

Generating Answers - Semantic Search

Special Topics: Generative AI-Driven Intelligent Apps Development

Ademilton Marcelo da Cruz Nunes (19679)

Links

Github:

<https://github.com/ademiltonnunes/Machine-Learning/tree/main/GenerativeAI/Fine-Tuning/Generating%20Answers>

Table of Content

- Introduction
- Introduction - Semantic Search
- Implementation
- Setting Cohere
- Input text
- Embeddings
- Generating Vector store database
- Semantic searching for articles in the vector store
- Getting Answers
- Getting Answers - Examples
- Conclusion

Introduction

This project exemplifies the use semantic search to generate answers. To do this project, I gave an example of the process of preparing data, embedding indexes to be used in the semantic search and how to generate answers.

Introduction - Semantic Search

Semantic search makes comparisons with embedded chunks of documents that are semantically close.

Embeddings is numerical representations of words or longer text that coordinates to words in a multidimensional space, vectorstore. There, similar sentences or phrases have embeddings indexes that are close semantically. The numerical coordinates capture relationships between words, providing a nuanced understanding of language.

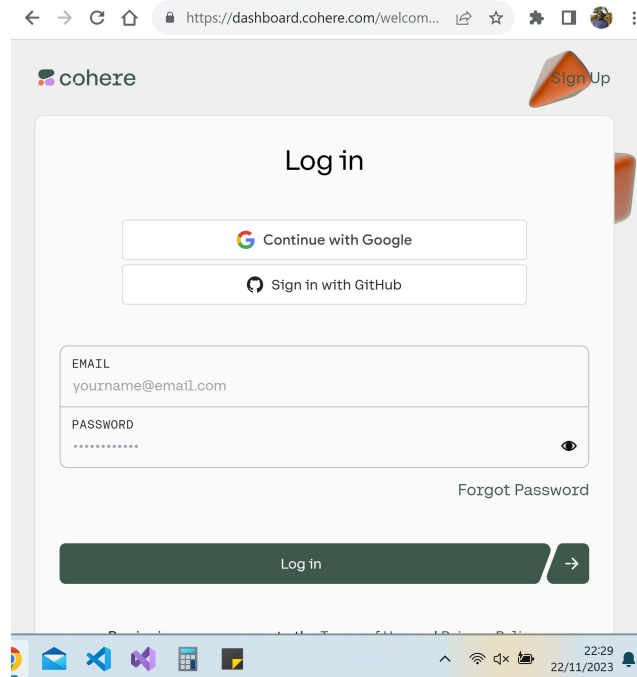
Implementation

This project implemented semantic search to generate answers. To be able to do this we used tools:

- **Cohere:** Cohere is a powerful library that provides features as 'embed' function. This function is designed to generate embeddings for words or phrases.
- **AnnoyIndex:** Annoy is a vector search library used for semantic search. Embed the query and find the nearest neighbors.

