Deep Learning for entity extraction in noisy environment

Information extraction from rich unstructured documents is a complex but key process in various businesses. For years, Analytics Consulting Datalab integrated within BNP Paribas CIB has been developing a platform that allows documents to be labeled but also to automatically retrieve the output of models. More than 20 businesses are currently integrated and use the platform.

In this context, the AI Lab developed state-of-the-art approaches based on pretrained language model transformers in order to provide relevant automatic extraction [1].
However, the team faces two industrial challenges:

- Onboarding new businesses and creating large dataset are costly
- The difficulty of integrating high quality dataset (some document can be partially labeled, some labels may be imprecise)

In this ecosystem, the intern will apply state-of-the-art **Active Learning** in the context of information extraction from documents. Meaning **optimizing the documents to label** in order to improve model performances [1].
During the project, the interns will conduct the following research:

- They will adapt **active learning scorers** to the task [2] and benchmark it on real world datasets
- in complement they will challenge the approach with a pipeline based on **documents clustering** [3] based on the assumption that having too many documents that are similar is less meaningful than having totally different documents
- They will explore the usage of **advanced loss functions in order to better handle noisy labels** [4]

The project will allow the intern to gain in expertise in various domain especially: document processing, advanced deep learning, computer vision and natural language processing.
This industrial research will have a direct impact on all the use cases of the platform in production.

References:

[1] LayoutLM: Pre-training of Text and Layout for Document Image Understanding Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, KDD 2020

[2] Scalable Active Learning for Object Detection, Elmar Haussmann, Michele Fenzi, Kashyap Chitta, Jan Ivanecky, Hanson Xu, Donna Roy, Akshita Mittel, Nicolas Koumchatzky, Clement Farabet, Jose M. Alvarez

[3] Self-supervised Document Clustering Based on BERT with Data Augment, Haoxiang Shi, Cen Wang

[4] Dice Loss for Data-imbalanced NLP Tasks, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, Jiwei Li

Practical Information:

Supervisor: RATNAMOGAN Pirashanth, Data Scientist, pirashanth.ratnamogan@bnpparibas.com

Location: Grands Moulins de Pantin, 9 Rue du Débarcadère, 93500 Pantin (Possible partial work from home)
Compensation: Competitive salary
How to apply: Please submit your application at paris.cib.analytics.consulting.careers@bnpparibas.com