	STAGE DE 1ERE ANNEE
Rédigé par Gustave Richter ING 1 GI	Créer le 30/06/2023
Référent : Karima el Gaoui	Tuteur : Xavier Bour

CY Tech

Stage de 1ère année au sein de IdIA-Tech

Table des matières

Remerciements

Je tiens à remercier mon maître de stage, Xavier Bour, pour m'avoir permis, tout d'abord d'effectuer ce stage, puis de m'avoir fait découvrir le webscrapping : son objectif, son fonctionnement et ses applications et puis de travailler sur une création nouvelle, un chatbot basé sur chatGPT afin de répondre à ses clients.

Merci pour l'accompagnement et le suivi tout au long du stage.

Introduction

La mission principale de mon stage était la création d'un chatbot pour aider et répondre à l'ensemble de ses clients mais en prenant en compte son activité de webscrapping.

L'intérêt du stage a été de découvrir un univers réel avec des techniques particulières puis surtout de créer de bout en bout un nouveau chatbot en prenant en compte les éléments et l'activité de la société.

Le stage s'est déroulé en 3 parties, tout d'abord la formation spécifique aux outils développés par IdIA-Tech, suivi de différentes missions de webscrapping pour des clients en cours, et enfin une fois la compréhension de l'entreprise intégrée, la création d'un chatbot « intelligent » à destination des nouveaux collaborateurs et de ses clients.

Tout d'abord, je vous présenterai la société, puis les missions que j'ai pu effectuer et je vous présenterai mon bilan et ma perception.

L'entreprise d'accueil

La société dans laquelle j'ai fait mon stage se nomme IdIA-Tech :

Il s'agit d'une jeune société : créée en 2020, elle est basée à Montpezat, à côté de Nîmes dans le Gard.

Juridiquement, il s'agit d'une société SASU : Société par Action Simplifiée à associé unique.

Elle est dirigée par Xavier Bour, qui est également le tuteur de mon stage.

Elle est consistée de 1 à 2 salariés et de plusieurs stagiaires qui tournent tous les 2 mois

Son activité principale est la programmation informatique (de manière générale) mais plus particulièrement spécialisée dans le Big Data dédié aux solutions de crawling et d'études des prix et tarifs présents sur le web.

Ses activités principales sont :

- Website crawling solutions
- Catalog import
- Tariff monitoring
- Artificial intelligence
- Amazon-like Solutions

Ses clients sont principalement des petites et moyennes entreprises, quelque fois des start-ups qui veulent démarrer des ventes en ligne de leurs produits sur leur propre site web et non pas via des market places ou autres revendeurs (type Amazon, Cdiscount, …)

Son fonctionnement :

- Pas d'abonnement : c'est justement ce qui intéresse des petits clients qui ne sont pas liés indéfiniment mais qui vont demander un travail en one-shot
- 3 modules principaux de prestations
    - INTELLIGENT IMPORTER :

Permet d'importer des catalogues externes en interprétant les pages web des sites et d'automatiser la future boutique d'un client

- PRICESTRACKER :
  Devenir le concurrent le moins cher et le mieux informé : mission récupérer l'ensemble des prix concurrentiels
- TOPFINDER
  Propose à vos visiteurs des suggestions de produits en fonction de leur comportement sur le site (derniers achats, produits visités, etc.) et améliore le moteur de recherche (correction orthographique...) Inspiré des fonctions d'Amazon : récupération des cookies

Il propose pour un module (PRICE TRACKER) différents niveaux de packs, de MEGA version, Diamond version, Gold version ou Silver version à des prix abordables : entre 150 et 450€ et des forfaits fixes pour les autres.

Les missions et les demandes sont variées selon les clients et leur type de produit. Par contre, les sites sur lesquels les analyses sont effectuées sont souvent les mêmes comme par exemple, Shopify, Amazon, Cdiscount, …

Les travaux sont principalement réalisés via le logiciel GRIMPORT : qui est spécialement utilisé pour faire du web scrapping.

Mes Missions – Le stage en détail

Mon stage s'est déroulé sur 8 semaines du 12 juin au 4 août 2023. Particularité de mon stage, il s'est déroulé entièrement en distanciel. Xavier étant basé à côté de Nîmes, il est habitué à recruter différents étudiants un peu partout en France.

Nous avons eu des points d'avancement régulièrement pour vérifier le travail effectué et surtout répondre à différentes questions que je pouvais me poser au cours des journées.

