Eindrapport

Uitbreiding Ricgraph: onderzoek naar het implementeren van publicatietopics en visualisatie

Auteurs: Alper Aydinhan
Pieter Mortier
Mirjam Mus
Christian Rotteveel
Gillian Schmitz
Rachel Yu

Datum: 26-06-2025
Versie: 1.0

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 context 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.

Inhoudsopgave

Inleiding

Dit eindrapport is opgesteld voor het Universiteit Utrecht-project tijdens het Innovation Semester van februari tot en met juni 2025. Voor dit project hebben 6 Hogeschool Utrecht studenten van verschillende HBO-ICT opleidingen samengewerkt om de software Ricgraph uit te breiden. Het document heeft als doel een overzicht te bieden over de opdracht, behaalde resultaten en aanbeveling.

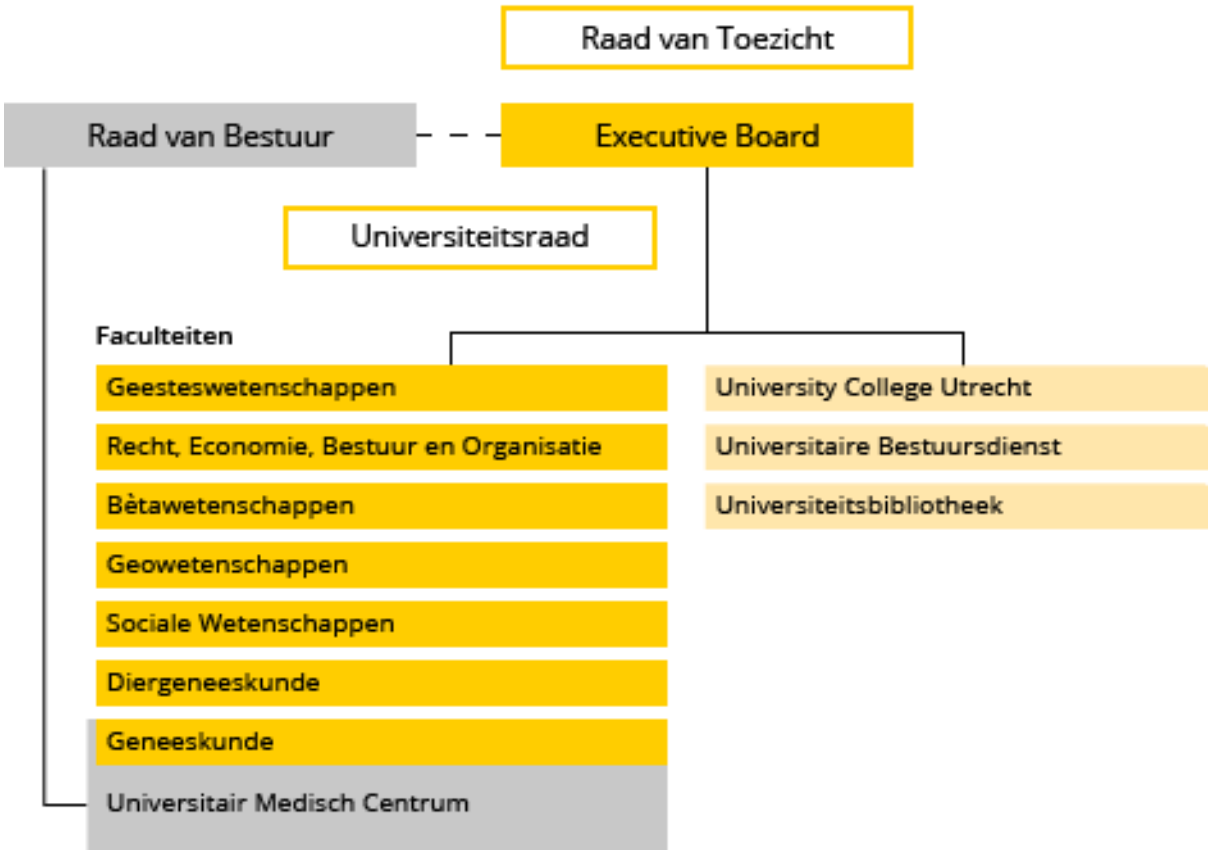
In hoofdstuk 1 wordt de organisatorische context van de Universiteit Utrecht omschreven. Hoofdstuk 2 behandeld de opdracht, waarbij gekeken wordt naar de kwestie, doelstelling en onderzoeksvragen. In hoofdstuk 3 worden de onderzoeksresultaten gepresenteerd. Daarnaast wordt in hoofdstuk 4 het resultaat getoond. Bovendien zal in hoofdstuk 5 de belangrijkste bevindingen van het project worden toegelicht. Het eindrapport wordt afgesloten met aanbevelingen in hoofdstuk 6.

1. Organisatorische context

De Universiteit Utrecht is een Nederlandse universiteit opgericht in 1636. De universiteit biedt meer dan 200 wetenschappelijke opleidingen aan op bachelor- en masterniveau. Naast het verzorgen van onderwijs verricht de universiteit ook wetenschappelijk onderzoek. Op de universiteit volgen meer dan 35.000 studenten onderwijs en zijn er ongeveer 8.600 medewerkers in dienst (Universiteit Utrecht, 2023).