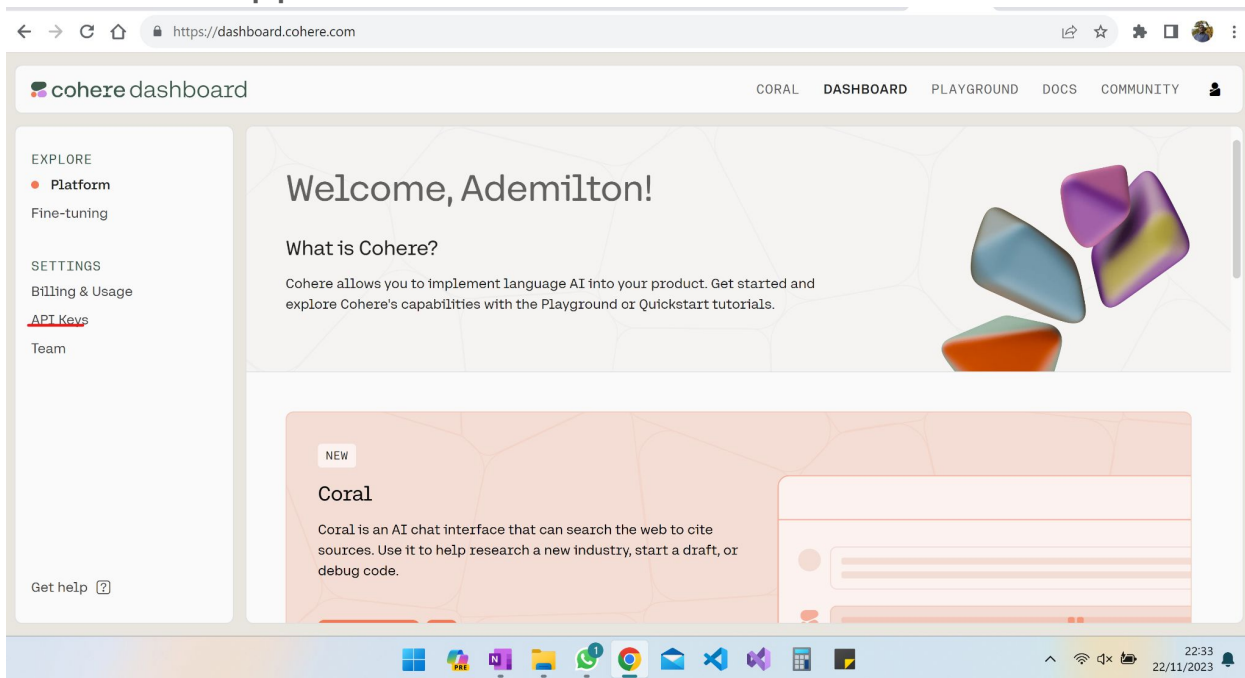
Setting Cohere

Cohere is accessed by the website: <https://docs.cohere.com/docs>. We can sign up using Google or Github accounts.

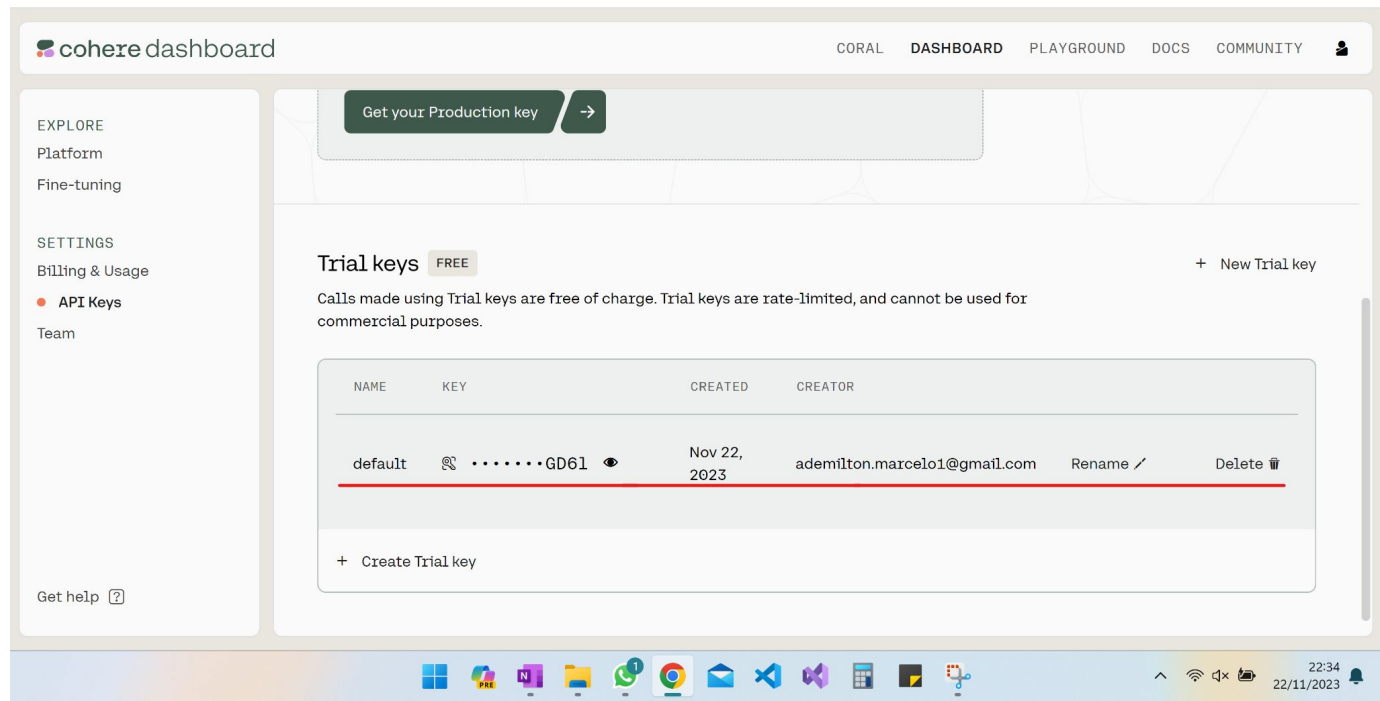


Setting Cohere

After creating our account, we can have access to our API key, this API we are going to use in our application.



