Coding Fest 2024

GPT-Powered News

The University of Sydney: Andrea Bosia, Michael Herbert, Peter Ka Wing Cheng, Dylan Leon, Axel Loen, Tasin Trirattanagool

Project ID: P23

Postgraduate

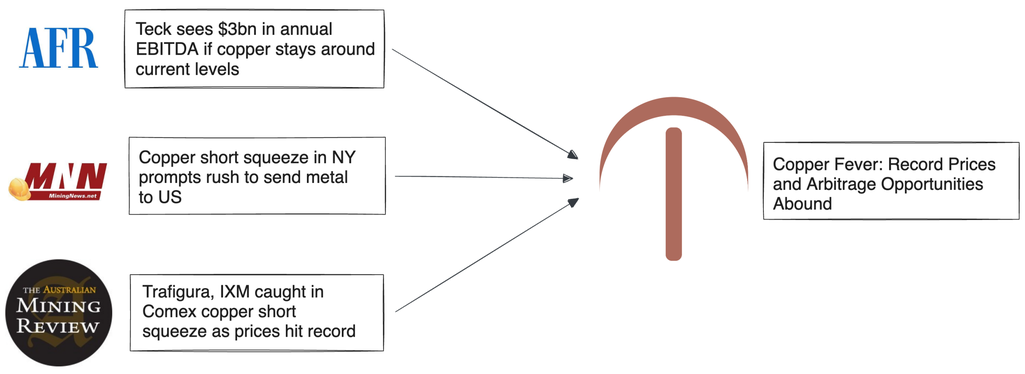
INTRODUCTION

The World Wide Web has provided unprecedented access to information. As the web grew, structuring data for easy retrieval and accessibility became crucial, leading to the development of the "semantic web." The semantic web enables processing engines to understand stored information, facilitating targeted searches and meaningful results. This project applies the semantic web paradigm to news search and aggregation, addressing the user's need for relevant information in an easily consumable format.

OBJECTIVE

This project aims to create a model that scrapes news articles, identifies and groups those that discuss the same story, and finally generates concise and representative summaries of those stories.

Web Scraping → Embedding and Semantic Search → Search and LLM → Serve Summary to Front-end



CHALLENGES

We identified two key challenges in developing this model:

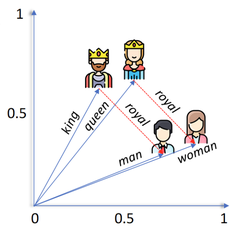
1. Semantic Similarity: Developing a method to identify related news based on their semantic content.
2. Sourcing Knowledge: Producing comprehensive summaries that integrate information from multiple articles.

Semantic Similarity

- To determine if two news articles are closely related based on content, we need a numerical representation that captures semantic meaning.
- This can be achieved using "**embeddings**" - a high-dimensional space where words and sentences are represented by dense numerical vectors.
- In the embedding space, the relative distance between vectors captures the semantic meaning of the words.
- By comparing the vectors of two articles in this space, we can determine how semantically similar or related they are in terms of content.

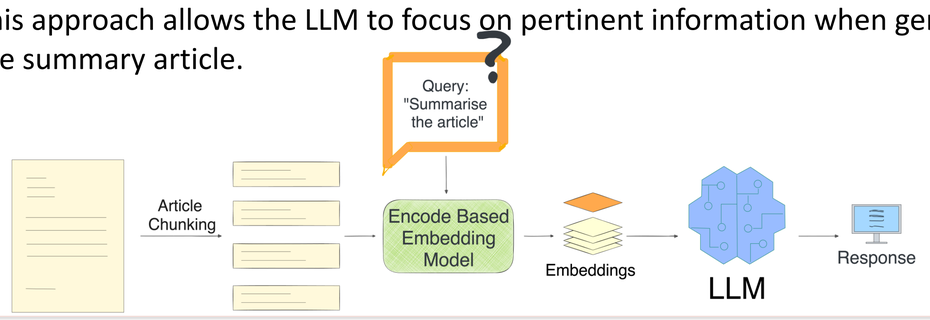
The graph exemplifies the high-level functioning of Embeddings. The semantic meaning of a word (and by extension a whole article) is captured enabling distance-based operations.