Opdrachtgever en ontwikkelaar van Ricgraph, Rik Janssen, is Business Analist aan de Universiteit Utrecht en onderdeel van de Directie Information and Technology Services (ITS). Deze directie valt onder de universitaire bestuursdienst en adviseert het college van bestuur over investeringen, de inrichting en gebruik van IT-infrastructuur. Zij formuleert daarnaast de hoofdlijnen van het universitaire informatiseringbeleid en is verantwoordelijk voor het programmamanagement van IT-investeringen en –vernieuwingen. Bovendien verzorgt de directie de IT-basisdiensten voor medewerkers, studenten en bezoekers van de Universiteit Utrecht.

Het organogram van de Universiteit Utrecht is hieronder afgebeeld:



Figuur 1: Het organogram van de Universiteit Utrecht

2. De opdracht

2.1 Kwestie

Ricgraph, ook wel bekend als Research in Context Graph, is software dat inzicht biedt bij het vinden van relaties tussen objecten zoals personen, hun vaardigheden, projecten, artikelen, datasets, enzovoort. Deze objecten kunnen worden verzameld uit meerdere organisaties en verschillende bronsystemen. Het doel van de software is om gebruikers inzicht te geven in relaties tussen objecten, die voorheen niet altijd duidelijk waren omdat deze bijvoorbeeld uit verschillende bronnen komen. Gezien wetenschap een zeer breed onderwerp is, helpt Ricgraph bij het vinden van relevante objecten. Zo kan Ricgraph een journalist bijvoorbeeld helpen om een onderzoeker te vinden voor een interview op basis van benodigde skills en publicaties. Ook kunnen onderzoekers elkaar vinden omdat zij bijvoorbeeld een gemeenschappelijke onderzoeksinteresse delen.

Voor deze opdracht spelen 2 kwesties. Ten eerste wil de opdrachtgever en tevens ontwikkelaar van Ricgraph, Rik Janssen, graag Ricgraph uitbreiden door topics van publicaties toe te voegen aan de software. Door de uitbreiding 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.

2.2 Doelstelling

De doelstelling van dit project is om Ricgraph uit te breiden met 2 onderdelen. Het eerste onderdeel zal het toevoegen van de publicatietopics aan Ricgraph zijn. Om deze uitbreiding te ontwikkelen zal er eerst onderzoek worden gedaan om te bepalen welke techniek benodigd is om de topics te extraheren en uit welke mogelijke bronnen deze geëxtraheerd kunnen worden.

Ook zal er een nieuwe vorm van visualisatie worden geïmplementeerd in Ricgraph, zodat de objecten en hun relaties op een intuïtieve manier bekeken kunnen worden door de eindgebruiker. Voor het ontwikkelen van de visualisatie is het doel om onderzoek te doen naar wat een geschikte vorm van visualisatie is doormiddel van interviews en literatuurstudie.

2.3 Onderzoeksvragen

De hoofdvraag van dit project is:

Hoe kan Ricgraph worden uitgebreid door middel van publicatie gerelateerde topics en gebruikersvriendelijke visualisatie om de gebruikerservaring te optimaliseren?

Om de hoofdvraag te kunnen beantwoorden zijn de volgende deelvragen geformuleerd:

1. Welke NLP-technieken zijn het meest geschikt voor de extractie van publicatie topic?
2. Welke publicatietopics zijn relevant om toe te voegen aan Ricgraph?
3. Welke vorm van visualisatie is het meest geschikt om objectrelaties in Ricgraph op een gebruikersvriendelijke manier inzichtelijk te maken?

3. Onderzoeksresultaten

3.1 Deelvraag 1

Welke NLP-technieken zijn het meest geschikt voor de extractie van publicatietopics?

3.1.1 Methode

Om publicatietopics aan Ricgraph toe te voegen is een Natural Language Processing techniek benodigd. Voor het beantwoorden van deze deelvraag wordt gebruik gemaakt van online deskresearch en een toepassingsgericht vergelijkend onderzoek.

3.1.2 Uitvoering

3.1.2.1 Geselecteerde technieken

Op basis van zijn de volgende technieken geselecteerd om te vergelijken:

TF-IDF

Term frequency – inverse document frequency (TF-IDF) is een statistische methode en meet hoe belangrijk een term is binnen een document in vergelijking met een verzameling documenten (corpus). De term frequency (TF) formule berekent hoe vaak een bepaalde term in een document voorkomt. De inverse document frequency (IDF) formule berekent hoe uniek een term is in het corpus door te bepalen in hoeveel documenten de term voorkomt. Vervolgens wordt de term frequency en vermenigvuldigd met de inverse document frequency wat resulteert in een TF-IDF score. Termen die frequent voorkomen in 1 document, maar zelden in het corpus krijgen een hogere TF-IDF score (Karabiber, z.d.).

keyBERT

Een techniek dat gebruik maakt van BERT (Bidirectional Encoder Representations from Transformers) modellen om trefwoorden te extraheren die het meest vergelijkbaar zijn met het document (Chiusano, 2022). Wat keyBERT uniek maakt vergeleken met andere keyword extraction technieken is dat KeyBERT rekening houdt met de semantische context van het document (Besbes, 2022). Bij andere technieken wordt vaak alleen gefocust op de statistische eigenschappen van de tekst.

Yake

YAKE! (Yet Another Keyword Extractor) is een unsupervised keyword extraction algoritme dat automatisch trefwoorden uit teksten kan extraheren. YAKE! ondersteunt teksten van verschillende groottes, domeinen en talen. Het unieke aan YAKE! Is dat het algoritme niet afhankelijk is van externe woordenboeken, thesauri of vooraf getrainde modellen. In plaats daarvan maakt het gebruik van kenmerken die uit de tekst worden geëxtraheerd (Mishra, 2022).