1ère partie : Formation

La première partie de mon stage a été la formation sur l'entreprise et sur les logiciels utilisés au sein de cette dernière.

Elle s'est déroulée sur deux semaines environ.

J'avais à regarder et étudier un guide de vidéos déjà préparé pour les nouveaux collaborateurs. Ces vidéos étaient accessibles uniquement en privé sur le site de la société IdIA-Tech : l'objectif était de suivre les vidéos et en même temps de réaliser les tests de programmation et d'effectuer les différents programmes.

Le programme le plus important à comprendre et intégrer est Grimport. Il s'agit d'un programme dérivé de Java. Pour le faire fonctionner j'ai dû mettre en place un site PrestaShop sur mon pc.

Grimport est un langage interprété couplé à un crawler. En d'autres termes, avec cet outil, il est possible de parcourir toutes les pages d'un site Internet, et d'exécuter sur chacune des pages un petit programme, appelé "script", codé en Grimport.

Grimport est un langage de programmation léger et simple qui dérive de Groovy. Il est donc très proche de Java et de Javascript.

Grimport est le langage de prédilection des data miners. Avec ce programme il est possible de rendre un site web ou les données du site interactifs. Les données peuvent changer ou évoluer en fonction des informations affichées sur des sites distants qui ne vous appartiennent pas nécessairement. Par contre il sera nécessaire dans certains cas d'avoir besoin ou non de l'autorisation des propriétaires de ces sites. Par exemple, lorsque nous exécutons un script Grimport sur le site de commerce électronique d'un de

des fournisseurs de nos clients, les descriptions de produits et autres informations ne sont pas protégées par le droit d'auteur et peuvent être importées librement (droit des données commerciales).

Ce langage de programme ne m'a pas paru très compliqué de manière générale car il est assez proche de Java.

Cependant, j'ai découvert ce que sont les « regex » : « regular expression ». Elles nous permettent dans Grimport de signaler de manière très précise les éléments que nous souhaitons récupérer, enregistrer puis utiliser.

Via Grimport nous récupérons les éléments directement sur le code source de la page nous utilisons donc une fonction « Select » qui fonctionne comme « ctrl + F » et avec beaucoup plus de restrictions exigées.

Afin de travailler en équipe mais en distanciel, nous utilisions un site privé créé par Xavier de type Dashboard sur lequel je me connectais régulièrement : ce site me permettait d'avoir accès aux autres scripts créés par l'équipe, de voir mes différentes missions à réaliser et que je pouvais commenter ou annoter.

Ces deux premières semaines furent riches en information avec un apprentissage soutenu à la découverte d'une entreprise et de son mode de fonctionnement.

2ème partie : Les Missions de Webscrapping

Une fois le logiciel intégré et utilisé pour des tests, Xavier m'a confié différentes missions pour des clients de son site.

Ma première mission consistait à utiliser le module PriceTracker : le client, Aérographe Discount, recherchait les niveaux de prix de deux de ses concurrents, Passion132 et 1001 Hobbies.

J'ai donc utilisé Grimport pour récupérer certains éléments de produits ciblés (aérosols) sur les deux sites concurrents. Les éléments recherchés étaient :

- Référence
- Nom
- Description
- Prix tarif

A l'aide de ces éléments, le client pouvait mieux positionner ses prix sur son propre site web et comparer son offre produit versus ses mêmes produits mais présentés sur des sites de revendeurs.

Un travail qui m'a occupé environ 2 à 3 jours pour finaliser l'ensemble de cette mission.

Autre mission, ou petit projet, comme le mentionnait Xavier : récupérer pour un client donné l'ensemble des pictogrammes d'un concurrent (durée de garantie, fonctionnement, forme ou design, … tout en fichier image) pour les enregistrer sur le site du client « en dur ». Ce client avait uniquement sur son site des liens web de ces images via le site de ses concurrents.

Le point un peu complexe était de comprendre les produits du client pour pouvoir rechercher un par un l'ensemble des images et les ré-injecter dans son site dans un dossier « Pictogramme ». Le site du client pouvait ainsi être totalement indépendant et fonctionner sans l'aide des sites autres.

Autre mission, récupérer et enregistrer toutes les discussions entre les clients et le site IdIA-Tech de type support technique et « SAV ». J'ai ainsi répertorié toutes les questions et réponses des 2 dernières années environ qui ont été posées via le site web. Cela représentait environ 350 discussions sur des sujets divers d'utilisation des modules et du logiciel.

