Team Project Progress Report

Project Title: Generative Search

Team Members:

- Team Captain: Mohammad Tamim (NetID: tamim2)
- Team Members:
  - Timothy Cannella (NetID: tdc5)
  - Luke Freitag (NetID: lukegf2)

Introduction

In the process of developing a Generative Search system, our team has organized our efforts into four distinct tracks, each playing a pivotal role in achieving our goal. In the planning section, we've outlined the details of our approach, from data collection and system design to results evaluation and reporting. Here we present the first progress report, showcasing our accomplishments and the path forward. With the foundation of meticulous planning and diligent execution, our project is well on its way to delivering a powerful Generative Search system that will undoubtedly meet our objectives.

Planning

Our team is currently working on developing a Generative Search system, and we've organized our efforts into four distinct tracks to ensure a smooth and efficient process:

1. **Track 1: Data Collection**
   a. <u>Transcript Collection:</u> In this phase, we're focused on gathering all the “CS 410 Text Information Systems” transcripts from Coursera. Our aim is to retrieve and store this textual information in a structured manner, including the lecture headings. We will save these transcripts in a text file.

   b. <u>Q&A Dataset Creation</u>: For this task, we're leveraging ChatGPT to generate questions and answers for each lecture. Our goal is to create a versatile dataset that includes 10 different paraphrases of the same question, all while ensuring that the answers to these questions remain identical. To illustrate, if we have a question like “What is embedding?”, we'll generate 10 other ways of asking the same question, such as 'Can you define embedding?' or “What do you mean by embedding?”. The final output will be a .csv file with two columns: “question” and “response”. This dataset will be invaluable in enhancing our system's ability to check for similarity during the document retrieval process within the Generative Search system.

   By breaking down our work into these two critical tracks within the Data Collection phase, we're making significant progress towards the successful development of our Generative Search system.

2. **Track 2: Design, Development, and Testing**
    a. Design the Generative Search system: Our first step involves designing the Generative Search system. we'll create a visual workflow that outlines how the system will function.

    b. Develop the Generative Search system:
        i. *Convert Questions to Embedded Vectors:* Using Google Colab and python, we will develop the code to take questions from the Q&A dataset and transform them into context-embedded vectors. The results will be stored in a .csv file with three columns: “question”, “response”, and “embedded_questions”.

        ii. *Online Generative Search Module:* This module is a crucial part of the system, which uses Retrieval Augmented Generation (RAG). In Google Colab, it accepts input from users, preprocesses the input by converting it into an embedded vector, and then performs a search to find the most similar question within the 'embedded_questions' using a similarity metric. Once the most similar question is identified, the system retrieves the corresponding response. It then utilizes a Large Language Model (LLM) along with prompt engineering techniques to generate an answer based on the gathered context.

    c. Testing: Testing is a critical aspect of our development process. We conduct rigorous testing to ensure the functionality and accuracy of the system. This phase ensures that the Generative Search system operates as intended and provides reliable results.

By following these steps in the Design, Development, and Testing phase, we are working towards building a robust and effective Generative Search system that will meet our objectives.

3. **Track 3: Results Evaluation**
    a. Evaluation Sample Set Creation: First, we'll create a set of 50 questions for evaluation purposes. Half of these questions (25) will be identical or semantically similar to questions from the Q&A dataset related to the course (e.g., “What is embedding?”) meant to assess false negatives. The other half (25) will consist of unrelated questions meant to assess false positives (e.g., “how is the weather today?”, “It's my birthday today?”, “I want to order food?”). The result of this step will be a .csv file with three columns: “question”, “generated_response”, and “accurate_response”. For the first 25 questions, we can obtain the accurate responses from the Q&A dataset. For the other 25 unrelated questions, we will copy and paste the following response in the “accurate_response” column: “I'm sorry, but I don't have expertise in this topic; my training data is limited to only “CS 410 Text Information Systems””.

