1 Résumé

Durant les 6 derniers mois, j'ai pu effectuer mon stage de fin d'étude à Sia Partners, en tant que consultant en data science. J'ai eu l'occasion de travailler sur 3 missions différentes :

- Webscrapping de sites d'achat pour Cleep (start-up dans le fond d'investissement de Sia Partners)
- Développement d'un moteur de recommandation pour Cleep
- Mise en place d'un nouveau processus de déploiement d'application pour Sia Partners et developpement fullstack de la palteforme de déploiement lié

1.1 Webscrapping pour Cleep

Cleep est une startup dont l'application éponyme repose sur l'enregistrement de produits vue sur des sites marchands via le lien Internet. Le bon fonctionnement de l'application passe donc par un enregistrement de l'ensemble des données liées aux produits sur internet. Cette récolte de donnée est appelé le webscrapping.

Cleep possédait déjà une solution automatisée pour faire cela, cependant elle n'était pas précise et l'utilisateur devait trier l'ensemble des informations que la solution de Cleep lui donnait. Le but de cette mission était de développer un script de webscrapping spécifique pour les 50 sites les plus populaires sur Cleep et donc agrandir leur base de donnée.

Les résultats de la mission sont concluant, j'ai pu scrapper les 50 sites ce qui a donné à Cleep une base de 220 000 produits scrappés. Cela a permis d'augmenter aussi leur taux de rétention car l'application est devenu moins "longue" à utiliser, car l'utilisateur ne doit plus trier les informations à chaque produit cleepé. En parallèle, j'ai pu aussi développer une méthode générique de scrapping d'image visant à améliorer le scrapping générique déjà en place. Bien qu'efficace, mon scrapper générique fourni tout de même trop de faux positifs (images ne correspondant pas à celle du produit). Cependant la méthode que j'ai employé peut s'avérer efficace en continuant le développement de cette dernière.

1.2 Moteur de recommandation

A la suite du webscrapping, Cleep disposait d'une quantité suffisante de donnée pour pouvoir développer un moteur de recommandation. J'ai pu donc me pencher sur ce sujet en ayant connaissance de l'ensemble des données présentes.

Le moteur de recommandation devenait nécessaire pour Cleep afin d'augmenter le taux de rétention (temps passé sur l'application). En effet l'application ressemblant beaucoup à Pinterest dans le visuel, la manque de moteur de recommandation empêchait les utilisateurs de naviguer sur le site et donc de découvrir de nouveaux cleeps (entité correspondant à un produit enregistré).

J'ai commencé alors par rechercher les différentes types de moteur de recommandation existants sur Internet pour m'interesser à celui de Pinterest. Je me suis fortement inspiré du papier de recherche de Pinterest à ce sujet pour développer le moteur de recommandation, et je l'ai adapté aux donnés de Cleep. Ce moteur repose sur une base de donnée en graphe, du fait de la connexion des données. Le principe repose alors sur le calcul d'un score entre le cleep original et le cleep à recommander, en prenant en compte l'ensemble des informations disponibles sur le chemin les séparant (en terme de traversée du grahe). Si le score obtenu est assez bon par rapport à un seuil, il est alors retenu comme recommandation.

Le moteur de recommandation que j'ai développé est en production et a permis d'augmenter le taux de rétention. Cependant la technique de calcul de score peut, et doit être revu afin de s'adapter au mieux à la population. Plusieurs effets de bords peuvent être perçu notamment à cause de la population de Cleep qui est beaucoup plus féminine faussant alors les recommandations pour les hommes.

1.3 Aide sur le sialab (plateforme et déploiement d'application)

Sia Partners a longtemps travaillé avec une multitude de bots (application avec interface graphique montrant par exemple les résultats d'un modèle en data science). Le développement et le déploiement de ces bots ne suivaient aucuns protocole précis. Ainsi, les bots étaient déployés sur une trentaines de serveur différents, sans contrôle de version et sans backup. Cette structure, en plus d'être inadapté pour le déploiement de bots à grandes échelles avec réplique et backups, impose un point individuel de défaillance (single point of failure) et des conflits internes entre dépendances de bots (librairies, services etc.). En effet les bots sont sur un même serveur et dans le même écosystème. Deux bots ayant des librairies conflictuelles peuvent mettre hors service l'ensemble des bots qu'hébergeait ce serveur.

C'est pour résoudre ces problèmes que j'ai aidé à la mise en place d'un nouveau processus de déploiement. Les consultants developpent maintenant leur bots en se basant sur un template leur permettant de conteneuriser leurs application (grâce à Docker).Le code de ces applications est ensuite versionné puis déployé sur Google Cloud Plateform grâce au module CI/CD (Continuous Integration, Continuous Deployment) de GitLab. De plus en parallèle, j'ai contribué au développement fullstack des plateformes sialab. Ces plateformes offre un moyen d'exposer des API via les serveurs Google Cloud de Sia Partners. Elles offrent aussi le moyen de lancer des scripts (tâches) de manière sécurisé.

Le processus de déploiement a fait ses preuves, et les applications (bots et plateformes sialab) sont déployés de manière continue à chaque changement. Il reste encore plusieurs améliorations à faire notamment au niveau de la lisibilité de développement vis à vis de ce nouveau processus. De plus la migration des anciens bots doit encore ce faire et c'est ce qui risque d'être assez long.

Mots : 880