INRIA Saclay,
Telecom Paristech,

https://project.inria.fr/dirtydata
https://dig.telecom-paristech.fr

Large-scale embedding of heterogeneous information

Research theme: Machine learning, data science
Keywords: Neural networks, knowledge graph embedding, text modeling, deep learning, dirty data, relational data
Duration & salary: 3 to 6 months, between 500 €and 800 € monthly
Research teams: Parietal & Soda (INRIA Saclay), DIG (Telecom Paristech)
Advisers: Gaël Varoquaux, Fabian Suchanek
Contact: gael.varoquaux@inria.fr
Application: Interested candidate should send CV and motivation letter

Context: For many data science problems, for instance in health or business, the most important part of the study is to assemble information about the objects at hand. For instance, a good prediction of housing prices requires assembling historical values of prices, but also various information about the neighboorhood –the access to education, transportation, parks, job, shops– more global trends of geographical growth... This information is available spread across multiple source, for instance on multiple internet pages.
From a knowledge-representation standpoint, these diverse information are formally represented using relational models. Knowledge graphs form large-scale efforts to formally represent as much information as possible so that it can be manipulated and queried by computers. Yago, for instance [4], assembles multiple sources, –such as wikipedia, geonames...– in a consistent relational structure that covers as much as possible of general knowledge.
The use of knowledge graphs in data science faces two challenges. The first challenge is that of data preparation: the information that contained in knowledge graphs must be extracted to be fed into a statistical modeling algorithm, such as a supervised learning model. Traditionally, this step is done manually, with the data-scientist crafting SQL or SPARQL queries and is very time consuming. The second problem is that integrating information across multiple sources, both to build a knowledge graph and to integrate information in a data-science analysis, faces the variability of how the shared entity are written: "Londres" in one dataset may need to be matched with "London" in another. This last problem is related to entity matching, in NLP and database research, or deduplication and record linkage in data management.

Proposed work: To solve the problems listed above, we have adapted knowledge-graph embedding techniques [2] to generate numerical representations (feature vectors) for all the entities that they represent, adapting the most recent models of this family of model [1]. However, a remaining challenge is to **deal with heterogeneity in the "surface form"** –the string representation– of the entities. For this, the proposed work during the internship is to **add a string-modeling layer** (using sequence modeling tools as developed for NLP), following ideas used in NLP [5].
Further work (to be continued in a possible PhD) will use the corresponding architecture to **extract vector representations of knowledge across multiple sources of internet data**, enriching wikipedia and Yago with tables from www.data.gov, data.gov.uk, www.data.gouv.fr. Indeed, these accumulate numerical information on objects of interest to many data-science problems: geographic and socio-economic information on cities, companies, or other administrative units. Formally, this model will ingest data from knowledge bases and tabular data to create an embedding model of the entities present in the data. The end-goal of these representations is to be used in analytic tasks, typically enriching a table to facilitate supervised learning on it.
A crucial point of this new approach is that it will be able to integrate information across such source without explicit matching, and thus it will open the door to large-scale data accumulation. As a result long term

research will focus on **even richer models** that create **representations for the tokens by capturing their context**, as with BERT in NLP [3].

Required skills:

- Knowledge of machine learning or applied maths background (mathematical optimization and statistics)
- Familiarity with fitting deep neural networks (typically pytorch)
- Programming skills in implementing algorithms (eg numerical computing, or matching).

[1] Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Mikhail Galkin, Sahand Sharifzadeh, Asja Fischer, Volker Tresp, and Jens Lehmann. Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework. *arXiv preprint arXiv:2006.13365*, 2020.

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *International Conference on Neural Information Processing Systems*, NIPS, page 2787, 2013.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

[4] Thomas Pellissier Tanon, Gerhard Weikum, and Fabian Suchanek. Yago 4: A reason-able knowledge base. In Andreas Harth, Sabrina Kirrane, Axel-Cyrille Ngonga Ngomo, Heiko Paulheim, Anisa Rula, Anna Lisa Gentile, Peter Haase, and Michael Cochez, editors, *The Semantic Web*, pages 583–596, Cham, 2020. Springer International Publishing.

[5] Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. Mimicking word embeddings using subword rnns, 2017.