus de web scraping.
- **ScraperAPI : voici** un autre service API PHP pour scraper des données sur internet qui fournit aussi des fonctionnalités qui évitent les blocages et CAPTCHAs. L'utilisation de ScraperAPI consiste à envoyer L'URL cible sous forme de requête HTTP à L'API avec une clé et cette dernière se chargera de scanner le document HTML afin de lui soutirer les informations. Pour obtenir cette clé d'authentification, il faut un compte, une fois la clé fournie, Nous devons à chaque demande l'inclure pour pouvoir bénéficier des services de cet outil.
- **QueryPath :** une bibliothèque pour recueillir à partir des pages HTML des informations avec PHP, offrant la possibilité d'utiliser des sélecteurs CSS et des expressions XPath. Cela facilite le processus d'extraction des données spécifiques à partir des balises.
- **PHP Web Scraper :** une bibliothèque simple utilisée par PHP pour des tâches de scraping, dotée d'une syntaxe facile à comprendre et à utiliser. Elle est idéale pour les développeurs qui recherchent une solution rapide et efficace pour extraire des données de pages web.
- **PHP Crawler :** une bibliothèque qui fait la corrélation entre le crawling et le scraping de sites web en PHP, avec des fonctionnalités pour extraire des liens et des données. Cela permet de parcourir plusieurs pages d'un site pour récupérer des informations à grande échelle.
- **WebCrawler :** une bibliothèque pour le scraping de pages web en PHP, conçue spécialement pour extraire des données structurées à partir de sites web. Cela facilite l'extraction de données organisées dans un format prévisible.

3.1.2.1. Web Scraping et JavaScript

JavaScript est un autre langage couramment utilisé pour le scraping web. Il intervient plus dans le développement frontend. Néanmoins, il dispose de technologies de développements

côté serveur. Concernant le moissonnage, il met à la disposition des programmeurs certains outils :

- **Cheerio : est** une bibliothèque JavaScript inspirée de jQuery qui permet de manipuler et de parcourir des documents HTML. C'est une excellente option pour le web scraping en JavaScript côté serveur (Node.js). Grâce à ses capacités de sélection, de « parsing » et de manipulation de pages web. Cheerio est largement utilisée pour effectuer du web scraping, extraire des données spécifiques, effectuer des transformations de données et bien plus encore dans des projets Node.js. Elle offre une approche simple et puissante pour l'analyse de pages web et l'extraction de données en JavaScript.
- **Puppeteer :** qui est un outil de test et d'automatisation de navigateur développé par Google. Il peut être utilisé pour le web scraping, car il permet d'émuler un navigateur complet, y compris l'exécution de JavaScript et l'interactivité avec des pages générées par JavaScript. En raison de ses fonctionnalités étendues de contrôle de navigateur, d'exécution de JavaScript, de capture d'écran et de génération de PDF. Puppeteer est particulièrement adapté aux projets de web scraping et d'automatisation de tests dans un environnement Node.js aussi. Il offre une approche puissante pour interagir avec des pages web et récupérer des données de manière dynamique.
- **Request-Promise et Axios : sont** deux bibliothèques JavaScript qui permettent d'effectuer des requêtes HTTP pour récupérer le contenu HTML d'une page web. Ils sont souvent utilisés en combinaison avec Cheerio ou d'autres bibliothèques de « parsing » pour le web scraping. Tant Request-Promise qu'Axios sont des bibliothèques très utilisées dans la communauté JavaScript pour effectuer des requêtes HTTP. Ils offrent des fonctionnalités similaires, mais leur choix dépend souvent des préférences personnelles du développeur et des exigences spécifiques du projet.
- **Node-crawler :** une bibliothèque de crawling et scraping pour Node.js qui offre des fonctionnalités asynchrones pour extraire des données de manière efficace et performante.
- **JSDom :** une bibliothèque de parsing HTML pour Node.js qui permet d'extraire et de manipuler des données à partir de pages web en utilisant des sélecteurs CSS, offrant ainsi une approche similaire au DOM en environnement serveur.
- **AxiosScraper :** une bibliothèque qui combine l'utilisation d'Axios avec des fonctionnalités de web scraping pour extraire facilement des données à partir de pages web, en profitant de la simplicité et de la puissance d'Axios pour effectuer les requêtes HTTP.

- **Osmosis :** une bibliothèque de scraping pour Node.js qui facilite l'extraction de données à partir de pages web en utilisant des sélecteurs CSS, permettant ainsi une approche flexible pour cibler les éléments spécifiques à extraire.
- **Horseman :** une bibliothèque de contrôle de navigateur pour Node.js, basée sur PhantomJS, qui offre des capacités avancées pour interagir avec les pages web et extraire des données. Cette bibliothèque permet de simuler des actions d'un utilisateur réel sur un navigateur, ce qui est utile pour extraire des données de sites web générés par JavaScript.

3.1.2.2. Web Scraping et JAVA

Quant à Java, c'est un langage de programmation polyvalent et très puissant utilisé dans une grande variété d'applications, y compris le scraping web. Java est appréciée pour sa fiabilité, sa performance et sa portabilité, ce qui en fait un choix populaire pour les projets de grande envergure. Il dispose de bibliothèques qui permettent de manipuler et de parcourir des documents HTML pour extraire des données :

- **Jsoup** : c'est une bibliothèque que java utilisée pour formater, manipuler des structures de pages dans le but d'obtenir des données web. En fait, elle est robuste pour le « parsing » et la manipulation de pages HTML. Elle est largement utilisée pour le web scraping et l'extraction de données à partir de pages web dans des projets Java. Avec sa syntaxe simple et ses fonctionnalités de manipulation du DOM, elle facilite la collecte et l'analyse de données à partir de sites web de manière efficace.
- **Selenium WebDriver :** cette bibliothèque java qui permet l'automatisation des tests web utilisable pour scraper des pages générées par JavaScript. Selenium WebDriver est puissante et polyvalente pour l'automatisation de tests de sites web, et elle peut également être utilisée pour l'extraction de données. Avec son contrôle complet du navigateur, ses fonctionnalités de manipulation du DOM et ses capacités d'interaction avec le contenu, c'est un choix populaire pour les projets Java qui nécessitent des scénarios avancés d'extraction de données à partir de pages web.
- **HTMLUnit :** voici une librairie java capable de faire du web scraping. Elle est utilisée pour diverses tâches d'automatisation, telles que les tests automatisés, le web scraping, l'extraction de données à partir de pages web, la simulation d'interactions utilisateur et bien plus encore. En raison de sa légèreté et de ses performances élevées, HTMLUnit est une excellente option pour les projets Java qui nécessitent un navigateur headless[8] pour effectuer des tâches d'automatisation sans avoir besoin d'une interface graphique.

[8] Navigateur web qui fonctionne sans interface graphique, sans entête

- **WebHarvest :** est un moteur java visant à automatiser des tests web et aussi à gratter des faits. Cet outil est très pratique pour des projets de web scraping et d'extraction de données où l'utilisation de fichiers de configuration XML est préférée pour définir les tâches d'extraction. Il offre une approche simple et pratique pour automatiser la collecte de données à partir de pages web sans avoir besoin d'écrire du code Java.
- **Abot :** une bibliothèque de crawling pour Java qui offre des fonctionnalités pour extraire des données à partir de pages web en utilisant des politiques de crawling personnalisables, ce qui permet de contrôler précisément le processus de scraping.
- **HtmlCleaner :** une bibliothèque de « parsing » HTML pour Java qui permet d'extraire des données à partir de pages web tout en nettoyant le contenu, ce qui facilite le processus d'extraction de données bien formatées.
- **Jauntium :** une bibliothèque pour le scraping de pages web en Java, qui propose des fonctionnalités avancées pour gérer les pages dynamiques et les formulaires, ce qui est utile pour extraire des données à partir de sites web qui utilisent du contenu généré par JavaScript.
- **WebCollector** : une bibliothèque de crawling et scraping pour Java conçue pour extraire des données à partir de sites web de manière efficace et parallélisée, ce qui permet d'accélérer le processus de scraping pour traiter de grandes quantités de données.

3.1.2.3. Web Scraping et PYTHON

Python est classé parmi les langages de programmation les plus célèbres. Il est largement considéré comme l'un des meilleurs langages de programmation pour le scraping web. D'ailleurs il suffit de demander à YouTube ou Google « *comment mettre en place un système de web scraping* » sans spécifier de langage, la majeure partie des résultats suggérés seront liés à ce dernier. Python dispose de nombreuses fonctionnalités intégrées pour le traitement du HTML, l'extraction et l'analyse de données et la manipulation des requêtes HTTP, ce qui en fait un choix populaire pour le scraping web. Il est doté également d'une très large gamme de bibliothèques et de Framework dédiés tels que :

- **Selenium :** qui est un outil open source très pratique dans l'automatisation des interactions avec des contenus web. Il permet de gratter des informations sur la toile. À la différence de BeautiFulSoup, Sélénium est capable de récolter des informations générées par JavaScript.
- **BeautiFulSoup :** c'est une bibliothèque python très utilisée dans l'atmosphère du grattage de données. BeautiFulSoup fournit des méthodes simples pour naviguer, rechercher et modifier le DOM. Il transforme un document HTML complexe en un arbre

d'objets Python et convertit aussi automatiquement le document en Unicode de sorte que vous n'aurez pas à penser à l'encodage.

- **Scrapy :** c'est un Framework python complet et puissant pour le harvesting. Avec ces fonctionnalités avancées, il assure correctement la gestion des données. Scrapy est largement utilisé pour l'extraction de données à partir de sites web complexes. Grâce à son architecture modulaire, sa gestion des requêtes HTTP, sa prise en charge de la pagination et des liens, ainsi que sa flexibilité, il est un choix populaire pour les projets de scraping Python qui nécessitent une solution complète et évolutive.
- **Requests-HTML** : encore une autre bibliothèque python pour collecter du contenu HTML de manière simple. C'est une excellente bibliothèque pour les projets de récolte de données en ligne scraping dont dispose Python. Grâce à son intégration étroite avec « requests », sa facilité d'utilisation et ses fonctionnalités pour le « parsing » de pages HTML, Requests-HTML offre une solution pratique et performante pour collecter des données à partir de sites web.
- **PyQuery :** une librairie python qui offre une interface similaire à jQuery pour parcourir, manipuler et extraire le contenu HTML et XML. Elle offre une syntaxe familière, une intégration aisée avec d'autres bibliothèques et une manipulation facile du DOM. Grâce à sa rapidité de « parsing » et à sa convivialité, PyQuery est largement utilisée pour collecter des données à partir de pages web dans des projets Python.
- **MechanicalSoup :** une bibliothèque pour le scraping de pages web en Python, qui facilite le remplissage de formulaires et l'extraction de données. Elle permet d'automatiser l'interaction avec les sites web, simulant le comportement d'un utilisateur.
- **Lxml :** une bibliothèque de parsing XML et HTML pour Python, qui offre des capacités avancées d'extraction de données à partir de pages web en utilisant des expressions XPath. Elle est rapide et efficace pour le traitement de grands volumes de données.
- **Pyppeteer :** une bibliothèque de contrôle de navigateur headless pour Python, basée sur Puppeteer. Elle permet d'interagir avec les pages web de manière similaire à un navigateur, ce qui est utile pour les sites web générés par JavaScript.
- **Requests :** une bibliothèque pour effectuer des requêtes HTTP en Python, souvent utilisée en combinaison avec BeautifulSoup pour le scraping. Elle facilite la récupération du contenu des pages web avant de les analyser.

Tableau 1 : Etude comparative des langages de programmations pour scraping

	Avantages	***Inconvénients***
Python	Syntaxe claire, simple, grande communauté et ressource disponible	Nécessite des compétences techniques avancées
	Plusieurs librairies et bibliothèques dédié au scraping	Rigoureux par rapport à l'indentation
	Bon support de traitement et d'extraction de données	Environnement de travail particulier
Java	Flexible, performance élevée, forte communauté	Syntaxe plus compliqué
	Portabilité grâce à la machine virtuelle (JMV)	Configuration initiale plus complexe par rapport à python
JavaScript	Utilisation courante dans le développement web, ce qui facilite l'intégration du scraping web dans des applications existantes	Nécessite souvent l'utilisation de bibliothèques tierces pour des fonctionnalités avancées
	Bon support pour la manipulation du DOM	Peut-être plus complexe à mettre en place et à configurer par rapport à Python
PHP	Intégration facile avec des projets PHP existants	Moins utilisé pour le scraping web par rapport à Python ou JavaScript
	Syntaxe familière pour les développeurs PHP	Écosystème de scraping web moins développé que celui de Python

3.1.3. Outils de Scraping prêts à l'emploi

En évidence, le scraping web n'est pas seulement l'affaire des connaisseurs en programmation informatique même s'ils sont les fournisseurs de solutions nous permettant d'effectuer efficacement cette tâche. Cela veut dire qu'il existe des outils non-codes [40, 41] disponibles pour pratiquer la collecte automatisée de données web. Ces APIs prêtent à l'emploi sont très habile dans la mission de rassemblement de contenu via internet. Sachant qu'il est relativement difficile de développer des systèmes très pratiques pour ça nous même, d'autres se sont substitués à nous en proposant des interfaces toutes faites avec des fonctionnalités

dédiées à la constitution d'agrégats. Dans les lignes qui suivent, nous allons exploiter quelques-unes :

- **Import.io[9] :** c'est une plateforme spéciale pour les entreprises d'e-commerce. Son slogan étant « *Enterprise scale eCommerce data to drive growth* » qui se traduit comme étant « *les données de commerce électronique à l'échelle de l'entreprise pour simuler la croissance* », elle permet de recueillir des informations sur nos concurrents de manière facile et simple puis de les structurer. Elle peut être exploitée en mode SaaS (Software As A Service) ce qui signifie que nous n'avons pas besoin de l'installer obligatoirement dans nos supports. Impor.io est un outil d'or pour la veille concurrentielle. D'ailleurs, pour se présenter la société laisse entendre ceci sur leur page d'accueil : « *Des millions de pages ? Des milliards de points de données ? Aucun problème. Nous fournissons les données Web dont vous avez besoin pour dynamiser votre entreprise avec des applications intuitives, des API puissantes et des services d'experts.* »

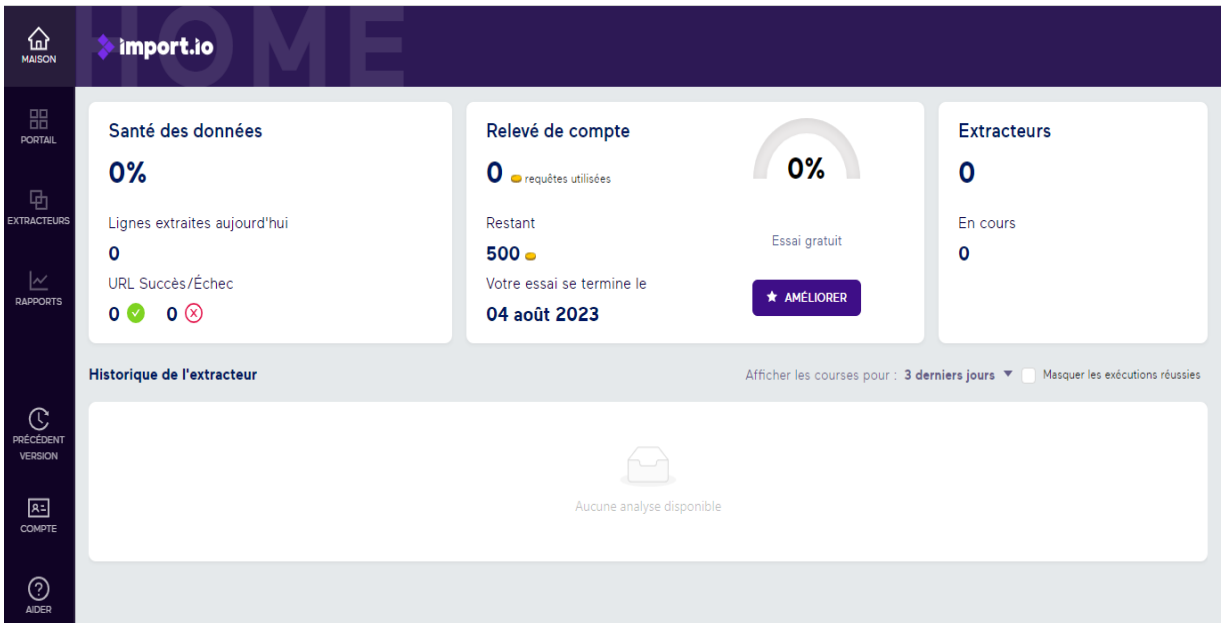


Figure 18 : ***Tableau de bord* de la plateforme *Import io***

- **Octoparse[10] :** octopasre est également une plateforme pour recueillir des informations web, il est utilisable en mode SaaS gratuitement. C'est aussi un outil téléchargeable et installable sur nos appareils. Il intervient dans différents domaines à savoir l'e-commerce, la crypto-monnaie, l'investissement, la bourse, le marketing, l'immobilier etc. Sa particularité est qu'il fournit un modèle de web scraping qui couvre les sites les plus populaires au monde (CF : Figure 19) avec la possibilité de personnaliser la

[9] https://www.import.io/
[10] https://www.octoparse.com

collection en fonction des besoins ressentis. Il est également très facile à utiliser et nécessite un abonnement. Pour scraper n'importe quel site, il suffit de renseigner son URL dans un formulaire et lancer puis il vous redirige vers un modal pour la personnalisation de vos données et pour terminer, vous validez et comme par magie les informations sont chargées dans un format que vous aurez à définir.