b. Quantitative Evaluation: In Python, we will perform a cosine similarity analysis between the “generated_response” and “accurate_response” for the first 25 questions related to the course. We will report the cosine similarity scores with 3 decimal points (e.g., 0.789). Responses with a cosine similarity score above 0.5 will be considered accurate. For the second set of 25 unrelated questions, we will assign a score of 1 if the “generated_response” is “I'm sorry, but I don't have expertise in this topic; my training data is limited to only “CS 410 Text Information Systems.””. Otherwise, we will assign a score of 0. The results will be saved in a .csv format, with four columns: “question”, “generated_response”, “accurate_response”, and “cosine similarity”.

c. Qualitative Evaluation: This phase involves a manual review of the 50 questions and responses. Team members will assess the accuracy and relevance of the “generated_response” in comparison to the “accurate_response”. A .csv file will be created with four columns: “question”, “generated_response”, “accurate_response”, and “is_generated_response_relevant?”. The “is_generated_response_relevant?” column will contain values of 0 or 1 to indicate whether the generated response is relevant.

By conducting these evaluations, we aim to comprehensively assess the performance of our Generative Search system, ensuring both quantitative accuracy and qualitative relevance in responses.

4. **Track 4: Reporting**

a. Prepare the Progress Report: We'll create an initial progress report for submission. This report will outline our project's status, achievements, and the direction we're heading.

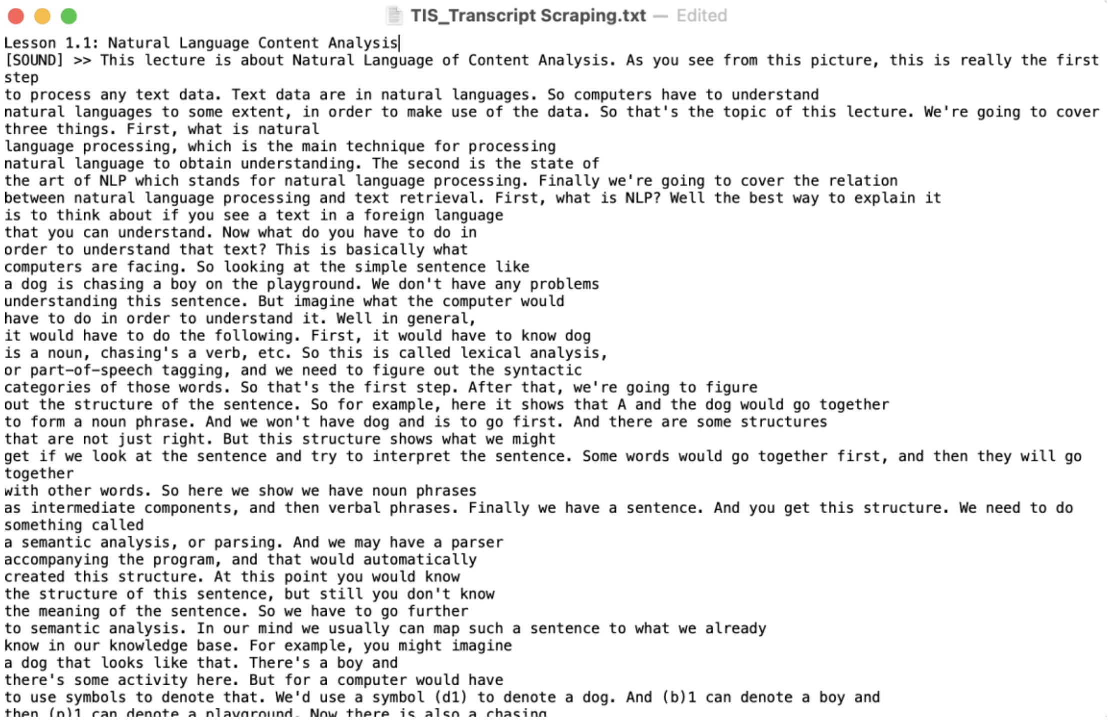
b. Create the Final Report: The final report is a comprehensive document that will include project details, the methodology employed, a summary of results, evaluation outcomes, and key recommendations.

c. Prepare the Presentation: In addition to the written reports, we will prepare a presentation to effectively communicate our project's goals, progress, findings, and recommendations.

