SOGECAP

STAGE DATA SCIENTIST NLP / COMPUTER VISION

Votre environnement

Sogécap, compagnie d'assurance vie du Groupe Société Générale, recherche un stagiaire *Data Scientist* pour travailler au sein du DataLab. Cette direction (13 personnes) réalise des études statistiques à haute valeur ajoutées auprès d'interlocuteurs variés (autres directions, filiales…). La direction développe également des solutions d'Intelligence Artificielle ayant pour but d'améliorer la connaissance client et l'efficacité opérationnelle des services de gestion de l'entreprise.
Quelques exemples de missions réalisées :

1) Construction et déploiement de modèles de scoring de connaissance client à destination des directions métiers (scores de fraude, scores d'appétence aux différents produits, score de *churn, ...*)
2) Construction et déploiement de solutions d'IA (***NLP, Computer vision et Speech Analytics*****)** pour optimiser les processus dans différents services de gestion : analyse automatique des verbatims clients, catégorisation automatique de emails, extraction d'informations dans des documents scannés, analyse du contenu audio des appels téléphoniques, ...)
3) Travaux R&D : transparence des algorithmes d'IA, travaux avec l'actuariat autour de l'utilisation de nouvelles données pour la tarification (analyse des données télématiques, données météo, modélisation de la probabilité de retard des vols, ...)

Votre rôle

Le but de ce stage est d'étendre les fonctionnalités des modèles d'Intelligence Artificielle existants qui permettent d'analyser et d'interpréter de façon automatique différents types de documents transmis par les clients : **emails, pièces jointes, documents pdfs et formulaires scannés nécessaires à la gestion des contrats.**

Le stage consistera à enrichir l'existant en s'appuyant sur les dernières avancées en **NLP/ Computer Vision.** Les tests seront à mener en particulier sur les modèles de *deep learning* de type ***Transformers,*** *les* modèles **multi modaux** (combinaison texte et image en input du modèle) et **multi tâches** (sorties de type différent en output du modèle). Les solutions à construire feront également usage des principales librairies Open Source sur le sujet (https://huggingface.co/, https://spacy.io/,..), et des services cognitifs de type OCR (Tesseract, solutions cloud, …)
Une étude des optimisations possibles du temps d'inférence des modèles proposés est également attendue au cours du stage.
Les deux sujets spécifiques suivants seront à traiter selon l'avancement du stage :

1) **Extraction de zones d'intérêt et d'informations spécifiques dans des documents/formulaires scannés**

L'objectif est de rajouter au modèle existant la capacité à détecter les casés cochées ou vides), la présence ou l'absence de signatures dans différents types de formulaires scannés. L'extraction et la normalisation d'informations spécifiques (dates, valeurs numériques, ...) est également attendu sur ce premier sujet.

2) **Catégorisation automatique des emails clients et typage des pièces jointes**

L'objectif est d'améliorer le modèle existant en termes de précision de la catégorisation automatique d'une part, et d'identifier d'autre part le type de chaque document fourni en pièce jointe par le client. En particulier, l'objectif est de pouvoir identifier et classifier chaque page contenue dans le pdf (un pdf unique pouvant contenir plusieurs documents distincts)

Votre profil

Etudiant en 2ième ou 3ième année ou d'une formation orientée data science / software engineering vous connaissez le cycle de vie d'un projet *data science* et avez une forte appétence pour le développement d'algorithmes de type NLP ou Computer Vision à l'Etat de l'art. Afin de déployer en production (API, batch) les modèles développés, une connaissance des « bonnes pratiques » du développement software (écriture de code modulaire et documenté, bonnes pratiques de collaboration et de *versioning*, tests unitaires, documentation)

Stack logicielle utilisée

- Requis (un des deux *a minima*) : Python et leurs librairies ML standards (xgboost, scikit-learn, Tensorflow, Transformers,
- Sont un plus : MLflow, Kedro, CI/CD Gitlab

Informations générales

Poste à pourvoir en stage pour une durée de 6 mois, basé à Paris La Défense (92). Télétravail possible 2 jours par semaine.