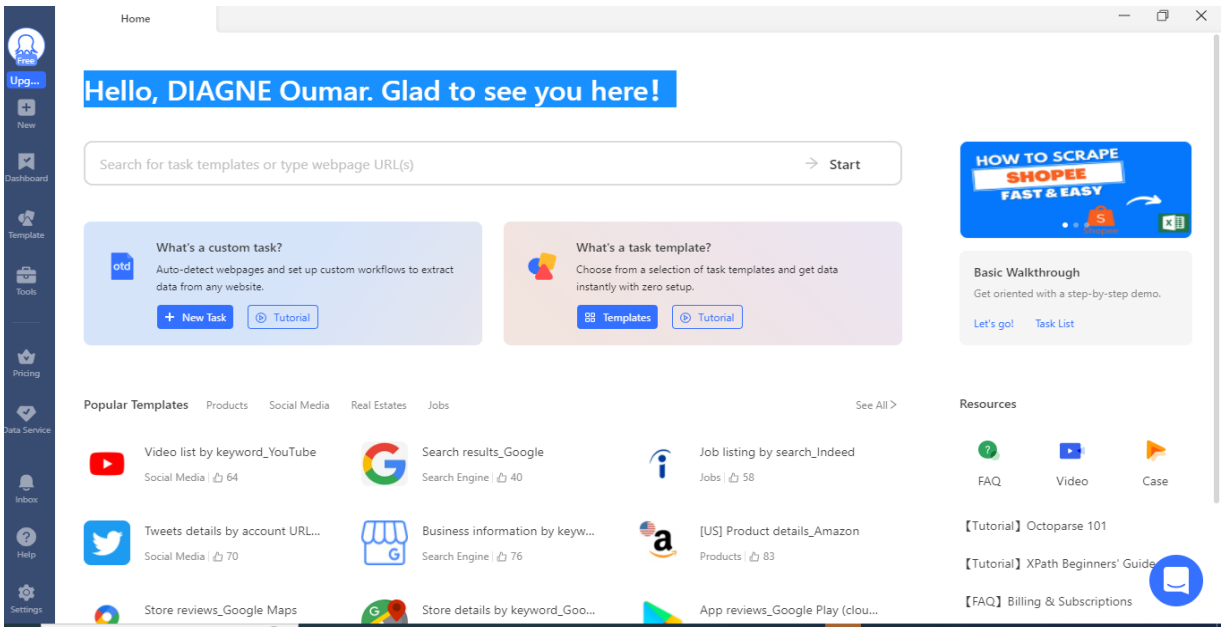


Figure 19 : ***Tableau de bord* de la plateforme *Octoparse***

- **Easy Web data Scraper :** c'est une extension Google qui permet d'extraire des données sur internet. Il suffit de se positionner sur un site cible puis nous lançons l'extension. Ainsi, il détecte automatiquement l'URL et commence à capturer les écrans un en un avec la possibilité aussi de paramétrer la sortie des données. Pour l'utiliser, il faut d'abord installer l'extension sur Google.

Figure 20 : ***Interface de l'outil Easy Web data Scraper***

- **Parsehub[11] :** cet outil est plus analyste de données et demande un certain niveau de compétence. Pour l'utiliser nous pouvons nous abonner à son service SaaS ou même l'installer dans nos appareils. Parsehub est favorable au web scraping car il dispose d'une fonctionnalité de rotation IP qui permet de changer votre adresse IP lorsque vous rencontrez des sites équipés de technique anti-scraping. C'est l'outil idéal pour le marketing mais aussi pour les chercheurs. A noter que pour avoir accès à ces fonctionnalités il faut d'abord se connecter.

[11] https://parsehub.com/

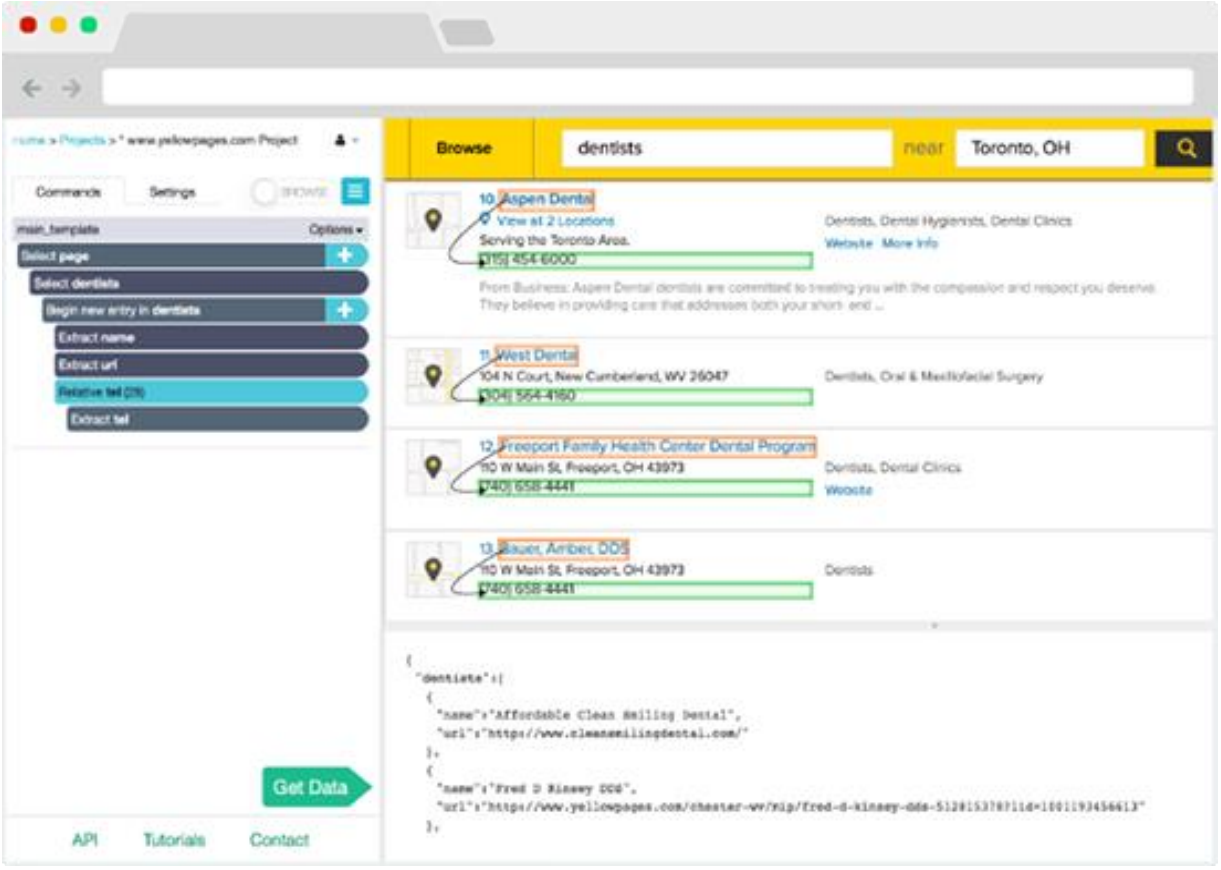


Figure 21 : ***Mode d'utilisation de la plateforme Parsehub***

- **ProWebScraper[12] :** il s'adresse à toutes les personnes physiques ou morales qui veulent acquérir facilement des données web sans écrire de code sur une interface très facile à prendre en main. Ces fonctionnalités les plus avancées sont payantes. C'est un outil qui permet de récupérer des données à grande échelle.

Figure 22 : ***Tableau de bord de la plateforme ProWebScraper***

[12] https://prowebscraper.com/

- **Monzanda[13] :** c'est un puissant outil de web scraping et d'automatisation de données qui permet aux utilisateurs d'extraire des informations à partir de sites web de manière efficace et facile. Cet outil est conçu pour les entreprises et les professionnels qui ont besoin de collecter des données à grande échelle pour la veille concurrentielle, l'analyse de marché, la génération de leads, la surveillance de prix, et bien d'autres applications. Son accompagnement ne s'arrête pas à l'extraction de données, Monzanda dispose aussi des fonctionnalités de visualisation de données qui pourraient beaucoup faciliter le travail d'un analyste de données par exemple. Il est installable sur ordinateur et aussi nous pouvons l'utiliser en ligne.

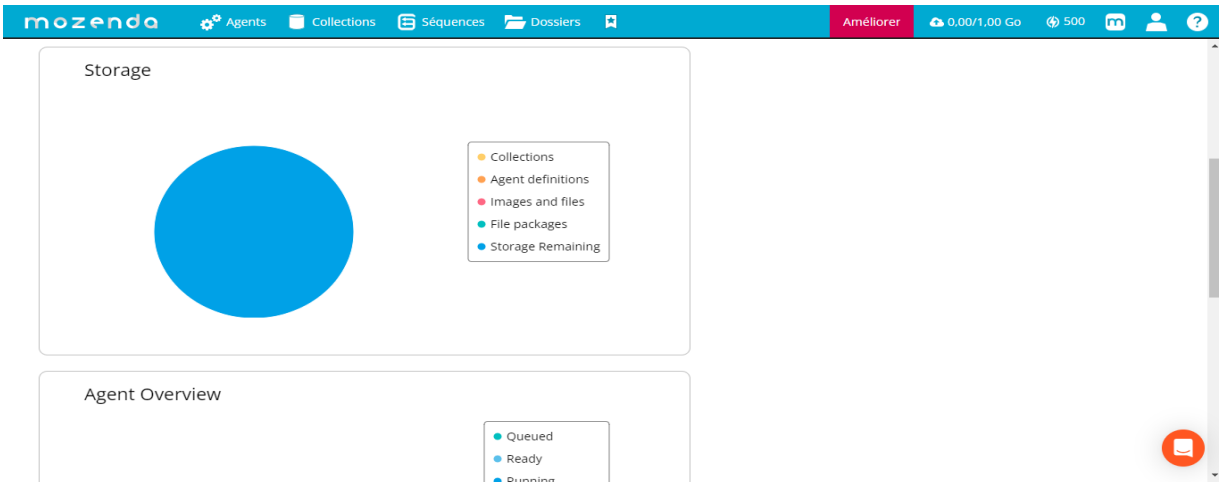


Figure 23 : ***Tableau de bord de la plateforme Monzanda***

- **Web Content Extractor[14]** : est un autre outil puissant de web scraping conçu pour collecter des données à partir de sites web de manière automatisée et efficace. Il est particulièrement utile pour les entreprises, les chercheurs, les professionnels du marketing et toute personne ayant besoin de collecter des informations spécifiques à partir de pages web.

[13] https://www.mozenda.com/
[14] https://www.webcontentextractor.com/

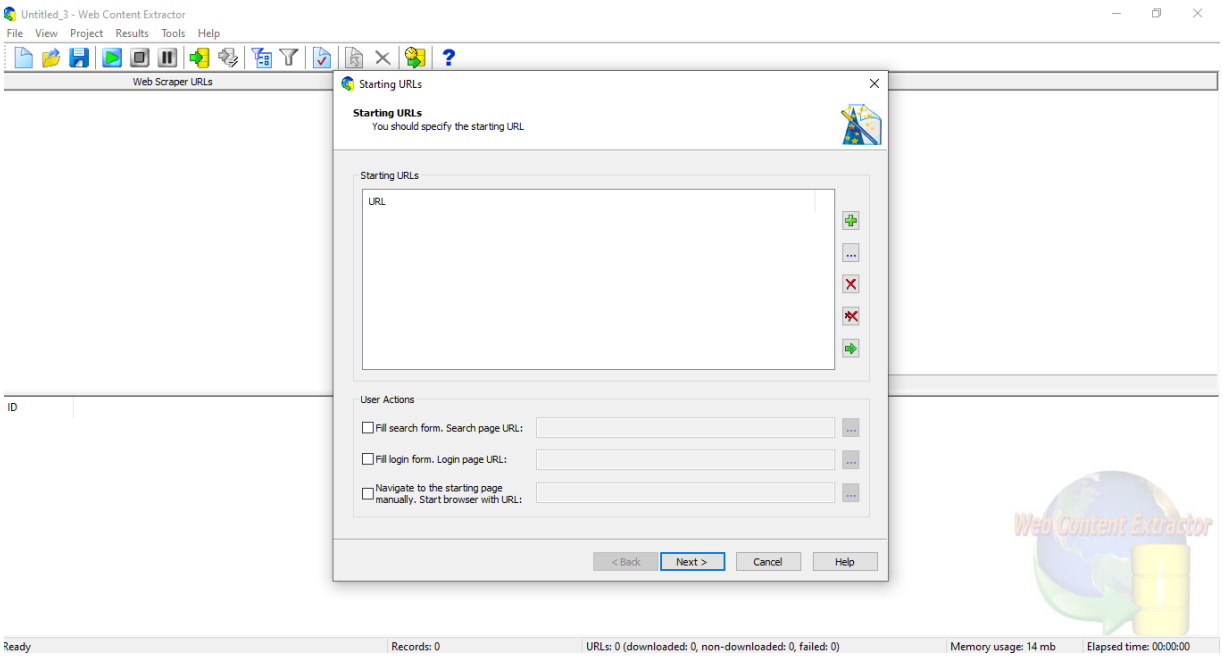


Figure 24 : ***Tableau de bord Web Content Extractor***

3.1.4. Web Scraping Vs Web Scraping Intelligent

L'étude comparative entre le web scraping traditionnel et le web scraping intelligent (également appelé web scraping avancé ou web scraping basé sur l'intelligence artificielle) permet de mieux comprendre les avantages et les inconvénients de chaque approche pour l'extraction de données à partir de sites web. Voici une comparaison des deux méthodes :

Tableau 2 : Web scraping traditionnel contre Web scraping avancée

Détails	**Web scraping traditionnel**	**Web scraping avancée**
Approche	Facile en mettre en place	Souvent Complexe
Efficacité	Limité	Pointu
Défis de collecte	Sensible au DOM, à l'anti-scraping	Forte capacité d'adaptation
Vitesse d'exécution	Lent	Rapide
Coûts	Moins coûteux	Coûteux

En définitive, le web scraping intelligent présente des avantages significatifs en termes d'efficacité, d'adaptabilité et de convivialité par rapport au web scraping traditionnel. Cependant, un choix entre les deux méthodes dépend des besoins spécifiques du projet, des compétences techniques de l'utilisateur et des contraintes budgétaires. Certaines tâches de scraping peuvent être bien adaptées au web scraping traditionnel, tandis que d'autres peuvent bénéficier des capacités avancées du web scraping intelligent.

