Offre de stage

Développement d'algorithmes de Natural Language Processing pour la recherche en oncologie

Contacts :

Pr. Xavier TANNIER

Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé

Sorbonne Université

xavier.tannier@sorbonne-universite.fr

Candidature au : https://www.welcometothejungle.com/fr/companies/aphp/jobs/stage-data-science-nlp-oncologie_paris?q=b25c911a484065242d84f2b1dca7a8fc&o=1444933

1 Contexte

L'Entrepôt des Données de Santé (EDS) de l'AP-HP regroupe les données cliniques collectées à l'occasion de la prise en charge de plus de 10 millions de patients. Plus de 200 projets de recherche sont actuellement en cours sur cette base pour réaliser des études épidémiologiques, développer et valider des algorithmes d'IA pour le soin, etc. Parmi les données de l'EDS, les documents cliniques regroupent de nombreuses informations d'intérêt mentionnées par les cliniciens (antécédents, comorbidités, facteurs de risque, symptômes, prescriptions médicamenteuses, etc.). Des algorithmes de *Natural Language Processing* (NLP) sont développés pour extraire automatiquement ces données et les mettre à disposition des chercheurs.

Le projet de recherche CovOnco analyse les données de l'EDS pour étudier les parcours de soins des patients pris en charge en oncologie à l'AP-HP, afin de les optimiser et d'améliorer ainsi l'organisation du système de santé. Si le parcours des patients au sein de l'hôpital est souvent renseigné sous la forme de données structurées, les analyses et actes

thérapeutiques qui ont lieu hors de l'hôpital ne sont souvent disponibles que sous la forme de texte libre à l'intérieur des comptes rendus cliniques. De premières analyses ont démontré la possibilité de détecter et de dater par une approche NLP certains de ces éléments du parcours de soin. L'objectif de ce stage est de poursuivre ces premiers travaux en développant, validant et appliquant des algorithmes de NLP afin de contribuer à diverses études en oncologie. Certaines propriétés des données de l'EDS pourront en particulier être exploitées à l'aide d'une approche par distant learning. Les algorithmes seront partagés en open source avec la communauté, par exemple dans la bibliothèque EDS-NLP (github.com/aphp/edsnlp).

2 Objectifs du stage

- Revue de littérature pour identifier les méthodologies de NLP/ML applicables (distant learning, labels bruités, etc.)
- Exploration des différents algorithmes dans le cas de la recherche CovOnco afin d'extraire différents éléments du parcours de soin (en priorité les dates de biopsie, d'imagerie et les mentions d'apparitions de métastase, en bonus si le temps le permet, les dates de traitement des thérapies orales antitumorales ou les résultats d'examens d'anatomopathologie)
- Intégration des algorithmes dans la bibliothèque open source EDS-NLP (github.com/aphp/edsnlp)
- Le cas échéant, contribuer à la rédaction d'une publication scientifique.

3 Profil recherché

- Ingénieur en fin d'études (spécialisé en *data science* / AI), ou Master de mathématiques appliquées en IA.
- Forte appétence pour le développement *Python*.
- Maitrise de *PyTorch* et *Pandas*.
- Expérience en NLP appréciée (maitrise de *spaCy* appréciée).

4 Modalités pratiques

Le stage sera rattaché au laboratoire LIMICS de Sorbonne Université. Une collaboration étroite avec les équipes de l'AP-HP permettra de bénéficier de l'expertise clinique des médecins référents et de s'inscrire dans la démarche de développement d'algorithmes open

source coordonné par l'équipe Sciences de Données de la Direction des Services Numériques.

Tuteur :

- Pr. Xavier Tannier

Lien fonctionnel :

- Dr. Emmanuelle Kempf, médecin oncologue réalisant une thèse de science en NLP appliqué aux données médicales, AP-HP
- Ariel Cohen, data scientist, Fondation AP-HP
- Romain Bey, responsable de l'équipe Sciences des Données, APHP

Lieu: mi temps LIMICS mi temps Hôpital Rothschild, 33 bvd de Picpus, 75012 Paris. Stage en 100% présentiel.

Gratification : environ 550€/mois

Date et durée : premier semestre 2023, 4 à 6 mois