Toutes ces discussions ont été sauvegardées en fichier « txt » pour une utilisation future que nous verrons dans la 3ème partie.

Cette partie du stage était très intéressante et surtout concrète : les cas clients ont été très différents et il s'agissait de faire une mission réelle : je suis passé de l'univers des aérosols à celui des équipements d'outillage (tournevis et autres outils de bricolage…)

Le point qui m'a le plus impressionné, c'est qu'il est extrêmement facile de récupérer aujourd'hui tout type d'information présentée sur un site web : on peut facilement récupérer l'ensemble des données et

en faire l'analyse (évidemment quand elles ne sont pas protégées par le droit des informations à caractère commercial).

C'est l'automatisation, la rapidité et la précision du langage Grimport pour ce type de tâche qui m'a surpris et fortement étonné.

3ème partie : Création d'un Chatbot

Cette dernière mission a constitué la plus grosse partie de mon stage, la deuxième moitié du stage, environ 4 semaines.

Le projet de créer un Chatbot avait déjà été démarré par Xavier mais n'avait pas encore abouti et le programme ne fonctionnait pas correctement.

Xavier m'a alors demandé de reprendre le projet et surtout de le faire fonctionner.

L'idée est de créer un chatbot sur le site web de IdIA-Tech pour permettre aux clients de poser un maximum de question et d'avoir directement des solutions proposées.

Le chatbot devait se baser sur :

- Toutes les discussions que j'avais récupérées en amont
- Et sur les documents informatifs de la société et des logiciels

La première étape a été de reprendre le projet et de comprendre ce qui ne fonctionnait pas et d'analyser pour essayer de le corriger.

La première version du projet était basée sur PrivateGPT : donc un chatbot existant pour lequel nous devions lui apprendre les différentes ressources que nous avions.

Le chatbot ainsi créé ne fonctionnait pas comme Xavier le souhaitait et ne répondait pas à son besoin : les résultats n'étaient pas concluants suite aux différents tests de PrivateGPT : les réponses n'étaient pas correctes ce qui signifiait que les informations données n'avaient pas été réellement intégrées ni comprises.

Nous sommes donc partis sur une recherche d'une autre solution différente de PrivateGPT.

J'ai commencé à rechercher d'autres solutions qui me semblaient plus viables, sur internet , via des forums de discussions, via YouTube, …. La recherche n'était pas vraiment évidente, je tombais principalement sur des solutions payantes. D'autres problèmes sont apparus comme la taille de nos documents qui ne pouvait convenir à ces outils déjà formatés et limités en taille.

J'ai testé de nombreux programmes de chatbot avec des essais gratuits jusqu'à tester le programme Microsoft de chatbot mais tout était à construire et surtout cette version n'utilisait pas l'intelligence artificielle : le programme était très complexe et je n'ai peut-être pas tout compris dans son utilisation.

A force de rechercher, j'ai trouvé une API LangChain qui m'a semblé intéressante et peut-être utilisable pour notre projet.

Cette API m'a permis de faire le lien entre ChatGPT et l'ensemble des documents (discussions et documents informatifs).

LangChain ne fonctionne qu'en langage Python ou Javascript/typescript. J'ai choisi le langage Python pour développer notre programme. Un des points essentiels auquel j'ai vraiment prêté attention a été de laisser de nombreux commentaires tout au long du programme.

Toutes les informations que j'ai pu trouver+ sur l'API LangChain étaient en anglais et j'ai donc précisé dans l'écriture du programme un maximum d'explications pour les futures améliorations du programme.

Un des problèmes que j'ai rencontrés et qui m'a bloqué pendant un certain temps était la trop grande diversité des ressources : certains éléments n'étaient pas utiles et induisaient en erreur le bot qui bloquait en boucle.

Afin de pouvoir finaliser et surtout présenter une version stable et surtout fiable, j'ai proposé à Xavier une solution « rapide » :

1. Dans un premier temps, supprimer des ressources toutes les discussions des clients récupérées sur le site
2. Et dans un deuxième temps, faire une analyse, une par une de l'ensemble des discussions pour les trier et supprimer celles qui étaient facteur d'erreur pour le bot.

Je suis donc parti sur la version 1 : j'ai pu tester et surtout finaliser avant la fin du stage la première étape du ChatBot en ne prenant en compte que les éléments informatifs du site web et des documents des logiciels.