3.1.5. Bibliothèques des langages de programmation Vs Outils prêts à l'emploi

Entre coder son propre système de collecte de données et utiliser les outils prêts à l'emploi pour la même tâche paraît ne nécessiter aucun effort dans le choix vu la facilité avec laquelle il est possible de constituer une banque de données avec certains outils mais, cela n'est pas toujours le cas. De ce fait, à travers une étude comparative, nous avons pu distinguer :

Tableau 3 : Bibliothèques des langages de programmation Vs Outils prêts à l'emploi

Détails	Bibliothèques	Outil prêt à l'emploi
Prise en main	Lent	Rapide
Utilisation	Compétences en programmation	Interface graphique
Résultats	Personnalisable entièrement	Moins que ça
Coûts	Open source (Gratuit)	Payant (Pour meilleur usage)
Outil	Flexible	Figer
Traitement de données	Contrôle total	En partie
Evolution	Maintenable et ajustable	Fixe

Bref, le choix entre l'utilisation de bibliothèques de programmation pour le scraping et l'utilisation d'outils de scraping prêts à l'emploi dépend comme précédemment du contexte, du projet, des compétences et du budget. Les bibliothèques de programmation offrent une plus grande flexibilité et une personnalisation avancée, tandis que les outils de scraping prêts à l'emploi sont plus accessibles et conviviaux pour les utilisateurs sans expérience en programmation.

3.2. Etat de l'art sur le scraping dans la veille concurrentielle

Le web scraping au service de la veille concurrentielle consiste à collecter de manière automatisée et systématique des informations pertinentes sur les activités, les produits, les services, les stratégies et les performances de ses concurrents à partir de sources en ligne, notamment les sites web des concurrents. Cette approche permet aux entreprises d'obtenir un avantage concurrentiel en surveillant de près leurs rivaux sur le marché. Grâce au web scraping, les entreprises peuvent obtenir des données en temps réel sur les prix des produits concurrents,

les promotions en cours, les avis des clients, les nouvelles annonces de produits, etc.

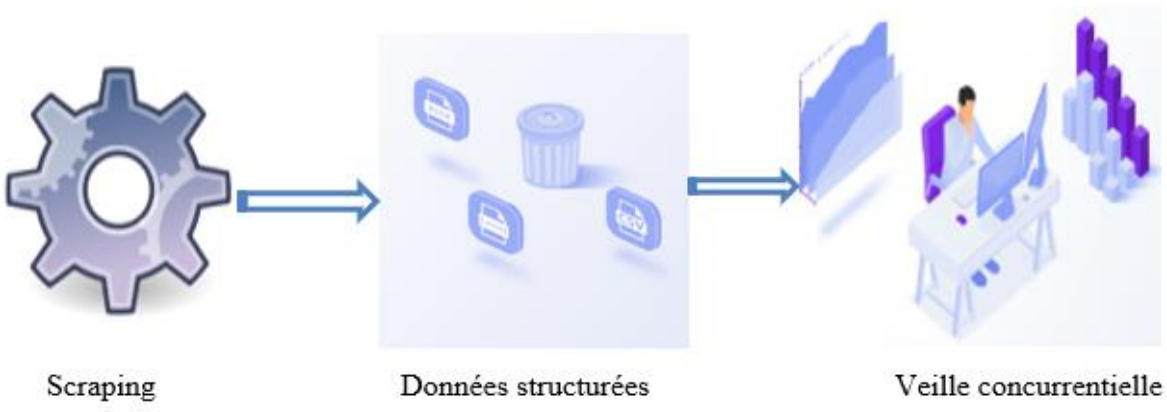


Figure 25 : ***Web scraping et Veille concurrentielle***

Chacun de ces domaines de la veille concurrentielle est essentiel pour prendre des décisions stratégiques éclairées et maintenir un avantage concurrentiel dans un environnement commercial dynamique et compétitif. A ce propos, le web scraping appliqué à la veille concurrentielle permet aux décideurs de manager leurs entreprises en se basant sur des informations probantes.

En somme, nous pouvons distinguer les formes d'analyse d'entreprise suivant deux paramètres : les Objectifs et les cibles.

Selon l'objectif, nous avons » [42,43] :

- V**eille concurrentielle passive :** elle consiste à collecter les informations sur ses semblables sans but précis. C'est une forme de veille sans trop d'impact car se contente de surveiller et de se renseigner sur des compétiteurs sans directement interagir avec les données dont elle dispose pour ne pas dire « s'informer pour s'informer ».
- **Veille concurrentielle active** : contrairement à la première, la veille active correspond parfaitement à celui soulevé par notre thème. Elle se veut proactive, cible des informations spécifiques provenant de diverses sources puis de les réutiliser pour produire de la connaissance et à partir de laquelle des actions concrètes sont menées dans le but de maintenir une entreprise donnée dans la bonne direction.

Selon la cible, nous avons [44] :

- **Veille concurrentielle directe** : elle consiste à collecter des données sur toute société qui propose des produits et services similaires à d'autres. Sont considéré comme des concurrents directs quand une entreprise X cible en même temps qu'une entreprise Y le même segment de marché. Dans ce combat, les organisations cherchent à attirer et fidéliser les mêmes clients.

- **Veille concurrentielle indirecte :** en revanche, les concurrents indirects sont ceux qui proposent des produits et services différents qui répondent à d'autres besoins ou désirs des clients.
- **Veille concurrentielle mixte :** cette approche combine à la fois des méthodes passives et actives, ainsi que des approches directes et indirectes pour recueillir des informations complètes et approfondies sur les concurrents.

A partir de ces deux axes, plusieurs combinaisons peuvent être mises sur place [45] :

- **Veille concurrentielle passive directe** : il s'agit de la collecte d'informations sur les concurrents de manière discrète et directe, en consultant leurs sites web, leurs publications, leurs documents publics, leurs rapports financiers, etc. sans interagir directement avec eux. Cette veille permet de suivre les informations de réels concurrents avec moins d'ambitions.
- **Veille concurrentielle passive indirecte** : dans ce cas, les informations sur les concurrents sont collectées de manière indirecte, en observant leurs actions sur le marché, leurs mouvements stratégiques, leurs annonces publiques, les commentaires des clients, etc., sans avoir de contacts directs avec eux.
- **Veille concurrentielle active directe** : elle implique une interaction directe avec les concurrents, tels que des sondages, des entretiens, des questionnaires, des échanges d'informations, etc. pour obtenir des informations de première main. C'est une notion qui sous-entend une espionnage active menée sur des concurrentes frontales histoires de mieux les connaitre pour finalement mieux les dominer.
- **Veille concurrentielle active indirecte** : ici, l'interaction avec les concurrents se fait de manière indirecte pour obtenir des informations sur leurs stratégies et leurs activités. Ce type de veille préconise, la récolte, le traitement et l'utilisation de la connaissance produite sur ces rivaux indirects pour prendre de l'avance sur eux au cas où ils seraient passés l'étape de concurrent direct.
- **Veille concurrentielle passive mixte** : il s'agit d'une combinaison d'approches de veille passive directe et indirecte, en recueillant des informations directes et indirectes sur les concurrents pour une vue d'ensemble plus complète. Pour cette manière de faire, les usagers s'informent uniquement pour s'informer sur tous leurs concurrents, sans exception.
- **Veille concurrentielle active mixte** : ce type de veille combine des méthodes actives directes et indirectes, en interagissant directement avec certains concurrents tout en obtenant des informations indirectes sur d'autres concurrents par le biais d'autres

sources. Dans ce type d'observation, la notion de concurrent qu'il soit direct ou indirect subit le même sort dans la manière dont les informations seront traitées et utilisées.

Il est important de noter que le web scraping peut être utilisé par toutes ces approches de veille concurrentielle. Tandis qu'elles sont distinctes par leurs cibles et leurs objectifs, ils utilisent le web scraping pour se renseigner de leurs vis-à-vis.

3.4 Positionnement

La mise en place d'un système performant dédié à la veille concurrentielle nécessite des compétences en programmation en tout cas dans une situation où nous rêvons d'un outil de scraping taillé sur mesure. Ce qui ne veut pas dire non plus que les outils de grattage d'informations web prête à l'emploi existant ne sont pas personnalisables. Ainsi, sachant que la gestion stratégique de l'information est devenue l'un des moteurs essentiels de la performance globale des entreprises, elle l'est encore beaucoup plus chez les sociétés spécialisées dans la vente en ligne pour deux raisons :

- Elles travaillent exclusivement avec des données
- La majeure partie de ces éléments sont disponible en ligne

Or, nous savons que ceci passe par la collecte, le stockage, la production de connaissance, la diffusion puis la prise de décision informée dans l'ultime objectif de maintenir dans le bon cap son organisation même dans une zone de turbulence. Cependant, le choix des outils appropriés est crucial pour garantir des résultats précis, une automatisation efficace et une maintenabilité du code. Dans cette étude, nous analyserons scientifiquement les raisons qui feront le choix de Python, de BeautiFulSoup et du scraping basé sur le DOM pour le Web Scraping avancé. Ceci étant, nous aurons comme :

- **Type de Veille concurrentielle** : **la veille concurrentielle active mixte**. En effet, cette astuce parmi les combinaisons de veille étudiée dans la sous partie 3.3 de ce chapitre nous donne le privilège non seulement d'effectuer une surveille proactive sur nos concurrents directs autrement et indirects.
- **Type de scraping : le Scraping basé sur le DOM**. En effet, c'est une approche robuste pour les sites dynamiques. Le scraping basé sur le DOM consiste à utiliser un navigateur (ou un navigateur headless) pour interpréter et exécuter le JavaScript présent sur la page web. Cette astuce est essentielle pour les sites web dynamiques qui utilisent du JavaScript pour afficher et modifier leur contenu. Le scraping basé sur le DOM permet de récupérer des données après que le JavaScript a rendu la page, garantissant une collecte précise des informations visibles à l'utilisateur. En effet, ce type de scraping est la méthode très fiable pour extraire des données à partir de sites web dynamiques. Il

permet de ralentir les problèmes de chargement asynchrone de contenu par JavaScript et assure une collecte précise des données même dans des scénarios complexes [46]. En plus il nous offre plus de flexibilité dans la mesure où nous aurons accès à toutes les autres formes de scraping qui passent nécessairement par le DOM pour agir.

- **Langage de programmation : le Python.** En effet, python est un langage polyvalent pour le Web Scraping. Il est devenu un choix populaire pour le Web Scraping grâce à sa simplicité, sa souplesse et son large écosystème de bibliothèques dédiées au scraping, comme Selenium, BeautiFulSoup, Scrapy, et Requests-HTML. Les avantages de Python incluent sa syntaxe claire et lisible, sa grande communauté de développeurs et ses performances satisfaisantes pour la plupart des tâches de grattage. Ce langage est également multiplateforme, offrant une portabilité pour les projets de scraping [47]. Par conséquent, Python est devenu le langage de prédilection pour le Web Scraping en raison de sa puissance, de son accessibilité et de son vaste écosystème de bibliothèques spécialisées, qui permettent d'automatiser efficacement la collecte de données sur le web.

- **Bibliothèque / librairie d'usage : le BeautiFulSoup.** En effet, BeautifulSoup est une bibliothèque Python conviviale pour l'analyse et le « parsing » de pages web. Elle permet aux développeurs de naviguer et de manipuler facilement le DOM (Document Object Model), ce qui facilite l'extraction d'informations spécifiques d'une page. Sa syntaxe intuitive basée sur des sélecteurs CSS et des expressions XPath offre une flexibilité dans l'identification des éléments ciblés. En fait, BeautifulSoup s'est avéré être une bibliothèque Python extrêmement utile pour le scraping grâce à sa simplicité et à sa facilité d'utilisation. Son approche basée sur le DOM permet aux développeurs d'extraire facilement des données intégrées et non intégrées à partir de pages web complexes [48].

Au regard de ce qui précède, notre étude nous a conduit à recommander Python comme langage principal pour le Web Scraping avancé, en raison de sa polyvalence, de sa communauté active et de son écosystème de bibliothèques spécialisées. Parmi ces bibliothèques, BeautifulSoup s'est avéré être l'outil idéal pour l'analyse et le « parsing » de pages HTML et XML, offrant une simplicité et une facilité d'utilisation appréciées par les développeurs. En ce qui concerne l'analyse des d'entreprises, nous avons opté pour une veille concurrentielle active mixte histoire de nous donner plus de marge de manœuvre. Enfin, le scraping basé sur le DOM est la méthode préconisée pour les sites web dynamiques, permettant une extraction précise des données après le rendu du JavaScript

Chapitre 4 : Etude de cas

4.1. Expression des besoins

Nous aurons besoin pour faire fonctionner ce dispositif un système de recueil très efficace. Ses données doivent être pertinentes, bien structurées et dans un format adapté pour faciliter leur analyse. Ce mécanisme doit fournir des fonctionnalités avancées pour l'extraction, le stockage, l'analyse et la visualisation des données, tout en prenant en compte les contraintes techniques et les réglementations en vigueur.

4.1.1 Cibles finales / Utilisateurs finaux

Les cibles de ce système incluent :

- Les acteurs du commerce en ligne ;
- Les professionnels du marketing et de la veille ;
- Les chercheurs ;
- Les commanditaires de sondages ;
- Les agrégateurs de données et les fournisseurs de services numériques.

4.1.2 Contraintes techniques

Nous pouvons être confrontés aux :

- Respect des politiques et des conditions d'utilisation des sites web ciblés ;
- Respect des réglementations en matière de protection des données et de confidentialité ;
- Contraintes de performances.

4.1.3 Besoins Fonctionnels

Le système doit être capable :

- D'extraire de manière automatiques les données de façon régulière sur différents sites ;
- De Fusionner les données extraites ;
- De traiter les doublons ;
- De s'adapter aux changements de structure ;
- De gérer la pagination pour atteindre le maximum de données ;
- De parcourir toutes les catégories de produits ;
- De traiter du JavaScript, d'extraire les données de sites dynamiques ;
- D'organiser et de représenter les données dans un bon format ;
- D'analyser et de visualiser et d'interpréter les résultats obtenus.

4.1.4 Besoins Technologiques

Pour faire fonctionner ce dispositif, nous aurons besoin de certains outils en conformité avec notre positionnement.

- **Anaconda** : Anaconda est une distribution de Python spécialement conçue pour les sciences des données et l'analyse de données. Elle inclut de nombreuses bibliothèques populaires telles que Pandas, NumPy, Matplotlib, Scikit-learn, etc., ce qui en fait un choix populaire pour les professionnels travaillant avec des données.
- **Pandas :** Pandas est une bibliothèque Python qui offre des structures de données puissantes et faciles à utiliser, telles que les Data Frames et les Séries, pour la manipulation et l'analyse de données. Elle est largement utilisée dans le web scraping pour traiter les données extraites et les préparer pour l'analyse.
- **BeautiFulSou**p : BeautiFulSoup est une bibliothèque Python qui facilite l'extraction des données à partir des pages web en analysant le code HTML et XML. Elle est très utile dans le web scraping pour parcourir et extraire des informations spécifiques à partir des sites web.
- **Power BI** : Power BI est un outil de business intelligence développé par Microsoft. Il permet de créer des rapports interactifs et des tableaux de bord à partir de données provenant de différentes sources, y compris des données extraites via le web scraping.
- **CSV** : CSV (Comma-Separated Values) est un format de fichier couramment utilisé pour stocker des données tabulaires, où les valeurs sont séparées par des virgules. Il est largement utilisé pour échanger des données et est souvent utilisé pour enregistrer les résultats du web scraping.
- **JSON** : JSON (JavaScript Object Notation) est un format de données léger et facile à lire et écrire, basé sur la syntaxe JavaScript. Il est largement utilisé pour représenter et échanger des données intégrées, y compris les résultats du web scraping qui peuvent être facilement traités par les langages de programmation.

4.2 Présentation des sources

Nous appelons ici source les boutiques en ligne. En effet, Le Sénégal compte plus d'une trentaine[15] de sites d'e-commerce. Etant donné que les sites de commerce électronique sont nos principales cibles, nous allons à travers cette liste en choisir trois comme échantillon d'étude : Jumia, Auchan et Expat Dakar.

