F/H - Post doctorants : Modélisation de thématiques multi-vues de comptes rendus médicaux

Localisation du poste : Marseille, Aix-Marseille Université (AMU)
Quotité du poste : 100%
Date du recrutement : 01/09/2023
Nature du recrutement : Poste ouvert aux titulaires d'un doctorat, CDD de 1 ans.
Mots clés : Clustering, Topic-Modeling, Multi-vues, Données textuelles, Analyse de dossiers médicaux, Oncologie thoracique, Cancer du poumon, Errance diagnostique

ENVIRONNEMENT ET CONTEXTE DE TRAVAIL :
Aix-Marseille Université (AMU) est une université pluridisciplinaire structurée autour de cinq grands secteurs disciplinaires (Arts, Lettres, Langues et Sciences Humaines ; Droit et Science politique ; Économie et Gestion ; Santé ; Sciences et Technologies et un secteur pluridisciplinaire). AMU forme 80 000 étudiants dans 18 composantes (facultés, écoles, instituts), réparties dans quatre départements (04, 05, 13 et 84) et 10 villes (en savoir plus : www.univ-amu.fr).

Contexte. Avec la quantité sans cesse croissante des données textuelles dans le domaine biomédical, la classification non supervisée (Clustering de textes), qui vise à regrouper un ensemble de textes en classes, est devenue plus critique que jamais. Or, pour regrouper des textes similaires, il est essentiel de saisir leur sens et pour cela de nombreuses représentations textuelles existent, des plus simples, comme Bag-Of-Words (BOW), aux plus complexes tels que les transformeurs récents (Bert, Roberta, etc.).
Objectif. Notre objectif sera de développer des modèles probabilistes multi-vues (qui peuvent considérer un nombre illimité de représentations textuelles), de comparer leur performance à plusieurs algorithmes récents que nous avons développés, et par-delà de développer une métrique d'homogénéité de prise en charge d'une part et une métrique d'errance diagnostique d'autre part.
Méthodologie. Nous mènerons dans un premier temps une étude rétrospective. Les comptes-rendus textuels des consultations de pneumologie du pôle hospitalier n°16 (cardiovasculaire et thoracique) de l'Assistance Publique - Hôpitaux de Marseille depuis 2010 ainsi que les actes PMSI et les pathologies CIM-10 (classification internationale des maladies) seront extraits de la base médico-administrative du département d'informatique médicale (DIM) de l'AP-HM. L'ensemble des données seront bien sûr anonymisées à la source dans le cadre du traitement par le DIM. Nous identifierons au travers du codage CIM-10 les patients qui, au terme de leur histoire médicale observée, relevaient en fait de l'oncologie thoracique. Cela inclura les patients avec un diagnostic avéré de cancer de la plèvre, du parenchyme pulmonaire, de l'œsophage, du thymus etc. Un modèle probabiliste multi-vues de détection des thématiques sera développé. En effet, plusieurs représentations (multi-vues) des comptes-rendus de consultation seront extraites à l'aide d'outils NLP (BioBERT, Biomedical-Entity-Linking, Sytora, etc) destinés aux données biomédicales. Ces représentations reflètent d'une part le sens sémantique contenu dans les comptes-rendus en considérant le contexte, et d'autre part des représentations des pathologies, symptômes, traitements, etc., apparaissant dans le texte des comptes-rendus. Enfin un modèle probabiliste de type topic-modeling recevra en entrée ces représentations multi-vues des rapports médicaux afin de découvrir de manière non-supervisée des classes pertinentes de comptes-rendus. Un clustering des actes PMSI et des pathologies selon la CIM-10 retenus pour chaque patient permettra de proposer une métrique de proximité conjointe de la sémantique des comptes-rendus, des actes réalisés, et des diagnostics retenus ouvrant la voie à l'identification et caractérisation de l'errance diagnostique et de l'homogénéité des prises en charge. Enfin, une métrique reflétant le pronostic médical des patients sera proposée (le décès et la date de décès des patients étant également accessibles et disponibles).

POSITIONNEMENT HIÉRARCHIQUE

Sous l'autorité hiérarchique de la responsable du Projet Rafika Boutalbi. Ce dernier garantira la bonne intégration de l'agent. Il supervisera les actions réalisées et la bonne avancée du programme défini dans les missions.

MISSIONS ET ACTIVITES PRINCIPALES

Missions :

- Revue de littérature des travaux réalisés précédemment pour l'analyse des rapports médicaux.
- Traitement des données : Nettoyage et extraction des différentes représentations utiles pour le modèle.
- Modélisation d'une approche multi-vues pour le clustering des rapports médicaux.
- Implémentation de l'approche précédemment modélisée.

Activités principales :

- Développement de la partie théorique de l'approche de modélisation des thématiques multi-vues.
- Implémentation de l'approche.
- Évaluation de l'approche mise en place à l'aide de mesures existantes et certaines à définir.

Conditions d'exercices :

Il/Elle sera affecté(e) à l'équipe QARMA du LIS, situé sur le campus de Châteaux Gambert
Il/Elle sera amené(e) à se déplacer régulièrement sur le site d'Aix-Marseille ou effectuer des missions dans le cadre de son activité.
Possibilité d'exercer en télétravail (2 jours/semaine) après 6 mois de contrat.

COMPÉTENCES REQUISES :

1. Compétences métiers et/ou techniques :

- Connaissances approfondies du traitement de texte et familiarité avec les différentes représentations de texte, comme les Transformers (BERT, RoBERTa, etc).
- Connaissances approfondies en Machine Learning, notamment en Topic Modeling et clustering.
- Publications de haut niveau dans des conférences et journaux internationaux (SIGIR, NIPS, ACL, KDD, etc)
- Connaissances concernant la détection et la désambiguïsation des entités nommées.
- Avoir été confronté(e), ou avoir travaillé sur des données de santé.

2. Compétences « transverses » :

- Avoir une très bonne maîtrise du langage de programmation Python.
- Être familiarisé avec les problématiques de santé.

3. Savoir-faire et Savoir être requis

- Curiosité scientifique et implication dans le projet
- Sociable et ouvert à la collaboration avec des chercheurs d'autres équipes notamment en santé et NLP.
- Publications de haut niveau dans des conférences et journaux internationaux.

Ce post-doc peut faire l'objet d'une prolongation si les résultats obtenus sont concluants et permettent d'atteindre les objectifs fixés.

Diplôme exigé

Formation : Niveau doctorat domaine Science des données, NLP, Machine Learning
Expérience souhaitée : Une expérience de 2 ans sera un plus.

Le dossier de candidature (CV, lettre de motivation et avec lorsque cela est possible, votre dernier entretien professionnel), devra être adressé, par voie électronique, à : rafika.boutalbi@univ-amu.fr, stephane.delliaux@univ-amu.fr