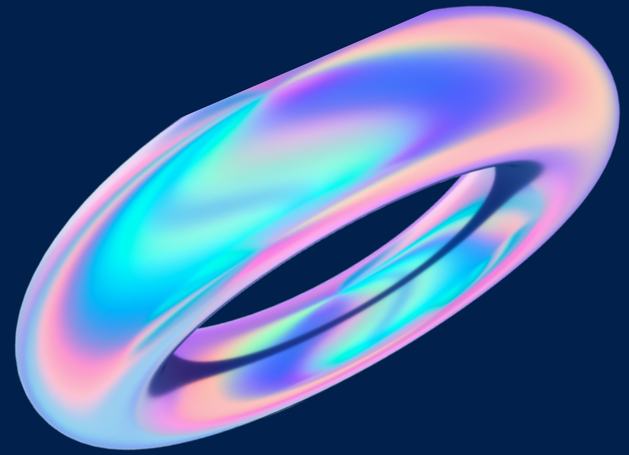
Schlumberger's New Year Hackathon



SHAASTRA 2023

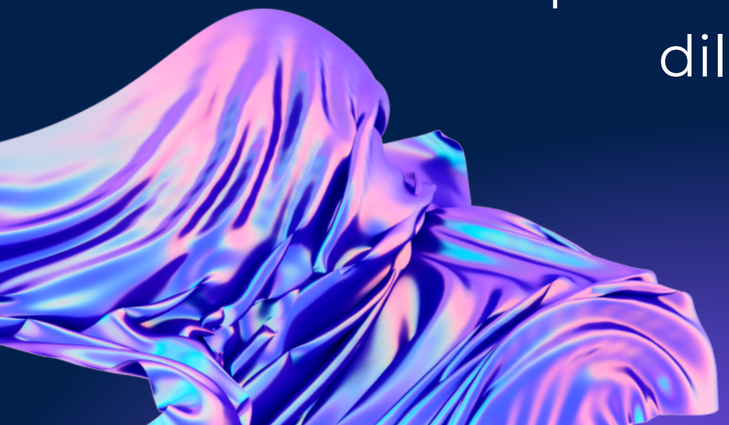
Indian Institute of Technology(IIT), Madras

Bharatfly_Coders

Together, we create amazing technology that unlocks access to energy for the benefit of all.

About Us

The Bharatfly_Coders team, comprising of three highly skilled members from the esteemed National Institute of Technology (NIT) Tiruchirapalli, participated in the highly competitive hackathon hosted by Schlumberger and organized by the renowned Indian Institute of Technology Madras. With a thorough analysis of the presented problem statement and a commitment to delivering the most optimal solution, we are excited to present the results of our diligent efforts. Let us begin the exploration.



Problem statement

In order to minimize manual curation and analysis of energy and carbon-related news and articles, a web scraping and crawling process will be implemented to collect data from various sources. This voluminous data will then be analyzed using advanced Natural Language Processing (NLP) techniques to extract a concise summary while preserving the integrity of the information. The data will be aggregated based on user-provided queries to ensure that it is tailored to their needs.

BASIC APPROACH

Crawl the websites, use Beautiful Soup to get article text, feed into the HuggingFace pipeline for inference, and set up a Flask/Fast API for backend. The problem with this approach is its scalability. Since a single model is loaded, even with batch inputs it won't be able to scale to larger requests or when hosted on machines with multiple GPUs.

```
from transformers import pipeline

classifier = pipeline("summarization")
classifier("Paris is the capital and most populous city of France, with an
## [{ "summary_text": " Paris is the capital and most populous city of Frai
```

Our Approach

1) Multi-Threaded Web Scrapping

Our data collection methodology includes the implementation of advanced web scraping techniques, leveraging the power of Beautiful Soup4 and multi-threading functionality. This approach enables us to scrape multiple websites and articles simultaneously, significantly reducing the overall scraping time.

Additionally, it makes it easier to scale as the FastAPI workers can handle multiple requests simultaneously, and each request, when threaded for various jobs, can be very time efficient. Furthermore, this approach also allows us to retrieve the data more quickly and efficiently, giving us the ability to process large amounts of data in a shorter amount of time.

Our data collection effort encompasses a thorough gathering of articles from the homepages of two reputable websites, specifically climate.mit.edu and electrical-engineering-portal.com. Additionally, we are also extracting relevant articles from the search query results of three highly credible websites, namely netl.doe.gov, iea.org and electrical-engineering-portal.com.

Furthermore, we have implemented a flexible control system for the number of articles we collect from each website. Currently, there is no limit on the number of articles we are collecting from the homepage websites, but we have set a maximum of 25 articles per website for the search query results. This number is adjustable and can be changed in the future as needed to ensure that we are collecting the most relevant and up-to-date data.