Rake

Rake (Rapid Automatic Keyword Extractor) is een algoritme dat trefwoorden extraheert uit teksten. Rake deelt de tekst op door scheidingstekens zoals punten en komma's. Vervolgens berekent het een score op basis van de frequentie woorden en hoe sterkt het samenhangt met andere woorden (graad). Deze score worden gebruikt om de belangrijkste trefwoorden te selecteren (Rose et al., 2010). Rake is simpel, efficiënt en vereist geen voorafgaande training.

Large Language Models

Large Language Models (LLM's) zijn foundationmodellen getraind op enorme hoeveelheden data. Dit maakt het voor de modellen mogelijk om natuurlijke taal te begrijpen en te genereren. Hierdoor kan een LLM een groot aantal taken uitvoeren, zoals het samenvatten van tekst, vertalen, en beantwoorden van vragen (IBM, 2023). Bekende LLM's zijn OpenAi's ChatGPT-4o en Google's Gemini. De LLM kan met behulp van prompts de relevante keywords uit de publicaties halen.

Jina Embeddings v3

Jina Embeddings v3 is een meertalig tekst-embeddingmodel dat is ontworpen voor verschillende NLP-toepassingen, zoals clusteren en classificeren (Jina, 2024). Het model kan tevens gebruikt worden voor het extraheren van trefwoorden uit teksten.

Topic Modelling

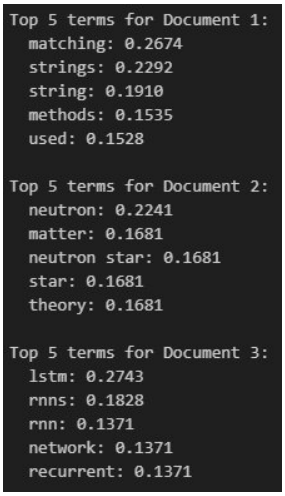
Topic modelling is een techniek om (verborgen) semantische patronen te ontdekken in een corpus en daarin aanwezige onderwerpen te identificeren. Topic modelling is een vorm van statistische modellering dat gebruik maakt van unsupervised machine learning om groepen gerelateerde termen te clusteren tot topics. Het kan bijvoorbeeld worden gebruikt (Pykes, 2019).

3.1.2.2 Resultaten geselecteerde technieken

Om de geselecteerde technieken uit te testen is gebruik gemaakt van een dataset bestaande uit 3 publicaties. Deze publicaties zijn door het projectteam volledig doorgelezen om te controleren of de geëxtraheerde topics overeenkomen met de inhoud van de publicatie. De resultaten van de testen zijn hieronder beschreven:

TF-IDF

TF-IDF is in staat om belangrijke topics te extraheren op basis van woordfrequenties. Wel is er sprake van beperkingen. Ten eerste moet bij deze techniek een vast aantal woorden gegeven worden per document. Dit zorgt ervoor dat een document bijvoorbeeld altijd 5 topics zal teruggeven. Ook is er sprake van redundante resultaten, zoals “strings” en “string”. Daarnaast kijkt de TF-IDF niet naar de semantische betekenis van woorden.



Figuur 2: resultaten TF-IDF

keyBERT

Helaas was keyBERT als losstaande techniek niet geschikt techniek voor het extraheren van topics. De techniek kijkt te veel naar hoe vaak een woord in de tekst voorkomt, waardoor er weinig echte topics uit komen. Mogelijk zou keyBERT in combinatie met een LLM betere potentie hebben.

RAKE

```
Phrase: generated using models like countvectorizer, Score: 20.166666666666664
Phrase: tr0009 airco 3f1 ",, Score: 16.0
Phrase: hm18 ups ac ",, Score: 16.0
Phrase: ad00 airco 2 ",, Score: 16.0
Phrase: matching techniques perform across, Score: 14.833333333333334
Phrase: solution leaves approximately 5, Score: 14.5
Phrase: method uses exact matching, Score: 14.333333333333334
Phrase: cleaned using multiple methods, Score: 13.666666666666666
Phrase: natural language processing, Score: 9.0
Phrase: comprehensive approach allows, Score: 9.0
```

Figuur 3: Resultaten RAKE

Rake is niet geschikt voor het extraheren van topics. Helaas geeft het algoritme lange zinnen terug en geen zelfstandige topics.

YAKE

```
Keyword: Airco, Score: 0.06916578968332032
Keyword: match code-like strings, Score: 0.07688100037666887
Keyword: Natural Language Processing, Score: 0.08030238674790854
Keyword: methods, Score: 0.08538495641895298
Keyword: study, Score: 0.08900218989903046
Keyword: strings, Score: 0.09030148907597056
Keyword: string comparison, Score: 0.10189968926708716
Keyword: code-like strings, Score: 0.10189968926708716
Keyword: matching, Score: 0.10675057434309546
Keyword: match code-like, Score: 0.12488775041680993
```

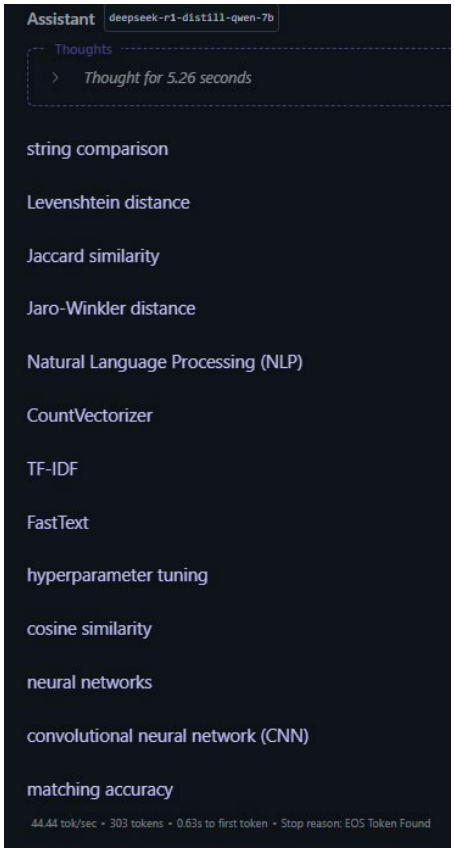
Figuur 4: YAKE resultaten

De techniek YAKE toont potentie om topics te extraheren. Tot tegenstelling van RAKE geeft YAKE daarentegen wel topics terug. Wel moet deze techniek gecombineerd worden met een LLM om topics te kunnen extraheren.