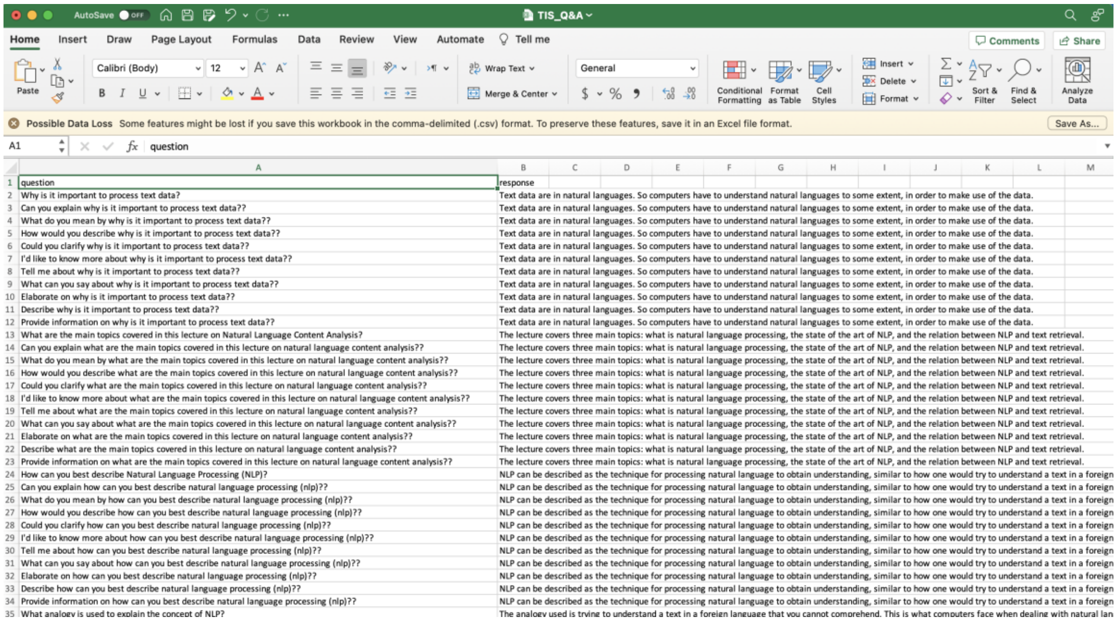
These reporting activities are crucial in documenting our project's journey, outcomes, and future steps, ensuring that our work is transparent and accessible to relevant parties.

Progress Overview

As of now, Track 1 has been successfully completed, encompassing both the Transcript Collection and Q&A Dataset Creation. We have gathered and stored the “CS 410 Text Information Systems” transcripts and generated a versatile Q&A dataset with paraphrased questions and identical answers. Below, we have also included a sample of the collected "CS 410 Text Information Systems" transcripts from Coursera (a few observations) alongside the paraphrased Q&A dataset (a few observations).



