ICT PHD

Research project for a PhD curriculum in ICT – Computer Engineering and Science

**Tutor**: prof. Sonia Bergamaschi

**(\*) Italian [Industrial] Co-tutor:** dott. Riccardo Martoglia (DataRiver)

 **(\*\*) Foreign Co-tutor:**

**Proposed Title of the research:** Exploring the Potential of Large Language Models for Multilingual Historical Document Analysis and Semantic Cataloguing in Digital Libraries.

**Keywords: (5)**

Digital libraries, Large Language Models (LLMs), distributed machine learning (ML) models, Semantic information extraction, Chat-GPT.

**Research objectives: --(max 10 rows)**

This research aims to explore the potential of Large Language Models (LLMs) based on transformers for analyzing content in multilingual historical documents. Specifically, it seeks to investigate the state of the art in the use of LLMs in Digital Libraries and their ability to extract semantic information for the development of semantic cataloguing systems. In particular, this research will face the challenge posed by the ITSERR project, where extracting semantic information from non-Latin documents presents a non-trivial task. The project will specifically assess the current limitations of chat-GPT applications and propose strategies to incorporate domain knowledge into the system. Furthermore, the study will go beyond GPT to investigate other LLMs, both proprietary and non-proprietary, such as BLOOM, to find possible areas of innovation. The proposed challenges will focus both on effectiveness and efficiency/scalability issues.

**Proposed research activity -- (max 10 rows)**

* Conduct an extensive review of the literature on the use of LLMs in Digital Libraries, with a focus on both proprietary and non-proprietary LLMs.
* Examine the present limitations of chat-GPT applications and propose methods for incorporating domain knowledge into the system (semantic information extraction).
* Using LLMs, create an effective and efficient framework for content analysis of multilingual historical documents, exploiting big data and distributed ML
* Apply the suggested approach to a wide range of multilingual historical documents to evaluate its effectiveness and find areas for improvement.
* Implement the proposed algorithms on platform for distributed processing of Big Data that provides high scalability, such as Apache Spark.

**Supporting research projects (and Department).**

The research will be conducted in the context of the PNRR project Italian Strengthening of Esfri RI Resilience (ITSERR) funded by the European Union – NextGenerationEU.

**Possible connections with research groups, companies, universities.**

International research group coordinated by "Fondazione per le scienze religiose di Bologna" of the "Big Data, Artificial Intelligence and Religious Studies" research line, in collaboration with Biblioteca La Pira, Palermo, and IDEO (Cairo).

 (\*) optional

(\*\*) optional/to be completed on the second year