Le programme fonctionnait mais avait encore un ou deux bugs à résoudre.

Mon objectif lors des derniers jours du stage a été de tester un maximum de configurations pour vérifier et éliminer tout problème. Deuxième objectif, laisser et transmettre au stagiaire qui allait me remplacer toutes les informations dont je disposais et les explications sur le fonctionnement du programme.

Je lui ai également transmis mes différentes suggestions d'amélioration du ChatBot que j'avais imaginé mais sans avoir eu le temps de les implémenter.

Cette dernière mission était vraiment très intéressante :

- J'ai appris à améliorer mes recherches sur le web, trier plus rapidement les différents résultats trouvés et d'en extraire les informations plus pertinentes et en lien direct avec ma recherche
- J'ai créé le chatbot de A à Z, ce qui a été un défi pour moi
- J'ai pu utiliser des nouvelles technologies comme LangChain (création en octobre 2022) et ChatGPT de manière plus précise
- Et au final, avoir un programme qui fonctionne !

Bilan

Le stage était intéressant et instructif : je le referai avec plaisir car cela m'a permis de découvrir un autre pan du Big Data avec des cas concrets de demande de clients.

J'ai également appris un nouveau type de logiciel dans un langage dérivé de Java.

La réalisation de prestation pour des clients « réels » a été attrayante et amenait du concret dans le quotidien.

Le point le plus complexe à gérer pour moi a été le full distanciel : que ce soit du début pour la formation et l'apprentissage, puis sur les journées du quotidien, il m'a été difficile de travailler seul toute la journée sans pouvoir partager soit des commentaires ou des questions / problèmes quelques fois simples mais vu par un œil extérieur peuvent être solutionnés rapidement.

Je pense que j'apprécie d'être en compagnie et d'échanger dans le cadre d'un travail, d'une mission. Le fait de faire partie d'une équipe est pour moi important.

Dans un futur métier, je choisirai un poste qui sera au sein d'une équipe avec un mix de présentiel et de distanciel.

Conclusion

Les 8 semaines de stage ont été réellement intéressantes : après les deux premières semaines d'apprentissage, j'ai pu commencer à travailler avec Grimport sur des missions clients et surtout à finaliser des demandes. La dernière partie du stage m'a permis d'explorer et de créer un chatbot à partir de rien et surtout de laisser un outil aux futurs collaborateurs qui est voué à s'améliorer et se perfectionner.

Le seul bémol, pour ma part, était le full distanciel, pas toujours facile à gérer au quotidien.

Ce stage autour du Big Data me conforte dans la poursuite de mes études : le domaine est si vaste et si plein de possibilités ; même en étant conscient des évolutions quotidiennes, je trouve cela toujours aussi fascinant.

Annexes

Annexe 1 : CV Gustave RICHTER

Annexe 2 : Présentation du Site web de IdIA-Tech

Annexe 3 : Exemple de webscrapping prestashop

Annexe 4 : Exemple de webscrapping sauvegardé en csv

Annexe 5 : Logiciel Grimport

Annexe 6 : Chatbot – code

Annexe 1 : CV Gustave RICHTER

Gustave Richter

Etudiant en 1er année ingénieur à CY-Tech

Étudiant studieux capable de s'autogérer efficacement lors de projets indépendants, ainsi que de collaborer au sein d'une équipe productive.

1 bis Avenue du Doyen Henry Vizioz, 64000 Pau, France
+33 06 52 31 32 33 gustave-richter
gustave.eloire.richter@gmail.com

Expériences professionnelles

Stage, *IdIA-Tech*
06/2023-08/2023 | Nîmes, France
- Web-scrapping réalisation de missions clients
- Création de chatBot pour interne-externe

Stage, *AKORDIA*
05/2022-07/2022 | Pau, France
- Conception ainsi que développement web et mobile

Stage, *ASSA ABLOY*
05/2020-07/2020 | Troyes, France
- Fabrication a la chaîne aux atelier presse et peinture

Stage, *Société ABCO DK*
06/2015-07/2015 | Copenhague, Danemark
- Entretien et installation de salle de cinéma ainsi que de studio cinéma et son
- Programmation et installation électronique

Projet

Barnys' House
2017-2018 | Pau, France
- Automatisation d'une maison à la campagne
- Modélisation de l'installation
- Programmation du système en Python