Large Language Models

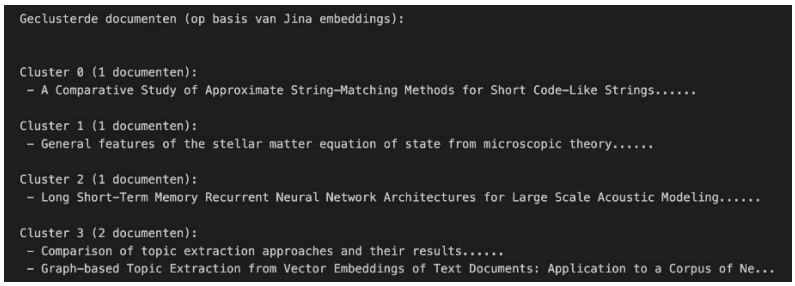
De LLM's tonen grote potentie voor het extraheren van topics. Voor het onderzoek zijn 3 LLM's getest, namelijk Gemma 3 12b it Q3, Granite 3.2 8B en deepseek-r1-distill-qwen-7b. Deze LLM's zijn uitgekozen omdat het kleine modellen zijn, zodat deze geschikt zijn voor lokaal gebruik zonder zware infrastructuur.

De eerste resultaten tonen aan dat de LLM's in staat zijn om topics uit publicaties te extraheren. Wel is verdere uitwerking van de LLM's benodigd om deze potentie te benutten. Zo moet er gewerkt worden aan de reproduceerbaarheid van de uitkomsten, bijvoorbeeld door consistente prompt strategieën.



Figuur 5: Resultaten van de deepseek-r1-distill-qwen-7b LLM

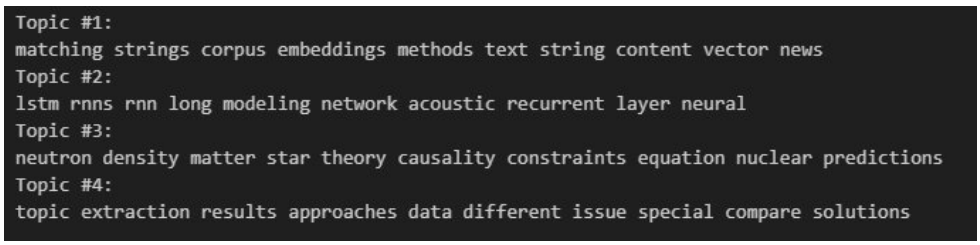
Jina v3



Figuur 6: Jina V3 resultaten

Het resultaat van de techniek Jina v3 voldoet niet aan de wensen omtrent het extraheren van topics. De techniek levert enkel clusters op. Ook bestaat de limiet dat er maar 1 cluster per document kan worden toegewezen. Als dit wordt gedaan op de splitsing van zinnen krijg je sommige goede clusters, maar ook zinnen die niets zeggen over de inhoud van de publicatie zullen clusters toegewezen krijgen. Hierbij ontstaat dus een nieuw probleem van bepalen welke clusters je wel, en welke je niet mee moet nemen.

Topic Modelling



Figuur 7: De topic modelling resultaten

Het resultaat van deze techniek bleek niet geschikt te zijn voor de kwestie omtrent het extraheren van publicatietopics. De methode groepeert woorden per onderwerp. Ook zijn er een vast aantal woorden nodig per topic wat de flexibiliteit beperkt. De techniek sluit dan ook onvoldoende aan bij wat nodig is voor het extraheren van publicatietopics.

3.1.3 Conclusie

Op basis van het vergelijkend onderzoek naar zeven NLP-technieken blijkt dat niet alle technieken geschikt zijn voor het extraheren van publicatietopics. Topic modelling en Jina gaven clusters terug, zonder expliciete en zelfstandige topics die direct bruikbaar zijn. Simpelere NLP-technieken zoals TF-IDF, RAKE en YAKE hadden dan weer verschillende limieten. Hierbij kan gedacht worden aan vaste parameters en redundantie. Large Language Models (LLM's) tonen de meest potentie als techniek voor het extraheren van publicatietopics. Zo kunnen zij zelfstandige en betekenisvolle topics extraheren die meteen bruikbaar zijn.

Voor het project zijn de LLM's dan ook uitgebreid verder onderzocht. Het resultaat hiervan is terug te lezen in het document *Onderzoek: Vergelijking van Large Language Models (LLM's) voor topic extraction*.

3.2 Deelvraag 2

Welke publicatietopics zijn relevant om toe te voegen aan Ricgraph?

3.2.1 Methode

Deze deelvraag wordt beantwoord op basis van online deskresearch en het uitvoeren van een vergelijkend onderzoek. Bij het extraheren van publicatietopics is het belangrijk dat consistentie en herbruikbaarheid gewaarborgd zijn. Om dit te kunnen garanderen is een controlled vocabulary benodigd. Voor deze deelvraag is gekeken naar een geschikte vocabulary die de topics kan extraheren voor Ricgraph.

