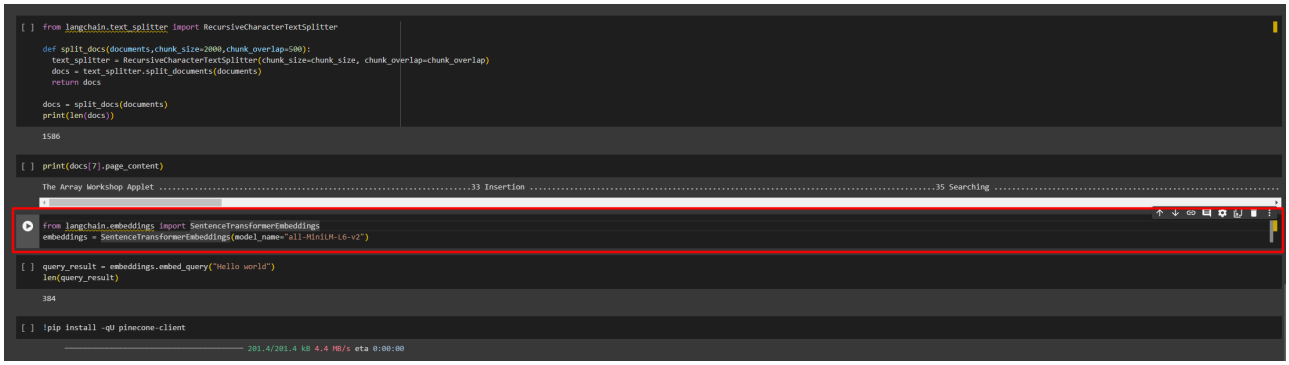
Variable

- hugging face api key = "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"

For code run in google colab

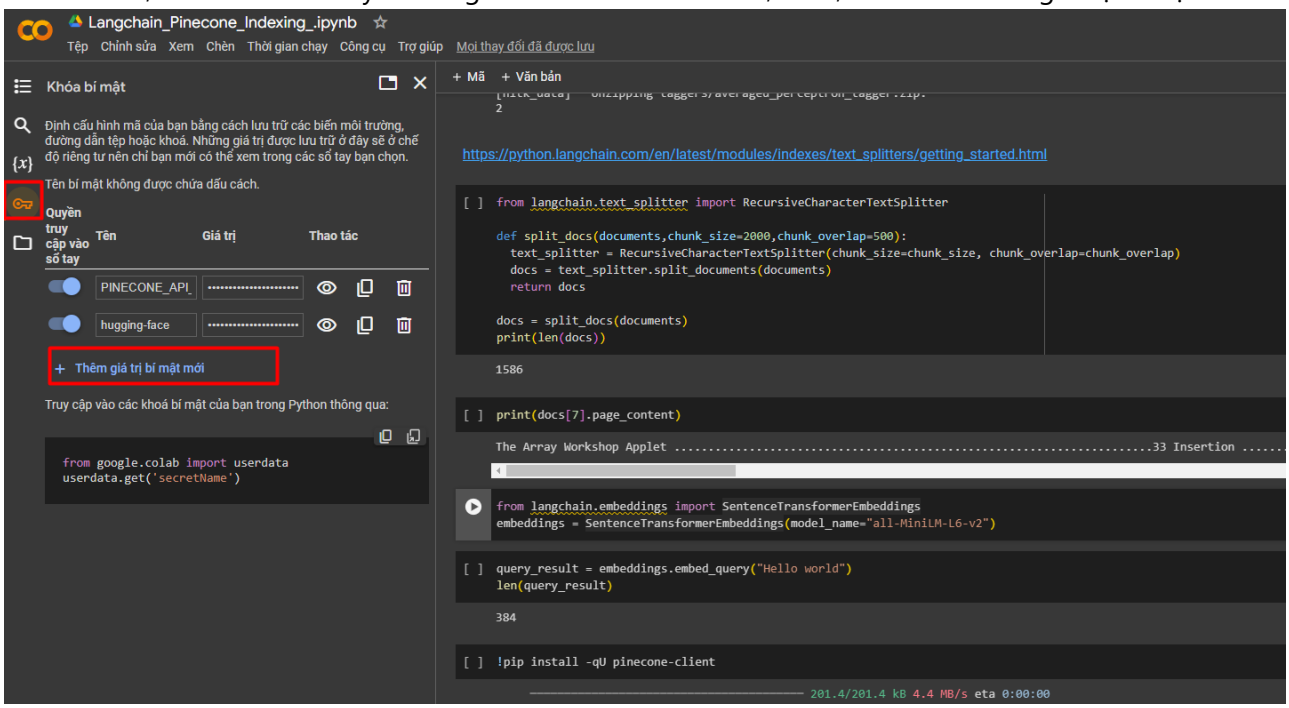
- This code is to convert data from documents into vectors and stored them inside an online vector-database called "Pinecone"