Organisation

Bénévole, Centre Social Gammes
2018-2020 | Montpellier, France
- Soutien scolaire et accompagnement de collégiens

Langages

Français - *Langue maternelle*
Anglais – *C1*

Loisirs

Loisirs (Jeux vidéos, Lecture SF, Manga, JDR)
Cinéma (Réalisation et montage vidéo)
Sports (Basket, Skateboard, Randonnée)

Education

Ingénieur informatique, *CY-Tech (ex EISTI)*
2021-présent | Pau, France
- Ecole internationale des sciences du traitement de l'information

Préparation technologique intégrée, *EPF*
2018-2021 | Montpellier, France
- Ecole d'ingénieurs

Bac scientifique STI2D, *Lycée Saint Nicolas*
2015-2018 | Paris, France
- Sciences de l'ingénierie
- Spécialité : Informatique des sciences numérique

Langage de Programmation

C ●●●●●
Python ●●●○○
HTML/CSS ●●●●○
JAVA ●●●○○

Soft Skills

Visualisation 3D - *envisager comment une personne ou un objet doit être disposé de manière adéquate dans un espace physique*

Réfléchir de manière créative - *Trouver des approches nouvelles et créatives à des situations ou des problèmes*

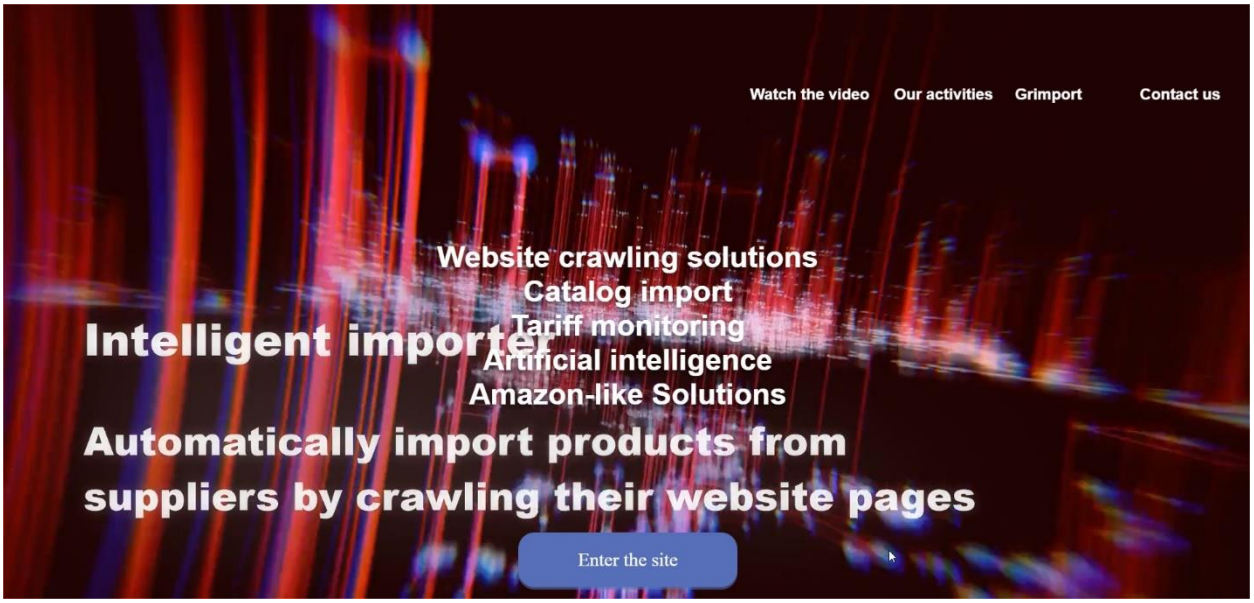
Prendre de nouvelles initiatives - *Accepter de nouvelles responsabilités et rechercher de nouveaux challenges*

