Grésillon Rémi M2 Web Analystes

Rapport projet de fin d'études Web Analyste

Université de Lille / 2020-2021

Introduction :

Il nous a été demandé de remettre pour notre fin d'études un projet faisant appel à toutes les compétences que nous avons développées principalement durant ces deux dernières années, cependant je pense pouvoir affirmer que l'année 2020 a profondément chamboulé l'essence même de l'enseignement. Nous sommes pourtant bel et bien arrivés à conclure ces deux années de Master, non sans difficultés que ce soit pour la progression de la courbe de niveau ou pour la motivation en générale.

Mais bien que la situation était complètement nouvelle pour nous tous, j'ai eu la chance de pouvoir intégrer dès le mois de mai, l'équipe d'Optimal Ways ce qui m'a permis en parallèle de la formation de développer des compétences supplémentaires mais aussi une première réelle expérience dans le domaine professionnel.

Cette dernière année m'a fortement marqué, bien que le confinement nous a empêché de pouvoir évoluer dans les meilleures conditions, cela a été pour moi l'une des années qui m'a le plus permis de changer.

Dans le cadre de ce projet de conclusion de master, je vais essayer de vous décrire au mieux la manière dont j'ai abordé les choses.

Plan du rapport :

1. Description du projet
2. Objectifs
3. Déroulement / Structure
4. Réussites / Échec
5. Evolution / Avenir
6. Conclusion

1. Description du projet

En poste chez Optimal Ways depuis mai, j'ai pu commencer à voir et comprendre les cas concrets que doit gérer une entreprise travaillant dans le domaine de la donnée en provenance du web.

C'est pour moi une réelle application de tout ce que j'ai pu apprendre en cours durant ces dernières années. Il nous a été conseillé, d'accorder notre sujet avec les problématiques de notre entreprise. J'ai donc tenté de trouver comment le scraping web pouvait être bénéfique à l'offre proposée par Optimal Ways. Avant de décrire mon projet, je vais vous expliquer ce qu'est un datalayer, la clé de mon projet, pour que l'on ait tous en tête la même définition.

datalayer : Objet javascript, qui contient tout un tas d'informations sur la page actuellement visitée. Ces informations peuvent être générées dans la page elle-même ou par des appels serveurs. Une fois le datalayer rempli, on peut transférer ces données dans des gestionnaires par exemple GTM (Google Tag Manager) ou encore Tag Commanders, ce qui permettra alors de traiter et d'analyser tout le site Web.

Chez Optimal Ways, une entreprise de consultant en digital analytics, chaque jour nous répondons aux besoins des clients et lorsqu'un client fait une grosse mise à jour de son site, ou bien qu'un nouveau client contacte l'entreprise; Un besoin se crée, "Il faut analyser leur datalayer", observer si les données que l'on va traiter sont correctes. L'on ne peut pas tirer de conclusion si les données sont faussées. Pour vérifier que ces données sont exactes, j'ai tenté d'utiliser le scrapping pour m'assurer que le datalayer soit correct sur chacune des pages importantes du site.

Optimal Ways travaille principalement dans le domaine du retail. Les pages que l'on peut catégoriser comme les plus importantes sur ces types de sites sont les pages produits. C'est pourquoi j'ai décidé de comparer le contenu de tous les produits d'un site et de les comparer avec son datalayer.

Quel site ?

-> J'ai décidé de me baser sur le site d'Electro Dépôt.

2. Objectifs

Mon objectif à travers ce projet est de créer un outil réutilisable dans le futur permettant de répondre aux attentes de mon entreprise.

"Le datalayer d'Electro depot est il erroné ?" Voici mon point de départ, une question que potentiellement je pourrais me poser à l'avenir. Cependant cette tâche

peut se comparer à rechercher une aiguille dans une botte de foin, il a trop de page pour détecter ces erreurs, actuellement hormis en voyant les données fausses remonter, il n'y a pas d'autre moyen de savoir si le datalayer est correct ou non. C'est pourquoi je tente d'automatiser cette recherche.

Il faut savoir qu'à la base de ce projet, je m'étais fortement éparpillé. J'avais d'autre objectifs secondaires voici deux d'entre-eux:

- Fonctionnalité de recherche de produit par autocomplétion en fonction des caractéristiques de produits.
- Recherche avancée à l'aide des avis des clients disponible en fin de page du site.

Cependant après avoir parlé avec les professeurs, je me suis rendu compte que je tentais de réaliser des objectifs qui cible une population complètement différentes qu'initialement, l'objectif principalement ciblant l'entreprise elle-même alors que ces recherches de produits seraient plus pour les consommateurs.

3. Déroulement / Structure

Dans un premier temps il fallait recueillir les données. Pour se faire sachant que je voulais récupérer uniquement les pages produits du site j'ai donc cherché et trouvé le sitemap du site Electro Depot.

https://www.electrodepot.fr/sitemap-index.xml

Pour la récolte et la génération de bases de données j'ai codé en python. Les principales librairies que j'ai pu manipuler sont beautifulSoup et Selenium pour la scraping, et mysql pour la création et l'insertion des données dans mon base de donénes

Pour récupérer le datalayer d'une page il faut une fois sur le navigateur et la bonne page souhaitée, aller dans le console pour renvoyer dans le cas de google, "datalayers" qui est comme expliqué précédemment un objet javascript. Il n'est pas possible via le scraping de directement récupérer cet objet. J'ai donc dû me tourner vers Selenium qui me permet sur chaque page de faire un "return datalayers" et de la stocker dans une variable python.