[15] https://gaynako.com/e-commerce/top-35-sites-vente-ligne-senegal/

4.2.1 Cible 1 : Jumia

Jumia est l'une des principales plateformes de commerce électronique au Sénégal qui a été lancée en 2012. Il offre aux consommateurs un large éventail de produits allant de l'électronique aux vêtements, en passant par les articles ménagers, les produits de beauté et bien d'autres catégories. La plateforme se positionne comme une solution pratique pour les achats en ligne, offrant un catalogue étendu et des options de livraison flexibles. D'une manière descriptive, Jumia a organisé les données sur sa plateforme ainsi :

Tableau 4 : Organisation des données Jumia

Libellé	Description
URL ciblée :	https://www.jumia.sn/
Informations sur les produits :	Image, Nom du produit, Prix compétitif, prix promotion, Évaluations et commentaires clients, Quantité stock.
Avis clients	Notes et commentaires clients
Données sur la livraison et les frais :	Boutique officielle
Catégories :	Une catégorie regroupe tous les articles de même nature. Exemple : Electronique (cette catégorie contient tous les variantes des articles de nature électronique)

4.2.2 Cible 2 : Expat-Dakar

Dans leur rubrique qui sommes-nous[16], le groupe Expat-Dakar se définit comme étant le plus grand site de petites annonces en ligne au Sénégal qui connecte facilement les vendeurs et les acheteurs. Expat-Dakar est donc un site de mise en relation entre propriétaire et consommateur. En détail, voici comment est structurée les données de son site :

[16] *https://www.expat-dakar.com/qui-sommes-nous*

Tableau 5 : Organisation des données Expat-Dakar

Libellé	Détails
URL Principal :	https://www.expat-dakar.com/
Informations sur les annonces :	Image, Titre de l'annonce, Prix, type (venant, occasion, nouveau), Date et heure de publication.
Informations sur les propriétaires :	Nom du propriétaire, Contact (WhatsApp, SMS, Appel)
Catégories :	Une catégorie regroupe toutes les annones de même nature. Exemple : Maison (cette catégorie contient toutes les variantes des annonces de nature maisson)

4.2.3 Cible 3 : Auchan

Auchan se décrit sur son site[17] comme le leader sur le marché de la grande distribution alimentaire. Ceci étant, les données utilisées par ce site sont regroupées ainsi :

Tableau 6 : Organisation des données Auchan

Libellé	Détails
URL Principale :	https://www.auchan.sn/
Informations sur les produits :	Image produit, Nom du produit, Prix compétitif, prix promotion, type (nouveau, occasion)
Catégories :	Une catégorie regroupe tous les produits de même nature. Exemple : Café (cette catégorie contient toutes les variantes des produits de nature café)

5.1.Modélisation des données

A partir de l'analyse effectuée sur la page principale Jumia, nous avons examiné en détail la structure pour comprendre la manière dont les données sont organisées. Cela nous a permis d'exprimer nos besoins comme suit : les catégories, la livraison, les articles (nom, image, prix (normal et promotion) et le stock (boutique).

[17] *https://www.auchan-retail.sn/fr/contenu/valeurs/auchan-senegal-introduction*

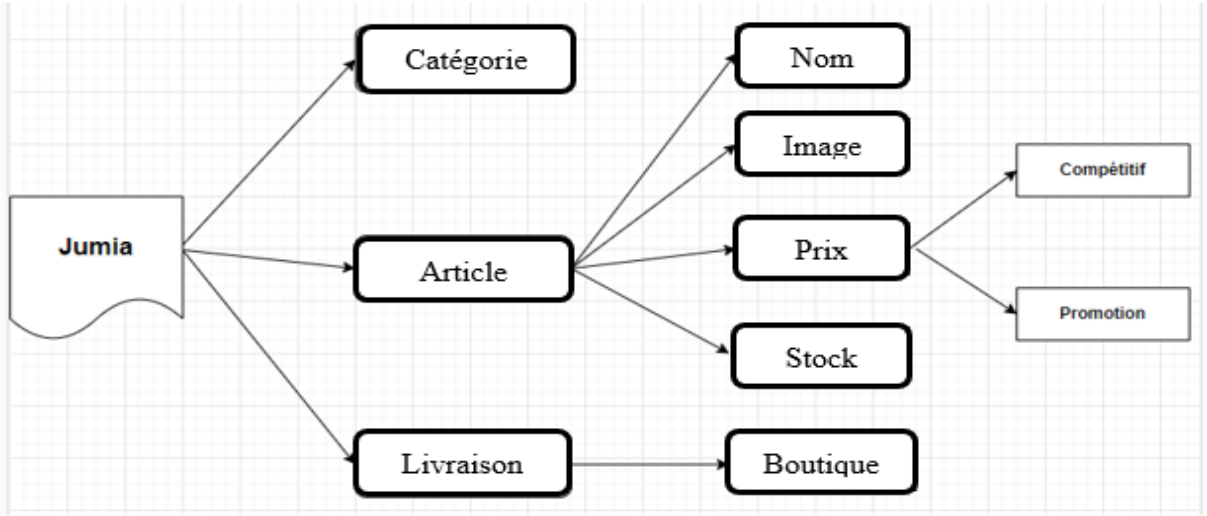


Figure 26 : ***Modélisation CSV des données Jumia***

En ce qui concerne Expat-Dakar nous avons exploré son contenu comme suit : catégorie, propriétaire, annonce (nom, image, type, prix).

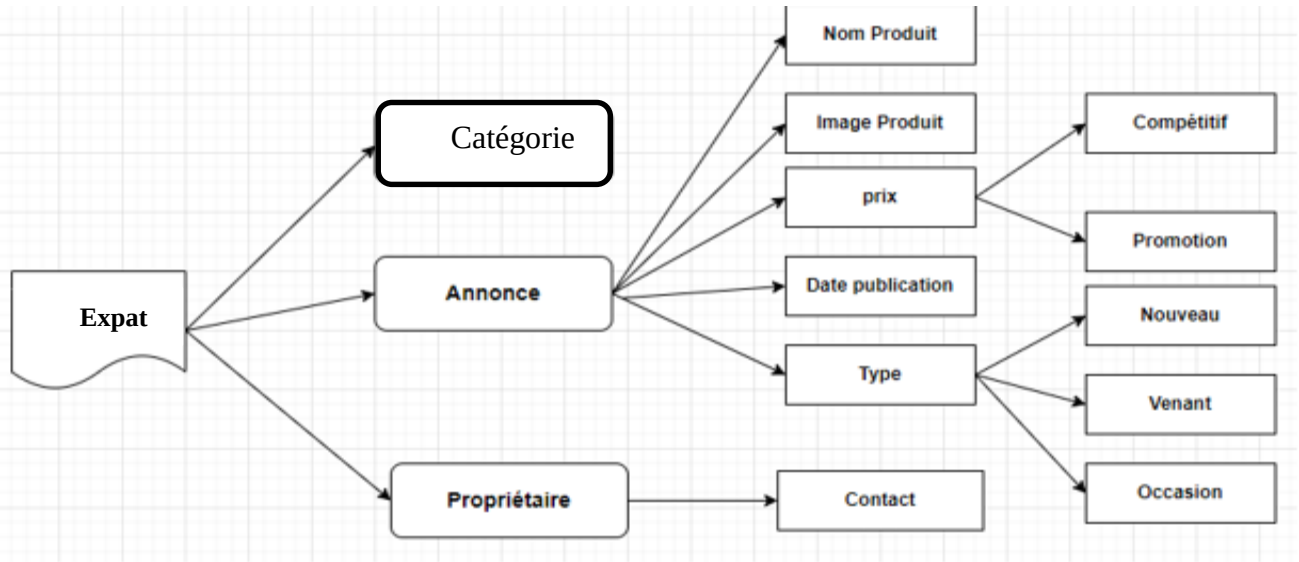


Figure 27 : ***Modélisation CSV des données Expat-Dakar***

Un examen approfondi de la modélisation des données sur le site Auchan s'impose. Comprendre la disposition et la structuration des informations sur cette plateforme est essentiel pour une collecte de données précise et une analyse pertinente. Ainsi nous avons :

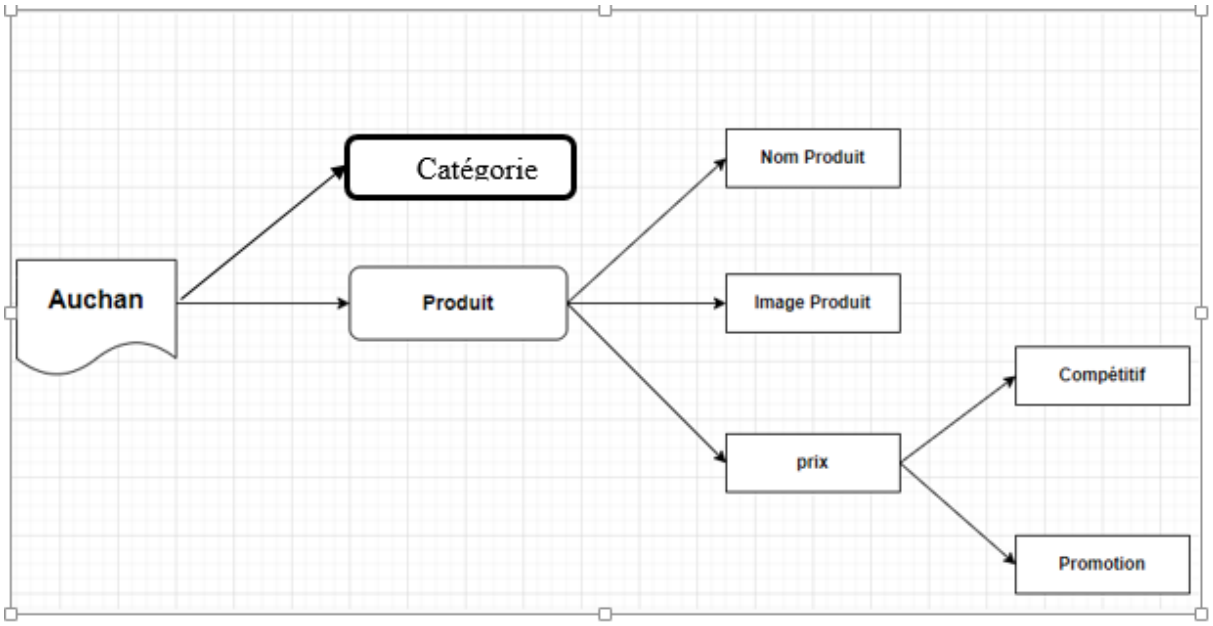


Figure 28 : ***Modélisation CSV des données Auchan***

Explication :

Dans cette partie de la modélisation des données, nous avons représenté les éléments ciblées chez les concurrents en fonction de nos besoins en se basant sur la manière dont elles sont organisées.

En fait, que ça soit Jumia, Expat ou Auchan, nous avons besoin de l'URL principale définie dans la partie présentation des sources (4.2) de ce chapitre ainsi que de la catégorie (qui regroupe l'ensemble des produits d'une même nature) des produits visés. Et dans chaque catégorie, nous allons collecter le nom, prix, propriétaire, contact, avis clients, lieu de livraison, prix normal, prix promotion et promotion.

Toutefois, si une propriété existe dans une source donnée plutôt que dans une autre, dans ce cas le programme va extraire les données disponibles.

Par exemple :

Url principale Jumia : https://www.jumia.sn/

Catégorie ciblée : électronique

Dans ce cas de figure, notre script va se pointer sur https://www.jumia.sn/electronique puis boucler sur l'ensemble des produits disponibles dans la catégorie électronique ou il va chercher toutes les données indiquées par notre modèle. A partir de là également, il peut parcourir toutes les catégories disponibles.

5.1.1. Modélisation des données des sources JSON

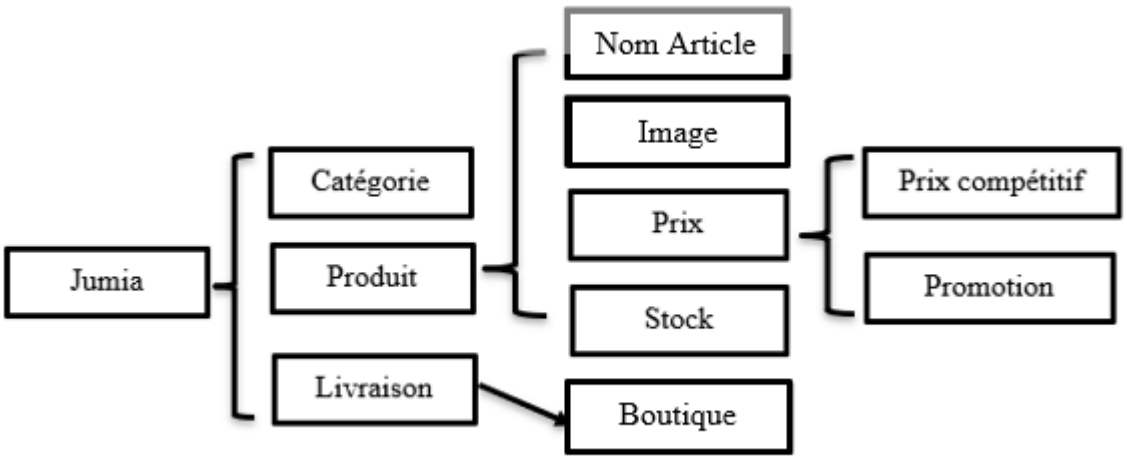


Figure 29 : ***Modélisation JSON des données Jumia***

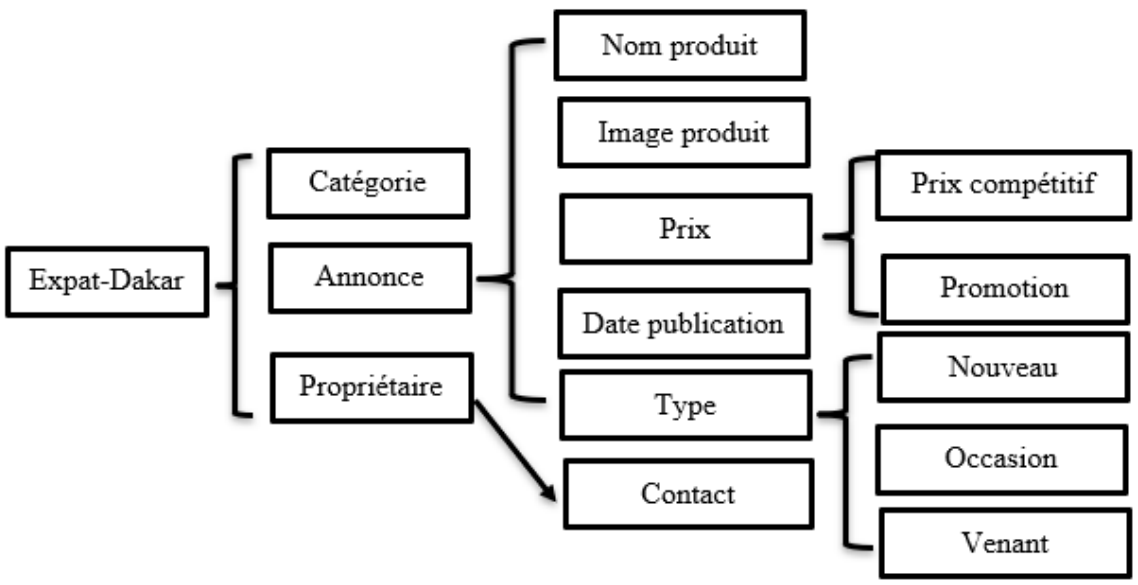


Figure 30 : ***Modélisation JSON des données Expat-Dakar***

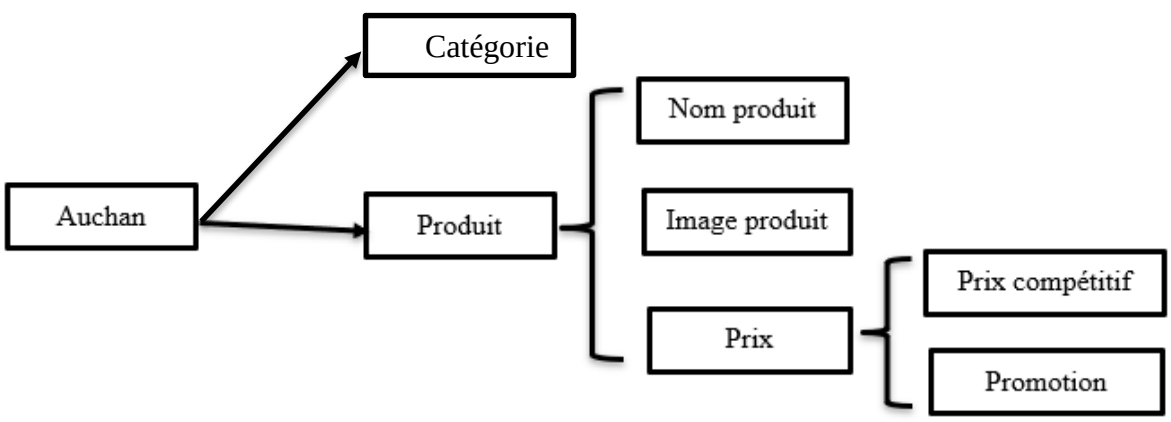


Figure 31 : ***Modélisation JSON des données Auchan***

NB :

Les explications de la modélisation CSV des données sources sont aussi valables pour cette partie. Sauf qu'ici la sortie des données sont sous format JSON.