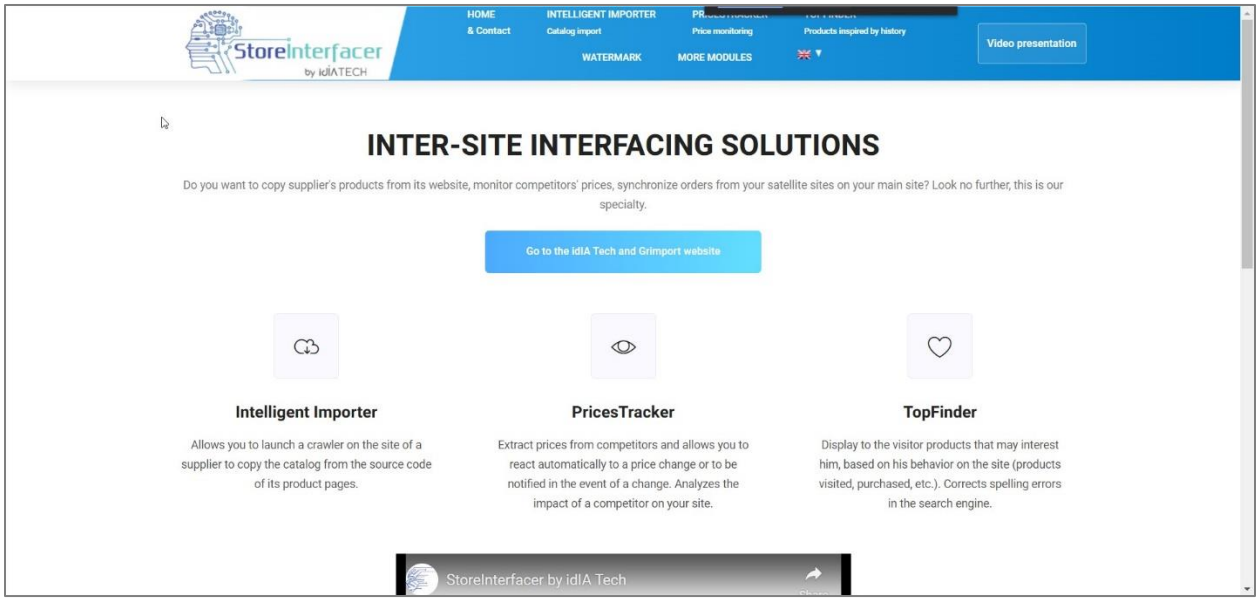
Teamplayer - *Développer la confiance au sein des équipes et renforcer les relations de travail afin d'atteindre des résultats*

