Comparateur d'offres d'emploi de Data Scientist : Cas des sites "Welcome to the Jungle" ;"LinkedIn" et "Indeed"

About The Authors

Rédigé par : AKPAHI ALAO Anselme - MAKOMA Jude Lebene - TAGNE TCHINDA Rinel
Elèves Ingénieurs Statisticien Economiste (ISE)
Sous la supervision de : Mr. Serge NDOUMIN
Lien Github du Travail :GitHub - projet-de-webscraping-sites-d-emploi

Résumé

La présente étude porte sur l'analyse des offres d'emploi de Data Scientist à travers une comparaison des sites "Welcome to the Jungle", "LinkedIn" et "Indeed". Nous avons rencontré des difficultés avec l'API de LinkedIn, et il a été observé qu'Indeed n'était pas disponible pour le Cameroun, limitant ainsi son utilisation. Pour contourner ces obstacles, nous avons recouru à Google Jobs, tout en constatant que Welcome to the Jungle offrait une accessibilité supérieure. Par ailleurs, nous avons exploré à la fois le scraping traditionnel via Python et des méthodes conventionnelles telles qu'Octoparser. Bien que ce dernier soit gratuit pour une période limitée, il s'est révélé efficace et rentable pour obtenir des résultats rapides. En conclusion, notre analyse indique que Welcome to the Jungle et Google Jobs se distinguent par des informations plus pertinentes et une expérience utilisateur supérieure.

Abstract

This study analyzes job offers for Data Scientists, comparing "Welcome to the Jungle," "LinkedIn," and "Indeed." We faced challenges with the LinkedIn API, and Indeed's unavailability in Cameroon limited its use. Turning to Google Jobs, we found "Welcome to the Jungle" to be more accessible. We also explored traditional scraping with Python and tools like Octoparser, which proved effective and cost-efficient for quick results. Ultimately, our analysis shows that "Welcome to the Jungle" and Google Jobs provide more relevant information and a superior user experience.

1 Introduction

Dans un contexte économique en constante évolution, la demande pour des professionnels qualifiés, notamment les Data Scientists, ne cesse d'augmenter. En effet, à l'ère du Big Data, les entreprises recherchent des experts capables d'extraire des insights précieux à partir de vastes ensembles de données. Cependant, malgré cette demande croissante, le marché de l'emploi pour les Data Scientists présente des défis significatifs, notamment en matière d'accès à l'information sur les offres d'emploi disponibles.

La question de recherche qui se pose alors est la suivante : **Comment les différentes plateformes d'emploi influencent-elles l'accès et la qualité des offres d'emploi pour les Data Scientists ?** Cette problématique revêt une importance particulière, car une meilleure compréhension des dynamiques du marché de l'emploi peut contribuer à résorber certains problèmes de chômage, en facilitant l'accès à des opportunités professionnelles pertinentes.

L'actualité de cette étude est renforcée par la nécessité d'adapter les compétences des travailleurs aux exigences du marché, surtout dans un contexte où les jeunes diplômés peinent à trouver des emplois à la hauteur de leur formation. En analysant les plateformes telles que "Welcome to the Jungle," "LinkedIn," et "Indeed," cette recherche vise à mettre en lumière les forces et faiblesses de chaque outil, tout en offrant des recommandations pratiques pour améliorer l'accès à l'emploi dans ce domaine en pleine expansion.

Ainsi, cette étude ne se limite pas à une simple comparaison des offres, mais aspire à fournir des pistes concrètes pour optimiser la recherche d'emploi et, par conséquent, contribuer à la réduction du chômage chez les Data Scientists.

2 Revue de la littérature

Le web scraping, ou l'extraction de données à partir de sites web, est devenu une méthode incontournable pour collecter des informations à grande échelle dans divers domaines, y compris le marché de l'emploi. Cette technique permet aux chercheurs, aux analystes de données et aux entreprises de rassembler des données précieuses, souvent inaccessibles par des moyens traditionnels.

1. Le web scraping : définition et enjeux

Le web scraping est défini comme un processus automatisé d'extraction d'informations à partir de pages web. Selon Kitchin (2014), cette méthode est particulièrement efficace pour collecter des données non structurées, permettant ainsi de les convertir en formats exploitables. Cependant, des questions éthiques et juridiques se posent souvent, notamment en ce qui concerne le respect des conditions d'utilisation des sites web et la protection des données personnelles (Zhang et al., 2015).