3.2.2 Uitvoering

De volgende controlled vocabularies zijn onderzocht en beoordeeld:

EuroSciVoc

EuroSciVoc is een meertalige taxonomie dat een groot scala aan wetenschappelijke termen vertegenwoordigd. De taxonomie bevat meer dan 1000 categorieën in 6 verschillende talen (Engels, Frans, Duits, Italiaans, Spaans en Pools). Elke categorie is weer verrijkt met relevante termen. De taxonomie wordt beheerd door het bureau voor publicaties van de Europese Unie (Publication Office of the EU, z.d.).

Voor het onderzoek is de dataset via de downloadpagina van EuroSciVoc gedownload. Het team ervaart de EuroSciVoc als beste van de 3 onderzochte vocabularies. Zo was er makkelijk mee te werken en bood de vocabulary een sterke wetenschappelijke representativiteit. Tevens wordt de vocabulary jaarlijks aangevuld met nieuwe termen, wat zorgt voor blijvende relevantie.

Library of Congress Subject Headings

De Library of Congress Subject Headings (LCSH) is een controlled vocabulary dat duizenden gestandaardiseerde termen bevat. Wereldwijd wordt het gebruikt door veel academische en openbare bibliotheken voor het indexeren en categoriseren van materiaal. De vocabulary wordt sinds de eerste publicatie in 1914 voortdurend bijgewerkt. De LCSH vormt de basis van andere vocabularies zoals FAST (Library of Congress, z.d.).

Voor het onderzoek is via downloadpagina LCSH de dataset in Turtle-bestandsversie (kleinste formaat) gedownload. Omdat deze dataset veel aanvullende gegevens bevatte, is een script geschreven om uitsluitend relevante topics te extraheren. Dit verkort de verwerkingstijd aanzienlijk. De LCSH was niet helemaal wat het team zocht. Het systeem is oorspronkelijk ontworpen voor handmatige categorisering, waardoor deze minder geschikt is voor geautomatiseerde topic-extractie. Dat komt voornamelijk omdat het extreem specifieke topics kan bevatten, wat het niet geschikt maakt om op te zoeken.

Faceted Application of Subject Terminology

Faceted Application of Subject Terminology (FAST) is een thesaurus met meer dan 1.7 miljoen trefwoorden die zijn verdeeld over verschillende facetten. FAST heeft als doel onderwerpscategorisering eenvoudiger en toegankelijker te maken. FAST is in eind 1998 ontwikkeld door OCLC Research en de Library of Congress, en is gebaseerd op de LCSH (OCLC, 2025).

Om de thesaurus uit de testen is via de downloadpagina van FAST de dataset gedownload. De FAST-dataset werkte tijdens het onderzoek niet naar behoren. Zo was er sprake van ruis waardoor de dataset moeilijk bruikbaar was voor topic extractie en niet geschikt was om te gebruiken als zoektermen.

3.2.3 Conclusie

Om de publicatietopics consistent te implementeren is een controlled vocabulary benodigd. Tijdens dit onderzoek zijn 3 verschillende controlled vocabularies getest, namelijk EuroSciVoc, LCSH en FAST.

Zowel LCSH en FAST waren niet wat het team zocht voor de kwestie voor de publicatietopics. De LCSH was minder geschikt voor geautomatiseerde topic-extractie. Bij FAST was er weer sprake van ruis.

Het team concludeert dan ook dat EuroSciVoc de meest geschikte controlled vocabulary is om te gebruiken. Zo ervaarde het team dat er makkelijk mee te werken was. Daarnaast sluiten de wetenschappelijke representativiteit goed aan bij Ricgraph. Bovendien wordt de taxonomie jaarlijks bijgewerkt zodat de topics relevant blijven.

3.3 Deelvraag 3

Welke vorm van visualisatie is het meest geschikt om objectrelaties in Ricgraph op een gebruikersvriendelijke manier inzichtelijk te maken?

3.3.1 Methode

Deze deelvraag wordt beantwoord op basis van zowel het testen van de softwarefunctionaliteit als veldonderzoek in de vorm van gestructureerde interviews met de opdrachtgever en een potentiële eindgebruiker. Deze combinatie leverde zowel technische als praktijkgerichte inzichten op.

3.3.2 Uitvoering

Interview Rik Janssen

Voor de uitvoering van deze deelvraag is als eerste een gestructureerd interview afgelegd met de ontwikkelaar van Ricgraph en tevens opdrachtgever van dit project, Rik Janssen. Dit interview was gericht op het verkrijgen van antwoorden over de achtergrond, (technische) wensen en toekomstvisie van Ricgraph. Uit dit interview kwamen de volgende requirements:

- De gebruikers moet meerdere topics kunnen selecteren en de resultaten moeten op basis van deze combinatie gefilterd kunnen worden.
- De visualisatie moet overzichtelijk blijven bij een groot aantal resultaten (50+).