Former et apprendre aux autres | Ecoute active

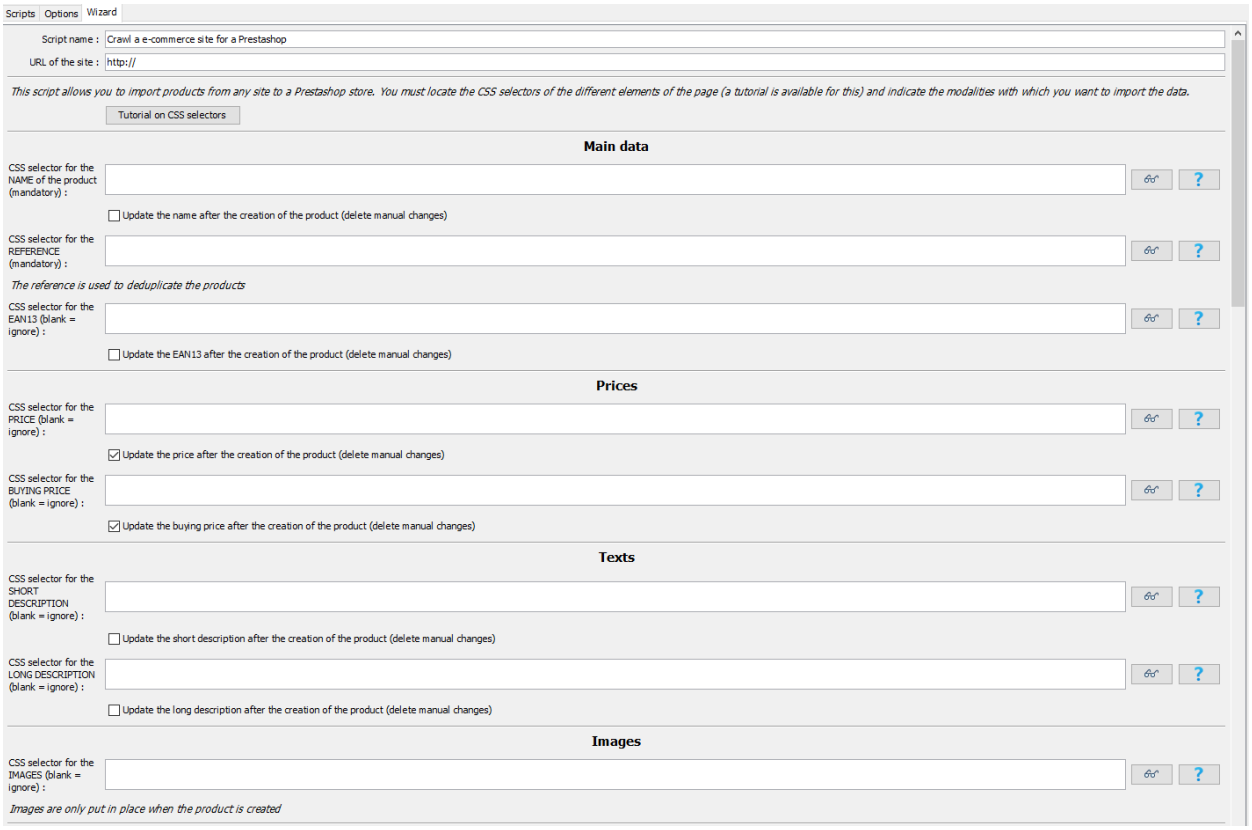
Ouvert d'esprit | Gérer les conflits

Annexe 2 : Présentation du Site web de IdIA-Tech

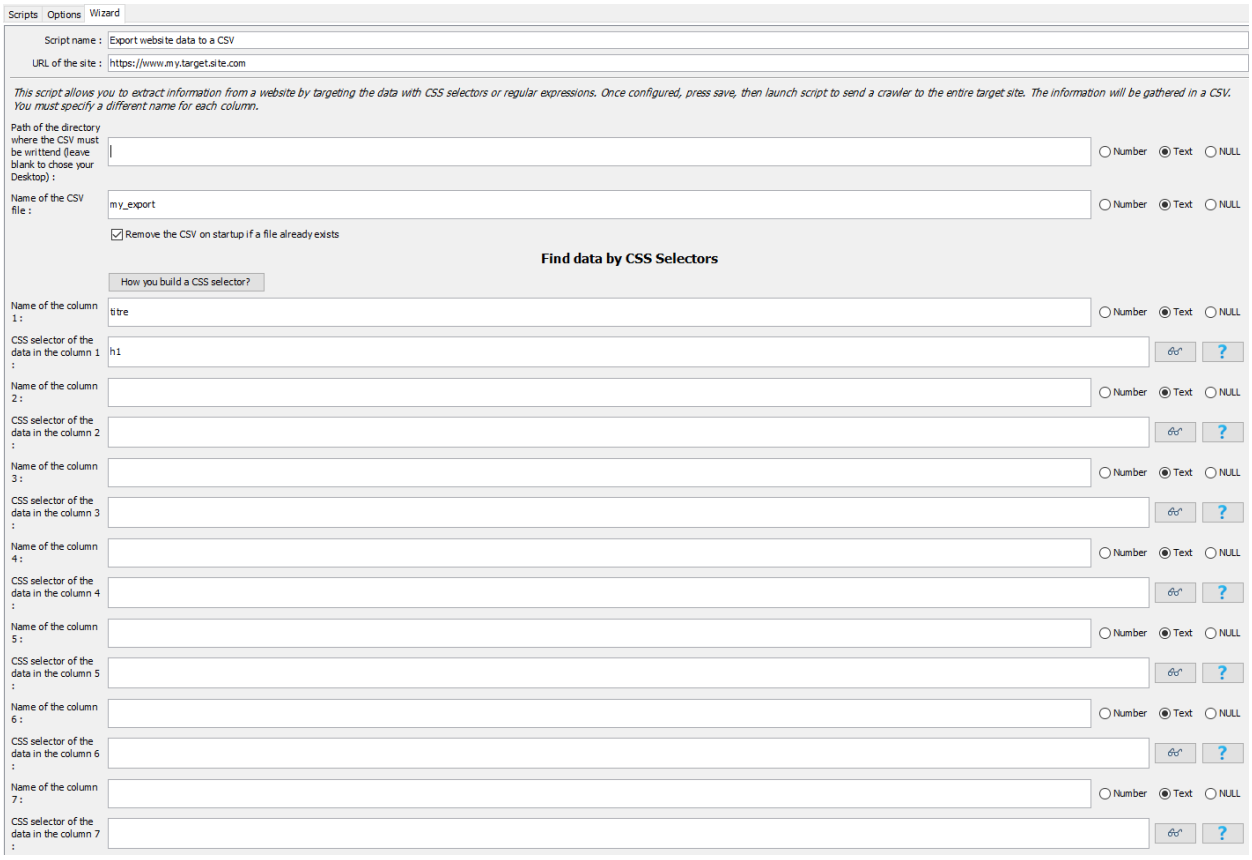



StoreInterfacer
by idIATECH
HOME
& Contact
INTELLIGENT IMPORTER
Catalog import
Price monitoring
Products inspired by history
WATERMARK
MORE MODULES
Video presentation
INTER-SITE INTERFACING SOLUTIONS
Do you want to copy supplier's products from its website, monitor competitors' prices, synchronize orders from your satellite sites on your main site? Look no further, this is our specialty.
Go to the idIA Tech and Grimport website
Intelligent Importer