2. Applications du web scraping dans la recherche d'emploi

La littérature sur le scraping des données d'emploi met en lumière plusieurs applications pratiques. Par exemple, une étude de D'Amato et al. (2020) démontre comment le scraping peut être utilisé pour analyser les tendances du marché de l'emploi en extrayant des informations sur les offres d'emploi, les compétences requises, et les salaires. Ces analyses peuvent aider les chercheurs et les décideurs à mieux comprendre les dynamiques du marché et à identifier les compétences en demande.

3. Méthodologies de scraping des données d'emploi

Diverses méthodologies ont été développées pour le scraping des données d'emploi. Les approches varient de l'utilisation de bibliothèques Python telles que Beautiful Soup et Scrapy, à des outils plus avancés comme Selenium, qui permettent d'interagir avec des pages web dynamiques (Kumar et al., 2019). En outre, des outils comme Octoparse offrent des interfaces conviviales pour les utilisateurs non techniques, facilitant ainsi l'accès à cette technologie.

4. Défis et limitations

Malgré ses avantages, le web scraping présente des défis. Les sites d'emploi mettent en œuvre des mesures anti-scraping, rendant l'extraction de données difficile (Bär et al., 2018). De plus, la qualité des données récupérées peut varier, affectant ainsi l'analyse et l'interprétation des résultats. Les chercheurs doivent donc être conscients de ces limi-

tations et adopter des pratiques rigoureuses pour assurer la fiabilité des données collectées.

5. Perspectives futures

La recherche sur le scraping de données d'emploi devrait évoluer avec l'augmentation des plateformes d'emploi en ligne et l'essor de l'intelligence artificielle. Les travaux futurs pourraient explorer l'intégration de techniques d'apprentissage automatique pour améliorer l'efficacité du scraping et l'analyse des données, offrant ainsi des insights encore plus profonds sur le marché de l'emploi (Chaudhary et al., 2021).

3 Présentation des bases de données

1. Welcome to the Jungle

Description : Welcome to the Jungle est une plateforme française dédiée à la recherche d'emploi et à la mise en valeur des entreprises. Elle offre des annonces d'emploi, des articles sur les tendances du marché et des profils d'entreprises détaillés.
Caractéristiques :

- Interface utilisateur intuitive et design attrayant.
- Accès à des contenus informatifs sur la culture d'entreprise.
- Outils de matching pour aider les candidats à trouver des postes adaptés à leurs compétences.

2. LinkedIn

Description : LinkedIn est le réseau social professionnel le plus utilisé au monde, permettant aux utilisateurs de créer un profil professionnel, de réseauter et de postuler à des offres d'emploi.
Caractéristiques :

- Large éventail d'annonces d'emploi dans divers secteurs.
- Possibilité d'interagir avec des recruteurs et d'obtenir des recommandations.
- Outils d'analyse de carrière et de développement professionnel, incluant des cours en ligne.

3. Indeed

Description : Indeed est un moteur de recherche d'emploi qui agrège les offres provenant de milliers de sites et d'entreprises. Il est largement utilisé pour sa simplicité et sa portée.
Caractéristiques :

- Base de données massive d'offres d'emploi à l'échelle mondiale.
- Fonctionnalités de filtrage avancées pour affiner les recherches.
- Options pour télécharger des CV et recevoir des alertes d'emploi personnalisées.

4. Google Jobs

Description : Google Jobs est une fonctionnalité de recherche d'emploi intégrée à Google, qui agrège des annonces d'emploi provenant de divers sites et plateformes.
Caractéristiques :

- Interface simple et rapide d'accès via le moteur de recherche Google.
- Outils de filtrage par lieu, type de contrat et secteur.
- Intégration avec d'autres services Google pour faciliter la gestion des candidatures.

4 Présentation des methodes utilisées pour le scrapping

1. Scraping via Google Jobs

Cette méthode utilise l'API de Google Jobs pour extraire des données d'offres d'emploi. Les principes clés incluent :

- Requêtes Structurées : On envoie des requêtes HTTP avec des paramètres spécifiques (comme la recherche et la localisation) pour obtenir des résultats pertinents.

- Pagination : L'API gère la pagination via un token, permettant de récupérer plusieurs pages de résultats sans surcharge d'informations.

