Stage H/F/X - Document Understanding Deep Learning et IA Documentaire

L'entreprise

Malakoff Humanis est un groupe de protection sociale paritaire, mutualiste et à but non lucratif. Nous sommes ainsi dirigés par des représentants de nos clients. Nous n'avons pas d'actionnaires à rémunérer, l'ensemble de nos bénéfices est donc réinvesti au profit de nos clients, en services, en accompagnement social ou pour soutenir des causes d'intérêt général que nous défendons. Nous sommes un groupe solide financièrement, soucieux d'une gestion rigoureuse et attaché au principe de mutualisation.

L'innovation est depuis toujours au cœur de notre action. Sa raison d'être : « Innover sans cesse au service de l'humain et en faire toujours plus pour protéger et accompagner ses clients entreprises, salariés et retraités. »

Les chiffres clés : N°1 de l'assurance collective (Santé et Prévoyance) en France, 10 millions d'assurés, 426 000 entreprises en assurance collective, 6,3 Milliards de Chiffres d'affaires.

Le Poste

Au sein de la direction Innovation Data et Digital (120 personnes, incubateur / accélérateur des projets pour tout le groupe Malakoff Humanis). Vous intégrerez l'équipe IA Documentaire en charge de développer des solutions IA autour du traitement automatique de documents numériques.

Malakoff Humanis reçoit de très grands volumes de documents de la part de ces assurés par exemple lors des affiliations ou pour des demandes de remboursement. Ces documents sont de formats et de natures très différents (image, pdf, word ; facture, devis, pièce d'identité). Ces documents ne sont pas exploitables directement car l'information présente dans ces derniers n'est pas structurée. Il convient donc d'utiliser des systèmes de parsing tels que les OCR pour détecter et extraire leurs contenus. Malheureusement ces systèmes sont très dépendants de la qualité de l'image (image retourné, présence de bruit, mauvais éclairage) et de leur contenu (manuscrit, tapuscrit, présence de tableau, langue du texte).

Notre ambition : développer de nouvelles approches d'extraction d'information en se passant d'OCR et en se basant sur des méthodes de Deep Learning tels que les CNN ou les Transformers. Dans cette optique, l'objectif de ce stage sera de développer d'améliorer les solutions déjà existantes pour traiter efficacement les documents numériques que reçoit Malakoff Humanis. Nous avons déjà des réalisations sur ce sujet et les résultats sont prometteurs. Ils ne vont pas sans nouveaux défis et nous vous proposons de venir explorer ce nouveau terrain de jeu avec nous.

Vos missions

Au sein du département Data Science et IA, vous rejoignez l'équipe produit IA Documentaire, avec quatre Data Scientists et un Product Manager. Vous interviendrez à différents niveaux du projet :

- Faire une revue de l'état de l'art et des modèles les plus performants pour le Document Understanding (Dit, LayoutLM, Donut)
- Prendre en main la plateforme dataiku et les outils utilisés en interne pour le développement des projets.
- Participer au développement de dataset de références et de vérités terrain (synthétique et/ou réel) qui seront utilisés pour comparer les résultats des différentes approches.
- Pour la construction du dataset, il existe différentes stratégies comme l'utilisation de modèle de génération de données synthétiques (eg Stable diffusion pour les images) ou la génération de texte avec les modèles de langages eg ChatGPT).
- Prendre en main les modèles sélectionnés, calibrer sur le dataset en explorant différentes stratégies de fine tuning.
- Optimiser le meilleur modèle en fonction des indicateurs de performances retenus selon les cas d'usages (temps d'exécution, mémoire occupée, etc).

Le profil

Nous recherchons une personne capable de développer de nouvelles solutions en s'inspirant de méthodes existantes ou qui imaginera de nouveaux algorithmes adaptés.

Vous préparez un bac+5 (ingénieur, master) en vision par ordinateur, machine learning, traitement du signal ou mathématiques appliquées. Vous avez une connaissance minimale de frameworks du domaine (OpenCV, Scikit-Learn, TensorFlow, PyTorch, Keras, ...).

Vous développez vos propres scripts en autonomie (idéalement Python). Nous utilisons Dataiku pour le prototypage et la mise en production des premières versions des produits. Nous travaillons en mode agile. Vous êtes créatif et vous souhaitez travailler sur un des défis du moment en machine learning.

Vous êtes curieux, vous avez envie d'apprendre. Vous avez du plaisir à utiliser des algos pour tester vos idées. Vous n'avez pas peur d'essayer et vous êtes toujours à la recherche du bon équilibre entre rigueur et exploration. Vous êtes autonome et vous pensez que la partie est plus belle quand on joue avec les autres.

Contrat : Stage de 6 mois

Contact : laziz.hamdi@malakoffhumanis.com

Localisation : Paris, 13ème arrondissement – proche BNF et Station F