Websites	Number of Articles
climate.mit.edu	No Limit
electrical-engineering-portal.com	No Limit
netl.doe.gov	25
iea.org	25
electrical-engineering-portal.com	25

2) Summarization:

There are two types of summarization:

1. Extractive: We extract meaningful sentences and phrases precisely as it exists.
2. Abstractive: We must interpret the context and reproduce the text, keeping core information intact.

For extractive, we can use either TextRank or LexRanker algorithm
In the case of abstraction, there are two choices:

1. Openai text-davinci model
2. T5/Flan-T5/BART/pegasus model (transformer).

Openai is a very accurate paid solution but is costly to scale and takes more time since it's a bigger model. In contrast, the transformer model is a free solution that can be hosted efficiently locally but requires high computing since, unlike openai, it is being run locally. Here we mainly propose the T5 like models.

2) Summarization:

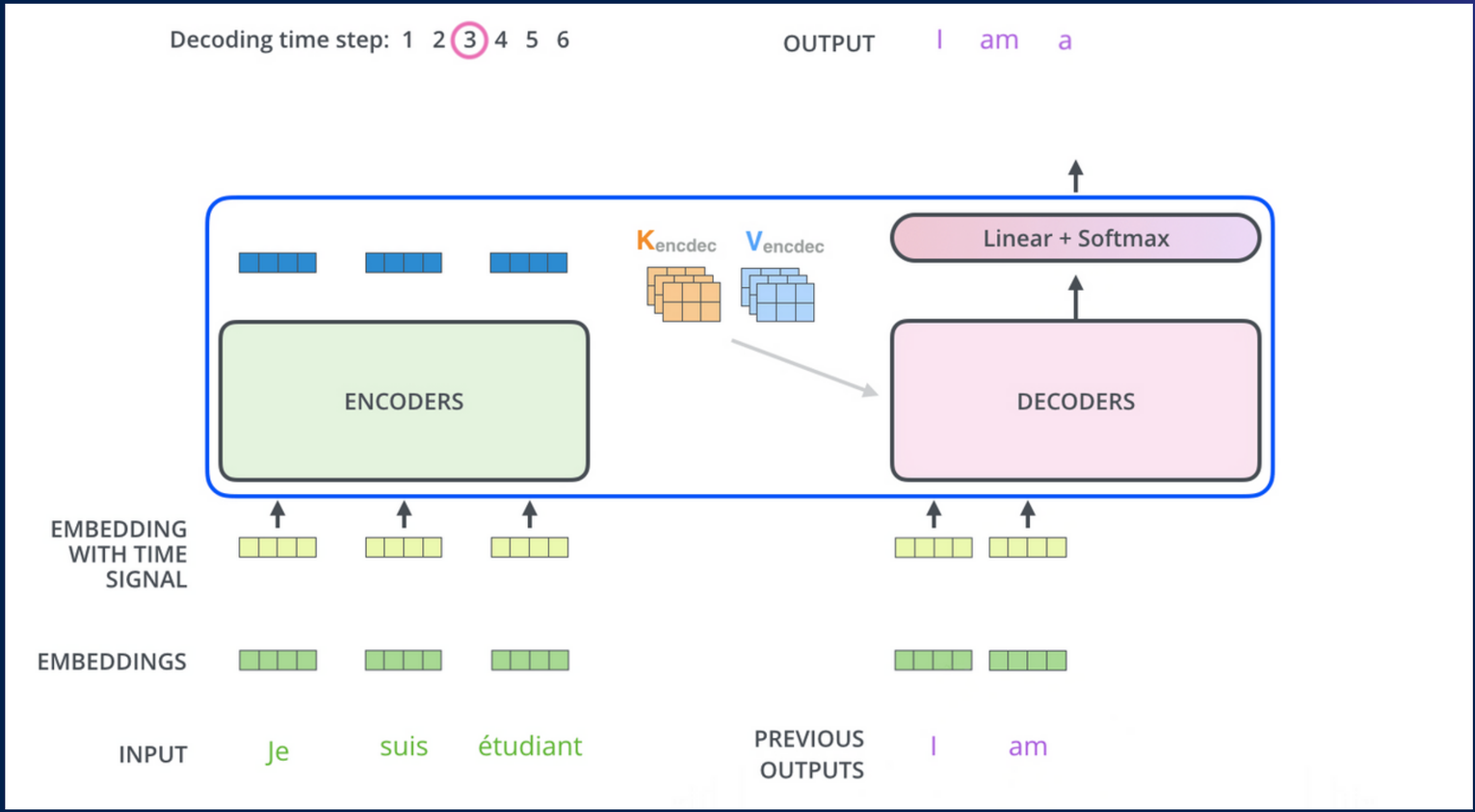
For summarization, we arer providing two kinds:

1) Individual Article Summarization
2) Combined Article Summarization

In the case of combined summarization, we show the overall summary for a search by combining all articles and passing it to the model. Since there is a limit for total input size for all models, including openai, we use lex ranker to get essential points from each article and then pass them combined into the model.