Interview Stefanie Ypma

Daarnaast is er ook een gestructureerd interview afgelegd met Stefanie Ypma, projectmanager aan de Universiteit Utrecht. Stefanie is een mogelijke potentiële eindgebruiker van Ricgraph die opzoek is naar experts voor onderzoek. Bij dit interview lag de focus op het verkrijgen van de wensen voor de visualisatie. Stefanie haar droomvisualisatie zou de volgende onderdelen bevatten:

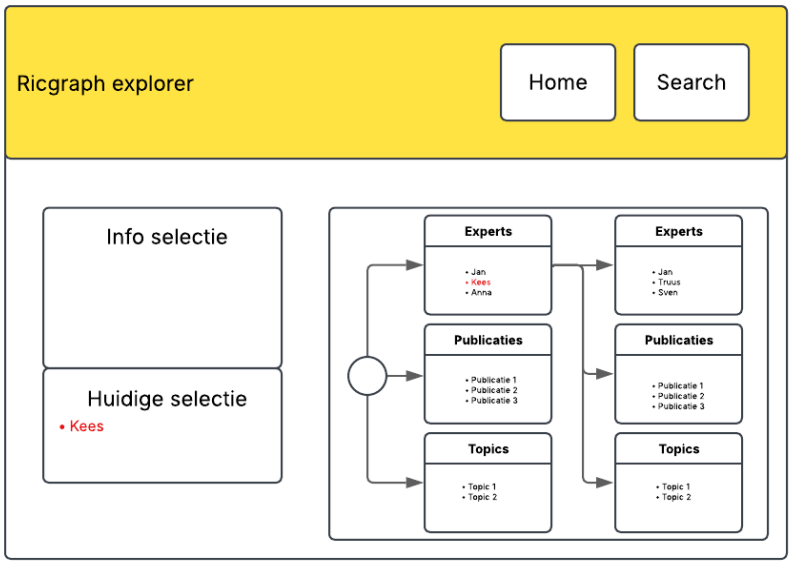
- Een filterfunctie dat experts kan vinden op basis van discipline. Hierbij zou vervolgens verder gefilterd kunnen worden op basis van expertises van deze experts.
- De gebruiker kan bij het zoeken verschillende informatie inzien over de experts. Hierbij wordt gedacht aan de organisatie waarvoor zij werkzaam zijn, vaardigheden en samenwerkingspartners.

Testen van softwarefunctionaliteit

Om inzicht te krijgen in de sterke en zwakke punten van de huidige visualisatie van Ricgraph is de software getest op basis van haar zoekfunctie. Daarbij zijn meerdere zoekresultaten uitgevoerd om te ervaren hoe dit gehele proces als eindgebruiker wordt ervaren. Een sterk punt van de huidige tabelweergave is dat de resultaten overzichtelijk en gestructureerd getoond worden. Of er nu 5 resultaten zijn of 200, het blijft behapbaar voor de eindgebruiker. Een zwakker punten van de huidige visualisatie is dat de zoekfunctie en de resultaten niet op dezelfde pagina getoond worden. Als de gebruiker een zoekinput invoert en deze fout is of niks gevonden kan worden, moet de persoon heen en weer met zoeken.

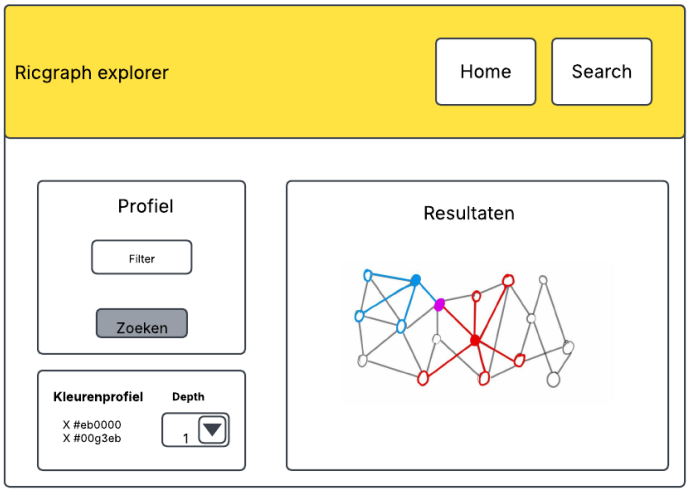
3.3.3 Uitwerking

Op basis van de interviews, het testen van de zoekfunctionaliteit van Ricgraph en eigen expertise heeft het projectteam 3 ideeën uitgewerkt om de huidige visualisatie van Ricgraph te verbeteren.



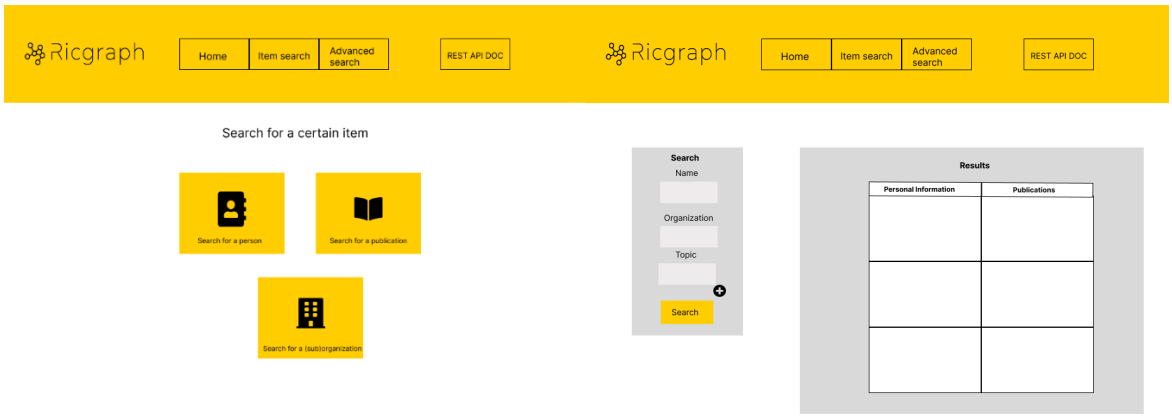
Figuur 8: Idee 1 visualisatie

Idee 1 gaat uit van objecten verkennen door middel van tabellen waarin interactief verdiept kan worden. De verkenning begint met het invoeren van een zoekterm, waarna tabellen verschijnen die overeenkomen met deze term. Vanuit deze tabellen kan de gebruiker zich verder verdiepen. Zo leidt een klik op bijvoorbeeld expert Kees tot aanvullende tabellen met gerelateerde objecten van deze expert. Dit proces gaat door totdat er geen gerelateerde objecten meer over blijven.