- Noted: The following code block can only run correctly after secret_key is setup inside Google Colab

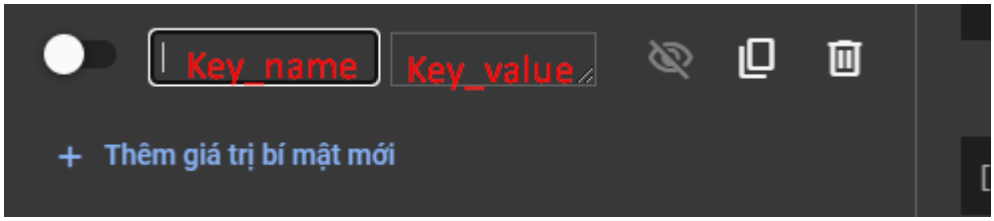


How to add secret key

1. On left side, click on icon "Key" belongs to left vertical sidebar, then, click on "Thêm giá trị bí mật mới"



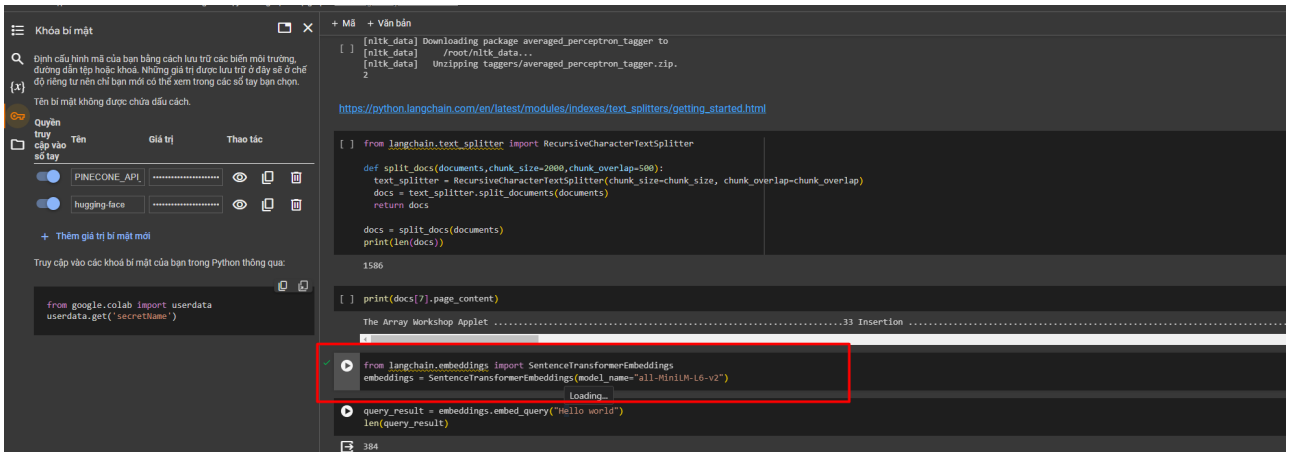
2. Then fill in key_name, key_value in the input field



3. Remember to toggle the button, that way your notebook can access the data inside Google Colab's secret key 

How to run the aforementioned code block

1. Open Secret key like aboved tutorial
2. Run the code block
3. Check if result is same as this image



Change Pinecode index

- You can change my Pinecone index to your own Pinecone index

- Change the following api key inside the code block to your own Pinecone database api key.

```
from pinecone import Pinecone, PodSpec
from langchain.vectorstores import Pinecone as Pn

pc = Pinecone(api_key="445535a3-d2a3-4ae1-9037-7322276f9a81")
index_name = "langchain-chatbot-pdf-demo"
spec = PodSpec(environment="gcp-starter")
# clean all indexes in project
for data in pc.list_indexes().names():
  pc.delete_index(data)

pc.create_index(
        index_name,
        dimension=len(query_result),  # dimensionality of text-embedding-ada-002
        metric='cosine',
        spec=spec
    )

# connect to new create index
index = pc.Index(index_name)
index.describe_index_stats()

text_field = "text"
vectorstore = Pn(
    index, embeddings.embed_query, text_field
)
# Batch insert the chunks into the vector store
batch_size = 100  # Define your preferred batch size
for i in range(0, len(docs), batch_size):
    chunk_batch = docs[i:i + batch_size]
    vectorstore.add_documents(chunk_batch)

print("Ok")
```

