Sujet de stage niveau M2

BPCE-LIPN (UMR CNRS 7030)

Zero/Few shot learning appliqué à la classification de texte (NLP)

— **Mots-clés :** NLP, Zero Shot Learning, Few shot Learning, Deep Learning, Azure Cloud, classification de documents, analyse de sentiment.
— Possibilité de poursuivre en thèse CIFRE avec BPCE et le LIPN.
— Lieux du stage : la BPCE,

1 Entreprise & Équipe

Le Groupe BPCE, deuxième groupe bancaire français issu de la fusion des Caisses d'Épargne et des Banques Populaires est un groupe diversifié regroupant, outre des activités de banque de détail, des filiales spécialisées, la banque de grande clientèle Natixis, des activités de gestion d'actifs, de banque privée et d'assurance, en France et à l'international.

Le Secrétariat Général du Groupe (SGG) rassemble les directions de la conformité, de la sécurité, du juridique, de la gouvernance, des contrôles permanents, des affaires publiques et enfin une direction transverse de l'Innovation et de la Data Intelligence.

Cette dernière est en charge des outils nécessaires aux domaines métiers relevant du SGG et de l'adoption des techniques les plus récentes de traitement des données et d'intelligence augmentée afin de permettre un meilleur accomplissement de ses missions.

Elle est composée de 3 pôles :

1. Intelligence Artificielle et Modèles
2. Data Factory
3. Pilotage Projets

2 Description du sujet

Le Secrétariat Général Groupe recrute un Data Scientist stagiaire pour le pôle Intelligence Artificielle et Modèles, avec la perspective de poursuite en thèse CIFRE. Cette équipe a pour mission d'identifier et de développer les solutions d'intelligence artificielle et les modèles statistiques relatifs à la conformité, aux contrôles permanents, au juridique, à la gouvernance et à la gestion des risques.

Dans le cadre des travaux en collaboration avec la Direction des Risques, l'équipe travaille sur un projet de création d'indicateurs pour la surveillance du risque de crédit des portefeuilles clients « corporate ». Il s'agit de générer de l'information pertinente à partir de données non structurées, notamment de données texte et de la restituer via une interface web.

Les technologies de traitement automatique du langage naturel sont actuellement très performantes, à condition de disposer d'une grande quantité de donnée annotées, ce qui est rare car la majorité des données disponibles ne le sont pas. Une approche courante pour traiter cette problématique est l'annotation manuelle, mais celle-ci peut être chronophage et est spécifique à un cas d'usage (non généralisable). Elle n'est donc pas convenable pour une utilisation «industrielle».

Afin de répondre à ce problème, nous souhaitons mettre en œuvre des méthodes de type Zero ou Few Shot Learning et les adapter à nos problématiques de classification et d'analyse de sentiment. Ces approches permettent de détecter/classer des documents à partir de seulement quelques exemples d'une classe et sans aucun réglage fin.

L'idée principale est inspirée des réseaux [SSZ17 ; Wan+20] qui apprennent une fonction de coût qui fait correspondre les images dans le cas du papier de recherche, dans un espace de représentation. Des prototypes sont calculés à partir des quelques exemples disponibles pour chaque classe et des scores de classification sont attribués à chaque observation d'entrée en fonction des distances entre son *embedding* et les prototypes.

Nous nous intéressons dans ce stage à deux cas d'applications : la classification d'articles de presse et l'analyse des sentiments associés à ces articles.

Le stage se déroulera en plusieurs phases :

1. Étudier l'état de l'art sur l'apprentissage Zero/Few Shot learning appliqué au langage naturel.
2. Explorer et analyser notre corpus de texte qui contient plus de 5 000 000 d'articles de presse en français et presque 42 000 000 d'articles en anglais.
3. Sur la base des études précédentes, implémenter un algorithme de classification basé sur le few-shot learning dans une logique « Human-in-the-loop ». D'autres méthodes de classification basées sur le few-shot peuvent être étudiées et appliquées pendant le stage.
4. Selon l'avancement, proposer un algorithme d'analyse plus fine des textes en proposant une modélisation jointe des deux tâches : classification des textes et analyse des sentiments.