5.1.2. Tableau de conformité

Pour à la fin disposer d'un fichier unifié où les données Jumia, Auchan et Expat-Dakar seront fusionnées pour une analyse croisée, nous allons uniformisés les renseignements extraites de nos différentes sources sur le tableau suivant :

Tableau 7 : Conformité des données de nos différentes sources

	Jumia	Auchan	Expat-Dakar	Données finales
Produits	Article	Produit	Annonce	Produit
Description	Nom Article	Nom produit	Titre annonce	Nom produit
Image	Image	Image	Image	Image
Prix	Prix	Prix	Prix	Prix
Promotion	Flash	Promotion	--	Promotion
Avis clients	Notes et commentaires	--	--	Avis clients
Livraison	Livraison	--	--	Livraison

5.1.3. Modélisation des données et du corpus final

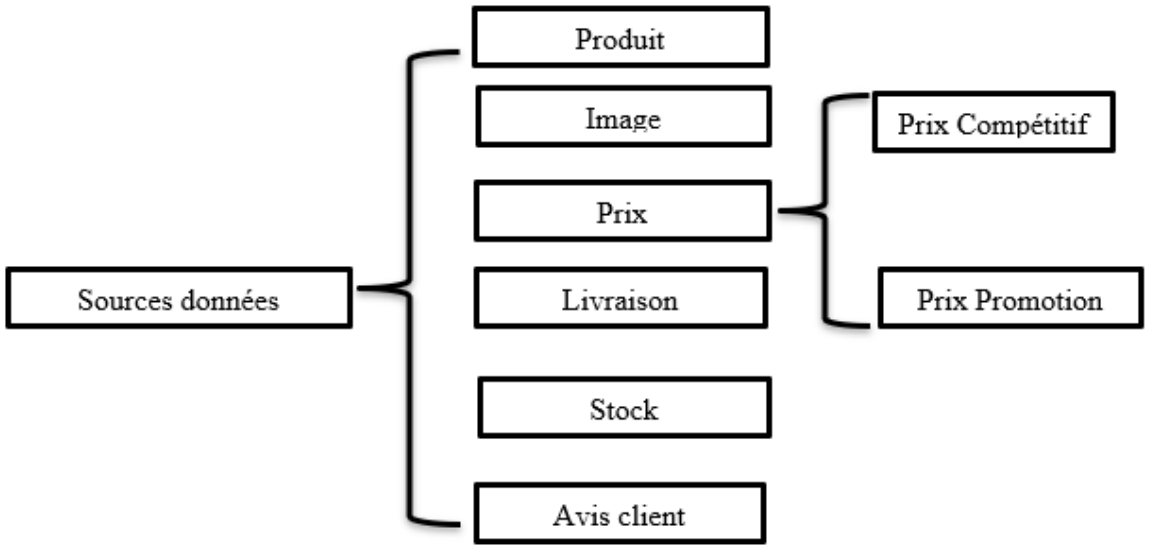


Figure 32 : ***Modélisation Finale des données de nos différentes sources***

5.2. Architecture générale

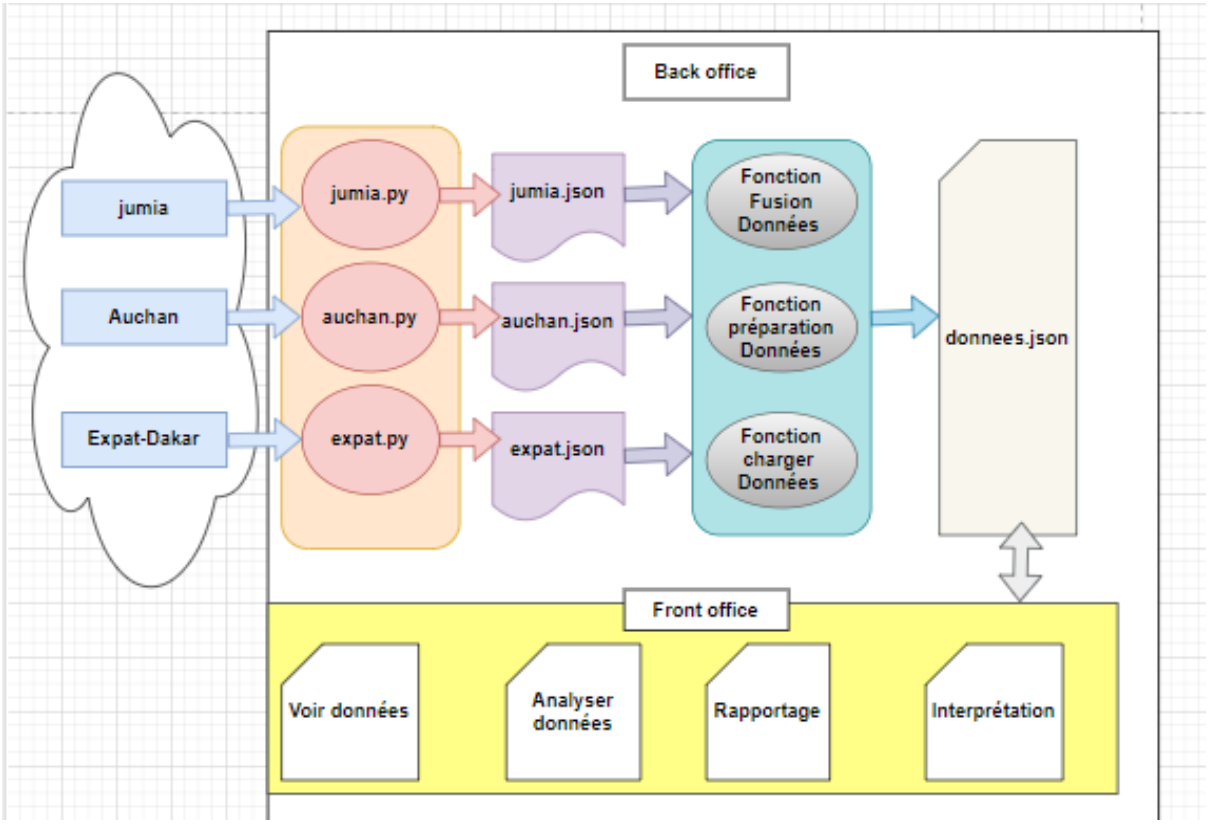


Figure 33 : ***Modélisation Architecture générale du système***

En arrière-plan, le système que nous proposons permet d'automatiser les processus de recueil de données à très grande échelle d'une manière pratique. Son mode de fonctionnement se dessine comme suit :

Le programme demande une URL, parcourt page par page, catégorie par catégorie et indexe toutes les données qui lui sont indiquées. Ensuite il poursuit son travail en les collectant avant de les passer au script de fusion qui se charge de synthétiser les données, en évitant les doublons. Pour finir, il charge à chaque nouveau traitement les données mise à jour dans la base de données. Concernant la partie front office, nous allons utiliser ces données collectées pour un traitement approprié, visionner et interpréter les résultats.

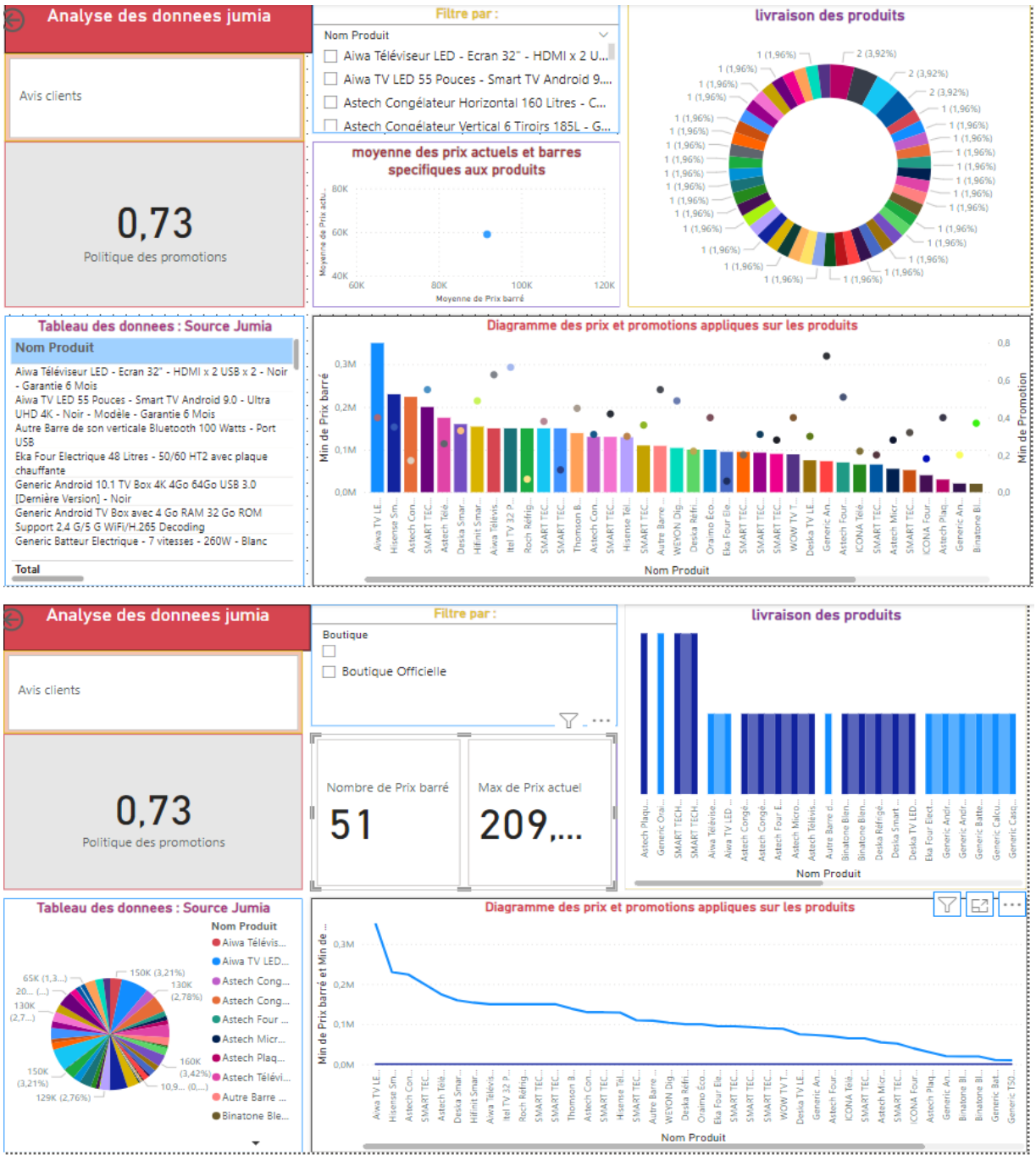


Figure 34 : ***Tableau de bord Analyse des données Jumia***

Sur Jumia, les paramètres d'analyse suivants sont utilisés :

• **Prix Moyen :** pour ajuster nos coûts ;

• **Lieu de Livraison des commandes** : chez jumia c'est la boutique officielle, nous pouvons choisir de livrer à domicile pour mettre à l'aise les clients ;

• **Taux par produit sur le total** : qui donne 1.96 % par produit sur 100 % ;

• **Taux maximum des promotions** : qui est égal à 73 % pour savoir jusqu'où plafonner nos offres spéciales ;

• **Les avis clients** : le résultat est positif car sur tous les produits, nous constatons une note

de 3.9 sur 5.

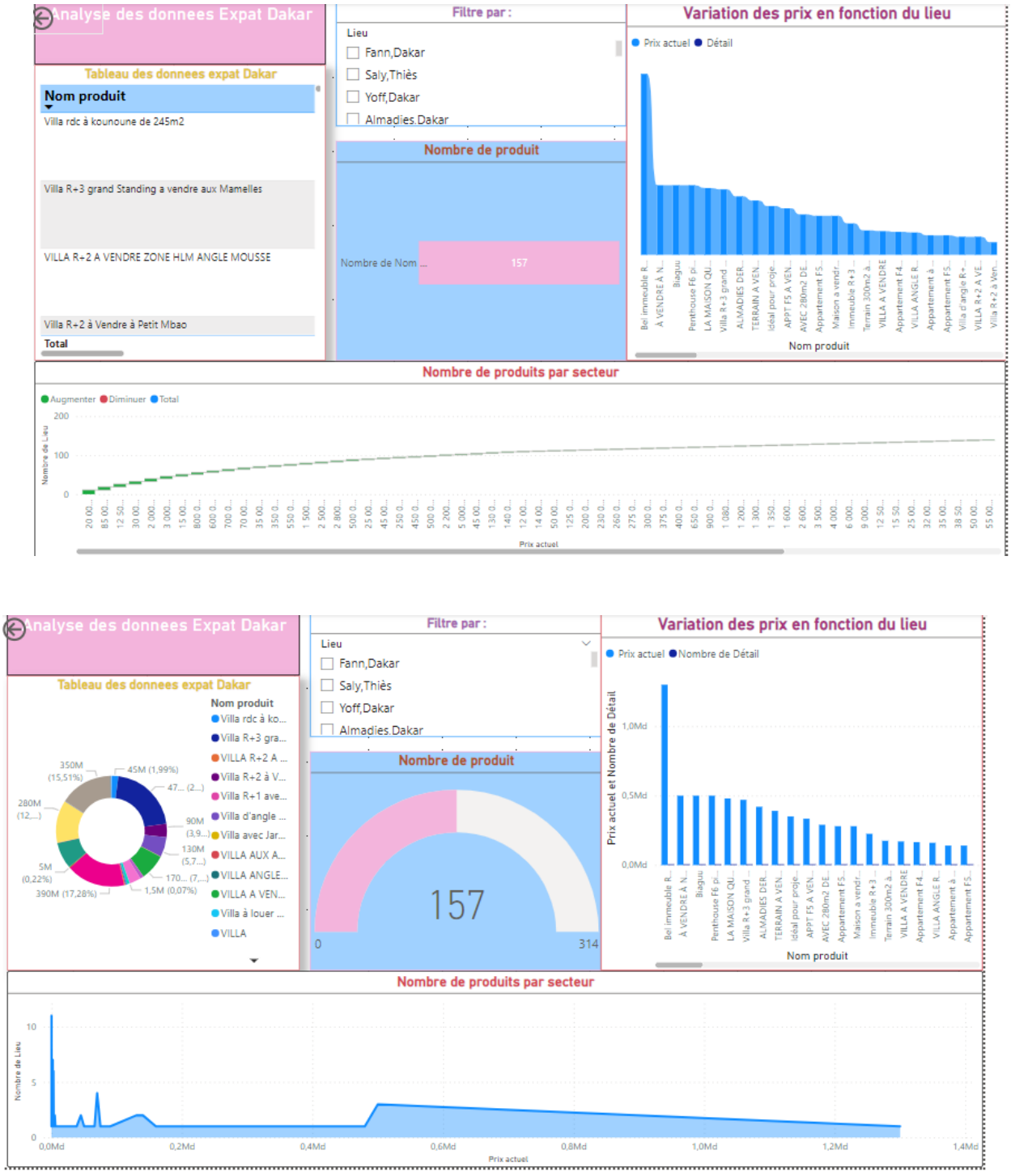


Figure 35 : ***Tableau de bord Analyse des données Expat Dakar***

Chez Expat, nous avons utilisé les paramètres d'analyse suivants :

