Deep Learning based Table Structure Recognition

Information extraction from rich unstructured documents is a complex but key process in various businesses. Within BNP Paribas CIB, Analytics Consulting Datalab has been developing a platform to automatically label documents with NLP models and retrieve user feedback. More than 20 businesses use this platform.

In this context, the team tries to get an **overall understanding of the document by extracting its layout** (tables, paragraphs, titles, headers, footers ...). While some preliminary work has been done on cutting-edge algorithms like **Faster R-CNN** [0] or **CascadeTabNet** [1], the goal of the internship will be to push the internal research on the subject.

The first objective will be to explore various other approaches in particular the usage of CascadeRCNN implemented in the package **deepdoctection** [2] and **pre-trained Transformer-based approaches** [3].

The second objective will be to package and implement a production ready micro service that could be used in the platform.

The internship will allow the student to gain in expertise in various domain especially: document processing, advanced deep learning, computer vision, natural language processing, python and java development.

This industrial research will have a direct impact on all the use cases of the platform in production.

References:

[0] Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, NIPS 2015
[1] CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents, Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, Kavita Sultanpure, CVPR2020 Workshop on Text and Documents in the Deep Learning Era
[2] https://github.com/deepdoctection/deepdoctection
[3] Dit: Self-supervised pre-training for document image transformer, J Li, Y Xu, T Lv, L Cui, C Zhang, F Wei, ACM Multimedia 2022

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