# Efficient AI models for Document Retrieval

In large organizations like BNP Paribas, many documents are created and made available for the employees. Being able to search this massive quantity of documents efficiently is essential for practical and for regulatory reasons. The corresponding task in **Natural Language Processing** (NLP) is called **Document retrieval**. It consists of proposing a list of ordered documents from a large corpus that answers the given query expressed in natural language.

In this context, Analytics Consulting CIB has developed a **Search** tool that responds to various search requests from employees and business. This tool is based on classical IR approaches such as **TF-IDF** and **BM25**[1] that works at syntax-level. However, the team faces two industrial challenges:

- Classical search methods fails to capture vocabulary misalignment. Using methods based on **Document Text Embedding** and **Deep Learning** could enhance the Search results.
- The difficulties related to doing search operation across multiple data sources or on ones that include very large files: Both cases can return sub-optimal results as proven by experience.

As part of the CIB Analytics Consulting team, you will have multiple objectives:

- Exploring state-of-the-art models including **Language Models and Transformer-based models** [2][3] and applying them to the **MS MARCO** [4][5] dataset for Document Retrieval tasks as well as internal datasets.
- Integrating these models to the **Search pipeline** starting by the simpler ones.
- Improving the overall search performances on difficult tasks including a better handling of multiple sources and very large files
- Contributing to the daily evolution of the Search tool

This internship combines **Academic Research** and **practical industrial development of AI tools**. By working on a new Document-retrieval model, the student will improve their theoretical and practical skills **in Advanced Deep learning** technique and **NLP**. As a member of the Search team, the student will also have the opportunity to gain expertise in the development and industrialization of AI models.

**References:**

[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.

[2] Ma, X., Guo, J., Zhang, R., Fan, Y., Ji, X., & Cheng, X. (2020). PROP: Pre-training with Representative Words Prediction for Ad-hoc Retrieval.

[3] Zhang, J., Liu, Z., Han, W., Xiao, S., Zheng, R., Shao, Y., Sun, H., Zhu, H., Srinivasan, P., Deng, D., Zhang, Q., & Xie, X. (2022). Uni-Retriever: Towards Learning The Unified Embedding Based Retriever in Bing Sponsored Search.

[4] MS MARCO: A Human-Generated MAchine Reading COmprehension Dataset

[5] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, & Ellen M. Voorhees (2020). Overview of the TREC 2019 deep learning track. CoRR, abs/2003.07820.

**Practical Information:**

**Supervisor:** HAJAIEJ Mhamed, Data Scientist, mhamed.hajaiej@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