King – Royal = man
Queen – Royal = woman



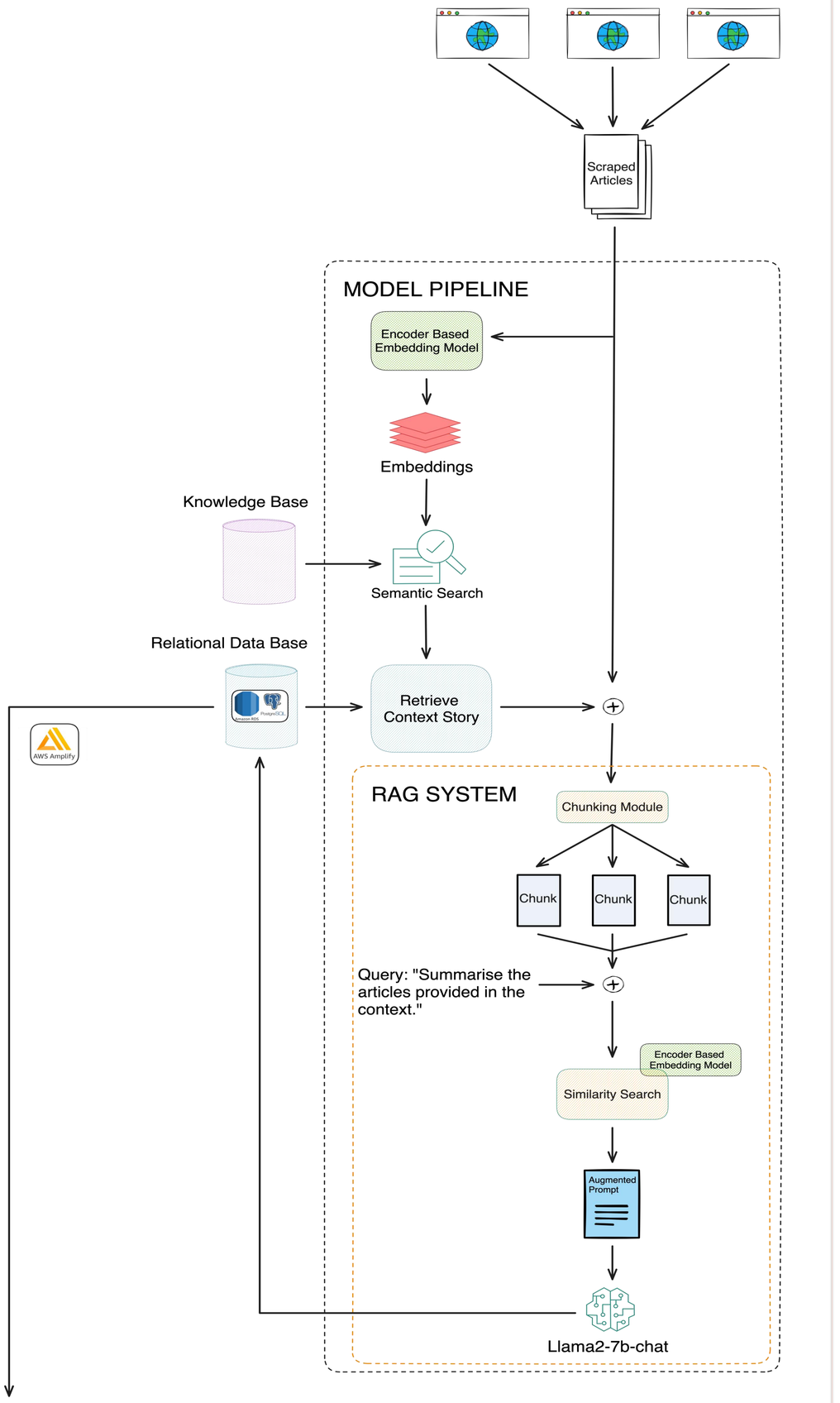
Sourcing Knowledge

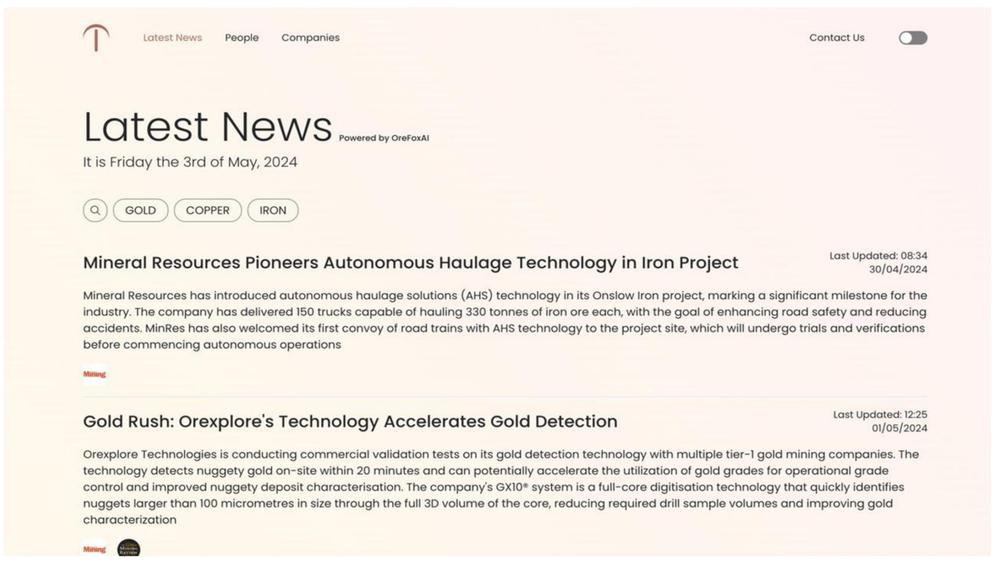
- When generating a summary article from pre-existing ones, we want the Large Language Model (LLM) to focus on relevant paragraphs, like how a human would.
- The Machine Learning community has addressed this task with Retrieval Augmented Generation (RAG) Systems. The process involves:
  1. Chunking the original article
  2. Producing an embedding for each chunk
  3. Creating an embedding for the query that expresses the task in natural language
- The chunks closest to the query embedding are retrieved, as they are deemed the most informative and relevant for the summary.
- This approach allows the LLM to focus on pertinent information when generating the summary article.



PIPELINE AND RESULTS

The web-scraper, embedding search and LLM summariser work in sequence as shown in the pipeline diagram below. A RestAPI is used to insert the summaries into a database, which are then retrieved by the front-end website and served to the user.





CONCLUSION AND FUTURE WORK

The scraping and LLM modules are scheduled to identify new stories each week, keeping the website up to date with the latest news. Future enhancements will involve richer metadata capture. Being able to tag specific corporations, geographies, or sentiments mentioned in the article will allow stories to be cross referenced with other datasets. This could develop into more predictive capabilities in the software architecture, such as of stock movements, commodity prices, or changes in the geopolitical landscape.