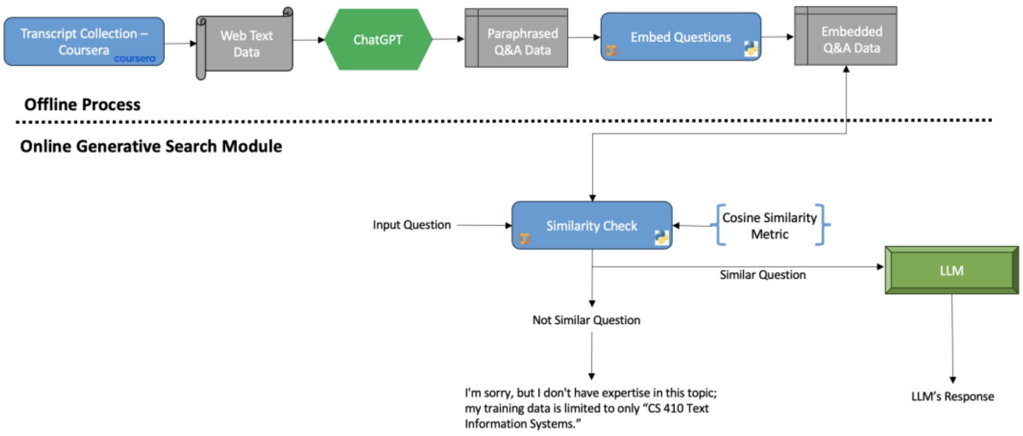
```
Lesson 1.1: Natural Language Content Analysis
[SOUND] >> This lecture is about Natural Language of Content Analysis. As you see from this picture, this is really the first
step
to process any text data. Text data are in natural languages. So computers have to understand
natural languages to some extent, in order to make use of the data. So that's the topic of this lecture. We're going to cover
three things. First, what is natural
language processing, which is the main technique for processing
natural language to obtain understanding. The second is the state of
the art of NLP which stands for natural language processing. Finally we're going to cover the relation
between natural language processing and text retrieval. First, what is NLP? Well the best way to explain it
is to think about if you see a text in a foreign language
that you can understand. Now what do you have to do in
order to understand that text? This is basically what
computers are facing. So looking at the simple sentence like
a dog is chasing a boy on the playground. We don't have any problems
understanding this sentence. But imagine what the computer would
have to do in order to understand it. Well in general,
it would have to do the following. First, it would have to know dog
is a noun, chasing's a verb, etc. So this is called lexical analysis,
or part-of-speech tagging, and we need to figure out the syntactic
categories of those words. So that's the first step. After that, we're going to figure
out the structure of the sentence. So for example, here it shows that A and the dog would go together
to form a noun phrase. And we won't have dog and is to go first. And there are some structures
that are not just right. But this structure shows what we might
get if we look at the sentence and try to interpret the sentence. Some words would go together first, and then they will go
together
with other words. So here we show we have noun phrases
as intermediate components, and then verbal phrases. Finally we have a sentence. And you get this structure. We need to do
something called
a semantic analysis, or parsing. And we may have a parser
accompanying the program, and that would automatically
created this structure. At this point you would know
the structure of this sentence, but still you don't know
the meaning of the sentence. So we have to go further
to semantic analysis. In our mind we usually can map such a sentence to what we already
know in our knowledge base. For example, you might imagine
a dog that looks like that. There's a boy and
there's some activity here. But for a computer would have
to use symbols to denote that. We'd use a symbol (d1) to denote a dog. And (b)1 can denote a boy and
then (p)1 can denote a playground. Now there is also a chasing
```



