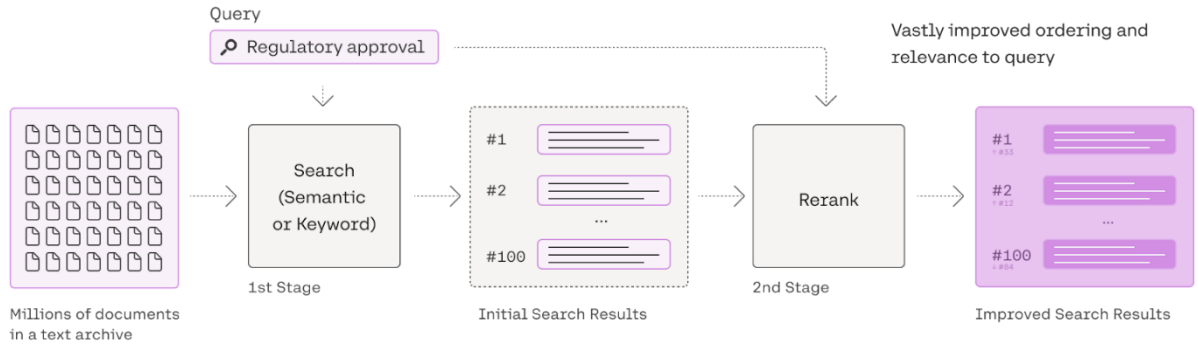
cohere	Systeemvereisten:	System agnostic
	Ontwikkeld door:	Cohere
Commerciële licentie	Contactpersoon:	Bert.Vanhalst@smals.be

Cohere Rerank is een service voor het verbeteren van de relevantie van zoekresultaten. De zoekresultaten die bekomen worden door een keyword search en/of vector search worden in een tweede stap verbeterd door de Rerank service. Dat gebeurt door elk zoekresultaat van de initiële zoekopdracht een score toe te kennen op vlak van relevantie ten opzichte van de gebruikersvraag (*query*). Het toekennen van dergelijke relevantiescores gebeurt op basis van geavanceerde taalmodellen.



Bron: https://cohere.com/blog/rerank

Naast een Engelstalig model biedt Cohere ook een meertalig model dat getraind is op meer dan 100 talen, waaronder Nederlands, Frans en Duits. Het rerank model van Cohere kan desgewenst gefinetuned worden om de resultaten nog te verbeteren in een specifiek domein.

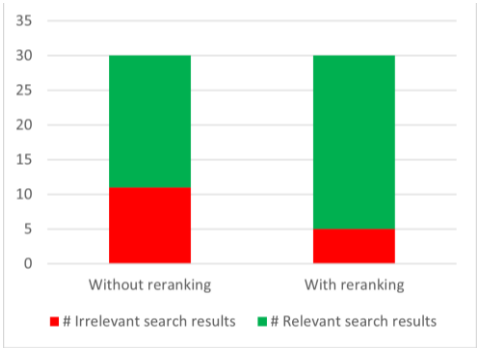
Uit onze testresultaten blijkt dat een reranking service als Cohere Rerank effectief een meerwaarde kan betekenen voor de kwaliteit van zoekresultaten. Het zorgt voor een krachtige semantische boost van de relevantie van de zoekresultaten. Cohere Rerank is eenvoudig toe te voegen aan een bestaande keyword search, vector search of hybrid search.

We pasten reranking toe op de retrieval component van een RAG (*Retrieval Augmented Generation*) systeem in het kader van een generatief vraag- en antwoordsysteem. Het doel is om de hybrid search (keyword + vector search) uit te breiden met een semantische reranking stap op basis van de Cohere Rerank API. De originele hybrid search bestond uit 12 zoekresultaten (6 voor keyword search en 6 voor vector search). Om geen relevante documenten te missen werd dat aantal uitgebreid naar 40 zoekresultaten (20 voor keyword search en 20 voor vector search). De resulterende documenten worden dan samen met de originele vraag van de gebruiker doorgestuurd naar de Cohere Rerank API. Als resultaat krijgen we voor elk van de documenten een relevantiescore terug. Vervolgens weerhouden we enkel de 10 documenten met de hoogste relevantiescore en geven die tenslotte mee als context aan het taalmodel voor het genereren van een antwoord.

Deze setup werd toegepast op een experimentele generatieve chatbot voor vragen rond de Mijn Geneesmiddelen app. De RAG-gebaseerde toepassing is ontwikkeld via de integratie van Cohere in Langchain. Het toevoegen van Cohere reranking aan de basis (hybrid) retriever is eenvoudig: we definiëren een *ContextualCompressionRetriever* met het Cohere Rerank model als *compressor*. Als reranking model kiezen we niet voor `rerank-english-v3.0`, maar voor `rerank-multilingual-v3.0` vanwege de meertalige context van de chatbot. Via het Cohere dashboard kan je een API key aanvragen.

```
compressor = CohereRerank(model="rerank-multilingual-v3.0", top_n=10)
compression_retriever = ContextualCompressionRetriever(base_compressor=compressor, base_retriever=retriever)
```

Om de meerwaarde van de reranking te evalueren werken we met de precision@3 metriek: het aandeel relevante documenten in de top 3 resultaten. We testten dit uit op 10 inputvragen, waarbij we per vraag dus telkens de top 3 zoekresultaten bekijken, wat neerkomt op een maximum score van 30. De standaard (hybrid) retrieval scoort hierbij 19 op 30. Als we de reranking hieraan toevoegen komen we aan een score van 25 op 30, wat toch een aanzienlijke verbetering is. Op 5 vragen scoort de retrieval mét reranking even goed als die zonder. Op 4 vragen scoort de retrieval met reranking beter. En verrassend genoeg is er ook één vraag waar de retrieval met reranking slechter scoort.



Maar globaal gezien kunnen we concluderen dat reranking de kwaliteit van de retrieval aanzienlijk kan verbeteren, wat ervoor zorgt dat de informatie die gevoed wordt aan het generatief taalmodel relevanter is. Dit laat het taalmodel toe om kwalitatievere antwoorden te genereren.

Cohere Rerank kan afgenomen worden als een SaaS API of via cloud services zoals OCI of AWS. De kost komt neer op 2$ per 1000 searches. Een search is een request met maximaal 100 documenten die moeten gerangschikt worden. Documenten die langer zijn dan 4096 tokens worden opgesplitst in meerdere chunks, waarbij elk van die chunks meetelt als apart document. Je kan een gratis trial API key aanvragen om de service uit te proberen. Het gebruik is dan wel beperkt tot 10 calls per minuut.