T5 and other transformer models consist of a encoder and a decoder. The output of the encoder for an article as input is a vector embedding representing the contents of an article in a n dimension space. Whereas the decoder decodes this vector back to natural language.

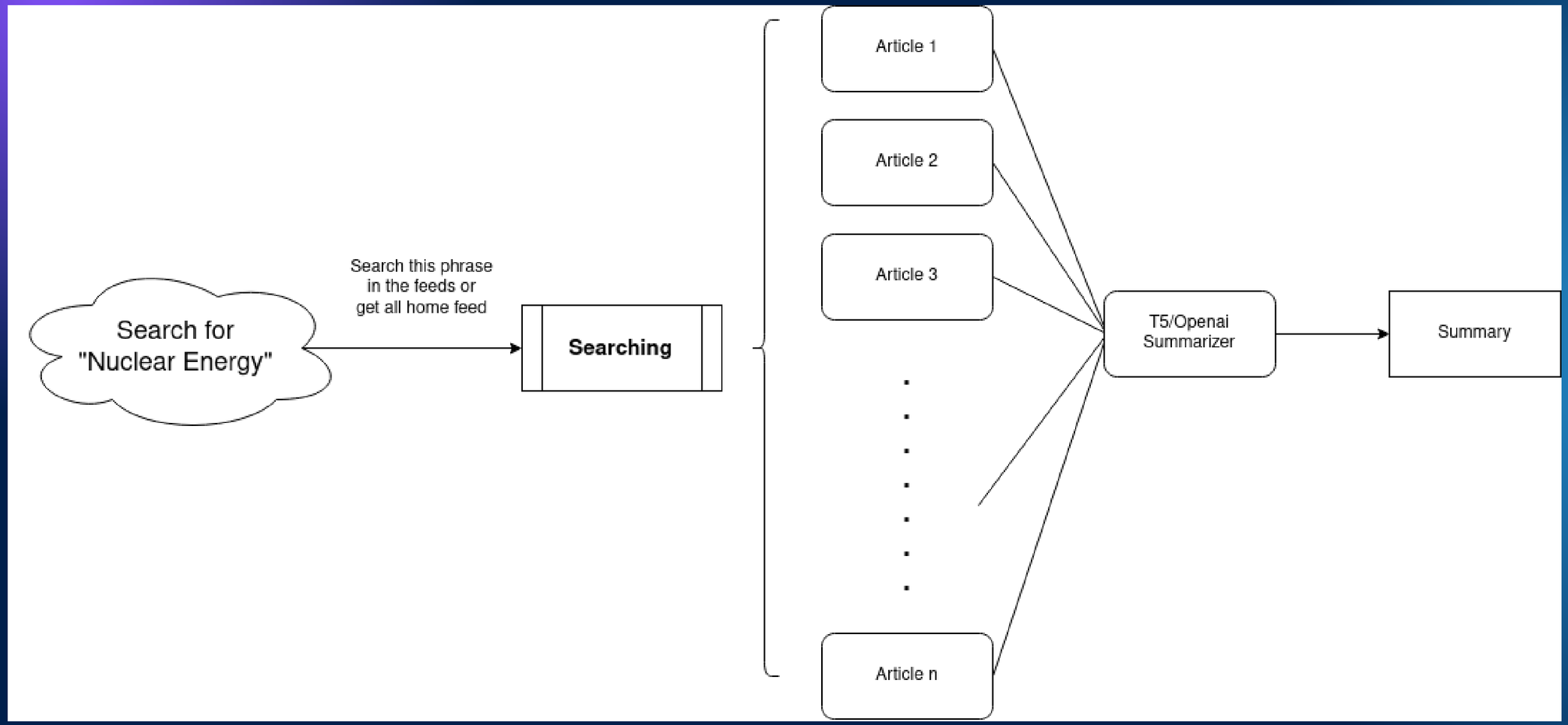
We make use of this encoded vectors to find keywords and to get articles based on queries.



Any of the encoder-decoder model can be used with our architecture

Model	ROUGE 1	ROUGE 2
Flan-T5	47.23	23.51
BART-RXF	40.45	20.69
ERNIE-GENLARGE	39.46	20.34
PEGASUS	39.12	19.86

BASIC SOLUTION



Why Stop here?

3) Query-Article Filtering:

Filtering articles based on direct/fuzzy matching of the keyword searched with article contents is easy. But in the case of query-based filtering, the algorithm needs to understand the query/question and show the articles based on the content/meaning of the query.