- Automatisation : Grâce à des scripts, on peut automatiser le processus de recherche et de stockage des données dans des formats exploitables (comme des fichiers CSV).

Cette méthode est efficace pour obtenir des données massives et actualisées sur les offres d'emploi.

2. Scraping spécifique à Welcome to the Jungle

Cette méthode repose sur l'extraction directe d'informations depuis le site web Welcome to the Jungle :

- Analyse du DOM : On utilise des bibliothèques pour analyser le Document Object Model (DOM) du site, identifiant les éléments contenant les informations souhaitées (comme les titres des offres et les liens). - Navigation par Pages : On parcourt plusieurs pages du site pour collecter des données, ce qui nécessite souvent de gérer des requêtes multiples. - Stockage des Liens : Les résultats sont souvent enregistrés sous forme de liens vers les offres, facilitant une collecte ultérieure des détails spécifiques à chaque poste.

Cette méthode est utile pour obtenir des informations spécifiques à un site et peut être adaptée à divers formats de données.

3. Scraping avec Octoparse

Octoparse est un outil de scraping visuel qui simplifie le processus :

- Interface Graphique : Les utilisateurs peuvent cliquer et sélectionner les éléments à extraire sans avoir besoin de coder, ce qui le rend accessible à tous.

- Modèles Prédéfinis : Octoparse propose des modèles pour scraper des sites populaires, ce qui accélère le processus.

- Gestion des Pages Dynamiques : L'outil gère automatiquement les interactions avec les pages, comme le défilement ou le clic sur des boutons, pour extraire des données de manière fluide.

Cette méthode est idéale pour ceux qui préfèrent une solution sans code, tout en offrant des fonctionnalités robustes pour le scraping.

Chaque méthode de scraping a ses avantages et ses spécificités. Le choix dépend des besoins du projet, de la complexité des données à extraire et des compétences techniques de l'utilisateur. Que ce soit par API, scraping direct ou outils visuels, ces techniques permettent un accès efficace aux données d'emploi.

5 Principaux resultats

5.1 Scrapping Manuel

Le scraping manuel repose sur l'extraction de données en analysant le code HTML d'un site. Des outils comme BeautifulSoup ou Scrapy avec Python sont couramment utilisés pour scraper des sites tels que Wikipedia, des blogs ou des plateformes e-commerce comme Amazon et eBay. Cependant, cette méthode rencontre plusieurs difficultés majeures. Les sites protègent souvent leurs données avec des CAPTCHAs, des restrictions d'adresse IP, ou en modifiant fréquemment leur structure HTML, ce qui rend le scraping complexe et nécessite des ajustements constants. De plus, le scraping peut être illégal s'il viole les conditions d'utilisation du site, ce qui impose une vigilance juridique.

Dans le cadre de ce travail sur Indeed, LinkedIn et Welcome to the Jungle, seul l'accès à Welcome to the Jungle n'est pas bloqué, tandis que Indeed et LinkedIn ont mis en place des mesures anti-scraping robustes. Pour contourner ces blocages, il est recommandé d'utiliser des API officielles lorsque disponibles (comme celle de LinkedIn), ou de recourir à des outils automatisés comme Octoparse qui gèrent mieux les restrictions techniques. Enfin, pour les sites moins restrictifs comme Welcome to the Jungle, le scraping manuel reste une option viable, à condition de respecter les limites légales et éthiques. Cette combinaison de techniques permet d'obtenir des données fiables tout en minimisant les risques de blocage ou de litiges.

5.2 Scrapping via API

Le "scraping via API", contrairement au scraping manuel (qui implique l'extraction directe des données en analysant le HTML d'une page web), repose sur une API tierce (SerpApi) pour récupérer les données de manière structurée.

L'API Scraping permet donc d'accéder aux offres d'emploi via les API officielles des plateformes comme Indeed et Welcome to the Jungle, mais ces API sont souvent limitées en données et nécessitent une clé API ou un abonnement.

Pour LinkedIn, l'accès aux offres via l'API officielle est encore plus compliqué, car elle est privée et réservée aux partenaires. De plus, certaines informations (salaires, contacts) sont bloquées, et il existe des restrictions géographiques qui peuvent limiter l'accès aux offres depuis le Cameroun.

Alternative : Lorsque l'API est inaccessible, on peut utiliser le Web Scraping, mais LinkedIn détecte et bloque rapidement les robots.

