Geospatially oriented language models

The data is one of the most important resource in the modern world and one type of it is geospatial data. This is data generated from a phenomenon that follows from an underlying geographic process, be it a walk in the park, a transit feed line, a demographic census, or a housing cost. A data instance from such a data set usually contains at least two kinds of information: about *what it is* (traditionally: features) and *where it is* (traditionally: geometry). This data can also occur with other modalities, such as natural language. It can be either a text representing a geospatial entity or a text including some geospatial context. Both of those cases require a model capable of incorporating geospatial information.

The project aims to innovate at the crossroads of natural language processing (NLP) and geographic information systems (GIS) by developing language models enriched with geospatial context. It is crucial because current language models struggle with tasks involving geospatial data. They have limited knowledge about the world and understanding of spatial relations. Existing works have shown that performance of language models on geospatial tasks is geographically biased. This is a result of the fact that developed countries are data-rich and therefore dominate in training data. That is why, we aim to develop models that are designed to understand geospatial data and generate responses that are factually correct regardless of the location in question.

This project comprises three pivotal tasks: creating a geospatially annotated dataset, developing a natural language model using an unsupervised encoder to incorporate geospatial context, and constructing a generative language model with geospatial context embedded in its decoder architecture.

Firstly, the project involves creating a diverse dataset that integrates natural language with precise geospatial annotations from sources like social media, news articles, and travel blogs. This foundational dataset will be annotated with geolocation metadata, ensuring a balanced representation across different regions and languages, thus providing a robust basis for training advanced models. Such dataset is crucial for building advanced machine learning models capable of understanding and generation of natural language.

The second task focuses on developing an unsupervised learning model that encodes geospatial context within its language representations. Unsupervised training means that we do not have to annotate every sample in a dataset with some labels. In unsupervised setting, model learns the underlying relations from raw data - in our case text and associated spatial information. By utilising unsupervised techniques, the model will learn to understand and utilise spatial information from the text, enhancing its ability to perform tasks such as location prediction and geospatial tagging without relying on extensive labelled data.

The third task aims to create a generative model that produces coherent and contextually relevant text based on geospatial inputs. Extensive evidence has been provided that generative language models fail at generating factual sequences, as factual sequences are rare and it is very improbable to select a lengthy sequence of low-probability tokens. This is even worse with geospatial context, where models - which have been fine-tuned to look attractively to the end-user, have not been subjected to optimisation procedures that produce a cost at failing geospatial constraints. With geospatial tasks including a significant part of the complexity of a real physical worlds, such as various kinds of distances between objects represented by language as named entities, that encode even more complex characteristics such as terrain or river barriers, or spatial correlations. The goal of this task is to build reinforcement learning-like procedures for training decoder models to be restrained in terms of the geospatial context through the cost function of the new optimisation model.

The integration of geospatial context into language models addresses a significant gap in current NLP technologies, which often overlook the spatial dimensions crucial for a comprehensive understanding of language. This research promises to have far-reaching implications across various fields, including urban planning, tourism, environmental monitoring, and disaster management, by providing tools for accurate and timely interpretation of spatial data in human-readable form. The project represents a significant advancement in NLP and GIS integration, promising to enhance various applications through the development of language models that inherently understand and generate text with embedded geospatial context.