In this scenario, we use T5 encoders to generate an encoded embedding vector for the query and use closing similarity w.r.t the article embedding and get a score between -1 and 1. We use this score to filter/sort the articles based on the query.

Since the article needs to be encoded in cases of summarization, we save and use those embedding for this filtering process, thus saving a lot of computing time which we couldn't have done with openai.

Also, since all these require a single model, all the articles can be batched and fed into the model parallelly using Pytorch.

clean energy

Dashboard

Complete Summary

Articles how to maximize profit?

China leading on world’s clean energy investment, says report

China is by far the largest force in global clean energy development and its firms are increasingly looking abroad for opportunities, a new report says.The report, released today by the US-based Insti....

Link **Openai Score: 71.96** **T5 Score: 68.35**

In-depth Q&A: The IPCC’s sixth assessment on how to tackle climate change

Limiting global warming to 1.5C or 2C would mean “rapid and deep” emissions reductions in “all sectors” of the global economy, says the latest report from the United Nations’ Intergovernmental Panel o....

Link **Openai Score: 72.32** **T5 Score: 68.35**

How should facilities optimize their power consumption due to the

T5 Openai TextRank LexRanker

openai Summary

limiting global warming to 1.5C or 2C would mean “rapid and deep” emissions reductions in “all sectors” of the global economy . emissions have continued to rise – albeit at a slowing rate – and it will be “impossible” to stay below 1.5C with “no or limited overshoot” without stronger climate action this decade .

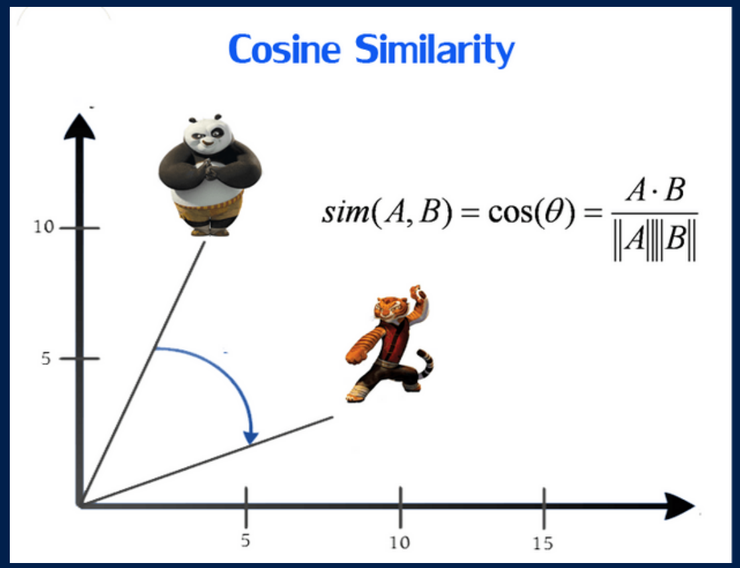
Keywords

warming reductions,reducing climate,cutting emissions,limiting warming,reducing emissions

4) Keyword Generation:

One standard method is to use TF-IDF to generate keywords based on the statistical composition of words in the entire document, which doesn't take the meaning of the article into account.

In this scenario, we again use the T5 encoder and pass each word or 1/2/n words (n-gram) as input for the encoder. Similar to the previous case, we use cosine similarity against the article and the generated word embedding. The highest-scored terms are considered the most contextually representing keywords for a given article.

