Overdrachtsdocument

Uitbreiding Ricgraph: onderzoek naar het implementeren van publicatietopics en visualisatie

Samenvatting

Van 10 februari tot en met 26 juni 2025 hebben 6 Hogeschool Utrecht studenten afkomstig van 4 verschillende HBO-ICT opleidingen samengewerkt voor het Innovation Semester project. In dit project stond Ricgraph centraal, software ontwikkeld door Rik Janssen van de Universiteit Utrecht.

Het project bestond uit het uitbreiden van Ricgraph doormiddel van 2 onderdelen. Het eerste onderdeel was het extraheren van publicatietopics om deze toe te voegen aan Ricgraph, zodat er nog meer resultaten gevonden kunnen worden. Het tweede onderdeel bestond uit het verbeteren van de visualisatie, om deze intuïtiever voor de eindgebruiker te maken.

Om de topics te kunnen extraheren zijn 7 verschillende NLP-technieken uitgetest. Uit dit vergelijkend onderzoek bleek dat Large Language Models (LLM's) de meeste potentie hebben voor de kwestie. De LLM's zijn vervolgens verder geanalyseerd in een diepgaand onderzoek waarbij ze worden vergeleken op basis van variatie en accuraatheid.

Daarnaast zijn er 3 controlled vocabularies vergeleken zodat de topics consistent geëxtraheerd kunnen worden. Hieruit is EuroSciVoc gekozen als de vocabulary met de meeste potentie, omdat deze goed aansluit bij de wetenschappelijke inhoud van Ricgraph en gemakkelijk in gebruik is.

Om een nieuwe visualisatie te ontwikkelen zijn verschillende interviews afgelegd en onderzoeken verricht. Op basis van deze informatie heeft het team 3 visualisatie ideeën uitgewerkt. In overleg is besloten om een uitgeklede versie van het idee met de zoekfunctionaliteit en resultaatweergave op dezelfde pagina uit te werken.

Op basis van diepgaande onderzoeken naar topic-extractie, verkenning van LLM's en optimalisatie van prompting is een geautomatiseerd proces ontwikkeld waarmee onderwerpen uit publicaties geëxtraheerd kunnen worden. De GraphDB is hierbij gevuld met geëxtraheerde topics van publicaties, zodat er in Ricgraph gezocht en gefilterd kan worden op topics.

Het resultaat voor de visualisatie is een extra zoekpagina binnen Ricgraph. Op deze pagina kunnen resultaten worden gevonden op basis van het filteren van topics. De objecten die voldoen aan de zoekopdracht worden op dezelfde pagina getoond. Bij de resultaten is het tevens mogelijk om relaties tussen andere objecten in te zien.

Om het resultaat te verbeteren zijn er verschillende aanbevelingen gedaan op het gebied van prompting, modelgrootte, controlled vocabularies en semantische zoekfunctionaliteiten. Zo zouden geoptimaliseerde prompts en grotere LLM's consistentere en nauwkeurige topics leveren. Een hiërarchisch opgebouwde vocabulaire en semantische zoekfunctie maken het zoeken specifieker en gebruikersvriendelijker.

Inhoud

Inleiding

Team 626 heeft in semester C en D van studiejaar 2024/2025 voor het Innovation Semester gewerkt aan project Ricgraph. Voor dit project hebben 3 Artificial Intelligence studenten, 1 Cybersecurity and Cloud student, 1 Software Development student en een Business IT and Management student samengewerkt om Ricgraph uit te breiden met 2 onderdelen. Dit overdrachtsdocument heeft als doel inzicht te bieden in wat het team heeft ontwikkeld in dit semester.

In hoofdstuk 1 zal de opdracht worden beschreven. Vervolgens zal in hoofdstuk 2 het resultaat worden. In hoofdstuk 3 worden er aanbevelingen gedaan. Tot slot zal in hoofdstuk 4

1. De opdracht

De opdracht voor dit project bestond uit 2 onderdelen. Ten eerste wilde de opdrachtgever Ricgraph uitbreiden door topics van publicaties toe te voegen aan de software. Bij topics moet gedacht. Door het toevoegen kunnen er nog meer relaties tussen objecten gevonden worden.

Daarnaast is er behoefte aan een intuïtieve vorm van visualisatie voor de resultaten die getoond worden in Ricgraph. De uitdaging die speelt is dat wanneer een persoon of object veel relaties heeft, de graaf zoveel punten bevat dat de visualisatie daarvan onoverzichtelijk wordt. Er is op huidig moment dan ook gekozen om de resultaten in Ricgraph te laten zien via een tabel, maar de opdrachtgever vindt dit niet de juiste oplossing en ziet dit graag anders.

De officiële opdrachtformulering wat opgesteld is door de opdrachtgever kan worden teruggelezen in bijlage A van dit document.

2. Resultaten

2.1 Publicatietopics

Om topics uit publicaties te kunnen extraheren zijn er 7 verschillende manieren van topic extraction uitgebreid getest. Deze manieren waren TF-IDF, RAKE, YAKE, topic modelling, LLM's, Jina v3 en keyBERT. De uitkomst van de resultaten waren divers, waarbij de LLM's de meeste potentie bieden om topics te kunnen extraheren.