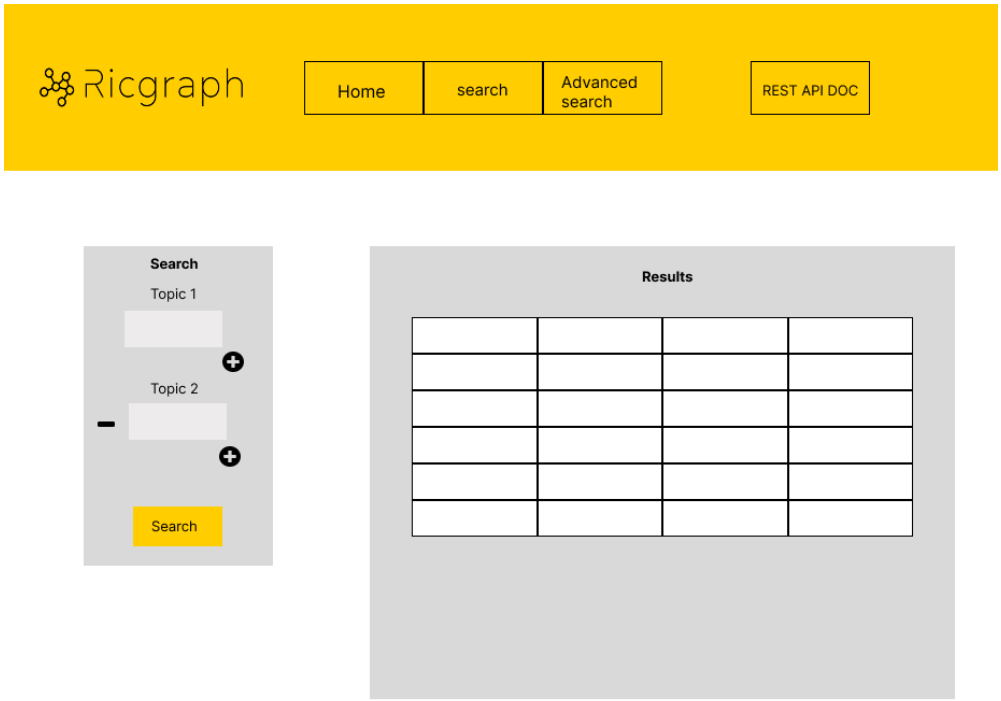
Figuur 9: Idee 2 visualisatie

Idee 2 verkent de objecten door middel van een interactieve graaf. Bij dit idee voert de gebruiker één of meer zoektermen in, waarna een interactieve graaf getoond wordt met relevante objecten. Als de gebruiker vervolgens een node aanklikt verschijnt er aanvullende informatie over deze node. De geselecteerde diepte bepaalt hoe ver de resultaten zich uitstrekken. Het kleurprofiel geeft visueel onderscheid aan het type object.



Figuur 10: Idee 3 visualisatie

Idee 3 gaat uit van verschillende zoekfuncties om resultaten te zoeken op basis van het type object (experts, publicaties of organisaties). Door specifiek op een object te zoeken kunnen verschillende use cases worden vervuld, bijvoorbeeld voor het zoeken van experts. Het voordeel van het zoeken is dat de filters zo aangepast kunnen worden wat past bij het type object. Daarnaast worden de resultaten bij elke zoekfunctionaliteit anders getoond is bij dit idee is de zoekfunctionaliteit toegevoegd aan dezelfde pagina waarbij de resultaten getoond worden om dit intuïtiever te maken.



In overleg met de opdrachtgever is uiteindelijk gekozen om een uitgeklede versie van idee 3 te ontwikkelen. Er zijn hierbij geen meerdere zoekfuncties voor verschillende objecten, maar één zoekfunctionaliteit waarbij gefilterd wordt op basis van publicatietopics. De eindgebruiker kan zelf het aantal topics kiezen om op te filteren, zodat er zo specifiek mogelijk gezocht kan worden.

3.3.4 Conclusie

Op basis van de afgelegde interviews, testresultaten en eigen analyse concludeert het projectteam dat een filter- en zoekfunctionaliteit waarbij de resultaten op dezelfde pagina worden weergeven, het meest geschikt is voor Ricgraph. De resultaten blijven net zoals in de voormalige situatie zichtbaar in een tabelweergave.

Het concept verhoogt het gebruikersgemak voor eindgebruikers, omdat zij niet langer meer hoeven te navigeren tussen pagina's om resultaten te zoeken en bekijken. Dit voorkomt ook frustratie van de eindgebruiker wanneer zij een (foutieve) zoekterm invullen waarbij geen resultaten worden gevonden. De gekozen structuur biedt bovendien de mogelijkheid voor toekomstige uitbreiding met extra filters

4. Projectresultaat

4.1 Topics

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 context van Ricgraph en biedt een brede dekking over diverse onderzoeksthema's. Het team vond tevens dat de vocabulary gemakkelijk was in gebruik.

Na het vinden van een geschikte controlled vocabulary kon gestart worden om de door LLM gegenereerde output automatisch te mappen op de controlled topics. 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.

4.2 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.