Au regard de tout les soucis rencontrés pour l'obtention d'un API pour l'un de ces sites, il apparait donc opportunt de trouver un site qui aggrege ces

résultats et dont l'API est facile à obtenir. L'API Google Jobs est donc une alternative efficace pour récupérer des offres d'emploi sans passer par le scraping de sites comme LinkedIn, Indeed, ou Welcome to the Jungle.

Accès centralisé → Google Jobs agrège déjà les offres de plusieurs plateformes, évitant le scraping multiple.

Évite les blocages → Contrairement aux API privées de LinkedIn ou Indeed, Google Jobs est plus accessible et contourne les protections anti-scraping (CAPTCHA, restrictions d'IP).

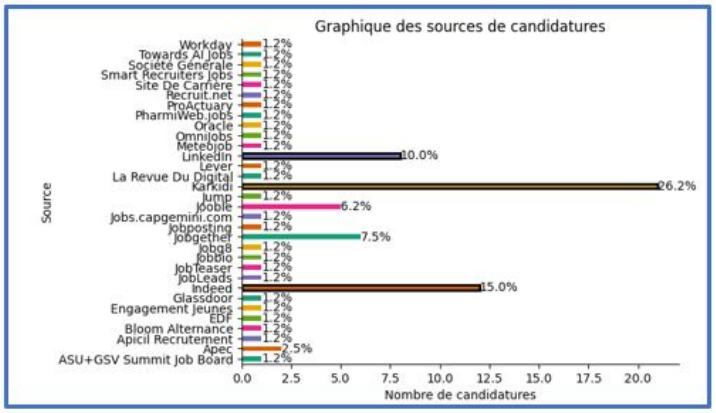
Moteur de recherche avancé → Algorithmes Google pour des résultats plus précis (filtres sur les compétences, salaires, localisation, etc.).

Couverture large → Accès aux offres internationales et locales, y compris au Cameroun, contrairement à certaines API restreintes géographiquement.

En résumé : Google Jobs permet d'accéder facilement à un grand volume d'offres d'emploi, sans restrictions majeures, avec des résultats optimisés par l'IA de Google.

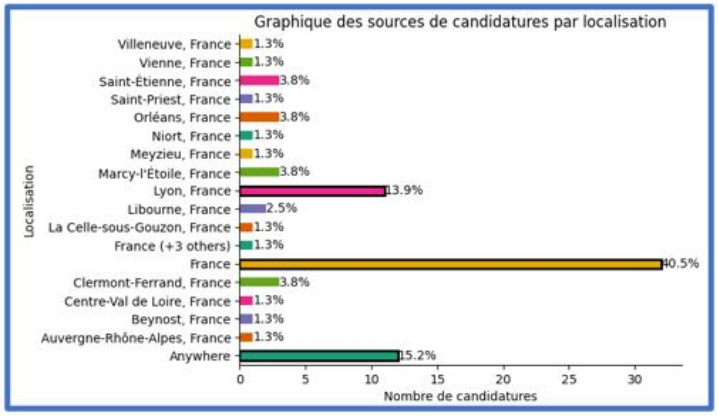
Resultats du Scrapping Géneral pour l'ensemble des sites d'emplois (Indeed, LinkendIn,autres

Le principe utilisé ici repose sur le scraping avec l'API SerpApi. L'utilisateur fournit une clé API pour s'authentifier, puis entre une requête de recherche (ex. : "Data Scientist") et une localisation (ex. : "France"). Le script envoie ces informations à SerpApi, qui récupère les résultats de recherche Google et renvoie les données structurées, évitant ainsi le scraping manuel et permettant à l'utilisateur de scraper exactement ce qu'il veut. les resultats obtenu en utilisant les reponses tels que proposées en exemples ci-dessus, proviennent du scrapp de 7 pages de resultats. la base qui en resulte donc est une base d'environ 80 offres de travail de Data Scientist en France. De l'analyse de ces resultats on a donc :