Ook heeft het team onderzocht hoe de prompts kunnen worden opgesteld om de resultaten te optimaliseren. Na deze onderzoeken zijn verschillende kleine open source LLM's (varianten van Deepseek, Gemma en llama) verkent om automatisch topics uit publicaties te extraheren. De modellen zijn geëvalueerd op basis van variatie en accuraatheid. Het gehele onderzoek kan gelezen worden in het document *Onderzoek: vergelijking van Large Language Models voor topic extraction*.

Om de gegenereerde output bruikbaar te maken is tekst-cleaning uitgevoerd. Hierbij zijn irrelevante resultaten verwijderd, zodat alleen belangrijke output overblijft.

Tijdens dit project is er ook onderzoek verricht naar controlled topics. Om de topics in Ricgraph consistent en herbruikbaar te maken was een controlled vocabulary essentieel. Na het onderzoeken van 3 verschillende controlled vocabularies, namelijk EuroSciVoc, LCSH en FAST, is EuroSciVoc gekozen. De vocabulary sluit goed aan op de wetenschappelijke inhoud van Ricgraph en biedt een brede dekking over diverse onderzoeksthema's. Het team vond tevens dat de vocabulary gemakkelijk in gebruik was.

Na het vinden van een geschikte controlled vocabulary kon gestart worden om de door LLM gegenereerde output automatisch te mappen op de controlled topcs. Dit gebeurde aan de hand van embeddings. . Elke controlled topic en topic uit de LLM worden omgezet naar embeddings a.d.h.v. de 'all-mpnet-base-v2' sentence transformer. Er zijn hier ook andere opties overwogen. De keuze is uiteindelijk gemaakt op kwaliteit van semantisch begrip. Zie in de onderzoekscode pipeline/generate_topics.ipynb voor een verduidelijking van dit algehele proces en beargumentering van het gekozen embeddings model. Nadat de tekst naar embeddings zijn omgezet wordt er een kNN search gedaan per embeddings van de LLM topic over de controlled embeddings om te kijken welke controlled topic het dichtste bij de output van de LLM ligt

Tot slot is een vulscript ontwikkeld om de GraphDB te vullen met de geëxtraheerde topics per publicatie. Dit maakt het mogelijk om te kunnen filteren en resultaten te vinden op basis van topics in Ricgraph.

2.2 Visualisatie

Voor de visualisatie is een Minimum Viable Product (MVP) opgeleverd in de vorm van een extra zoekpagina binnen Ricgraph. Op deze zoekpagina kan er op een gebruikersvriendelijke manier gefilterd worden op basis van topics. De topicselectie heeft een autosuggestie, waardoor gebruikers efficiënter en gebruikersvriendelijker kunnen zoeken. Door het filteren op topics wordt het voor de gebruikers mogelijk om specifieker te zoeken naar gewenste objecten.

Op dezelfde pagina worden ook de gevonden resultaten getoond. Op deze resultaten kan vervolgens weer geklikt worden waardoor relaties van deze objecten verder onderzocht kunnen worden. De visualisatie brengt hierbij een balans tussen zoekresultaten vinden en daarbij hun relaties te kunnen bekijken.

3. Aanbevelingen

Voor dit hoofdstuk zijn de aanbevelingen opgesplitst in 2 onderdelen. Het eerste gedeelte gaat over aanbevelingen omtrent de kwestie. Het tweede onderdeel betreft aanbevelingen voor het Innovation Semester.

3.1 aanbevelingen kwestie

3.1.1 Prompting

Prompting kun je eerder zien als een iteratief proces. Met het ontwikkelen van de huidige oplossingen hebben we over meerdere modellen en met meerdere prompts gekeken wat de output hiervan is. Deze outputs hebben we geanalyseerd en op basis van de criteria hebben we gekeken hoe goed deze prompts presteerde. Dat betekent dus dat de prompt is gebaseerd op basis van de algemene performance over alle modellen. Prompts kunnen vrijwel altijd nog iets beter omdat er zo veel verschillende manieren zijn om iets te vragen. Het is dus nodig om nog een optimale prompt te zoeken voor het gekozen LLM. Dit zal een algemene betere beschrijving geven voor elke topic die het LLM uit de publicatie kan halen. Daarmee kan de output dus beter worden gemapt op de controlled topics.

3.1.2 Een grotere LLM

In onze oplossing nemen we voornamelijk kleine modellen mee, dit komt beide doordat deze open source waren en omdat we de grotere modellen niet kunnen runnen. Een mogelijke verbetering hierin zit in het gebruiken van een commercieel beschikbare LLM zoals ChatGPT. Het is nu enkel een aanname, maar waarschijnlijk kunnen grotere modellen beter en consistenter variërende en accurate topics halen uit de publicaties. De grootste nadelen hierbij zijn de kosten voor de API-calls en dat veel van de commercieel beschikbare LLM’s niet open source zijn.