Setting Cohere



The screenshot shows the Cohere dashboard interface. At the top, there's a navigation bar with the Cohere logo and links to CORAL, DASHBOARD, PLAYGROUND, DOCS, and COMMUNITY. A user profile icon is on the right. On the left, a sidebar contains 'EXPLORE' (Platform, Fine-tuning) and 'SETTINGS' (Billing & Usage, API Keys, Team). The main content area has a 'Get your Production key' button. Below this, the 'Trial keys' section is highlighted with a 'FREE' tag. It includes a '+ New Trial key' link and a table of existing keys. The table has columns for NAME, KEY, CREATED, and CREATOR. One key is listed: 'default' with a masked key '.....GD61', created on 'Nov 22, 2023' by 'ademilton.marcelo1@gmail.com'. Action links 'Rename' and 'Delete' are next to the key. A '+ Create Trial key' button is at the bottom of the table.

cohere dashboard

CORAL DASHBOARD PLAYGROUND DOCS COMMUNITY

EXPLORE
Platform
Fine-tuning

SETTINGS
Billing & Usage
● API Keys
Team

Get help ?

Get your Production key →

Trial keys **FREE** + New Trial key

Calls made using Trial keys are free of charge. Trial keys are rate-limited, and cannot be used for commercial purposes.

NAME	KEY	CREATED	CREATOR
defaultGD61	Nov 22, 2023	ademilton.marcelo1@gmail.com

+ Create Trial key

22:34 22/11/2023

Input text

To do an example of semantic search, data is needed to include in the vector store. In this example, we extract data from these sources, that talk about “How to Build a Career in AI”:

- <https://www.deeplearning.ai/the-batch/how-to-build-a-career-in-ai-part-1-three-steps-to-career-growth/>
- <https://www.deeplearning.ai/the-batch/how-to-build-a-career-in-ai-part-2-learning-technical-skills/>
- <https://www.deeplearning.ai/the-batch/how-to-build-a-career-in-ai-part-3-choosing-projects/>

Input text

With the site's content extracted, I needed to "clean up" the text, removing empty spaces and new lines.

```
# Split into a list of paragraphs
texts = text.split('\n\n')

# Clean up to remove empty spaces and new lines
texts = np.array([t.strip(' \n') for t in texts if t])

texts[:3]
```

array(["The rapid rise of AI has led to a rapid rise in AI jobs, and many people are building exciting careers in this field. A career is a decades-long journey, and the path is not always straightforward. Over many years, I've been privileged to see thousands of students as well as engineers in companies large and small navigate careers in AI. In this and the next few letters, I'd like to share a few thoughts that might be useful in charting your own course.\n\nThree key steps of career growth are learning (to gain technical and other skills), working on projects (to deepen skills, build a portfolio, and create impact) and searching for a job. These steps stack on top of each other:\n\nInitially, you focus on gaining foundational technical skills.\n\nAfter having gained foundational skills, you lean into project work. During this period, you'll probably keep learning.\n\nLater, you might occasionally carry out a job search. Throughout this process, you'll probably continue to learn and work on meaningful projects.\n\nThese phases apply in a wide range of professions, but AI involves unique elements. For example:\n\nAI is nascent, and many technologies are still evolving. While the foundations of machine learning and deep learning are maturing – and coursework is an efficient way to master them – beyond these foundations, keeping up-to-date with changing technology is more important in AI than fields that are more mature.\n\nProject work often means working with stakeholders who lack expertise in AI. This can make it challenging to find a suitable project, estimate the project's timeline and return on investment, and set expectations. In addition, the highly iterative nature of AI projects leads to special challenges in project management: How can you come up with a plan for building a system when you don't know in advance how long it will take to achieve the target accuracy? Even after the system has hit the target, further iteration may be necessary to address post-deployment drift.\n\nWhile searching for a job in AI can be similar to searching for a job in other sectors, there are some differences. Many companies are still trying to figure out which AI skills they need and how to hire people who have them. Things you've worked on may be significantly different than anything your interviewer has seen, and you're more likely to have to educate potential employers about some elements of your work.\n\nThroughout these steps, a supportive community is a big help. Having a group of friends and allies who can help you – and whom you strive to help – makes the path easier. This is true whether you're taking your first steps or you've been on the journey for years.\n\nI'm excited to work with all of you to grow the global AI community, and that includes helping everyone in our community develop their careers. I'll dive more deeply into these topics in the next few weeks.\n\nLast week, I wrote about key steps for building a career in AI: learning technical skills, doing project work, and searching for a job, all of which is supported by being part of a community. In this letter, I'd like to dive more deeply into the first step.\n\nMore papers have been published on AI than any person can read in a lifetime. So, in your efforts to learn, it's critical to prioritize topic selection. I believe the most important topics for a technical career in machine learning are:\n\nFoundational machine learning skills. For example, it's important to understand models such as linear regression, logistic

