Offer #2023-06970

Internship Research on Geo-coding of textual input (GEO-ReSeT 3/3)

Contract type: Internship agreement
Level of qualifications required: Graduate degree or equivalent
Function: Internship Research
Level of experience: Recently graduated

About the research centre or Inria department

The Inria centre at Université Côte d'Azur includes 37 research teams and 8 support services. The centre's staff (about 500 people) is made up of scientists of different nationalities, engineers, technicians and administrative staff. The teams are mainly located on the university campuses of Sophia Antipolis and Nice as well as Montpellier, in close collaboration with research and higher education laboratories and establishments (Université Côte d’Azur, CNRS, INRAE, INSERM ...), but also with the regiona economic players.

With a presence in the fields of computational neuroscience and biology, data science and modeling, software engineering and certification, as well as collaborative robotics, the Inria Centre at Université Côte d'Azur is a major player in terms of scientific excellence through its results and collaborations at both European and international levels.

Context

The work will be embedded in a project in collaboration between Université de Paris Cité (team LIPADE, Paris) and Inria (team EVERGREEN, Montpellier).

By using location on the Earth’s surface as the common link between different modalities, a geo-spatial foundation model would be able to incorporate a variety of data sources, including remote sensing imagery, textual descriptions of places, and features in maps. Leveraging the large amounts of available unlabeled geo-spatial data from these different sources, the GEO-ReSeT (Generalized Earth Observation with Remote Sensing and Text) ANR project has the objective to learn a better representation of any geo-spatial location and convey a semantic representation of the information.

Such a foundation model has the potential to revolutionize Earth observation by allowing for few or zero-shot solutions to classical problems such as land-cover and land-use mapping, target detection, and visual question answering. It will also be useful for a wide range of applications with a geo-spatial component, including environmental monitoring, urban planning and agriculture.

By leveraging several data modalities, this foundation model could provide a more comprehensive and accurate understanding of the Earth’s surface, enabling more informed decisions and actions. This will be particularly valuable for new potential users in sectors such as journalism, social sciences or environmental monitoring, who may not have the resources or expertise to collect their own training datasets and develop their own methods, thus moving beyond open Earth observation data and democratizing the access to Earth observation information.

Assignment

The work to be conducted during the proposed M2 internship will contribute to the ambition of the GEO-ReSeT ANR project by linking textual descriptions of places, found online, to their approximate geo-location, a task known as geocoding.

This text-location link will then be used along the project in combination with other geospatial data modalities, such as those stemming from remote sensing sensors, in order to train multi-modal models that are aware about the way in which people describe locations.

This will be done by first combining information stemming from different databases containing geographic named entities, such as Open Street Maps, Wikipedia and gazetteers, such that geographic
Main activities

In this work, our objective is to develop a pipeline to automatically link any piece of text describing a place with its most likely geographical footprint. The work to be performed in this internship will lead to the following three contributions:

- **Contribution A**: a first pipeline to allow querying a variety of databases that include geographic named entities, such that, given a name, a list of possible geographic footprints, either in the form of points or polygons, is obtained.
- **Contribution B**: a second pipeline in which LLM are used to determine if a piece of text does contain geographic information and proposes potential named entities that would be associated to it. These proposals will be used to query the pipeline developed in the previous contribution in order to obtain candidates for the geographic footprint that is relevant to the text. An additional module will be developed in which LLM will be once again used to determine which of the proposed footprints is the most likely.
- **Contribution C**: the developed pipelines will be used to build a large dataset of text and the corresponding geographic footprints. The candidate will propose a methodology for evaluating the obtained dataset.

Skills

- Python programming
- Deep Learning with Python (preferably with Pytorch)
- Experience with NLP

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

General Information

- **Theme/Domain**: Data and Knowledge Representation and Processing Statistics (Big data) (BAP E)
- **Town/city**: Montpellier ou Paris
- **Inria Center**: Centre Inria d'Université Côte d'Azur
- **Starting date**: 2024-01-01
- **Duration of contract**: 6 months
- **Deadline to apply**: 2024-01-31

Contacts

- **Inria Team**: ZENITH
- **Recruiter**: Marcos Gonzalez Diego / diego.marcos@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

The keys to success

We are looking for someone with strong competences in Python programming and Deep Learning, ideally with experience with geospatial data and NLP.
Warning: you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Applications must be submitted online on the Inria website. Collecting applications by other channels is not guaranteed.

Defence Security:
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST). Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy:
As part of its diversity policy, all Inria positions are accessible to people with disabilities.