SIREA number:

Title: Postdoctoral researcher in artificial intelligence and natural language processing (SCAI/BnF research program)

Body: Postdoctoral contract (12 months, renewable)

Attachment: UMR 7222 ISIR

Keywords: machine learning, explainability, databases, computer science, applied mathematics, statistics, natural language processing, recommendation

The activities that make up the job description are called upon to evolve according to knowledge of the profession and the needs of the service.

About us

Sorbonne University is a multidisciplinary research university created on January 1, 2018 by the merger of Paris-Sorbonne and UPMC universities. Deploying its training to 54,000 students, including 4,700 doctoral students and 10,200 foreign students, it employs 6,300 teachers, teacher-researchers and researchers and 4,900 library, administrative, technical, social and health staff. Its budget is €670 million. Sorbonne University has top-notch potential, mainly located in the heart of Paris, and is extending its presence to more than twenty sites in Île-de-France and in the regions. Sorbonne University has an original organization into three Faculties of Letters, Science & Engineering and Medicine which have significant autonomy in implementing the university's strategy within their perimeter on the basis of a contract of objectives and means. University governance is primarily dedicated to promoting the university's strategy, steering, developing partnerships and diversifying resources.

Presentation of the structure

In a national and international context marked by competition around artificial intelligence, Sorbonne University has created the "Sorbonne Center for Artificial Intelligence" (SCAI), which brings together in a single place, located in the heart of the Latin Quarter, a strategic range of disciplines of modern artificial intelligence. SCAI's ambition is to contribute significantly to the excellence of interdisciplinary research in artificial intelligence by promoting exchanges between teacher-researchers, researchers, teachers, students and industrialists.

The research project described below is part of the strategic partnership between Sorbonne University and the BnF, which brings together in this specific context the expertise of the MLIA team from ISIR at the BnF in order to develop joint research work on the subject of recommender systems.

The National Library of France (BnF) is one of the largest heritage libraries in the world. Its mission is to collect, catalogue, preserve, enrich and communicate the national documentary heritage. Engaged for many years in ambitious programs to digitize its collections, to which is now added the massive entry of natively digital collections, the BnF continues to enrich its digital heritage. Its mass, diversity and rate of growth require new processing and consultation tools. To allow as many people as possible to discover and appropriate this heritage, the BnF has been involved for several years in artificial intelligence (AI) technologies.

Mission and main activities

Project description
Gallica, the BnF's digital library, brings together more than 10 million digitized documents freely accessible online (18.5 million visits per year). However, most users do not know that Gallica contains printed documents, but also photographs, sound recordings, videos or even 3D objects. In satisfaction surveys, only a minority consider that the search engine's answers are relevant and a majority would like to be better guided in their searches. In a context where the BnF has just adopted a new entity-relationship oriented data model to describe its collections (the LRM model) and is also considering an overhaul of the Gallica search engine, the matter is to question how the use of an appropriate language model can help users find their way around the mass of collections and improve the visibility of the least known content. In this project, the BnF undertakes to adopt a resolutely ethical approach. The potential exploitation of user logs must respect their privacy and guarantee both the relevance and transparency of the algorithms, avoiding the risk of filter bubbles. Three lines of thought emerge:

1) based on available information, including both user logs and document data (full text by OCR and descriptive metadata), how to assess similarity between documents?
2) how to disambiguate user queries and make them appropriate suggestions thanks to a conversational agent using artificial intelligence?
3) how to establish user confidence in the design and auditing of algorithms?

Main missions
This project consists of working on access to information in the Gallica library, from the point of view of machine and deep learning techniques. The research axes concern (1) the analysis and indexing of textual documents as well as (2) the analysis of user traces and (3) recommendation systems. We will focus in particular on multimodal techniques that allow to contextualize a document or a request from user interactions.

The successful candidate will be responsible for:

- Implement models to learn the semantics of textual data in order to vectorize them.
- Develop algorithms based on representation learning methodologies to effectively mix text and user traces.
- Use a large language model to generate relevant questions for the user.
- Report and present development work in a clear and effective way, both to discuss with BnF experts and to write machine learning publications.

The collection of printed books will be primarily targeted by the program described above, but an extension to other collections equipped with textual descriptors (in particular iconographic collections) could be envisaged.

Training:
A doctoral degree in computer science or equivalent is required, as well as a solid scientific background, particularly in NLP and/or Recommender Systems and/or Information Retrieval. Experience in international research projects and SHS applications would be an asset.

General informations:
Places: Campus Pierre and Marie Curie of Sorbonne University and Datalab of the BnF
Contract: fixed term of 12 months with possibility of extension
Expected date of employment: as soon as possible
Work shift: full time
Desired experience: 1 to 3 years
Salary according to experience

Main contacts:
Laure Soulier, MCF in computer science at Sorbonne University, MLIA team, ISIR.
Vincent Guigue
Lucie Termignon, data and artificial intelligence project manager at the BnF.
Jean-Philippe Moreux, Scientific expert of Gallica at the BnF.

Team management: NO
Project management: YES

Knowledge and skills

A solid background in natural language processing or text analysis is essential, and good programming skills are required. Experience with recommender systems is assumed. An understanding of the ethical issues of such systems is also expected.
Language: knowledge of French not mandatory but strongly desired

Applications (CV + motivations + possible references) should be sent by email to
xavier.fresquet@sorbonne-universite.fr with copy to philippe.chevallier@bnf.fr and laure.soulier@isir.upmc.fr