3.1.3 Betere set aan controlled topics

In het begin zijn er meerdere verschillende datasets van controlled topics uitgeprobeerd. Hieronder vallen LCSH, FAST Topical en EuroSciVoc. De controlled topics hebben meerdere eisen waaraan ze moeten doen. Hij moet namelijk regelmatig worden aangevuld met meerdere disciplines naarmate die ontstaan over de tijd heen. Daarnaast moeten ze ook alle disciplines dekken.

Uiteindelijk hebben we voor EuroSciVoc gekozen omdat dat onder de gevonden datasets de beste was en makkelijkste om mee te werken. Ook wordt deze elk jaar aangevuld met nieuwe topics.

Hierin zit nog wel een nadeel; het zijn allemaal losse topics, zonder relatie aan elkaar. In de onderzoeken is er duidelijk te zien dat de huidige topics juist te veel specifieke dingen dekken. Veel sub-topics waarop je juist zou willen zoeken in Ricgraph is niet mogelijk met deze dataset. Om specifieker te kunnen zoeken op onderwerp zouden er dus meer topics onder het hoofd-topic moeten vallen. Met een boomstructuur zou je daarop dus ook meteen de overkoepelende topics mee kunnen nemen. Bijvoorbeeld; als je 'sentiment-analysis' als topic vindt, krijg je daarbij meteen 'natural language processing'. Op die manier kun je meer en accuratere topics krijgen om op te zoeken.

3.1.4 Semantische zoekfunctie

Dit idee bestaat uit 2 delen; 1 deel voor het zoeken op topics, en de andere op de resultaten o.b.v. andere documenten. Dit maakt allemaal gebruik van document search en is redelijk eenvoudig zelf te implementeren.

3.1.4.1 Semantische topic search

Dit idee kwam voort tijdens het zoeken naar de topics. Soms weet je namelijk niet precies welke topic je zoekt, daarnaast hebben veel topics overlap met elkaar. Het idee is dus om de resultaten van het zoeken/selecteren van de topics om te zetten naar embeddings, dit maakt het mogelijk om te kijken welk van de embeddings dicht bij je zoekopdracht ook relevant zijn op wat je getypt hebt. Ook kleine typfouten waardoor de zoekbalk geen suggesties meer geeft bij de huidig mogelijke selectie aan topics zou hiermee kunnen worden opgelost. Het grootste voordeel hiermee zal zijn dat je ook kan beschrijven wat je zoekt in een zin, zonder dat je de precieze topic hoeft te weten.

3.1.4.2 Semantische document search

Nadat er topics zijn ingevuld en door de resultaten wordt gezocht kan de gebruiker publicaties bookmarken die relevant zijn aan zijn/haar onderzoek. Het idee achter de semantische document search is dat nadat de gebruiker publicaties heeft gebookmarkt, er soortgelijke documenten worden getoond a.d.h.v. de embeddings. Op deze manier krijg je naast dat je op topics zoekt, ook nog soortgelijke resultaten op de publicaties die interessant zijn voor wat de gebruiker zoekt.

3.1.5 Weging van topics

In de huidige implementatie is er alleen te zien welke topic aanwezig is. Maar niet elke topic is in werkelijkheid even relevant. Soms is het 80% van 1 topic, en maar 20% van de andere. Daarom zou het een goede uitbreiding op de topics zijn om een weging aan elke topic toe te voegen. Dit brengt wel wat vraagstukken met zich mee: hoe kom je achter de wegingen van elke topic, en hoe ga je de wegingen opslaan in de database?

3.2 Innovation Semester project

Om het Innovation project zo goed mogelijk te laten verlopen raden we de volgende aanbevelingen aan.

3.2.1 SCRUM-methodiek

De eerste aanbeveling is om SCRUM-methodiek goed te hanteren tijdens dit project. Niet alleen word je hiervoor beoordeeld voor het Innovation semester, maar draagt het ook bij aan een duidelijke werkstructuur en regelmatige voortgangsmomenten. Bekijk ook zeker de feedback van de teambegeleider op tijd.

Tevens raden wij jullie aan om verschillende methoden voor de sprint retrospectives te gebruiken. Wij hebben herhaaldelijk de start-stop-continue methode gebruikt, wat zorgde dat het begon te vervelen. De zeilboot was voor ons een erg effectief methode.

3.2.2 Roadmap

Ook raden wij jullie aan om een roadmap te maken in de beginfase van het project. Een roadmap geeft niet alleen een structuur, maar helpt ook om lange termijndoelen op te delen in overzichtelijke tussenstappen. Dit kan dan weer helpen bij gericht en zo efficiënt mogelijk werken.

4. Overdracht

Alle relevante documentatie en code zijn samengevoegd in een ZIP-bestand. Neem contact op met Rik om toegang te krijgen.

Bijlagen

Bijlage A – Officiële opdrachtformulering

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

In this challenge, 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 challenge 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 challenge, 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 and LLM 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 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 HU or their researchers learn from these concepts? How can it be used to improve/expand research and education at HU?

What does the ideal 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.

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