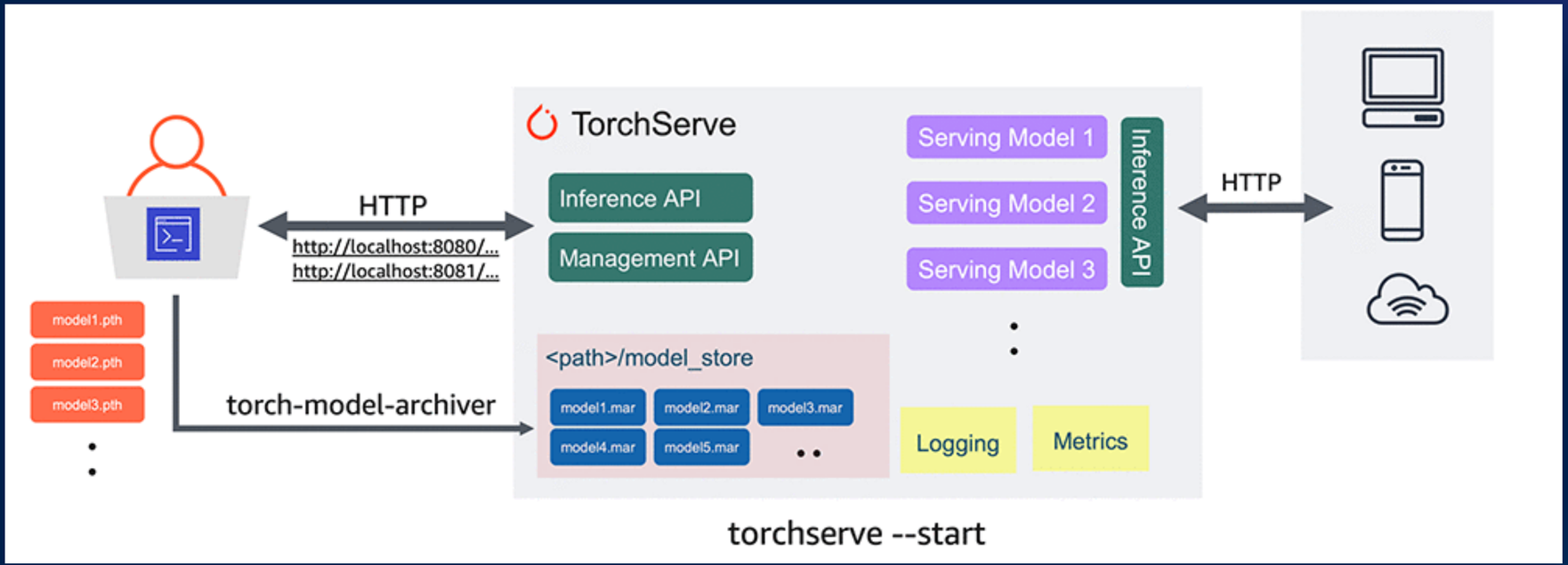


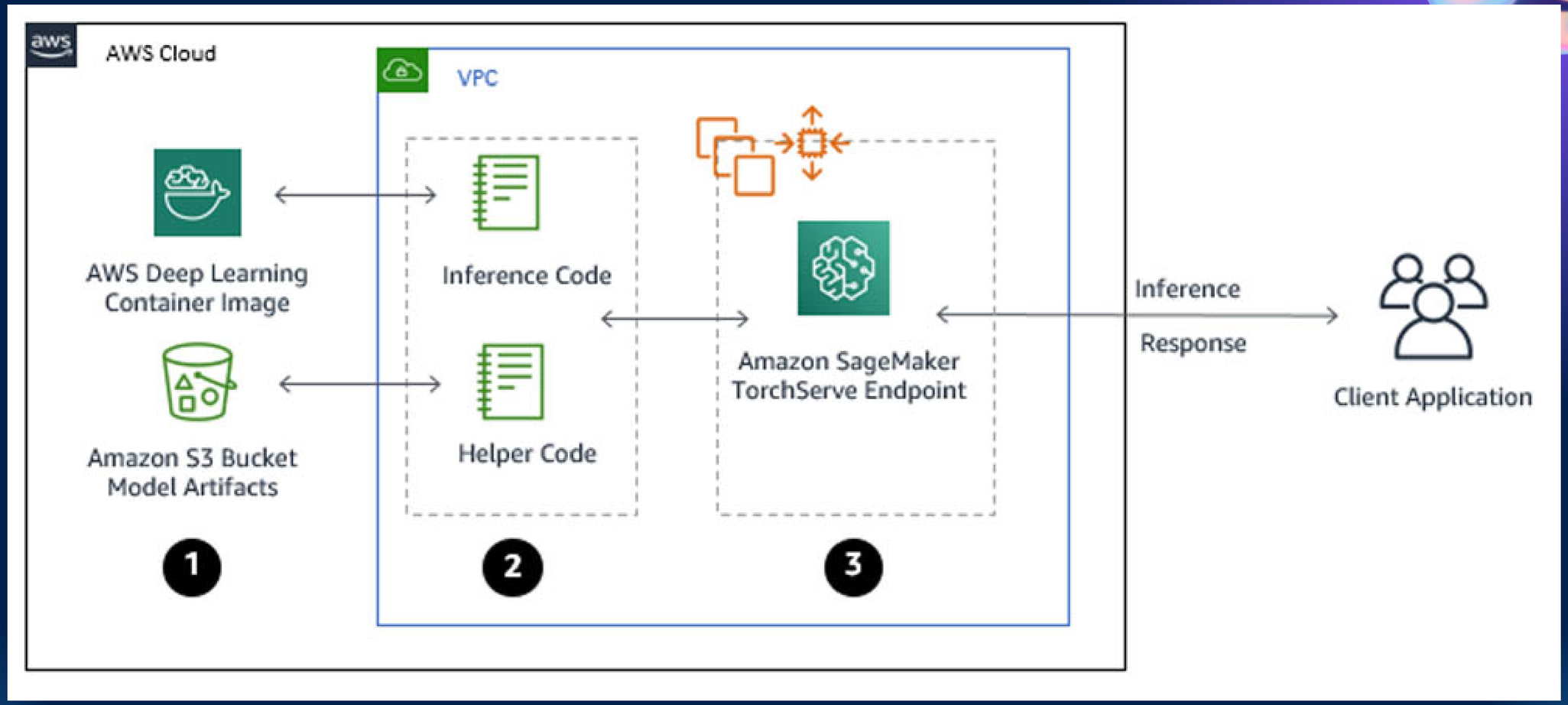
We can use these keywords as new search queries to further search more articles and continue this entire search/crawl nested, allowing us to crawl over broader.

