Classification d'expériences professionnelles LinkedIn par les techniques "few-shot learning"

1. Contexte

LinkedIn compte plus de 25 millions de membres en France, ce qui représente plus de 80% de la population active française. Cette proportion est probablement encore plus importante pour les emplois qualifiés, faisant de LinkedIn une source de données extrêmement intéressante pour les chercheurs en sciences sociales qui étudient les marchés du travail, notamment ceux des individus fortement qualifiés (Bac +5 et plus).

Afin d'étudier les conséquences de l'intelligence artificielle sur la société, le CNRS et un consortium d'universités dont CY Cergy Paris Université ont entrepris de créer la première base de données européenne des travailleurs dans le domaine de l'intelligence artificielle. Dans un premier temps, un premier échantillon de 2500 profils LinkedIn ayant au moins une expérience professionnelle en IA a été constitué en utilisant un bot de scraping. Un des problèmes rencontrés (outre les restrictions de LinkedIn) est que nombre des expériences professionnelles des individus sont inintéressantes pour notre étude (par exemple les utilisateurs renseignent leur job d'été ou leur job étudiant). Nous devons ainsi recourir à des méthodes de classification automatique pour déterminer si une expérience professionnelle peut-être considérée comme étant dans le domaine de l'IA ou non (classification binaire).

Différentes méthodes de classification des expériences professionnelles LinkedIn ont déjà été testées dans ce projet : à partir des mots-clés, à partir des modèles basiques de Machine Learning, et en utilisant des modèles de Deep Learning avec architecture Transformers (Bert et dérivés). Ces derniers modèles sont très satisfaisants du point de vue des résultats obtenus (accuracy supérieure à 95%, F1-score proche de 90%), mais nécessitent un grand nombre de données d'exemple (1500 expériences LinkedIn) pour constituer le training set, et donc un effort de labellisation conséquent. Si nous nous intéressons à d'autres emplois ou compétences que ceux dans l'intelligence artificielle, cela nécessitera de constituer et de labelliser à nouveau un nombre très important d'exemples.

Par conséquent, nous aimerions tester d'autres modèles d'intelligence artificielle qui peuvent donner des résultats satisfaisants mais nécessitant un effort de labellisation bien moins important. Les techniques dites de « few-shot learning » permettraient éventuellement d'atteindre cet objectif. C'est ce que vous allez devoir implémenter dans ce projet.

2. Objectifs

Les objectifs de ce projet sont les suivants :

- Réaliser une revue de la littérature des techniques few-shot learning
- Mettre en œuvre ces techniques sur la base de données composée d'environ 10 000 expériences professionnelles déjà labellisées. L'utilisation du langage de programmation Python est fortement recommandé.
- Interpréter les résultats obtenus et les comparer avec ceux donnés par les autres méthodes de classification déjà testées.

Vous serez éventuellement amené à mettre en pratique ces techniques sur une autre base de données-fournie par un autre client, après avoir signé un accord de non divulgation sur l'entreprise et le contenu de la mission.

3. Partenaires et Equipe

Le laboratoire de CY Cergy Paris Université en Economie (Thema) ainsi que le CNRS sont partenaires de ce projet. L'équipe est composée de :

- Olha Nahorna, ingénieure recherche CNRS
- François Maublanc, enseignant-chercheur en Economie à CY Cergy Paris Université

Vous serez éventuellement amené à interagir avec le client cité.