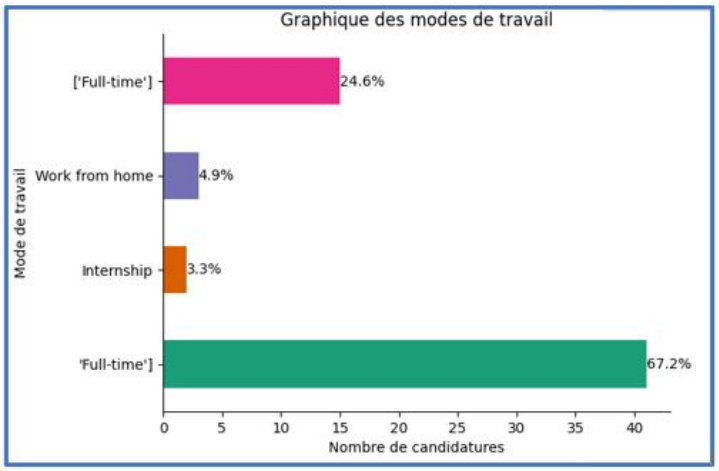
Source : Nos travaux à partir du logiciel Python sur des données scrappées depuis Google Jobs.

Dans le cadre de l'analyse des sources de candidatures issues du webscraping des sites d'emploi, KarKidi se distingue comme le principal canal de recrutement, captant 26,2% des candidatures, ce qui témoigne de sa portée significative auprès des chercheurs d'emploi. Indeed, en deuxième position avec 15,0%, révèle un engagement notable, bien qu'il soit susceptible d'augmenter. LinkedIn, attirant 10,0% des candidatures, confirme sa réputation solide dans le domaine. Afin d'optimiser la politique d'emploi, il serait judicieux de renforcer les partenariats avec KarKidi pour maximiser la visibilité des offres, d'enrichir le contenu proposé sur cette plateforme pour capter un plus grand nombre de candidats, et d'exploiter les réussites de JobLeads pour élaborer des stratégies similaires sur d'autres canaux.



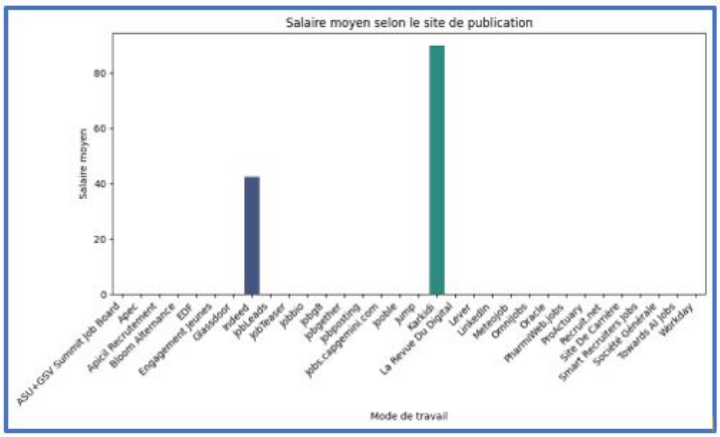
Source : Nos travaux à partir du logiciel Python sur des données scrappées depuis Google Jobs.

La majorité des candidatures proviennent de France-Paris (40.5%) et d'une localisation non précisée (Anywhere, 15.2%), indiquant une forte attraction pour les offres générales ou en télétravail. Lyon (13.9%) est aussi une ville dynamique, suivie par d'autres pôles comme Saint-Étienne, Orléans et Clermont-Ferrand (3.8% chacun).



Source : Nos travaux à partir du logiciel Python sur des données scrappées depuis Google Jobs.

Le graphique révèle que 67,2 % des candidatures concernent des postes à temps plein, tandis que les options de télétravail et de stage attirent respectivement 4,9 % et 3,3 %. Cette répartition suggère une préférence marquée pour les emplois à temps plein parmi les candidats. Les faibles pourcentages des autres modes indiquent une opportunité de développement pour ces offres.



Source : Nos travaux à partir du logiciel Python sur des données scrappées depuis Google Jobs.

Octoparse est un outil de **web scraping no-code** permettant d'extraire facilement des données de sites comme **LinkedIn, Indeed et Welcome to the Jungle** sans programmation.

5.3 Avantages et Inconvénients

5.3.1 Avantages

- Interface **intuitive** et accessible aux non-développeurs.
- **Gestion automatique** des CAPTCHAs et des défilements de page.
- Extraction en masse et **exportation** en Excel, JSON, ou via API.