- If you want to change the name of the index, you can change this

```
[ ] from pinecone import Pinecone, PodSpec
    from langchain.vectorstores import Pinecone as Pn

    pc = Pinecone(api_key="445535a3-d2a3-4ae1-9037-7322276f9a81")
    index_name = "langchain-chatbot-pdf-demo"
    spec = PodSpec(environment="gcp-starter")
    # clean all indexes in project
    for data in pc.list_indexes().names():
      pc.delete_index(data)

    pc.create_index(
            index_name,
            dimension=len(query_result),  # dimensionality of text-embeddi
            metric='cosine',
            spec=spec
        )

    # connect to new create index
    index = pc.Index(index_name)
    index.describe_index_stats()

    text_field = "text"
    vectorstore = Pn(
        index, embeddings.embed_query, text_field
    )
    # Batch insert the chunks into the vector store
    batch_size = 100  # Define your preferred batch size
    for i in range(0, len(docs), batch_size):
        chunk_batch = docs[i:i + batch_size]
        vectorstore.add_documents(chunk_batch)

    print("Ok")

    /usr/local/lib/python3.10/dist-packages/langchain_community/vectorstor
```

For ui code:

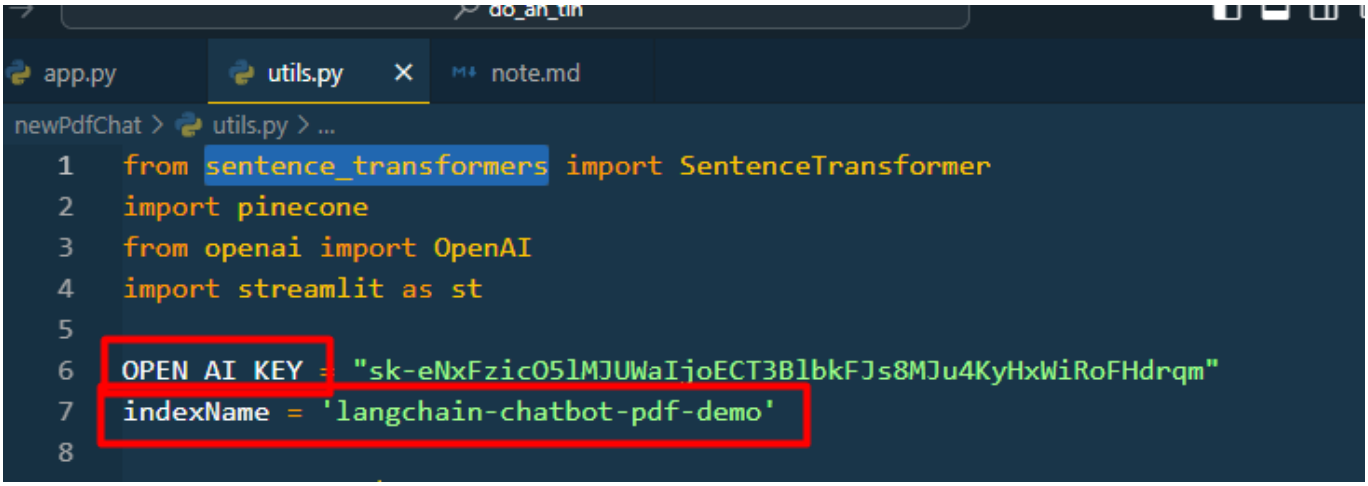
- 2 files: app.py and utils.py, with app.py is the main file.

Library to install

- used the following command

```
pip install langchain langchain-openai openai tiktoken pinecone-client streamlit
sentence-transformers
```

Variables to be notified:



- index_name is the name of the database in Pinecone, this one must be correct or else the code can't run
- Same with openapi key.