✓ 12s completed at 9:38PM

Embeddings

Using Cohere, I transformed the inputted text into embedding indexes. Below are the dimensions of the vector store

```
response = co.embed(texts=texts.tolist()).embeddings

#Check the dimensions of the embeddings
embeds = np.array(response)

print("Dimensions of embeddings:", embeds.shape)
print("X-axis of embeddings:", embeds.shape[0])
print("Y-axis of embeddings:", embeds.shape[1])
```

```
} default model on embed will be deprecated in the future,
Dimensions of embeddings: (1, 4096)
X-axis of embeddings: 1
Y-axis of embeddings: 4096
```

✓ 0s completed at 10:13 PM

22:13
23/11/2023

Generating Vector store database

Using AnnoyIndex, I built a vector store database and saved it locally as "test.ann".

```
✓ 0s [24] #Create the search index, pass the size of embedding (vector)
```

```
search_index = AnnoyIndex(embeds.shape[1], 'angular')
```

```
✓ 0s [25] # Add all the vectors to the search index
```

```
for i in range(len(embeds)):  
    search_index.add_item(i, embeds[i])
```

```
# 10 trees
```

```
search_index.build(10)  
search_index.save('test.ann')
```

✓ 0s completed at 10:15PM



22:15
23/11/2023



Semantic searching for articles in the vector store

By accessing the vector store database and including queries, AnnoyIndex will search and retrieve all articles that are semantically close to this query.

```
[15] def search_andrews_article(query):  
      # Get the query's embedding  
      query_embed = co.embed(texts=[query]).embeddings  
  
      # Retrieve the nearest neighbors  
      similar_item_ids = search_index.get_nns_by_vector(  
          query_embed[0],  
          10,  
          include_distances=True)  
  
      search_results = texts[similar_item_ids[0]]  
  
      return search_results
```

✓ 0s completed at 10:15 PM



22:23
23/11/2023



Semantic searching for articles in the vector store

Example

```
✓ [27] results = search_andrews_article("Are side projects a good idea when trying to build a career in AI?")  
0s  
print(results[0])
```

default model on embed will be deprecated in the future, please specify a model in the request.

The rapid rise of AI has led to a rapid rise in AI jobs, and many people are building exciting careers in this field. A career is a decades-long journey, and the path is

Three key steps of career growth are learning (to gain technical and other skills), working on projects (to deepen skills, build a portfolio, and create impact) and sear

Initially, you focus on gaining foundational technical skills.

After having gained foundational skills, you lean into project work. During this period, you'll probably keep learning.

Later, you might occasionally carry out a job search. Throughout this process, you'll probably continue to learn and work on meaningful projects.

These phases apply in a wide range of professions, but AI involves unique elements. For example:

AI is nascent, and many technologies are still evolving. While the foundations of machine learning and deep learning are maturing – and coursework is an efficient way to Project work often means working with stakeholders who lack expertise in AI. This can make it challenging to find a suitable project, estimate the project's timeline and While searching for a job in AI can be similar to searching for a job in other sectors, there are some differences. Many companies are still trying to figure out which A Throughout these steps, a supportive community is a big help. Having a group of friends and allies who can help you – and whom you strive to help – makes the path easier

I'm excited to work with all of you to grow the global AI community, and that includes helping everyone in our community develop their careers. I'll dive more deeply int

Last week, I wrote about key steps for building a career in AI: learning technical skills, doing project work, and searching for a job, all of which is supported by bein

More papers have been published on AI than any person can read in a lifetime. So, in your efforts to learn, it's critical to prioritize topic selection. I believe the mo

Foundational machine learning skills. For example, it's important to understand models such as linear regression, logistic regression, neural networks, decision trees, c Deep learning. This has become such a large fraction of machine learning that it's hard to excel in the field without some understanding of it! It's valuable to know the Math relevant to machine learning. Key areas include linear algebra (vectors, matrices, and various manipulations of them) as well as probability and statistics (includi Software development. While you can get a job and make huge contributions with only machine learning modeling skills, your job opportunities will increase if you can als This is a lot to learn! Even after you master everything in this list, I hope you'll keep learning and continue to deepen your technical knowledge. I've known many machi

✓ 0s completed at 10:26 PM



22:26

23/11/2023



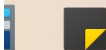
Getting Answers

Obtaining all articles retrieved from the vector store database, we will use them as context to generate the response to the input query. Using Cohere, we will generate the response to the input query. We can configure how many responses we want to receive.