Keywords
warming reductions,reducing climate,cutting emissions,limiting warming,reducing emissions

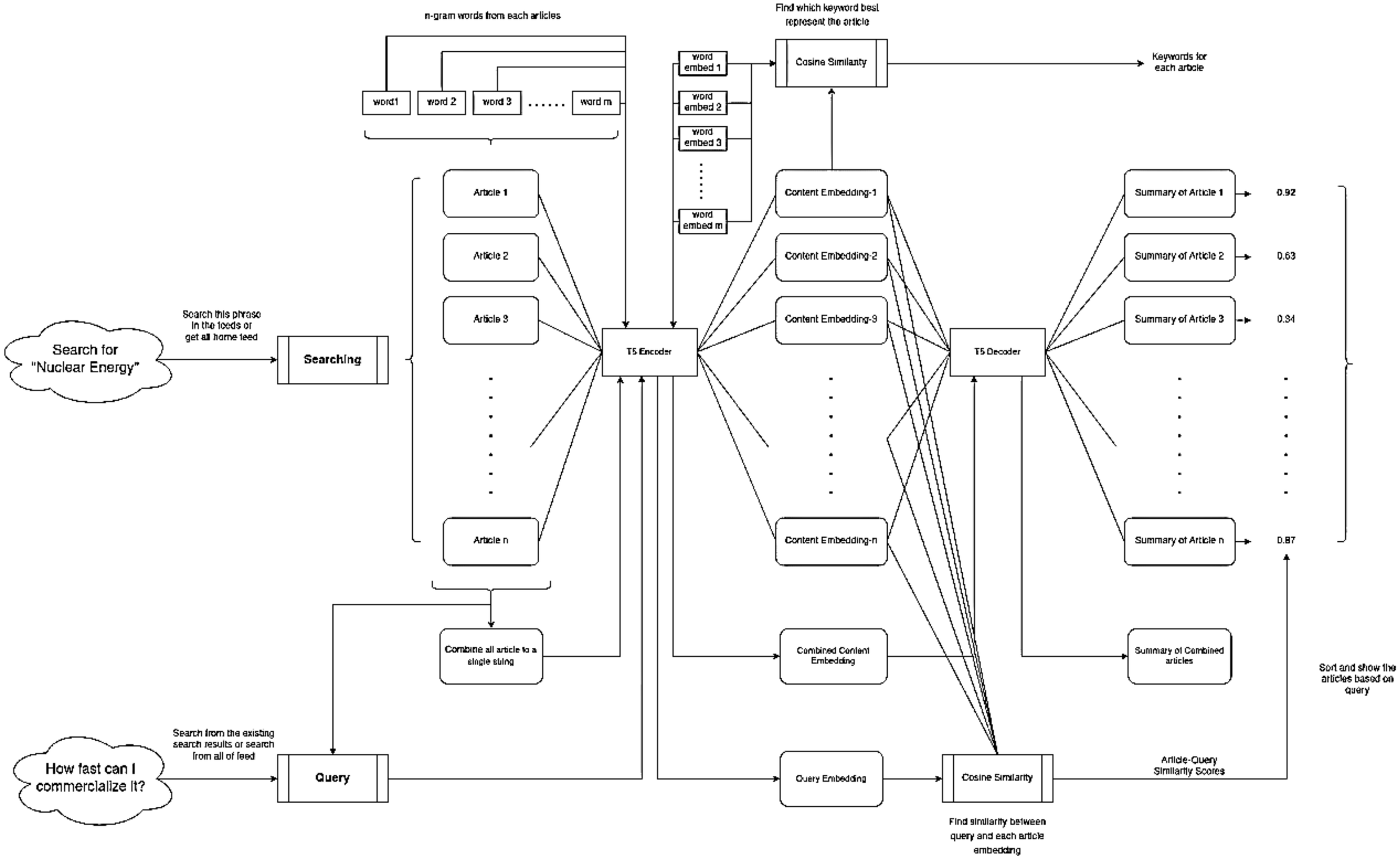
5) Pytorch Serve

We are using PyTorch Serve, which allows us to scale and infer the transformer-based models much more at a production level. It can provide automatic batching of multiple requests, handling multiple GPUs, model allocation/management, health check, and, more importantly, make the API more reliable to serve. It is also directly deployable in cloud services such as AWS.

