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

Le descriptif de l’offre ci-dessous est en Anglais

Type de contrat : Convention de stage
Niveau de diplôme exigé : Bac + 5 ou équivalent
Fonction : Stagiaire de la recherche
Niveau d'expérience souhaité : Jeune diplômé

A propos du centre ou de la direction fonctionnelle

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.

Contexte et atouts du poste

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.

Mission confiée

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 points or polygons can be linked to each entity.

In a second step, a pipeline will be developed to obtain the most likely geographic named entities that are referred to in any piece of text that describes a place.

In order to avoid restricting us to cases where entities' names appear exactly as in the databases, we will leverage pre-trained Large Language Models (LLM) to resolve ambiguities and gather evidence towards the most likely entities that are being described in the text.

**Principales activités**

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.

**Compétences**

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

**Avantages**

- 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

**Informations générales**

- **Thème/Domaine :** Représentation et traitement des données et des connaissances Statistiques (Big data) (BAP E)
- **Ville :** Montpellier ou Paris
- **Centre Inria :** Centre Inria d'Université Côte d'Azur
- **Date de prise de fonction souhaitée :** 2024-01-01
- **Durée de contrat :** 6 mois
- **Date limite pour postuler :** 2024-01-31

**Contacts**

- **Équipe Inria :** ZENITH
- **Recruteur :** Marcos Gonzalez Diego / diego.marcos@inria.fr

**A propos d'Inria**

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.
L'essentiel pour réussir

We are looking for someone with strong competences in Python programming and Deep Learning, ideally with experience with geospatial data and NLP.

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

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

Sécurité défense :
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.