Ces travaux pourront s'appuyer sur l'interactivité de l'interface web actuellement déployée à une panel d'analystes afin d'obtenir des annotations.

Les résultats obtenus pendant le stage peuvent conduire à des contributions à des logiciels libres, voire à une publication scientifique, en fonction des compétences et de la motivation du/de la stagiaire.

3 Environnement technique

Les développements seront menés sur une plateforme cloud (Azure/Microsoft), dans l'environnement Azure machine learning, en langage Python.

Ce stage permettra au candidat d'acquérir des connaissances à la fois techniques (NLP, Machine Learning/Deep Learning, Python, Cloud Azure) et fonctionnelles ainsi qu'une bonne vision de l'environnement bancaire.

4 Collaboration avec le laboratoire de recherche LIPN

Le sujet du stage possède deux aspects : le premier est un aspect fondamental qui consiste à développer un nouveaux algorithmes Zero ou Few Shot learning voir les adapter aux problématiques de classification et d'analyse de sentiment des sources textuelles du Groupe BPCE. Le second est un aspect applicatif qui entre dans le cadre des activités du Groupe BPCE. Ces deux aspects permettront des collaborations et des échanges avec l'entité R&D et l'équipe encadrante universitaire du laboratoire LIPN (UMR CNRS 7030).

Par conséquent, l'autre objectif de ce stage est de permettre au Groupe BPCE d'approfondir les compétences dans les domaines de l'apprentissage et du deep learning pour l'analyse de données textuelles. Le laboratoire LIPN, particulièrement les deux équipes encadrantes RCLN et A3 seront impliquées dans le suivi et la définition de la direction de recherche. L'équipe RCLN (Représentation des Connaissances et Langage Naturel) s'intéresse au langage pour son pouvoir expressif et à la Représentation des Connaissances pour le Traitement Automatique des Langues. L'équipe A3 (Apprentissage Artificiel et Applications) développe des algorithmes dans le domaine d'apprentissage automatique et de l'intelligence artificielle pour la fouille de grandes masses de données de natures différentes (continues, binaires, mixtes).

Le sujet de stage sera suivi potentiellement d'une proposition de poursuite des travaux dans le cadre d'une thèse CIFRE financée par le Groupe BPCE en collaboration avec le laboratoire LIPN UMR CNRS 7030, selon le profil du candidat.

5 Profil et compétences requises

Vous êtes étudiant de dernière année d'une formation scientifique de type école d'ingénieur ou universitaire, avec une spécialisation en statistiques, Data Science ou Big Data.

Vous faites preuve de curiosité et d'une appétence pour la recherche, d'une grande rigueur et d'un sens critique développé, et disposez de fortes capacités analytiques. Les problématiques bancaires vous intéressent.

Vous disposez d'un bon relationnel et d'un très bon niveau de synthèse rédactionnelle.

6 Postuler

Vous souhaitez postuler ? Envoyez par mail votre candidature (CV, Lettre de Motivation & Relevés de notes) à l'adresse de l'équipe (tomeh@lipn.fr, mustapha.lebbah@univ-paris13.fr,charnois@lipn.univ-paris13.fr) en rappelant la référence de l'offre de stage [Zero/Few Shot Learning NLP-22] dans l'objet du courriel.

Références

[SSZ17] Jake Snell, Kevin Swersky et Richard Zemel. "Prototypical Networks for Few-shot Learning". In : *Advances in Neural Information Processing Systems*. Sous la dir. d'I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett. T. 30. Curran Associates, Inc., 2017. URL : https://proceedings.neurips.cc/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf.

[Wan+20] Yaqing Wang, Quanming Yao, James T Kwok et Lionel M Ni. "Generalizing from a few examples : A survey on few-shot learning". In : *ACM Computing Surveys* 53.3 (2020), p. 1-34.