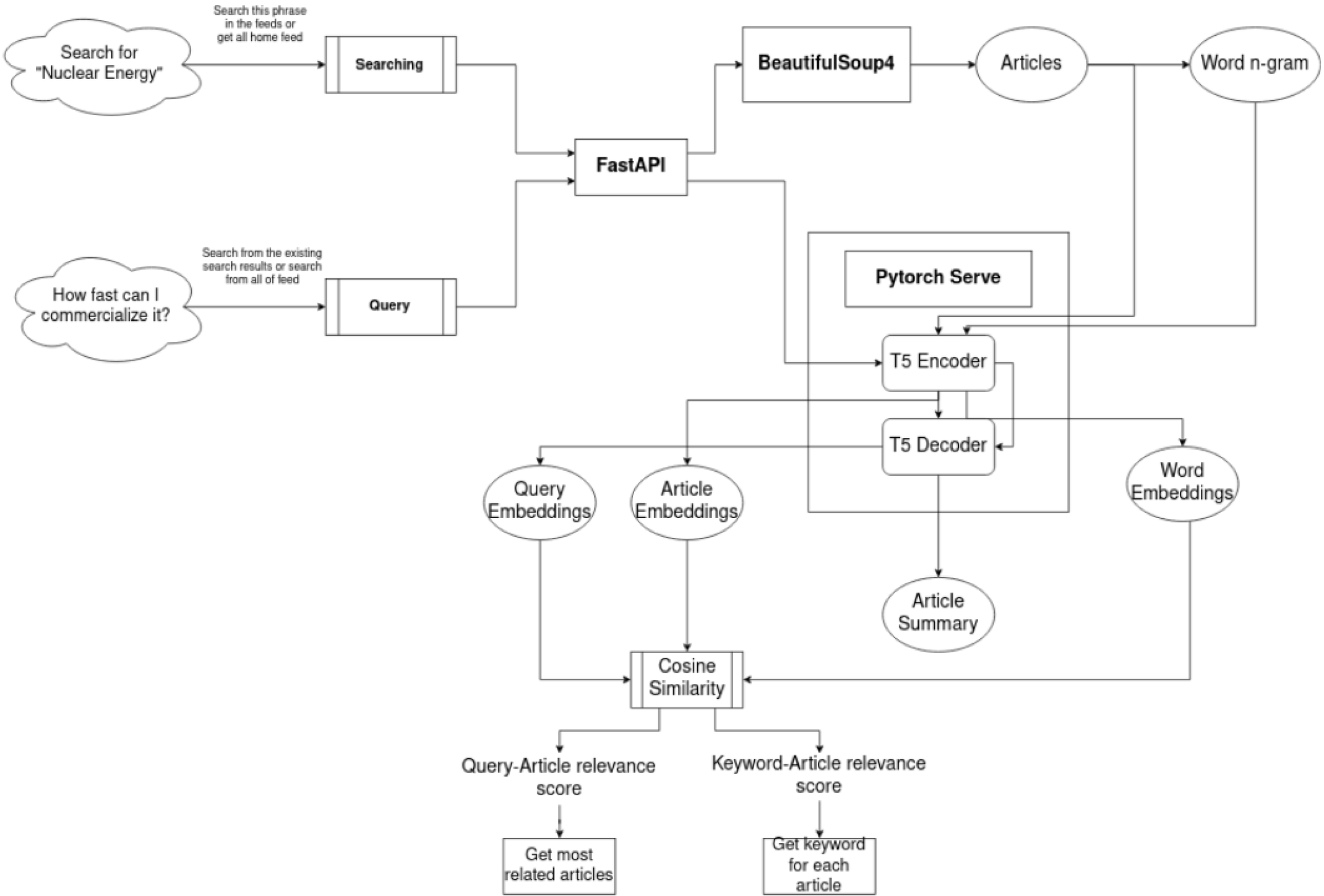



AWS Cloud
VPC
AWS Deep Learning Container Image
Inference Code
Amazon S3 Bucket Model Artifacts
Helper Code
Amazon SageMaker TorchServe Endpoint
Inference
Response
Client Application
1
2
3

