Natural Language Access to Decentralised Data for Disaster Management
– Selecting and Filling SPARQL Templates from Natural Language Statements

Background
As experiences from the past pandemic shows, a completely top-down approach to managing a crisis is not always favourable for society. Authorities making decision, relying on only their own data, failing to engage in interactions with citizens and societal communities creates trust issues and spurs reactions, such as hesitancy to follow recommendations and even the rise of conspiracy theories. However, collecting and integrating data from citizens is not an easy task due to data quality issues, trustworthiness etc. However, even disregarding all these issues, and assuming that data can be understood and trusted, even the mere access to citizen collected data is a challenge. In some cases citizens have spontaneously started using geodata platforms, such as Google maps and Open Streetmap to gather common data, such as indicating places that provided home delivery during the pandemic. While open platforms can hold certain kinds of data, other types of data require much more secure and restricted sharing mechanisms. One such platform for secure and privacy preserving data sharing is the Solid standards and ecosystem (see https://solidproject.org/). Based on core Web standards, Solid adds a layer of security and restricted data access on top of the Web. Hence, in research at IDA (LiU) we are currently exploring how this platform can enable secure, confidential and privacy preserving data sharing in various use cases, including the management of future crises.

Bild från Räddningstjänsten Skåne Nordväst

However, access to such data by decision makers is a challenge. Since decision makers, e.g. employees at public authorities and disaster management centres are rarely computer scientists, and therefore formulating and submitting SPARQL (a query language for the Web) over a set of distributed Solid pods with data, is out of reach for them. What these users need is easy access to the data, but still in a manner that is precise, i.e. more to the point than a mere keyword search. Therefore, the proposed thesis project will investigate the addition of an important component to this platform, namely natural language interaction with the data. More in detail, the thesis project is intended to leverage existing tools and solutions from implementation of Large Language Models (LLMs) for Natural Language Understanding tasks, to the specific task of translating a natural language question to a SPARQL query, and its result again back to natural language.

Thesis Outline
The aim of the thesis is analysis and fine-tuning LLMs for natural language understanding tasks in the area of providing decentralized data access over a set of Solid pods, e.g. applicable in a disaster management setting. A first goal of the work is to prepare a specialised dataset for disaster management, together with query templates and samples of unstructured text descriptions. The next step will be to analyse available LLMs (chatGPT, Llama, Bard, etc) with respect to their performance on natural language understanding tasks, such as translation from natural language questions to formal queries, and comparing to other classical approaches (regex, transformer-based models etc). At this point, the students should select one feasible approach to test – feasible meaning both in terms of accessibility and size of the LLM, in addition to mere

performance on the task, and potentially fine tune it using a part of the collected dataset. Then the solution will be evaluated on a pipeline to create SPARQL requests over RDF data, based on a finite set of SPARQL templates and understanding the context of an analysed natural language question in relation to the underlying data. Finally, the result of the thesis is expected to be both an implementation of this pipeline, i.e. a tech product which provides an interface for human request as natural text for a decentralized knowledge base hosted in Solid pods, but even more importantly an analysis of its performance, with careful analysis of benefits and drawbacks, future development needs etc.

Your profile
We are looking for a student with a background in Natural Language Processing, and/or Knowledge Representation and AI. Relevant specific background knowledge is machine learning, data structure and algorithms, and skills in service development, as well as Python programming.

Your Workplace
You will be working on this project at the Department of Computer and Information Science (IDA) at LiU, but in close (online) collaboration with researchers at the University of Ghent, Belgium, who are experts on the Solid platform, and Kharkiv National University of Radio Electronics, Ukraine, who provide existing implementations of LLMs. The work is placed within the research group on Semantic Web technologies at IDA (see https://liu.se/en/research/semantic-web) and related to the SecurityLink network in the area of decision support and emergency response.

More Information and Contact
If you are interested in this thesis topic, please contact Assoc. Prof. Eva Blomqvist (eva.blomqvist@liu.se). To apply for the topic please enclose a CV and your transcript of records (LADOK-transcript) that shows your background, and also motivate why you want to do this thesis. Selection is made continuously, as soon as the right applicant is found