- **Nombre de produits par lieu** : cela nous permet de savoir dans quels secteurs les activités sont plus concentrées ;
- **Prix Maximum par secteur** : Les offres sont plus chères dans le centre-ville de Dakar ;

- **Prix Minimum par secteur** : Les offres sont moins chères dans la banlieue Dakaroise ;
- **Les variétés de produits** : à la recherche de nouveaux produits ;
- **Nombre de produits** : plus de 300 et en moyenne 157 ;

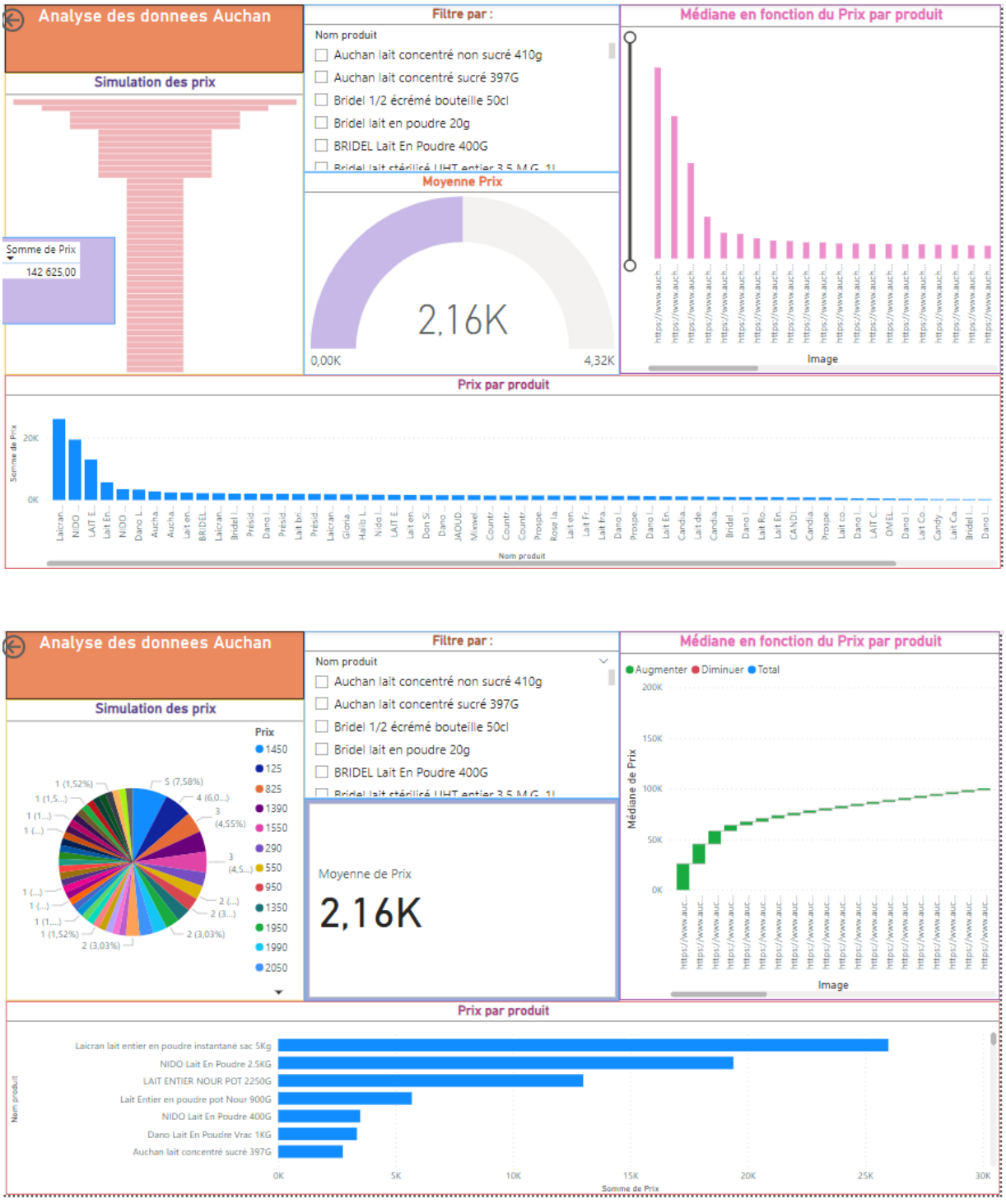


Figure 36 : ***Tableau de bord Analyse des données Auchan***

Chez Auchn, nous avons utilisés les paramètres d'analyse suivant :

- **Médian prix** : nous en déduisons que 50 % des produits coûtent moins de 2016 francs et les 50 % restants vont au-delà de ce prix moyen ;

- **La moyenne des prix pratiqués :** 2016 francs en moyenne pour comprendre leur politique des prix ;
- **Nombre de produits :** une large gamme de produits, plus de 400 pages par catégorie ;

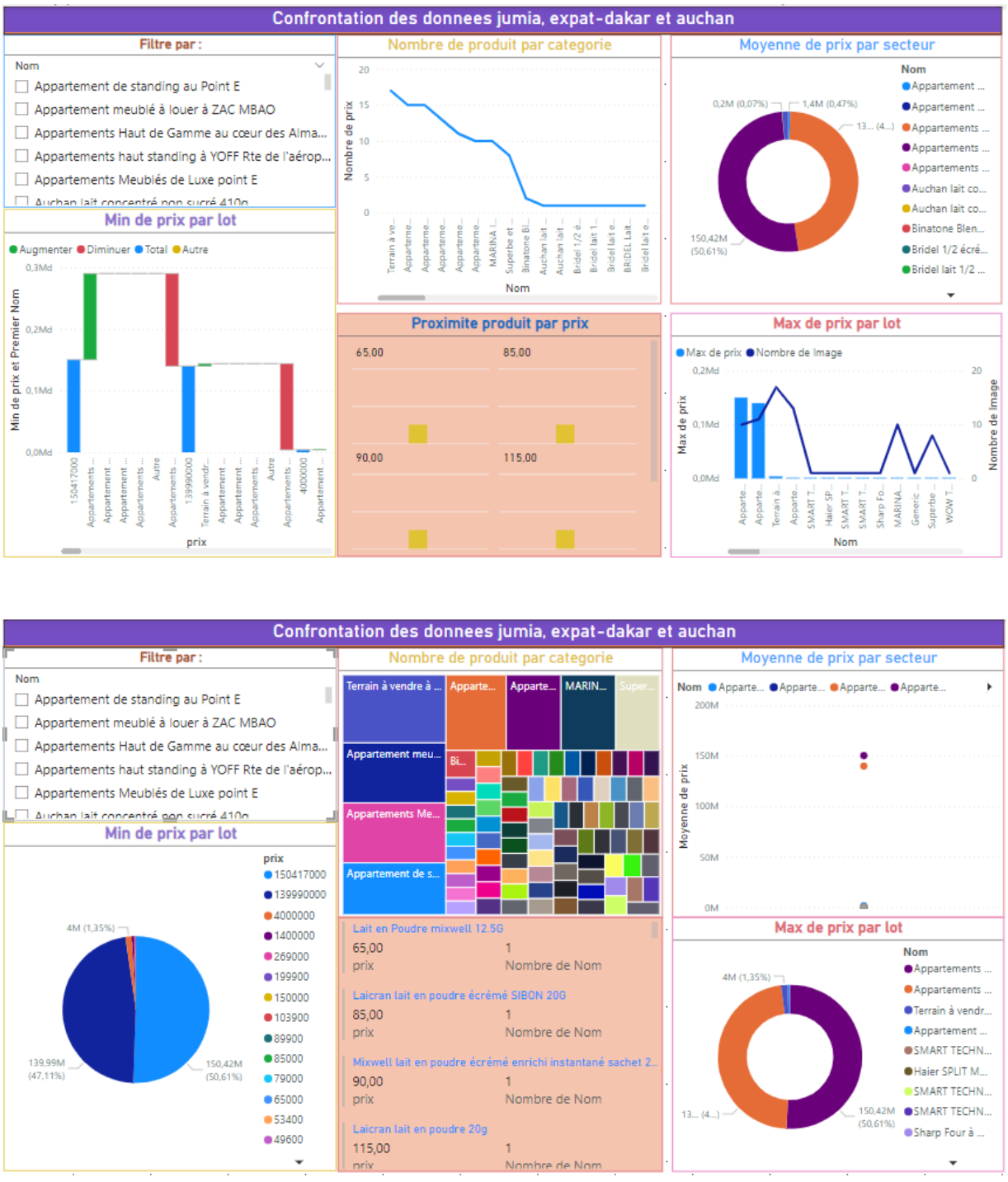


Figure 37 : ***Tableau de bord Analyse des données Fusionnées***

NB :

Ici nous avons appliqué les mêmes paramètres d'analyse utilisés sur les données spécifiques à chaque source (Jumia, Auchan et Expat) sur les données croisées à notre disposition pour avoir un aperçu global sur l'environnement concurrentiel.

5.3. Interprétation des résultats obtenus

Depuis la phase de la collecte d'informations sur nos concurrents en passant par l'analyse des résultats jusqu'à leurs visualisations, nous avions qu'un seul objectif : Produire de la connaissance utile à la maîtrise de l'environnement de compétitions et d'en servir pour se démarquer par rapport aux vis-à-vis avec une meilleure part de marché. Ainsi, nous en déduisons :

Analyse des Prix :

- Sur Jumia, nous avons observé une tendance intéressante qui acte le fait que les produits ont tendance à être moins chers (pour les mêmes produits) par rapport aux autres sites. Cela pourrait être dû à une stratégie de prix compétitive visant à attirer et fidéliser les clients dans cette catégorie populaire. Toutefois, le calcul de la moyenne des prix effectué sur la catégorie électronique en atteste.
- Auchan, en revanche, se démarque par des prix exceptionnellement bas pour les produits alimentaires. Cela peut indiquer une stratégie axée sur l'offre de produits alimentaires à un prix attractif pour fidéliser les clients. D'autant plus qu'il est spécialisé dans la distribution des denrées de premier nécessité, leur stratégie basée sur le low-cost attire plus d'un.
- Expat-Dakar a tendance à proposer des produits de niche à des prix plus élevés. Cela suggère que le site cible un public spécifique prêt à payer un peu plus cher pour des produits spécifiques, il peut s'agir des consommateurs qui sont dans le besoin.

Analyse des Promotions :

- A ce niveau, Jumia par rapport à Auchan et Expat-Dakar est de loin plus présent. D'ailleurs, parmi tous les produits scaper, nous avons constaté que Jumia a appliqué une promotion et cette dernière peut aller jusqu'à son niveau maximum qui est de 70% comme l'indique notre tableau de bord. Cette politique consistant à appliquer des remises sur tous les produits attire aussi beaucoup de clients.

Analyse des Images :

- Les images de produits de haute qualité sur Jumia peuvent jouer un rôle clé dans la conversion des achats. Les consommateurs ont tendance à être plus enclins à acheter un produit lorsqu'ils peuvent le voir clairement.
- Auchan a certes des images de produits avec moins de qualité que Jumia. Il pourrait améliorer la qualité de ses images pour rendre ces marchandises plus attrayantes et accroître le taux de conversion.
- Expat-Dakar qu'en a lui doit impérativement améliorer la qualité des images, surtout si elle veut attirer un public international (Tourisme) qui s'attend à une expérience de

magasinage en ligne de haute qualité.

Analyse des Commentaires et des Évaluations :

Les commentaires et les évaluations des clients jouent un rôle crucial dans la prise de décision des consommateurs. Sur les trois sites, il est essentiel de maintenir un bon niveau de service client pour obtenir des critiques positives.

- Jumia bénéficie de sa grande base de clients pour générer plus de commentaires, mais cela signifie également qu'ils doivent gérer davantage de commentaires négatifs. Cela nécessite une gestion proactive de la réputation en ligne. D'après nos analyses, sur une note de 5, Jumia parvient à obtenir une moyenne de 3.9 ce qui donne une idée sur comment les clients trouvent leurs produits
- Auchan a l'avantage de recevoir des commentaires positifs pour ses prix compétitifs. Cela peut être un point fort à mettre en avant. Cela informe aussi, de la pertinence de ce dernier dans sa politique des prix.
- Expat-Dakar peut capitaliser sur les commentaires positifs aussi par rapport au service rendu que les consommateurs trouvent très pratique.

Analyse des Tendances :

- L'importance de suivre les tendances saisonnières et les fluctuations des ventes permet d'ajuster les stocks et le marketing en conséquence. Par exemple, pendant les périodes de fête, Jumia est plus actif en publicité et en promotion.
- Auchan planifie des campagnes de marketing relatif aux produits alimentaires en fonction de leur saison, en mettant en avant des offres spéciales.
- Expat-Dakar concentre ses efforts sur les festivals locaux en proposant des produits spéciaux ou des réductions pour attirer davantage de clients lors de ces événements.

Analyse Concurrentielle :

- Pour maintenir un avantage concurrentiel, Jumia devrait continuer à élargir sa gamme de produits tout en maintenant des prix compétitifs. Des offres spéciales régulières peuvent également aider à attirer davantage de clients.
- Auchan peut capitaliser sur sa réputation de prix abordables en mettant l'accent sur la qualité des produits alimentaires et en élargissant sa gamme de produits.
- Expat-Dakar peut renforcer sa position en tant que fournisseur de produits uniques en ajoutant de nouvelles catégories de produits de niche et en offrant un service client exceptionnel.

Remarque :

Ces connaissances obtenues sur nos challengers grâce au web scraping et la veille concurrentielle stimulent que l'environnement dans lequel nous évoluons est dense vue les géants qui le composent. Toutefois, des recommandations très pointues en découlent pour nous

permettre de s'imposer comme leader. Pour se faire :

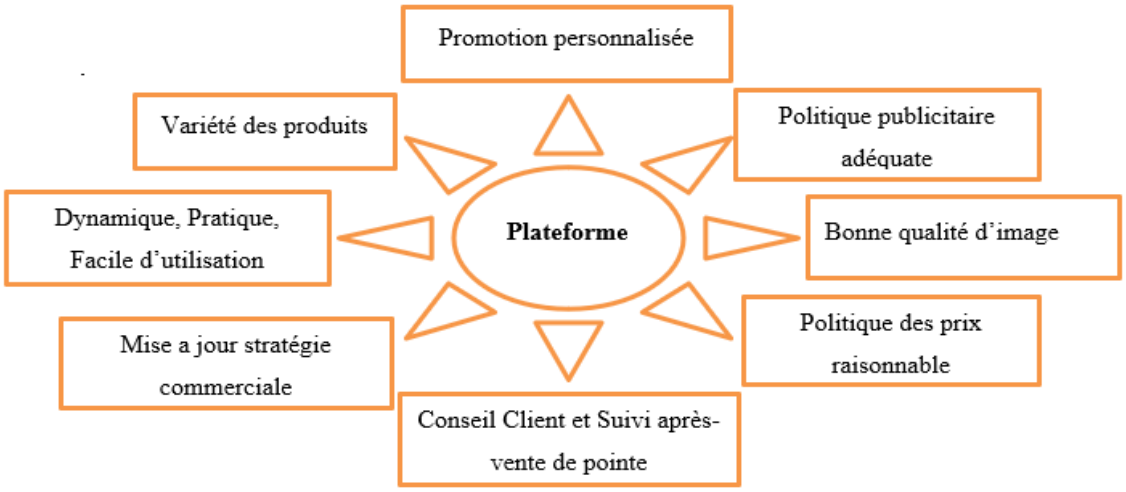


Figure 38 : ***Orientation commerciale***

Commentaire :

Les analyses effectuées sur les prix, les images, les concurrents, les promotions et les avis clients etc. nous ont permis d'en déduire :

- **La personnalisation des promotions** : fait référence à l'adaptation des promotions et des offres commerciales pour répondre de manière spécifique aux besoins, aux préférences et au comportement d'un individu ou d'un groupe de clients
- **Une politique publicitaire adéquate** : c'est une stratégie bien planifiée et exécutée pour promouvoir un produit, service ou entreprise auprès du public cible. Cette approche implique la sélection judicieuse des canaux de publicité, la création de messages pertinents et attrayants, et la gestion efficace des budgets publicitaires. Une politique publicitaire réussie doit prendre en compte les caractéristiques du marché, les comportements des consommateurs et les objectifs commerciaux
- **Une politique des prix raisonnables :** consiste à fixer des prix qui reflètent la valeur perçue par les clients tout en assurant la rentabilité de l'entreprise. Une politique des prix raisonnables peut impliquer des tactiques telles que la tarification basée sur la valeur, la tarification concurrentielle ou la tarification par coût plus marge bénéficiaire. Il est essentiel de trouver un équilibre entre la maximisation des profits et la satisfaction du client.
- **Un conseil client et suivi après-vente :** Le conseil client implique de fournir des informations approfondies et utiles aux clients afin de les guider dans leurs décisions d'achat. Cela peut inclure des recommandations de produits, des comparaisons, des

explications techniques et des conseils personnalisés en fonction des besoins spécifiques du client.