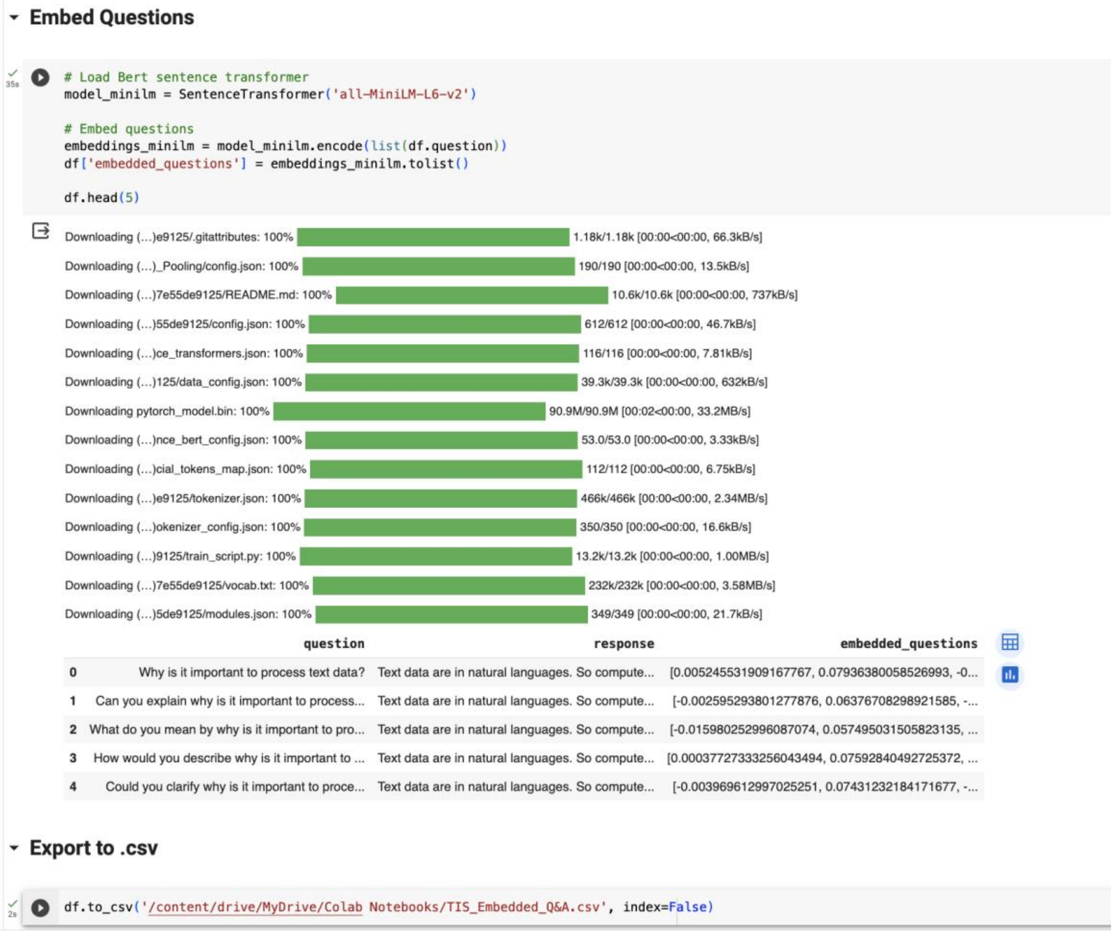
Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

