Using AI and LLM techniques to cluster and visualize large amounts of research information

Rik D.T. Janssen, December 19, 2024

Innovation project description for University of Applied Sciences Utrecht, Education program HBO-ICT, for multidisciplinary teams with third-year HBO-ICT students specializing in Cloud Services and Security, Business IT and Management, Artificial Intelligence, Software Development, and Technical Informatics.

Description

In this innovation project, we use AI and LLM techniques to cluster and visualize large amounts of research information. Research information is about anything related to research: research results, the persons in a research team, their collaborations, their skills, projects in which they have participated, as well as the relations between these entities. Examples of research results are publications, data sets, and software.

The tool to use in this innovation project is Ricgraph (www.ricgraph.eu), also known as Research in context graph. Ricgraph is software that is about relations between items. These items can be collected from various source systems and from multiple organizations. Insights can be obtained by combining information from various source systems, arising from new relations that are not present in each separate source system.

In this innovation project, we use research information from multiple organizations and from multiple source systems (e.g. from the Research Information System Pure, OpenAlex, Yoda, Research Software Directory, profile pages of an organization, the research information system from the HU, Publinova, etc.).

What problem or innovation is being addressed in this innovation project?

Usually, authors assign keywords to their publications. These keywords may be biased in some way. Using AI (Artificial Intelligence) and LLM (Large Language Model) techniques to assign "concepts" to publications means that there is a kind-of common "authority" assigning these concepts over the set of all research outputs, which probably gives a more consistent labeling. This approach also guarantees that a concept is assigned to every research output (sometimes data sets or software lack keywords).

The advantage of using AI and LLM techniques is that research outputs cluster in some way, and these clusters have a meaning, since they also group the collaborators of research outputs into knowledge fields. E.g. researchers may use this information to find colleagues who work on similar subjects, but who they do not know yet. This might be interesting for future collaborations, such as co-authoring future research.

Challenges are:

- How to get reproducible results (the "concepts") from AI/LLMs? How many concepts should be assigned to a research output?
- How to make LLMs summarize research outputs into a useful vocabulary? Should it be a controlled vocabulary? How many items should the vocabulary contain to be able to cover all research fields? How to prevent hallucinations?
- Which LLM should be used? Why? Does it matter anyway?
- Should the vocabulary be hierarchical? If so, how should it be represented?
- Are the clusters meaningful for researchers? What can be learned from these clusters?
- How to visualize the concepts found? What do users need?
- How can this process be done in an efficient way? How to store the (intermediate) results?
- How to implement this in Ricgraph? How should the user interface look like?
- Is this a good idea anyway? What are better approaches?

- What can the University of Applied Sciences Utrecht or their researchers learn from these concepts? How can it be used to improve/expand research and education at University of Applied Sciences Utrecht?

What does the ideal end result look like?

- Software, preferably in Python, that presents research outputs to a AI or LLM, gets concepts, and stores these in Ricgraph.
- A meaningful visualization and user interface that is tested on its usefulness for researchers in Ricgraph Explorer.

To which technologies and/or concepts will the students be exposed?

- AI, LLM.
- Concepts, vocabularies, controlled vocabularies.
- Graphs.
- Visualization.
- Reasoning techniques.
- Software development, software efficiency.
- Working with large data sets.
- User interface design, user interfaces.