- **Une variété de produits :** C'est est une stratégie clé dans le domaine commercial. Cela implique de proposer une gamme diversifiée de produits qui répondent à différents besoins et préférences des clients.

Conclusion et perspectives

Les entreprises d'aujourd'hui concourent à développer leurs capacités à se renseigner efficacement sur des concurrents, des clients, des fournisseurs et d'autres facteurs du marché. Ce processus connu sous l'appellation de veille concurrentielle, veille stratégique ou encore analyse d'entreprise passe par le web scraping pour constituer sa matière première [49]. C'est-à-dire que l'acquisition de données marque le point de départ de la pratique de surveillance régulière des semblables. Ces informations peuvent être utilisées pour ajuster leur propre stratégie marketing, identifier les opportunités et les menaces, prendre des décisions éclairées et rester compétitives sur le marché [50]. Le moissonnage offre un moyen efficace et rapide de collecte de données à grande échelle, permettant ainsi aux entreprises d'obtenir une vision complète et détaillée de leurs concurrents sans avoir à effectuer des recherches manuelles fastidieuses. Toutefois, il peut être confronté à des limitations techniques, telles que la gestion de la complexité des sites web, les protections anti-scraping mises en place par certains sites, et les changements fréquents dans la structure des pages web. Malgré ces défis, il reste un outil puissant et précieux pour les organisations engagées dans la veille concurrentielle, leur permettant de rester informées et réactives dans un environnement commercial dynamique et compétitif.

En fait, la possibilité d'anticiper les mouvements du marché, qu'il s'agisse d'opportunités émergentes ou de menaces potentielles, donne aux entreprises un avantage significatif [51]. Au cœur de cette démarche se trouve la compréhension approfondie des clients, rendue possible par les données sur leurs comportements, leurs préférences et leurs habitudes d'achat. Cette connaissance intime du public cible permet une personnalisation plus précise des produits, services et campagnes marketing, renforçant ainsi la relation client. De plus, la surveillance continue des activités des concurrents, y compris les variations de prix, les stratégies promotionnelles et les lancements de produits, est cruciale pour rester compétitif [52]. Les données en temps réel alimentent la réactivité des entreprises aux changements du marché, favorisant une adaptation rapide des stratégies. En optimisant les opérations internes et en stimulant l'innovation basée sur les données, la veille commerciale propulse les entreprises vers une croissance durable et une position stratégique dans un paysage commercial dynamique.

Par ailleurs, la combinaison du web scraping avec l'intelligence artificielle représente une avancée stratégique significative. Cette synergie offre la possibilité de créer une banque d'informations sophistiquée, capable de décrypter les signaux complexes et de guider les entreprises à maîtriser des environnements commerciaux denses, renforçant ainsi leur positionnement sur le marché [53]. Elle représente une avancée significative dans le domaine du web scraping, intégrant des caractéristiques intelligentes et évolutives. Cette approche se

manifeste par la capacité du système à apprendre en temps réel. En d'autres termes, si un site web modifie sa structure, le système peut ajuster ses règles de scraping de manière autonome, réduisant ainsi la dépendance à des mises à jour manuelles constantes.

Enfin, dans le contexte de la veille concurrentielle, l'intégration du web scraping et du web crawling crée un système de collecte exceptionnellement efficace [54]. Cette approche permet de suivre en temps réel les mises à jour des sites concurrents, facilitant ainsi un processus d'extraction d'informations instantanées et adaptatives. Autrement dit, lorsque le bot d'indexation capte un événement de mise à jour des données du site ciblé, cela pourrait déclencher le programme de collecte à saisir ces informations et de cette manière le processus sera entièrement automatisé.

Référence

1. ZINAOUI, T. (2020). Rôle de la veille concurrentielle dans la création de la richesse dans l'entreprise. *Revue Marocaine de la Prospective en Sciences de Gestion*, (3).
2. Isaac, F., Hamon, T., Fouqueré, C., Bouchard, L., & Emirkanian, L. (2001). Extraction informatique de données sur le web. *Revue DistanceS*, *5*(2), 195-210.
3. https://welovedevs.com/fr/articles/web-scraping-definition/ visité le 20/05/23
4. https://www.sales-hacking.com/post/web-scraping visité le 20/05/23
5. Clements, D. W. G. (1968). Flux d'informations, systèmes de contrôle et gestion d'entreprise. *Management International Review*, 27-29.
6. https://ledigitalizeur.fr/growth-hacking/web-scraping/ consulté le 30/05/23
7. https://kinsta.com/fr/base-de-connaissances/web-scraping/ consulté le 01/06/23
8. Calof, J., & Skinner, W. (1998). La veille concurrentielle: le meilleur des mondes pour les gestionnaires. *Optimum, La revue de gestion du secteur public*, *28*(2), P43.
9. Calof, J., & Skinner, W. (1998). La veille concurrentielle: le meilleur des mondes pour les gestionnaires. *Optimum, La revue de gestion du secteur public*, *28*(2), P47.
10. https://www.manager-go.com/intelligence-economique/veille-concurrentielle.html consulté le 10/06/23
11. Rallet*, A. (2001). Commerce électronique et localisation urbaine des activités commerciales. *Revue économique*, *52*(7), 267-288.
12. PELET, J. É. (2018). e-COMMERCE.
13. Jain, VIPIN, Malviya, BINDOO et Arya, SATYENDRA (2021). Un aperçu du commerce électronique (e-Commerce). *Journal of Contemporary Issues in Business and Government* , *27* (3), 665-670.
14. Fensel, D., Ding, Y., Omelayenko, B., Schulten, E., Botquin, G., Brown, M. et Flett, A. (2001). Intégration des données produits dans le e-commerce B2B. *Systèmes intelligents IEEE* , *16* (4), 54-59.
15. Park, JH et Kim, IH (2014). Effet supplémentaire du probiotique Bacillus subtilis B2A sur la productivité, le poids des organes, la microflore intestinale de Salmonella et la qualité de la viande de poitrine des poussins de chair en croissance. *Science avicole* , *93* (8), 2054-2059.
16. DAHOU, A. (2022). Economie numérique et veille stratégique.

17. Noyé, D. (2004). *Pour fidéliser les clients*. INSEP éditions.

18. https://fourweekmba.com/fr/veille-concurrentielle/ consulté le 3/06/23

19. Isaac, H., & Volle, P. (2014). *E-commerce: de la stratégie à la mise en œuvre opérationnelle*. Pearson Education France.

20. Mkadmi, Abderrazak et Imad Saleh. Bibliothèque numérique et recherche d'informations. 2008, p. 24, 26, 27.

21. Mkadmi, Abderrazak et Imad Saleh. Bibliothèque numérique et recherche d'informations. 2008, p. 24, 26, 27.

22. Kembellec, G. (2016). Le web de données en contexte bibliothécaire. *I2D–Information données & documents*, *53*(2), 30-31.

23. Espinasse, B., Fournier, S., & de Freitas, F. L. G. (2007). AGATHE: une architecture générique à base d'agents et d'ontologies pour la collecte d'information sur domaines restreints du Web. In *CORIA* (pp. 367-384).

24. Faheem, M., & Senellart, P. (2014). Crawl intelligent et adaptatif d'applications Web pour l'archivage du Web. *ISI , 19* (4).

25. Karthikeyan, T., Sekaran, K., Ranjith, D. et Balajee, JM (2019). Extraction de contenu personnalisé et classification de texte à l'aide de techniques de grattage Web efficaces. *Journal international des portails Web (IJWP) , 11* (2), 41-52.

26. Stroppa, N. (2005). *Définitions et caractérisations de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles* (Doctoral dissertation, Télécom ParisTech).

27. https://www.redhat.com/fr/topics/digital-transformation/what-is-machine-learning consulté le 28/08/23

28. https://www.ibm.com/fr-fr/topics/machine-learning consulté le 06/09/23

29. Ramesh, TR, Lilhore, Royaume-Uni, Poongodi, M., Simaiya, S., Kaur, A. et Hamdi, M. (2022). Analyse prédictive des maladies cardiaques avec des approches d'apprentissage automatique. *Journal malaisien d'informatique* , 132-148.

30. Miclet, C. A. L., & CornuØjols, A. (2010). Apprentissage Artificiel. *Concepts et algorithms, 2e edition, Eyrolles*.

31. Scarnò, M., & Seid, Y. (2018). Use of artificial intelligence and Web scraping methods to retrieve information from the World Wide Web. *Journal of Engineering Research and Application*, *8*.

32. Van der Aalst, W. M., Bichler, M., & Heinzl, A. (2018). Robotic process automation. *Business & information systems engineering*, *60*, 269-272.

33. Flores, L. (2008). Web 2.0 et Etudes de marché: vers une nouvelle génération d'études de marché?. *Revue française du marketing*, (220), 7-16.

34. LABONNE, M., OLIVEREAU, A., & ZEGHLACHE, D. (2018). Automatisation du processus d'entraînement d'un ensemble d'algorithmes de machine learning optimisés pour la détection d'intrusion.

35. Suganya, R., Krupasree, R. S., Gokulraj, S., & Abinesh, B. (2022). Product Review Analysis by Web Scraping Using NLP. In *Smart Data Intelligence: Proceedings of ICSMDI 2022* (pp. 427-436). Singapore: Springer Nature Singapore.

36. https://www.cscience.ca/2020/12/03/quelques-mots-sur-lapprentissage-automatique/ consulté le 25/08/23

37. Scarnò, M., & Seid, Y. (2018). Use of artificial intelligence and Web scraping methods to retrieve information from the World Wide Web. *Journal of Engineering Research and Application*, *8*.

38. Dilts, R. (2004). Modéliser avec la PNL. *P.: InterEditions Dunod*.

39. https://www.octoparse.fr/blog/le-meilleur-langage-pour-web-scraping consulté le 10/09/23

40. https://www.octoparse.fr/blog/top-30-des-logiciels-de-web-scraping-gratuits-en-2021 visité le 26/09/23

41. https://www.sales-hacking.com/post/16-meilleurs-outils-de-scraping-pour-extraire-des-donnees-du-web consulté le 26/09/23

42. CONCURRENTIELLE, E. D. L. V., STRATEGIQUES, E. C., & ANGLO-FRANÇAISE, U. E. C. (2008). Jamie SMTH. *Benchmark européen de pratiques en intelligence économique*, 181.

43. Sage, E. (1999). *La concurrence par comparaison (Yardstick Competition): théories et application. Une proposition pour les secteurs de l'eau en France* (Doctoral dissertation, Paris 9).

44. BOZEC, S. Présentation d'une méthodologie de recherche combinée: téléservices et avantage concurrentiel. *ACTES DES COMMUNICATIONS*, 15.

45. DELECROIX, B., GUILLEMIN-LANNE, S., & SIX, A. (2004). Veille concurrentielle et veille stratégique: deux applications d'extraction d'information. *Veille Scientifique et Stratégique, Toulouse*, 117-128.

46. Rahmatullah, A., & Gunawan, R. (2020). Web scraping with html dom method for data collection of scientific articles from google scholar. *IJIS*, *2*(2).

47. Lawson, R. (2015). *Web scraping avec Python* . Packt Publishing Ltd.

48. Hajba, GL et Hajba, GL (2018). Utiliser une belle soupe. *Scraping de site Web avec Python : Utilisation de BeautifulSoup et Scrapy* , 41-96.

49. https://research.aimultiple.com/ai-web-scraping/ consulté 11/09/23

50. Lesca, H., & Lesca, E. (1994). Veille stratégique. *La méthode LE SCAning*.

51. Drevon, E., Maurel, D., & Dufour, C. (2018). Veille stratégique et prise de décision: une revue de la littérature. *Documentation et bibliothèques*, *64*(1), 28-34.

52. Nivol, W. (1993). *Systèmes de surveillance systématique pour le management stratégique de l'entreprise: le traitement de l'information brevet: de l'information documentaire à l'informationn stratégique* (Doctoral dissertation, Aix-Marseille 3).

53. https://ai.plainenglish.io/pioneering-the-future-of-web-scraping-with-intelligent-ai-agents-unleash-the-power-of-autogen-222aa73daad6 consulté 14/09/23