IN DEPTH ARCHITECTURE


n-gram words from each articles
word1
word 2
word 3
word m
Find which keyword best represent the article
word embed 1
word embed 2
word embed 3
word embed m
Cosine Similarity
Keywords for each article
Search for "Nuclear Energy"
Search this phrase in the feeds or get all home feed
Searching
Article 1
Article 2
Article 3
Article n
T5 Encoder
Content Embedding-1
Content Embedding-2
Content Embedding-3
Content Embedding-n
T5 Decoder
Summary of Article 1
Summary of Article 2
Summary of Article 3
Summary of Article n
0.92
0.63
0.84
0.87
Combine all article to a single string
Combined Content Embedding
Summary of Combined articles
How fast can I commercialize it?
Search from the existing search results or search from all of feed
Query
Query Embedding
Cosine Similarity
Article-Query Similarity Scores
Find similarity between query and each article embedding
Sort and show the articles based on query

IMPLEMENTATION ARCHITECTURE


Search for "Nuclear Energy"
Search this phrase in the feeds or get all home feed
Searching
BeautifulSoup4
Articles
Word n-gram
FastAPI
How fast can I commercialize it?
Search from the existing search results or search from all of feed
Query
Pytorch Serve
T5 Encoder
T5 Decoder
Query Embeddings
Article Embeddings
Word Embeddings
Article Summary
Cosine Similarity
Query-Article relevance score
Keyword-Article relevance score
Get most related articles
Get keyword for each article

Advantages

We use SOTA models and give results based on the content of the article rather than the traditional statistical methods, which tend to be inacccurate.

In this scenario, we use the T5 model to cover all the features and can use the encoded vectors for multiple jobs. This helps save much computing, allows for efficiently parallelizing the computation, and is extremely easy when combined with torch serve.

T5 is a free model, making it cheaper when scaled to a larger target audience than the openai model. The Openai model can still be used in case of personal use-only circumstances.

Working Application

Demo

clean energy

Dashboard

Complete Summary

Articles how to maximize profit?

China leading on world's clean energy investment, says report

China is by far the largest force in global clean energy development and its firms are increasingly looking abroad for opportunities, a new report says.The report, released today by the US-based Insti....

Link Openai Score: 71.96 T5 Score: 68.35

In-depth Q&A: The IPCC's sixth assessment on how to tackle climate change

Limiting global warming to 1.5C or 2C would mean "rapid and deep" emissions reductions in "all sectors" of the global economy, says the latest report from the United Nations' Intergovernmental Panel o....

Link Openai Score: 72.32 T5 Score: 68.35

How should facilities optimize their power consumption due to the

T5 Openai TextRank LexRanker

openai Summary

limiting global warming to 1.5C or 2C would mean "rapid and deep" emissions reductions in "all sectors" of the global economy . emissions have continued to rise – albeit at a slowing rate – and it will be "impossible" to stay below 1.5C with "no or limited overshoot" without stronger climate action this decade .

Keywords

warming reductions,reducing climate,cutting emissions,limiting warming,reducing emissions

We are able to generate both types of summaries including keywords which can be used for future nested searches

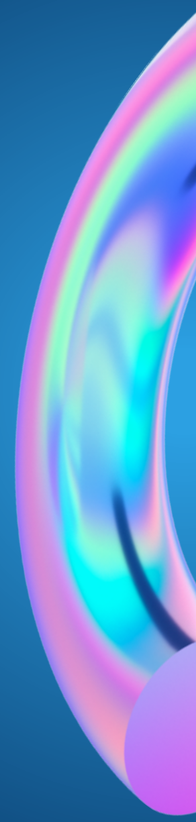
T5 | Openai | TextRank | LexRanker

t5 summary

- These resources of energy can be naturally replenished and are safe for the environment.
- Examples of renewable sources of energy are: Solar energy, geothermal energy, wind energy, biomass, hydropower and tidal energy.
- The resources that cannot be renewed once they are consumed are called non-renewable sources of energy.
- Natural Sources of Energy
- Following are the sources of energy that are renewable:
- What are the advantages and disadvantages of wind power?
- It is one of the clean sources of energy.
- Nuclear energy can be used to create electricity, but it must first be released from the atom.
- A nuclear reactor, or power plant, is a series of machines that can control nuclear fission to produce electricity.
- Because nuclear fuel can be used to create nuclear weapons as well as nuclear reactors, only nations that are part of the Nuclear Non-Proliferation Treaty (NPT) are allowed to import uranium or plutonium, another nuclear fuel.

Keywords

energy resources,resources energy,renewable resources,energy sources,renewable resource



Demo Video:

https://drive.google.com/file/d/14tntrE3LSb9CngHxyfdgoHXkG0ah5M0q/view?usp=share_link

Github Repo link

https://github.com/FrozenWolf-Cyber/Schlumberger-SHAASTRA

Thank You

We acknowledge our sincere thanks to the hackathon organisers who gave us an opportunity to work hands on, in web scraping and NLP techniques.

Wherever the drill goes, Schlumberger goes