



Results of web scraping from blog posts are shown below.

"\n\n\n\n\n\n\n\n\n\nSocial lisetening Social Listening 2023 Social Listening Analytics Social Listening Tool Trend Marketing 2023 Wisesight,Zocial Eye \nNew era social listening via KIRIN A.I engine

About us (<https://wisesight.com/about-us/>)

This section, Using the same method as blog posts section but web pages does not return the content if the request does not include headers "User-Agent" as shown

```
r = requests.get("https://wisesight.com/about-us/")
r.text
```

✓ 0.2s

```
'403 - Forbidden | Access to this page is forbidden.\n'
```

After include headers "User-Agent". web pages normally return the content. You can seen the implementation on the code-based

This page has returned a lot of whitespaces on text output. To fix the issue by split the text to a list of text and join it back to single text with one spacebar.

Results of scraping are shown below.

"WISESIGHT - Thailand's No.1 Social Media Analyzing Service Skip to content PlatformZocial EyeInfluencer DirectorySocial MetricOmni Channel SolutionWarroomCommand CenterServiceResearchData ConsultingCh

## Data Cleaning

In this section, To remove special characters on raw data such as \n, \r etc. by replace it with empty characters apply this to 3 sources that mentioned before.

### Results of Data Cleaning.

[1] วิชาชีพที่คิดจะละเลย Series 8 มูลค่ากว่า 7 ล้านเหรียญสหรัฐ จีนมี 2 พาร์ทเนอร์ยักษ์ใหญ่ กรุ๊ปซี ฟินโนเวต และ เทค แมทริคซ์ เร่งพัฒนาโซลูชันขยายตลาด มุ่งสู่การเป็นแพลตฟอร์มระดับโลก มุ่งสู่การเป็นแพลตฟอร์มระดับโลก มุ่งสู่การเป็นแพลตฟอร์มระดับโลก กรุงเทพฯ - 17 สิงหาคม 2564 : วิชาชีพ Social Listening Social Listening 2023 Social Listening Analytics Social Listening Tool Trend Marketing 2023 Wisesight, Zocial Eye New era social listening via KIRIN A.I.engine for data analysis find "WISESIGHT - Thailand's No.1 Social Media Analyzing Service Skip to content PlatformZocial EyeInfluencer DirectorySocial Media MetricOmni Channel SolutionWarroomCommand CenterServiceResearchData Consulting

## Preprocess for RAG

Document data after Data cleaning can be used directly with RAG but it will slow down the response time as huge text input requires more compute time and token but it can be split into smaller text (chunk). To reduce compute time, token and cost per question. RAG framework has tools for doing the task.

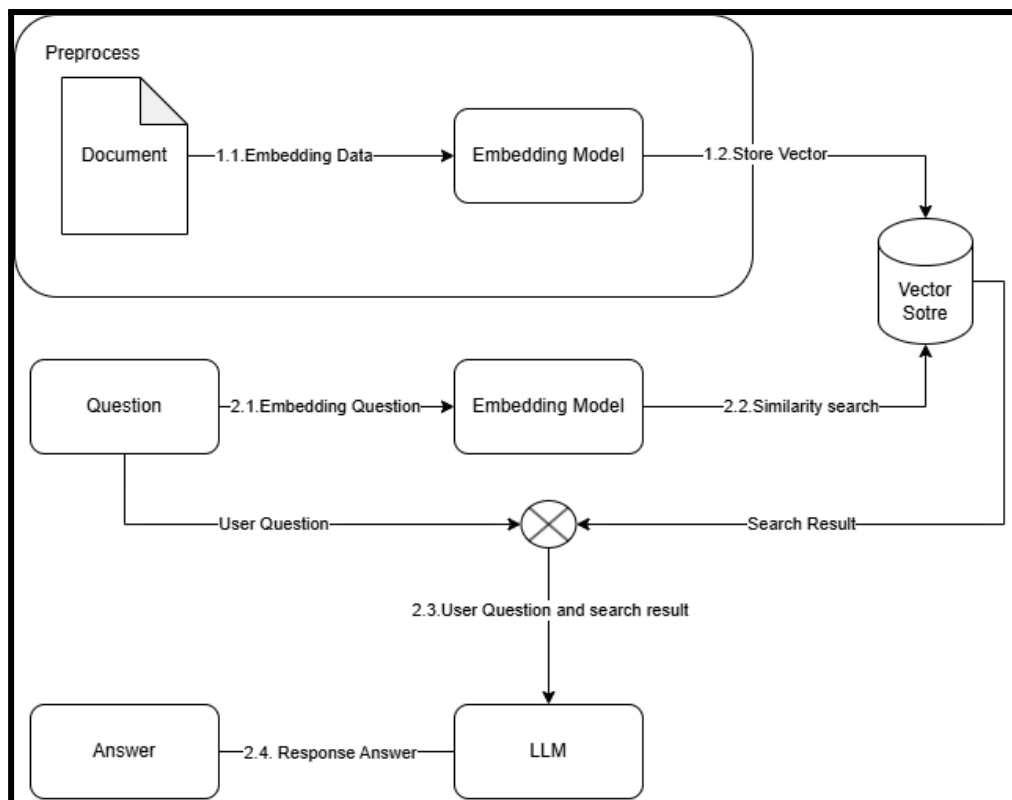
As the tools to be used are mainly built for the English language that separate word by spaces but for Thai language separate word by sense of reader. To fix the problems, by using pythainlp word tokenize library by default. pythainlp using “newmm” to generate English like the Thai language that separates words by spaces. Result after preprocess.