5. Conclusie

Tijdens dit project is gezocht naar antwoord op de vraag: '*Hoe kan Ricgraph worden uitgebreid door middel van publicatie gerelateerde topics en gebruikersvriendelijke visualisatie om de gebruikerservaring te optimaliseren*?'. Om deze vraag te beantwoorden is online deskresearch, veldonderzoek en vergelijkende onderzoeken uitgevoerd

Uit het vergelijkend onderzoek blijkt dat LLM's de meeste potentie hebben om automatisch topics te kunnen extraheren uit publicaties. Echter tonen zij nog enkele beperkingen, voornamelijk op het gebied van nauwkeurigheid bij specifiekere en gedetailleerdere onderwerpen. Om betere resultaten te bereiken is verdere verfijning van de LLM's en aanpassingen aan de temperatuurinstellingen essentieel. Er wordt aangeraden om de aanbevelingen in dit eindrapport te bestuderen.

Daarnaast blijkt uit online deskresearch dat een controlled vocabulary kan helpen bij het waarborgen van consistente en herbruikbare topics. Uit het vergelijkend onderzoek blijkt dat de EuroSciVoc vocabulary het beste past bij Ricgraph. Deze vocabulary sluit namelijk goed aan bij de wetenschappelijke context van de software en biedt een breed scala aan topics omtrent onderzoeksgebieden. Tevens wordt elk jaar de vocabulary aangevuld met nieuwe topics voor relevantie.

Nadat een geschikte vocabulary is gevonden zijn de geëxtraheerde topics gemapt met de controlled topics. Vervolgens is een vulscript ontwikkeld om de topics per publicatie toe te voegen aan de GraphDB. Hierdoor is het nu mogelijk om binnen Ricgraph te zoeken en filteren op topics, wat de relevantie van de resultaten vergroot.

Voor de visualisatie is gekozen om een extra zoekpagina aan Ricgraph toe te voegen. Op deze zoekpagina kunnen resultaten gefilterd worden op basis van topics. De resultaten worden vervolgens op dezelfde pagina in tabelweergave weergeven. Hierbij bestaat de optie om op een resultaat te klikken en hierbij gerelateerde objecten te bekijken.

Al met al vormt dit project een basis waarmee Ricgraph uitgebreid wordt met de toevoeging van publicatietopics en een gebruikersvriendelijke zoekfunctionaliteit. Deze uitbreidingen dragen dan ook bij aan een de optimalisatie van de gebruikerservaring.

6. Aanbevelingen

Er zijn veel punten waar nog kan worden uitgebreid en worden verbeterd. Hieronder per onderwerp een toelichting wat de oorsprong, het probleem en mogelijke oplossingen zijn hiervoor.

6.1 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.

6.2 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.

6.3 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.

6.4 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.

6.4.1 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.

6.4.2 Semantische document search

Nadat er topics zijn ingevuld en door de resultaten wordt gezocht kan het een goede functionaliteit zijn om de gebruiker publicaties te laten 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.

6.5 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?

Bronnen

AWS. (z.d.). *What is Machine Translation? - Neural Machine Translation Explained - AWS*. Amazon Web Services, Inc. https://aws.amazon.com/what-is/machine-translation/

Besbes, A. (2022, 6 januari). How to Extract Relevant Keywords with KeyBERT - TDS Archive - Medium. *Medium*. https://medium.com/data-science/how-to-extract-relevant-keywords-with-keybert-6e7b3cf889ae

Chiusano, F. (2022, 12 maart). Keyword and keyphrase extraction with KeyBERT. *Medium*. https://medium.com/nlplanet/two-minutes-nlp-keyword-and-keyphrase-extraction-with-keybert-a9994b06a83

Eppright, C. (2021, 25 maart). *What is Natural Language Processing (NLP)?* https://www.oracle.com/ng/artificial-intelligence/what-is-natural-language-processing/

Glover, E. (2024, 13 februari). *Machine Translation: How It Works and Tools to Choose From*. Built In. https://builtin.com/artificial-intelligence/machine-translation

IBM. (2023a, augustus 24). *What is Sentiment Analysis?* https://www.ibm.com/think/topics/sentiment-analysis

IBM. (2023b, november 2). *Large Language Models*. https://www.ibm.com/think/topics/large-language-models

Instituut van de Nederlandse Taal. (2016, 3 december). *Chatbot*. https://anw.ivdnt.org/article/chatbot

Jina. (2024, 11 oktober). *Jina Embeddings v3: A Frontier Multilingual Embedding Model*. https://jina.ai/news/jina-embeddings-v3-a-frontier-multilingual-embedding-model/#parameter-dimensions

Karabiber, F. (z.d.). *TF-IDF — Term Frequency-Inverse Document Frequency*. https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/

Library of Congress. (z.d.). *Process for Adding and Revising Library of Congress Subject Headings*. https://www.loc.gov/aba/cataloging/subject/lcsh-process.html

Mishra, A. (2022, 14 mei). Keyword Extractor YAKE! - Aditya Mishra - Medium. *Medium*. https://medium.com/@adityamishra.rishu/keyword-extractor-yake-35870de21a0d

OCLC. (2025, 24 maart). *FAST: Subject terminology schema*. https://www.oclc.org/en/fast.html

Publication Office of the EU. (z.d.). *European Science Vocabulary (EuroSciVoc) - EU Vocabularies - Publications Office of the EU*. EU Vocabularies. https://op.europa.eu/nl/web/eu-vocabularies/euroscivoc

Rose, S., Engel, D., & Cramer, N. (2010). *Text Mining: Applications and Theory*. https://doi.org/10.1002/9780470689646.ch1