54. https://www.scrapingbee.com/blog/scraping-vs-crawling/ consulté 15/09/23

Table des matières

ANNEXE 1 : Scraper du site Jumia

jumia.py M

jumia.py > toutesLesUrls

```python
import json
import requests
from datetime import datetime
from urllib.parse import urlparse
from bs4 import BeautifulSoup
from beautifultable import BeautifulTable

# Charger les données JSON depuis un fichier
def charger_json(fichier_json="donneesJumia.json"):
    try:
        with open(fichier_json, "r", encoding="utf-8") as lire_fichier:
            base_de_donnees = json.load(lire_fichier)
            return base_de_donnees
    except:
        base_de_donnees = dict()
        return base_de_donnees

# Enregistrer les données dans un fichier JSON
def enregistrer_donnees_scrappees_en_json(donnees, fichier_json="donneesJumia.json"):
    with open(fichier_json, "w", encoding="utf-8") as fichier_obj:
        json.dump(donnees, fichier_obj, indent=4, ensure_ascii=False)

# Initialisation des données scrapées existantes
def initialisation_donnees_scrappees_existantes(json_db):
    donnees_scrappees = json_db.get("donnees_scrappees")
    if donnees_scrappees is None:
        json_db['donnees_scrappees'] = dict()
    return None
```

```python
# Obtenir l'heure actuelle pour le scraping
def heure_scraping():
    maintenant = datetime.now()
    chaine_dt = maintenant.strftime("%d/%m/%Y %H:%M:%S")
    return chaine_dt

# Effectuer une requête HTTP GET pour obtenir le contenu HTML
def requete_url(website_url):
    requetes_data = requests.get(website_url)
    if requetes_data.status_code == 200:
        soup = BeautifulSoup(requetes_data.text, 'html.parser')
        return soup
    return None

# Générer les URLs pour toutes les pages d'une catégorie
def toutesLesUrls(cat_url):
    nosPages = []
    for pages in range(1, 10):
        url_page = f"https://www.jumia.sn{cat_url}?page={pages}#catalog-listing"
        nosPages.append(url_page)
    return nosPages

# Traiter les données extraites avec BeautifulSoup
def traitement_donnees_beautiful_soup(soup):
    liste_produits = soup.find_all('div', class_='col16 -pvs')
    liste_de_donnees = []
    for produit in liste_produits:
        titre = produit.find('div', class_='name')
        image_element = produit.find('img')
        prix_element = produit.find('div', class_='prc')
        old_price_element = produit.find('div', {'data-oprc': True})
        remise_element = produit.find('div', class_='bdg _dsct')

        donnees = {
            "Nom": titre.text.strip() if titre else "",
            "Image": image_element.get('data-src', "") if image_element else "",
            "Prix": prix_element.text.strip() if prix_element else "",
            "Ancien_prix": old_price_element['data-oprc'] if old_price_element else "",
            "Promotion": remise_element.text.strip() if remise_element else ""
        }
        liste_de_donnees.append(donnees)

    return liste_de_donnees
```

```python
# Parcourir les pages et récupérer les données
def pagination(cat_url):
    lesPages = toutesLesUrls(cat_url)
    donnees_scrappees_liste = []

    for page in lesPages:
        soup = requete_url(page)
        if soup:
            donnees_page = traitement_donnees_beautiful_soup(soup)
            print(f"Scrapping de {page}")
            donnees_scrappees_liste.extend(donnees_page)
        else:
            print(f"Impossible d'accéder à {page}")

    return donnees_scrappees_liste

# Vérifier les doublons et ajouter les nouvelles données
def verifier_doublons(base_de_donnees, alias, nouvelle_donnees):
    if alias in base_de_donnees['donnees_scrappees']:
        anciennes_donnees = base_de_donnees['donnees_scrappees'][alias]['donnees']
        nouvelles_donnees_uniques = [donnees for donnees in nouvelle_donnees if donnees not in anciennes_donnees]
        base_de_donnees['donnees_scrappees'][alias]['donnees'].extend(nouvelles_donnees_uniques)
    else:
        base_de_donnees['donnees_scrappees'][alias] = {
            "url": "",
            "scrape_a": "",
            "statut": True,
            "domaine": "",
            "donnees": nouvelle_donnees
        }

    return base_de_donnees
```

```python
# Boucle principale du programme
while True:
    print("""================ Bienvenue dans ce programme de scraping =============
    ==>> Appuyez sur 1 pour voire l'historique
    ==>> Appuyez sur 2 pour scraper un site web
    ==>> Appuyez sur 3 pour quitter
    """)

    choix = int(input("|==>> Veuillez entrer votre choix   :"))

    base_de_donnees_json = charger_json()
    initialisation_donnees_scrappees_existantes(base_de_donnees_json)

    if choix == 1:
        # Affichage de l'historique
        table = BeautifulTable()
        table.columns.header = ["Sr no.", "Alias", "Domaine du site web", "URL", "Scraped at", "Status"]
        table.set_style(BeautifulTable.STYLE_BOX_DOUBLED)

        local_json_db = charger_json()
        count = 0
        for alias, data in local_json_db['donnees_scrappees'].items():
            count += 1
            table.rows.append([count, alias, data['domaine'], data['url'], data['scrape_a'], data['statut']])

        if not local_json_db['donnees_scrappees']:
            print('===> Aucune donnée existante trouvée !!!')
        print(table)
    elif choix == 2:
        cat_url = input("Entrez le chemin de la catégorie à scraper (ex: /electronique, /maison-bureau-electromenager, ...):
        url_pour_scraper = f"https://www.jumia.sn{cat_url}"
        print(f"Vous scrappez la catégorie : {cat_url}")

        base_de_donnees_json = charger_json()
        initialisation_donnees_scrappees_existantes(base_de_donnees_json)

        donnees_scrappees_liste = pagination(cat_url)

        alias_pour_stocker_donnees = input("Entrez un alias pour enregistrer les données scrapées: ")

        base_de_donnees_json = verifier_doublons(base_de_donnees_json, alias_pour_stocker_donnees, donnees_scrappees_liste)
        base_de_donnees_json['donnees_scrappees'][alias_pour_stocker_donnees]['url'] = url_pour_scraper
        base_de_donnees_json['donnees_scrappees'][alias_pour_stocker_donnees]['scrape_a'] = heure_scraping()
        base_de_donnees_json['donnees_scrappees'][alias_pour_stocker_donnees]['domaine'] = urlparse(url_pour_scraper).netloc

        enregistrer_donnees_scrappees_en_json(base_de_donnees_json)

        print("\n===> Données enregistrées avec succès dans donneesJumia.json !!!\n")
    elif choix == 3:
        print('Merci d\'avoir utilisé ce programme !!!')
        break
    else:
        print("Entrez un choix valide.")
```

ANNEXE 2 : Module de fusion des données de nos sites cibles

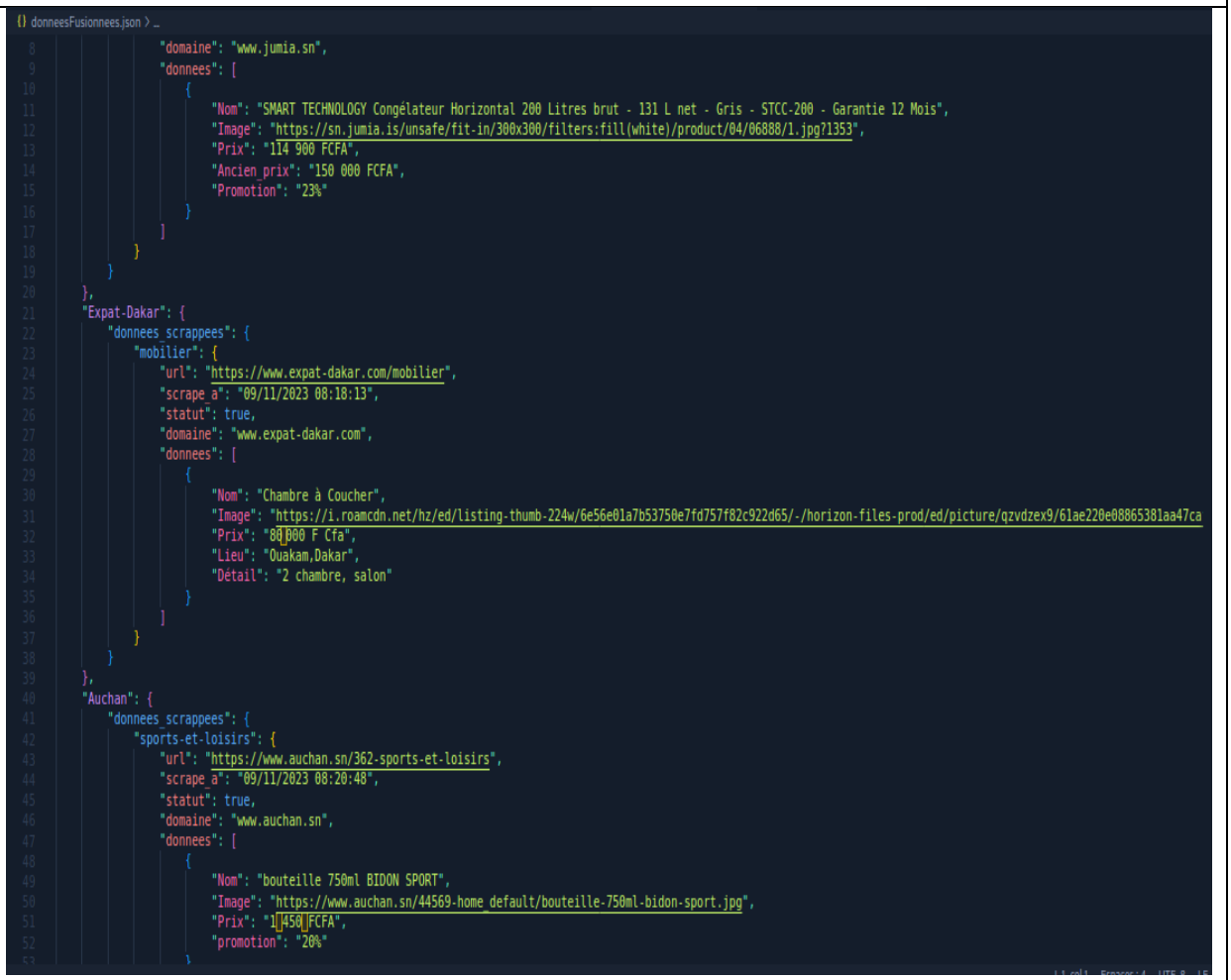
fusion.py

```python
import json

# Charger les données JSON depuis les fichiers
def charger_json(fichier_json):
    try:
        with open(fichier_json, "r", encoding="utf-8") as lire_fichier:
            donnees = json.load(lire_fichier)
            return donnees
    except:
        return []

# Charger les données de chaque fichier
donnees_jumia = charger_json("donneesJumia.json")
donnees_expat = charger_json("donneesExpatDakar.json")
donnees_auchan = charger_json("donneesAuchan.json")

# Créez un dictionnaire pour stocker les données fusionnées
donnees_fusionnees = {}

# Ajoutez les données de chaque source au dictionnaire
donnees_fusionnees['Jumia'] = donnees_jumia
donnees_fusionnees['Expat-Dakar'] = donnees_expat
donnees_fusionnees['Auchan'] = donnees_auchan

# Enregistrez les données fusionnées dans un fichier JSON
with open("donneesFusionnees.json", "w", encoding="utf-8") as fichier_obj:
    json.dump(donnees_fusionnees, fichier_obj, indent=4, ensure_ascii=False)

print("Données fusionnées enregistrées dans donneesFusionnees.json avec encodage UTF-8
```

ANNEXE 3 : Sortie des données Jumia en JSON


```json
{
    "donnees_scrappees": {
        "electronique": {
            "url": "https://www.jumia.sn/electronique",
            "scrape_a": "09/11/2023 08:15:47",
            "statut": true,
            "domaine": "www.jumia.sn",
            "donnees": [
                {
                    "Nom": "SMART TECHNOLOGY Congélateur Horizontal 200 Litres brut - 131 L net - Gris - STCC-200 - Garantie 12 Mois",
                    "Image": "https://sn.jumia.is/unsafe/fit-in/300x300/filters:fill(white)/product/04/06888/1.jpg?1353",
                    "Prix": "114 900 FCFA",
                    "Ancien_prix": "150 000 FCFA",
                    "Promotion": "23%"
                },
                {
                    "Nom": "Fer à Repasser à Vapeur CAC Petit Modele - 1400W / 220-240V - CA820454",
                    "Image": "https://sn.jumia.is/unsafe/fit-in/300x300/filters:fill(white)/product/12/507121/1.jpg?1589",
                    "Prix": "6 000 FCFA",
                    "Ancien_prix": "20 000 FCFA",
                    "Promotion": "70%"
                },
                {
                    "Nom": "Sharp Four à Micro-Ondes Avec Grille 34 Litres - 1000W - Garantie 6 Mois - Gris",
                    "Image": "https://sn.jumia.is/unsafe/fit-in/300x300/filters:fill(white)/product/80/817/1.jpg?6530",
                    "Prix": "84 900 FCFA",
                    "Ancien_prix": "100 000 FCFA",
                    "Promotion": "15%"
                },
                {
                    "Nom": "ICONA Fer A Repasser A Sec 1200W - Diil-101PLA/PA- Multicolore - Garantie 6 mois",
                    "Image": "https://sn.jumia.is/unsafe/fit-in/300x300/filters:fill(white)/product/36/451021/1.jpg?1976",
                    "Prix": "4 200 FCFA",
                    "Ancien_prix": "8 500 FCFA",
                    "Promotion": "51%"
                }
            ]
        }
    }
}
```

ANNEXE 4 : Sortie des données fusionnées : Jumia, Expat Dakar et Auchan




```
{} donneesFusionnees.json > …
                "domaine": "www.jumia.sn",
                "donnees": [
                    {
                        "Nom": "SMART TECHNOLOGY Congélateur Horizontal 200 Litres brut - 131 L net - Gris - STCC-200 - Garantie 12 Mois",
                        "Image": "https://sn.jumia.is/unsafe/fit-in/300x300/filters:fill(white)/product/04/06888/1.jpg?1353",
                        "Prix": "114 900 FCFA",
                        "Ancien prix": "150 000 FCFA",
                        "Promotion": "23%"
                    }
                ]
            }
        }
    },
    "Expat-Dakar": {
        "donnees_scrappees": {
            "mobilier": {
                "url": "https://www.expat-dakar.com/mobilier",
                "scrape_a": "09/11/2023 08:18:13",
                "statut": true,
                "domaine": "www.expat-dakar.com",
                "donnees": [
                    {
                        "Nom": "Chambre à Coucher",
                        "Image": "https://i.roamcdn.net/hz/ed/listing-thumb-224w/6e56e01a7b53750e7fd757f82c922d65/-/horizon-files-prod/ed/picture/qzvdzex9/61ae220e08865381aa47ca
                        "Prix": "80 000 F Cfa",
                        "Lieu": "Ouakam,Dakar",
                        "Détail": "2 chambre, salon"
                    }
                ]
            }
        }
    },
    "Auchan": {
        "donnees_scrappees": {
            "sports-et-loisirs": {
                "url": "https://www.auchan.sn/362-sports-et-loisirs",
                "scrape_a": "09/11/2023 08:20:48",
                "statut": true,
                "domaine": "www.auchan.sn",
                "donnees": [
                    {
                        "Nom": "bouteille 750ml BIDON SPORT",
                        "Image": "https://www.auchan.sn/44569-home_default/bouteille-750ml-bidon-sport.jpg",
                        "Prix": "1 450 FCFA",
                        "promotion": "20%"
                    }
```

Résumé

Ce mémoire traite la contraignante problématique liée à l'application du Web scraping pour une veille concurrentielle optimale. En effet, l'art d'**extraire des données depuis un site web** a un nom : c'est le **web scraping**, aussi appelé harvesting. Cette technique permet de récupérer des informations d'un site, grâce à un programme ou un logiciel et de les réutiliser ensuite. En automatisant ce processus, nous évitons ainsi de devoir récolter les données manuellement, nous gagnons du temps et nous accédons à un fichier unique et structuré. Le **web scraping** est une technique informatique qui a de nombreux usages. D'autres **applications du web scraping** sont particulièrement utiles dans le cadre de la prospection ou de la veille concurrentielle d'une entreprise. Nous pouvons collectés les données d'un site concurrent pour surveiller ses variations de prix ou bien l'évolution de ses offres. Des données, comme les prix pratiqués par la concurrence, les différentes gammes de produits proposées ou encore celles qui sont le plus mises en avant peuvent par exemple être des indicateurs précieux pour adapter son positionnement. Toutefois, lors du processus de collecte, il est important de bien préparer les données pour les rendre propres. Cela va permettre d'éviter les doublons, les données aberrantes et tout autre risque capable de biaiser les résultats d'analyse et du traitement. Dans ce contexte, le contenu extrait peut subir différentes manipulations à partir desquelles des informations clés peuvent en découler pour aider les décideurs à éclairer leur vision et mieux adapter leurs stratégies. Cette pratique permet de récupérer les données externes pour les confronter avec ses propres données et ainsi dégager des axes d'amélioration ou avoir une meilleure compréhension de son environnement.

Mots clés : Veille concurrentielle, Web scraping, Web Crawling, E-Commerce, Machine Learning

---

Abstract

This thesis addresses the challenging issue related to the application of web scraping for optimal competitive intelligence. Indeed, the art of extracting data from a website has a name: it's web scraping, also known as harvesting. This technique allows for the retrieval of information from a site, using a program or software, and then reusing it. By automating this process, we thus avoid having to collect data manually, saving time and accessing a unique and structured file. Web scraping is a computer technique with numerous uses. Other applications of web scraping are particularly useful in the context of prospecting or competitive intelligence for a company. We can collect data from a competitor's site to monitor its price variations or the evolution of its offers. Data such as prices practiced by the competition, the different product ranges offered, or those highlighted the most can be valuable indicators, for example, to adjust positioning. However, during the collection process, it is important to prepare the data well to make it clean. This will help avoid duplicates, aberrant data, and any other risks that could bias the results of analysis and processing. In this context, the extracted content can undergo various manipulations, from which key information can emerge to help decision-makers clarify their vision and better adapt their strategies. This practice allows for the retrieval of external data to confront it with one's own data and thus identify areas for improvement or gain a better understanding of the environment.

Keywords: Competitive monitoring, Web scraping, Web Crawling, E-Commerce, Machine Learning