[1] ไชยสิทธิ์ บิด คือ ระดมทุน Series B มูลค่ากว่า 7 ล้านเหรียญสหรัฐ จีเอ็มเอ็ม 2 พาร์กเกอร์ ยักษ์ใหญ่ กรุ๊ปพีเอ็ม โน เวล และ เทค แมทริคซ์ เร่งพัฒนา โซลูชัน 5 ขวบ ขยาย ตลาด มุ่งสู่ การ เป็น เทค สตาร์ทอัพ  
 "Social listening Social Listening 2023 Social Listening Analytics Social Listening Tool Trend Marketing 2023 Wisesight, Social Eye New era social listening v  
 "Wisesight - Thailand's No.1 Social Media Analyzing Service Skip to content Platform Social Eye Influencer Directory Social Metricomi Channel Solution Warroom Command

# Architecture and Approach

Using Retrieval-augmented generation (RAG) to provide specific data and context to Generative AI to generate answers to users and reduce hallucination from the model. Frameworks used in the assignment are **llama index**.

## Architecture



### High Level Design of RAG

1. This step is a preprocess to ingest documents with embedding to be stored for query and provide context to LLM as input with user questions.
2. When the user input the question. The system will query from the vector store. The top result (Number of top results can be configured but more input will be affected response time and cost). with user questions will fetch to LLM to generate the answer for the user.

You can change the LLM model to be any favor you want or to be matched with your budget and requirement. It can be OpenAI api, Ollama, Gemini etc. In report, using **Ollama with llama3.2:3b model running locally**.

Embedding model is **BAI/bge-m3** on HuggingFace.

## Challenges faced and solutions

- To read PDF documents to text. Which library is suitable for reading Thai language?
  - Solution: Do research. Try all of the available libraries and compare the result.
- RAG tools are written for English that separate words by whitespace but Thai language.
  - Solution: tokenize thai document by pythainlp to be using RAG tools.

## Analyze, Performance and Observations.

### Question and answer from system.

- **Q: Who is wiselight CEO?**
  - **A: Kla Tangsuwan**
- **Q: wiselight ทำอะไร**
  - **A: Wiselight ช่วยให้คุณสามารถวิเคราะห์และเข้าใจข้อมูลจากโซเชียลมีเดียได้อย่างเหมาะสม และช่วยให้ธุรกิจของคุณสามารถตัดสินใจได้ดีขึ้นโดยใช้ข้อมูลที่ถูกต้องและเป็นประโยชน์**
- **Q: wiselight ระดมทุน Series B จากใคร?**
  - **A: กรุงศรี ฟิน โน เวต และ เทค แมทริกซ์**
- **Q: ต้องการติดต่อ wiselight**
  - **A: คุณสามารถติดต่อ WISELIGHT ได้ผ่านอีเมลโดยการเข้าไปที่เว็บไซต์ของพวกเขาและคลิกที่ส่วน "ติดต่อเรา" หรือสามารถส่งอีเมลโดยตรงไปยัง contact@wiselight.com เพื่อรับการตอบกลับ**
- **Q: KIRN A.I. ทำอะไรได้บ้าง?**
  - **A: KIRN A.I. ช่วยสรุปภาพรวมของการเติบโตของแคมเปญผ่านช่องทางต่างๆ และให้ข้อมูลการใช้งาน Social Listening เพื่อหาความเข้าใจลูกค้าและเพิ่มประสิทธิภาพในการตลาด online**

## Analysis and Observations

Embedding model and Similarity search are working great for selecting a part of a document to be fetched to the LLM model. You can see the evidence in the evaluation section on jupyter notebook. From the answer above, The LLM model works great when asking close-ended questions but on open-ended questions some it has hallucination and answers out of context. It might be solved by using a bigger model to be better at language understanding.

Another issue, Have found is that the response output has space between words for Thai answers. This issue can be from word tokenization preprocess. It can be fix by have other method to handle the chunk spilt.

## Performance metrics

As a NLP with generative tasks, It depends on your favor or user favor for the answer. It can be tuned to match with your favorite style of answer. It is quite hard to measure because it depends on humans.