Allows you to launch a crawler on the site of a supplier to copy the catalog from the source code of its product pages.
PricesTracker
Extract prices from competitors and allows you to react automatically to a price change or to be notified in the event of a change. Analyzes the impact of a competitor on your site.
TopFinder
Display to the visitor products that may interest him, based on his behavior on the site (products visited, purchased, etc.). Corrects spelling errors in the search engine.
StoreInterfacer by idIA Tech

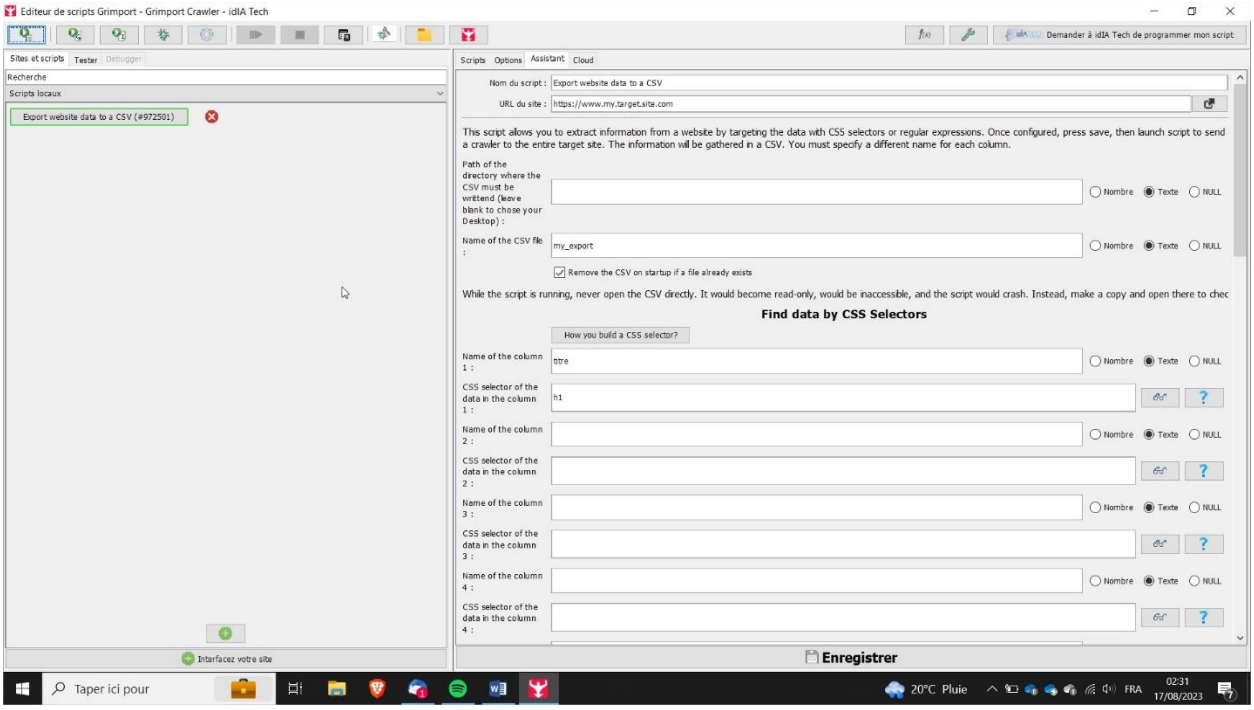
Annexe 3 : Exemple de webscrapping prestashop



Annexe 4 : Exemple de webscrapping sauvegardé en csv



Annexe 5 : Logiciel Grimport



Annexe 6 : Chatbot – code