	A	B
1	question	response
2	Why is it important to process text data?	Text data are in natural languages. So computers have to understand natural languages to some extent, in order to make use of the data.
3	Can you explain why is it important to process text data??	Text data are in natural languages. So computers have to understand natural languages to some extent, in order to make use of the data.
4	What do you mean by why is it important to process text data??	Text data are in natural languages. So computers have to understand natural languages to some extent, in order to make use of the data.
5	How would you describe why is it important to process text data??	Text data are in natural languages. So computers have to understand natural languages to some extent, in order to make use of the data.
6	Could you clarify why is it important to process text data??	Text data are in natural languages. So computers have to understand natural languages to some extent, in order to make use of the data.
7	I'd like to know more about why is it important to process text data??	Text data are in natural languages. So computers have to understand natural languages to some extent, in order to make use of the data.
8	Tell me about why is it important to process text data??	Text data are in natural languages. So computers have to understand natural languages to some extent, in order to make use of the data.
9	What can you say about why is it important to process text data??	Text data are in natural languages. So computers have to understand natural languages to some extent, in order to make use of the data.
10	Elaborate on why is it important to process text data??	Text data are in natural languages. So computers have to understand natural languages to some extent, in order to make use of the data.
11	Describe why is it important to process text data??	Text data are in natural languages. So computers have to understand natural languages to some extent, in order to make use of the data.
12	Provide information on why is it important to process text data??	Text data are in natural languages. So computers have to understand natural languages to some extent, in order to make use of the data.
13	What are the main topics covered in this lecture on Natural Language Content Analysis?	The lecture covers three main topics: what is natural language processing, the state of the art of NLP, and the relation between NLP and text retrieval.
14	Can you explain what are the main topics covered in this lecture on natural language content analysis??	The lecture covers three main topics: what is natural language processing, the state of the art of NLP, and the relation between NLP and text retrieval.
15	What do you mean by what are the main topics covered in this lecture on natural language content analysis??	The lecture covers three main topics: what is natural language processing, the state of the art of NLP, and the relation between NLP and text retrieval.
16	How would you describe what are the main topics covered in this lecture on natural language content analysis??	The lecture covers three main topics: what is natural language processing, the state of the art of NLP, and the relation between NLP and text retrieval.
17	Could you clarify what are the main topics covered in this lecture on natural language content analysis??	The lecture covers three main topics: what is natural language processing, the state of the art of NLP, and the relation between NLP and text retrieval.
18	I'd like to know more about what are the main topics covered in this lecture on natural language content analysis??	The lecture covers three main topics: what is natural language processing, the state of the art of NLP, and the relation between NLP and text retrieval.
19	Tell me about what are the main topics covered in this lecture on natural language content analysis??	The lecture covers three main topics: what is natural language processing, the state of the art of NLP, and the relation between NLP and text retrieval.
20	What can you say about what are the main topics covered in this lecture on natural language content analysis??	The lecture covers three main topics: what is natural language processing, the state of the art of NLP, and the relation between NLP and text retrieval.
21	Elaborate on what are the main topics covered in this lecture on natural language content analysis??	The lecture covers three main topics: what is natural language processing, the state of the art of NLP, and the relation between NLP and text retrieval.
22	Describe what are the main topics covered in this lecture on natural language content analysis??	The lecture covers three main topics: what is natural language processing, the state of the art of NLP, and the relation between NLP and text retrieval.
23	Provide information on what are the main topics covered in this lecture on natural language content analysis??	The lecture covers three main topics: what is natural language processing, the state of the art of NLP, and the relation between NLP and text retrieval.
24	How can you best describe Natural Language Processing (NLP)?	NLP can be described as the technique for processing natural language to obtain understanding, similar to how one would try to understand a text in a foreign
25	Can you explain how can you best describe natural language processing (nlp)??	NLP can be described as the technique for processing natural language to obtain understanding, similar to how one would try to understand a text in a foreign
26	What do you mean by how can you best describe natural language processing (nlp)??	NLP can be described as the technique for processing natural language to obtain understanding, similar to how one would try to understand a text in a foreign
27	How would you describe how can you best describe natural language processing (nlp)??	NLP can be described as the technique for processing natural language to obtain understanding, similar to how one would try to understand a text in a foreign
28	Could you clarify how can you best describe natural language processing (nlp)??	NLP can be described as the technique for processing natural language to obtain understanding, similar to how one would try to understand a text in a foreign
29	I'd like to know more about how can you best describe natural language processing (nlp)??	NLP can be described as the technique for processing natural language to obtain understanding, similar to how one would try to understand a text in a foreign
30	Tell me about how can you best describe natural language processing (nlp)??	NLP can be described as the technique for processing natural language to obtain understanding, similar to how one would try to understand a text in a foreign
31	What can you say about how can you best describe natural language processing (nlp)??	NLP can be described as the technique for processing natural language to obtain understanding, similar to how one would try to understand a text in a foreign
32	Elaborate on how can you best describe natural language processing (nlp)??	NLP can be described as the technique for processing natural language to obtain understanding, similar to how one would try to understand a text in a foreign
33	Describe how can you best describe natural language processing (nlp)??	NLP can be described as the technique for processing natural language to obtain understanding, similar to how one would try to understand a text in a foreign
34	Provide information on how can you best describe natural language processing (nlp)??	NLP can be described as the technique for processing natural language to obtain understanding, similar to how one would try to understand a text in a foreign
35	What analogy is used to explain the concept of NLP?	The analogy used is trying to understand a text in a foreign language that you cannot comprehend. This is what computers face when dealing with natural lan

In Track 2 Design, Development, and Testing, we've made significant progress. We have successfully designed the Generative Search system with a visual workflow, which is presented below.