Getting Answers

```
[17] def ask_andrews_article(question, num_generations=1):  
  
    # Search the text archive  
    results = search_andrews_article(question)  
  
    # Get the top result  
    context = results[0]  
  
    # Prepare the prompt  
    prompt = f"""  
    Excerpt from the article titled "How to  
    Build a Career in AI"  
    by Andrew Ng:  
    {context}  
    Question: {question}  
  
    Extract the answer of the question from  
    the text provided.  
    If the text doesn't contain the answer,  
    reply that the answer is not available."""  
  
    prediction = co.generate(  
        prompt=prompt,  
        max_tokens=70,  
        model="command-nightly",  
        temperature=0.5,  
        num_generations=num_generations  
    )  
  
    return prediction.generations
```

✓ 0s completed at 10:26 PM



22:28

23/11/2023




Getting Answers - Examples

Case 1:

I generated a query that asks "Are side projects a good idea when trying to build a career in AI?". I configured to get an single response.

```
#####  
# Step 4.1 Generating Answers - Test Case 1  
#####  
results = ask_andrews_article("Are side projects a good idea when trying to build a career in AI?")
```

✓ 0s completed at 10:43 PM



Getting Answers - Examples

Case 1:

Response:

“ Yes, side projects are considered a great idea when trying to build a career in AI. Side projects provide opportunities to work on small projects in spare time. Their success leads to incremental benefits such as stretching your skills, building confidence in tackling bigger projects, and providing opportunities to connect and collaborate with like-minded individuals. Side projects are a common trend in”

Getting Answers - Examples

Case 2:

I generated a query that asks "Are side projects a good idea when trying to build a career in AI?". I configured to get three response.

```
[19] #####  
# Step 4.2 Generating Answers - Test Case 2  
#####  
results = ask_andrews_article("Are side projects a good idea when trying to build a career in  
    num_generations=3  
)  
  
for gen in results:  
    print(gen)  
    print('--')
```

✓ 0s completed at 10:43 PM



22:47
23/11/2023



Getting Answers - Examples

Case 2:

Responses:

1. Yes, side projects are considered a great idea when trying to build a career in AI. They can help strengthen your skills and be a pathway to mastering greater technical complexity. Side projects are a great way to stir creative juices and oftentimes can turn into something significant, and supplementary to your career journey and building your skill sets.
2. Yes, side projects are considered a great idea when trying to build a career in AI. Side projects provide opportunities to develop skills and interests outside of one's regular job or studies. They can also serve as a creative outlet and a chance to explore new ideas and domains. In the field of AI, side projects can be particularly valuable since the field is
3. Yes, side projects are considered a great idea when trying to build a career in AI. They can help cultivate passion projects outside of work, stretch creative muscles, stimulate creativity, and provide a constructive outlet for brainstorming. Side projects may also offer opportunities to explore niche areas of AI that individuals would not otherwise have exposure to within a structured work setting.

Getting Answers - Examples

Case 3:

I generated a query that asks "What is the most viewed televised event?". I configured to get five response. This question is not related to the text input and should not answer it.

```
#####  
# Step 4.3 Generating Answers - Test Case 3  
#####  
results = ask_andrews_article(  
    "What is the most viewed televised event?",  
    num_generations=5  
)  
  
for gen in results:  
    print(gen)  
    print('--')
```

✓ 0s completed at 10:43 PM



22:52

23/11/2023



Getting Answers - Examples

Case 3:

Responses:

1. Sorry, the answer to this question is not available in the text provided.
2. Sorry, the answer to this question is not available in the text provided.
3. Unfortunately, the provided text does not contain information regarding the most viewed televised event.
4. Sorry, the answer to this question is not available in the text provided.
5. Sorry, the answer to this question is not available in the text provided.

Conclusion

In this project, powered by Cohere's text embedding and AnnoyIndex's vector storage, represents a advancement in intelligent information retrieval. The meticulous embedding of context into vectors, facilitated by Cohere, imbued our system with a sophisticated understanding of the semantic nuances within textual data. This, combined with AnnoyIndex's efficient vector storage, enabled swift and accurate searches, exemplifying the potential of semantic search technologies.

During testing, the system consistently delivered pertinent answers when queried about learned contexts. Importantly, its honesty in admitting a lack of knowledge when faced with unrelated queries underscored its discerning reliance on semantic understanding. As a cohesive whole, this project underscores the practicality and potential of semantic search systems, showcasing their prowess in distilling meaningful insights from vast textual landscapes. The fusion of advanced technologies in this endeavor lays the groundwork for future innovations in the realm of intelligent information retrieval.