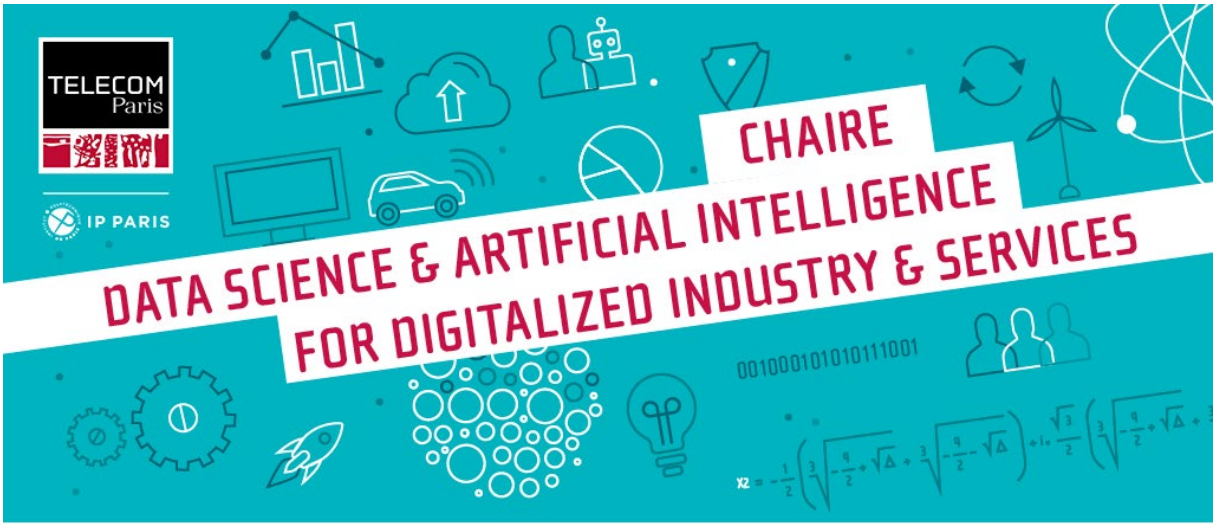

Offre de stage

Signaux Faibles et Données Textuelles

Possibilité de poursuivre sur une thèse

A propos de la chaire

La Chaire Data Science and Artificial Intelligence for Digitalized Industry and Services, portée par Florence d'Alché-Buc, enseignante-chercheuse dans le département Image, Données, Signal de Télécom Paris, la chaire DSAIDIS réunit cinq partenaires industriels : Airbus Defence & Space, Engie, Idemia, Safran et Valeo. Son objectif général est de développer, en liaison étroite avec les partenaires, une formation et une recherche de niveau international.

Ses quatre principaux axes de recherche sont :

1. Analyse et prévision de séries temporelles (Predictive Analytics on Time Series) ;

2. Exploitation de données hétérogènes, massives et partiellement étiquetées (Exploiting Large Scale and Heterogeneous, Partially Labelled Data) ;

3. Apprentissage pour une prise de décision robuste et fiable (Learning for Trusted and Robust Decision) ;

4. Apprentissage dans un environnement dynamique (Learning through Interactions with a Changing Environment).

Plus d'information sur la chaire : **www.telecom-paris.fr/dsaidis**

Description du stage

Encadrement

Stephan Clémençon, Matthieu Labeau

Lieu et dates du stage

Adresse : Télécom Paris, 19 Place Marguerite Perey, 91120 Palaiseau

Date de début du stage : début 2020

Équipe d'accueil de la thèse

Département IDS, équipe S2A

Mots clés

Evénements rares, théorie des valeurs extrêmes, traitement automatique des langues, plongement lexical

Sujet détaillé

L'analyse des données textuelles est un domaine de recherche en plein essor, poussé d'une part, par les enjeux applicatifs du domaine (recommandation, e-reputation, gestion de la relation client, cybersécurité), et d'autre part, par la multiplication des plate-formes d'expression des citoyens et des médias sur le web. Dans ce contexte, des techniques de représentation de l'information textuelle ont été récemment développées dans la perspective d'automatiser certaines tâches (e.g. reconnaissance du type d'opinion exprimée, traduction automatique). Les méthodes les plus populaires sont basées sur la réduction de dimension appliquée aux statistiques de co-occurrences des mots dans un texte [5,6]. Il existe des modélisations visant à généraliser ces méthodes afin d'étudier les propriétés géométriques de ces représentations de mots ('plongements lexicaux')[7]. En pratique, il semble que l'utilisation des statistiques de co-occurrences conduit les algorithmes à encoder d'autres informations que sémantiques. Ainsi, les composantes principales de ces plongements encodent simplement la fréquence des mots représentés [8].

C'est un problème puisque sur le plan applicatif, un des enjeux majeurs est de pouvoir détecter les signaux faibles. L'objectif du stage sera de développer des méthodes de représentation de mots de manière à faciliter la détection de signaux faibles sur des données textuelles. Le point de départ sera de s'inspirer des approches et critères fondées sur la théorie des valeurs extrêmes ayant permis d'étendre les techniques d'apprentissage supervisé (e.g. classification) ou non supervisé, voir [1], [2], [3] ou [4].

Profil du candidat

Etudiant titulaire d'un master 2 recherche

- Apprentissage statistique / reconnaissance des formes

- Traitement du langage naturel

- Bon niveau en programmation (Java, C/C++, Python)

- Bon niveau d'anglais

Candidatures

A envoyer à matthieu.labeau@telecom-paris.fr, stephan.clemencon@telecom-paris.fr

- Curriculum Vitae

- Lettre de motivation personnalisée expliquant l'intérêt du candidat sur le sujet (directement dans le corps du mail)

- Relevés de notes des années précédentes

- Contact d'une personne de référence

Les candidatures incomplètes ne seront pas examinées.

Références

[1] A Multivariate Extreme Value Theory Approach to Anomaly Clustering and Visualization. S. Clémençon, M. Chiapino, V. Feuillard and A. Sabourin. 2019, In Computational Statistics.

[2] On Binary Classification in Extreme Regions. S. Clémençon, H. Jalalzai and A. Sabourin. In the Proceedings of NIPS, Montréal, Canada, 2018.

[3] Anomaly Detection in Extreme Regions via Empirical MV-sets on the Sphere. S. Clémençon, A. Thomas, A. Sabourin & A. Gramfort. In the Proceedings of AISTATS 2017, Fort Lauderdale, USA.

[4] Sparse Representation of Multivariate Extremes with Applications to Anomaly Detection. S. Clémençon, N. Goix & A. Sabourin. In Journal of Multivariate Analysis, 2017.

[5] Distributed Representations of Words and Phrases and their Compositionality. T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado & J. Dean. In Neural Information Processing Systems, 2013.

[6] GloVe: Global Vectors for Word Representation. J. Pennington, R. Socher & C. D. Manning. In mpirical Methods in Natural Language Processing (EMNLP), 2014.

[7] A Latent Variable Model Approach to {PMI}-based Word Embeddings. A. Sanjeev, L. Yuanzhi, L. Yingyu, M. Tengyu & R. Andre. In Transactions of the Association for Computational Linguistics, 2016.

[8] All-but-the-top : Simple and Effective Post-processing for Word Representations. J. Mu & P. Viswanath. In the International Conference on Learning Representations, 2018.