En parallèle je scrape chaque élément de la page associé au datalayer, je m'explique, dans les variables que j'ai souhaité analyser dans ce datalayer, prenons par exemple le prix, je vais visiter le site avec Selenium mais aussi directement avec BeautifulSoup pour récupérer le prix affiché sur la page. L'objectif étant de trouver si il y a des différences entre ce qui est présent par rapport aux données que l'on remonte pour les analyses. Un mauvais exemple ici serait un téléphone qui aurait un

prix de 300 euros dans le datalayer mais 280 euros sur le site. Dans une analyse des revenus générés sur ce site, une différence de 20 euros remonterait à chaque achat.

Une fois les données scrapées je les insères dans ma base de données que j'ai créé au préalable, à la base (j'en reparlerai dans la partie réussite/échec) mais cette base de données avait pour objectif de récupérer les avis des clients présent en bas de page pour affiner la distinction des produits dans une partie analyse des composantes du site.

Pour ce qui est du rendu visuel, j'ai conçu une application web sur deux pages.

La première voulant répondre à mon objectif principale ("Le datalayer d'Electro depot est il erroné ?"), sur cette page on peut retrouver :

- Une partie scoreboard, relatant des différentes caractéristiques du scraping
    - A terme cette partie peut largement évoluer par l'ajout de nouveaux scorecards permettant de donner plus de sens aux informations

- Une recherche de produits précise par dimension qui unique
    - Si on a une ID de produit en tête on peut directement aller voir ce que donne les éléments du datalayer associé et les données scrapées

- Une comparaison des pages du site correctes et erronées par dimension avec des liens référant aux pages
    - A l'aide du scoreboard cité précédemment on peut directement aller chercher l'erreur la plus commune et tenter de comprendre d'où elle vient pour ainsi tenter de la corriger.

Sur la deuxième page, j'ai voulu créer une partie visualisation de données sur les produits disponible sur le site, à l'aide de chartjs et de php j'ai créé des builders de graphique permettant de naviguer entre les différentes dimensions disponibles dans la bases de données.

J'ai rajouté une fonctionnalité d'export de données sous fichier CSV pour permettre de retravailler les données ou tout simplement de pouvoir créer un réel historique de données dans le temps.

4. Réussites / Échec

a. réussites

La plus grande de mes peurs après m'être lancé ce projet était de n'avoir aucune page du site d'Electro dépôt erroné, il était possible que ce soit le cas. Alors on aurait pu se demander si j'avais fait mon projet dans le sens où je n'aurai rien eu de particulier à montrer.

Cependant après avoir scrappé leur site j'ai trouvé des erreurs, cette crainte a donc disparue. Une fierté s'est donc créée car ayant montré les résultats de cette recherche à Optimal Ways on m'a directement demandé de réutiliser cette application pour d'autres clients.

A termes, j'aimerai donc créer de nouvelles versions de ce projet permettant de changer de sources de données ainsi que d'analyser les données d'un site ou d'un autre. Mais aussi dans le temps après chaque mise à jour majeure d'un site. On pourra créer un réel historique des données.

b. échecs

Même après avoir abandonné le côté consommateur, je voulais toujours récupérer les avis des clients pour agrémenter mes graphiques de données, cependant il m'a été impossible de récupérer ces avis de clients, ils étaient appelés via l'API de bazaarvoice, et bien que j'ai réussi à retrouver le json à la main dans le partie Sources du mode développeur je n'ai pas trouvé de moyen de l'automatiser

Je ne me sers pas non plus des caractéristiques que j'ai pourtant scrap et inséré dans la base de donnée, bien que j'ai tenté de trouver un intérêt pour l'entreprise.

Enfin dans la partie visualisation de données mise à part les deux types de graphiques que j'ai mis en place et l'export de données que je trouve intéressant, Je me sens gêné de ne pas avoir autant de représentation visuelle des données que ce que j'aurais voulu produire.

5. Evolution / Avenir

A termes j'ai réussi à créer un template de scraping par catégorie de page par exemple produit / catalogue produit / panier / confirmation d'achat etc etc
ainsi lorsque je me tenterai à scraper un nouveau site je n'aurai qu'à tout simplement modifier ces fonctions.

Avec une forte inspiration de data studio (encore un produit google décidément) j'aimerai donner la possibiliter de changer de sources de données, de temporalité, etc.. en ayant pour objectif d'historiser les données

Une chose est sûre, je vais ré-utiliser cette application.

6. Conclusion

Ce projet m'a bien fait comprendre l'étendu des principes que j'ai pu apprendre durant mon cursus à la fac, cependant je vois bien que derrière tout ça j'ai encore beaucoup à apprendre, et pas mal de sujet à approfondir, j'ai hâte de voir ce que les prochaines années mon donner. Je tiens à remercier tous les professeurs que j'ai pu côtoyer durant mon master, j'ai réellement eu un second souffle de découvrir un monde que j'aime vraiment, je tiens à remercier aussi tous les professeurs qui ont pu m'aider pour la réalisation de ce projet que ce soit pour me donner des idées, m'aiguiller sur mes recherches ou tout simplement m'entendre me plaindre. Je tiens à remercier tous mes collègues pour avoir pris le temps de me donner leur avis sur mon avancée et enfin Nicolas qui bien que étant supposé travailler pour les besoins de l'entreprise m'a laissé des périodes ou je pouvais me donner à fond pour mon projet.

Merci d'avoir lu ce rapport.
Rémi