Additionally, we've completed the conversion of questions to embedded vectors and stored them in a .csv file.



Embed Questions

```
# Load Bert sentence transformer
model_minilm = SentenceTransformer('all-MiniLM-L6-v2')

# Embed questions
embeddings_minilm = model_minilm.encode(list(df.question))
df['embedded_questions'] = embeddings_minilm.tolist()

df.head(5)
```

```
Downloading (…)e9125/.gitattributes: 100%  1.18k/1.18k [00:00<00:00, 66.3kB/s]
Downloading (…)_Pooling/config.json: 100%  190/190 [00:00<00:00, 13.5kB/s]
Downloading (…)7e55de9125/README.md: 100%  10.6k/10.6k [00:00<00:00, 737kB/s]
Downloading (…)55de9125/config.json: 100%  612/612 [00:00<00:00, 46.7kB/s]
Downloading (…)ce_transformers.json: 100%  116/116 [00:00<00:00, 7.81kB/s]
Downloading (…)125/data_config.json: 100%  39.3k/39.3k [00:00<00:00, 632kB/s]
Downloading pytorch_model.bin: 100%  90.9M/90.9M [00:02<00:00, 33.2MB/s]
Downloading (…)nce_bert_config.json: 100%  53.0/53.0 [00:00<00:00, 3.33kB/s]
Downloading (…)cial_tokens_map.json: 100%  112/112 [00:00<00:00, 6.75kB/s]
Downloading (…)e9125/tokenizer.json: 100%  466k/466k [00:00<00:00, 2.34MB/s]
Downloading (…)okenizer_config.json: 100%  350/350 [00:00<00:00, 16.6kB/s]
Downloading (…)9125/train_script.py: 100%  13.2k/13.2k [00:00<00:00, 1.00MB/s]
Downloading (…)7e55de9125/vocab.txt: 100%  232k/232k [00:00<00:00, 3.58MB/s]
Downloading (…)5de9125/modules.json: 100%  349/349 [00:00<00:00, 21.7kB/s]
```

	question	response	embedded_questions
0	Why is it important to process text data?	Text data are in natural languages. So compute...	[0.005245531909167767, 0.07936380058526993, -0...
1	Can you explain why is it important to process...	Text data are in natural languages. So compute...	[-0.002595293801277876, 0.06376708298921585, -...
2	What do you mean by why is it important to pro...	Text data are in natural languages. So compute...	[-0.015980252996087074, 0.057495031505823135, ...
3	How would you describe why is it important to ...	Text data are in natural languages. So compute...	[0.00037727333256043494, 0.07592840492725372, ...
4	Could you clarify why is it important to proce...	Text data are in natural languages. So compute...	[-0.003969612997025251, 0.07431232184171677, -...

Export to .csv

```
df.to_csv('/content/drive/MyDrive/Colab Notebooks/TIS_Embedded_Q&A.csv', index=False)
```

Currently, our team is actively working on developing the “Online Generative Search Module”. After this is done, we will focus on testing the functionality of this module. Then, we will work on Track 3, which is ‘Results Evaluation’.

In terms of Track 4, we have prepared the First Progress Report as part of this submission. We are working on building the final report and preparing the presentation as well.

Our project is progressing exceptionally well, and we are confidently on course to meet exceed our objectives well before the deadline. With remarkable efficiency and a well-coordinated team, we have encountered no roadblocks or significant challenges along the way. Our progress is not just steady but remarkable, affirming our commitment to delivering exceptional results.

Conclusion

In conclusion, our team has made substantial progress in the development of our Generative Search system. We have successfully completed the critical tasks outlined in Track 1, ensuring the availability of essential data and a versatile Q&A dataset. Additionally, our progress in Track 2 is commendable, with the design phase accomplished, and developments made in converting questions to embedded vectors. As we move forward, our focus remains on delivering a Generative Search system that aligns with our project objectives, providing reliable and efficient document retrieval capabilities.