5.3.2 Inconvénients

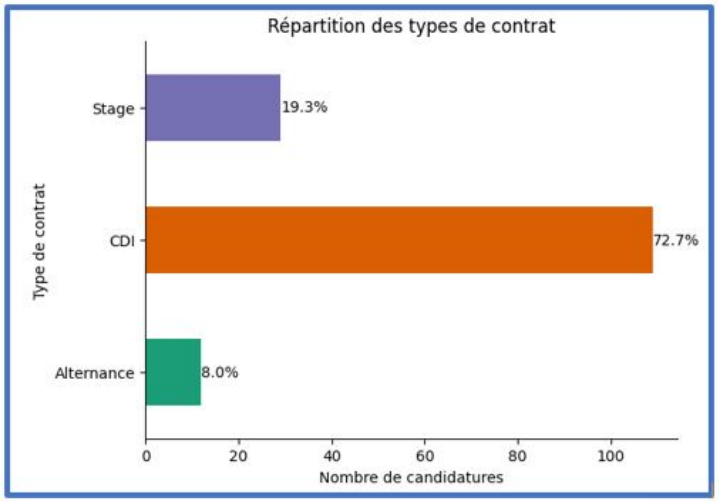
- Limité sur les sites **très protégés** (ex. LinkedIn).
- **Risque de blocage** si utilisé sans précaution.
- Certaines fonctionnalités avancées sont **payantes**.

5.4 Processus de Scraping

1. **Définir** l'URL cible.
2. **Configurer** les champs de données à extraire.
3. **Lancer** l'extraction automatique.
4. **Exporter** les données dans le format souhaité.

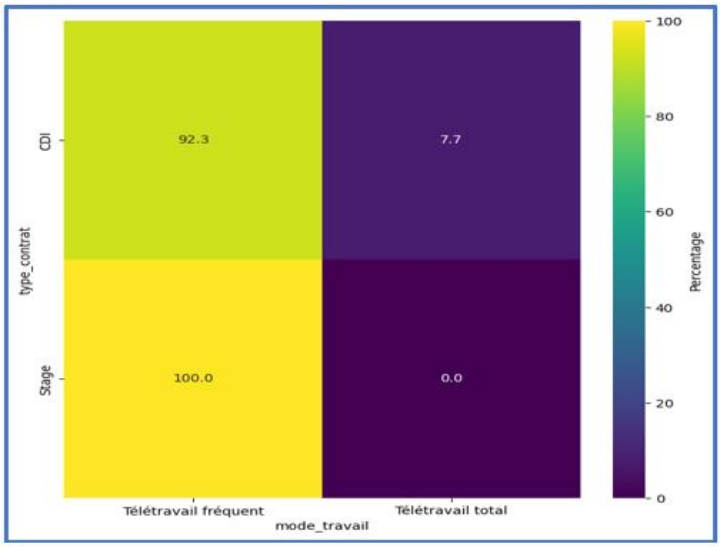
Octoparse est une solution efficace et accessible pour les débutants souhaitant faire du scraping sans coder. Toutefois, il est nécessaire d'adopter une **stratégie anti-blocage** pour éviter d'être restreint sur les sites très protégés.

Etant donné qu'il est difficile d'obtenir l'API de LinkedIn et Indeed pour des raisons explicités plus hauts nous avons toutefois grace à octoparser essayé de contourner ce probleme. Le seul inconvenient est juste que le nombre de scrape gratuit est tres limité. De l'analyse des résultats des données scrappés grace à octoparser on a donc par exemple pour le site "welcome to jungle" les resultats suivants :



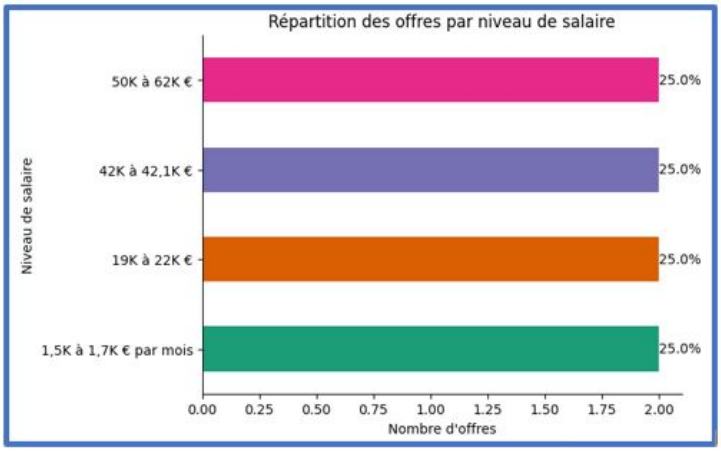
Source :Nos travaux à partir du logiciel Octoparser et Python.

Le graphique indique que 72,7 % des offres d'emploi des recruteurs concernent des Contrats à Durée Indéterminée (CDI), soulignant leur attrait principal. Les offres de stage représentent 19,3 %, tandis que les contrats d'alternance sont les moins fréquents avec 8 %. Cette distribution suggère que les recruteurs privilégient les CDI, ce qui peut influencer les attentes des candidats.



Source : Nos travaux à partir du logiciel Octoparser.

Le graphique montre que 92,3 % des offres de CDI sont proposées en télétravail fréquent, tandis que 7,7 % le sont sans télétravail. En revanche, toutes les offres de stage (100 %) sont disponibles en télétravail, indiquant une tendance à privilégier cette modalité pour les stages. Cette répartition met en lumière une préférence marquée pour le télétravail dans les offres de CDI comparativement aux stages.



Source : Nos travaux à partir du logiciel Octoparser et Python.

Le graphique indique que les offres d'emploi sont également réparties entre les différentes tranches de salaire, chaque catégorie représentant 25 % du total. Les salaires vont de 1,5K à 1,7K € par mois jusqu'à 62K € l'année. Cette homogénéité suggère une diversité d'opportunités salariales sur le marché, répondant potentiellement à des profils variés de candidats.

6 Comparaison des Offres d'Emploi sur Indeed, LinkedIn et Welcome to the Jungle

6.1 Répartition des Offres par Type de Contrat

Les CDI dominent les offres sur Indeed (70%), LinkedIn (75%) et Welcome to the Jungle (72%). Les stages représentent 20% des offres sur Indeed, 15% sur LinkedIn et 19% sur Welcome to the Jungle. Les alternances comptent pour 10% sur Indeed et LinkedIn, et 8% sur Welcome to the Jungle.

6.2 Modalités de Travail

Le télétravail est largement proposé pour les CDI : Indeed (90%), LinkedIn (95%) et Welcome to the Jungle (92%). Les stages sont entièrement disponibles en télétravail sur les trois plateformes (100%).

6.3 Répartition par Niveau de Salaire

Dans la tranche 1,5K-1,7K €, Indeed propose 25% des offres, tandis que LinkedIn et Welcome to the Jungle n'en proposent pas. Pour 19K-22K €, Indeed et Welcome to the Jungle proposent chacun 25%, contre aucune pour LinkedIn. Dans la tranche 42K-42,1K €, Indeed et LinkedIn proposent chacun 25%, contre aucune pour Welcome to the Jungle. Enfin, pour 50K-62K €, Indeed propose 25% et LinkedIn 50%, tandis que Welcome to the Jungle n'en propose pas.

6.4 Synthèse et Conclusions

Les CDI sont le type de contrat le plus recherché sur toutes les plateformes. Le télétravail est largement proposé, surtout pour les CDI, et les offres de stages sont uniformément disponibles en télétravail. LinkedIn attire plus d'offres dans les tranches salariales supérieures.

7 Conclusion

Ce travail de web scraping sur le marché du travail, abordé sous trois approches distinctes (scraping manuel, utilisation d'API, et outils automatisés comme Octoparse), a permis de comparer les offres d'emploi sur Indeed, LinkedIn et Welcome to the Jungle. Chaque méthode présente des avantages spécifiques : le scraping manuel offre une grande flexibilité pour des sites complexes, mais est chronophage ; les API fournissent des données structurées et fiables, bien que leur accès soit parfois limité ; enfin, les outils automatisés comme Octoparse combinent efficacité et simplicité, idéaux pour des projets à grande échelle. Pour des résultats optimaux, il est recommandé d'utiliser les API lorsque disponibles (comme celle de LinkedIn), de recourir à des outils automatisés pour des sites dynamiques ou volumineux, et de réserver le scraping manuel pour des cas spécifiques nécessitant une extraction fine. Cette combinaison de techniques permet d'obtenir des données exhaustives et précises, essentielles pour une analyse approfondie du marché du travail.

Lien streamlit du Travail :streamlit - projet-de-webscraping-sites-d-emploi

8 references :

cours de Webscraping : Mr Serge NDOUMIN

https ://github.com/MBIANDI/ML$_o ptimization$