Langchain Chat with your Data:

Vector Stores and Embedding

By:
Melanie Meby Olisah

Process

Step 1: Study the key ideas of Vectorstores and Embedding

1. Load documents

2. Split the documents into small, semantically meaningful chunks

3. Create an index for each chunk by embeddings

> **The index is created by embeddings which are numerical representations of text.**
>
> **Text with semantically similar content has similar vectors in this numeric space.**

4. Store these index in a vector stores for easy retrieval when answering questions

5. Search answer of a question.

Both should have similar index

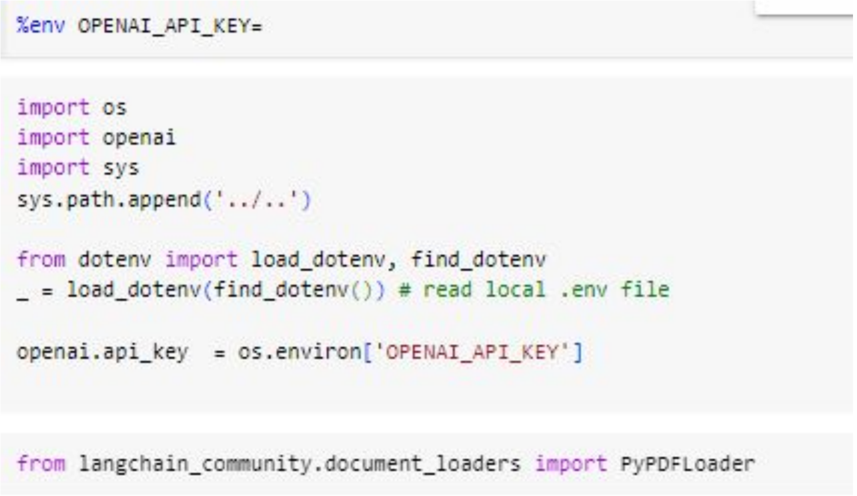
6. Edge Cases - Failure

2 types of failures in similarity search

> **Diversity**
>
> **Specifity**

Solved by Advanced Retrieval

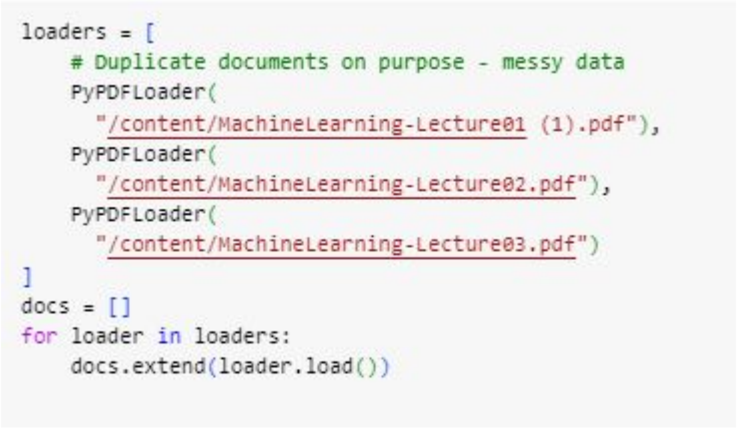
LOAD DOCUMENTS



```
%env OPENAI_API_KEY=
```

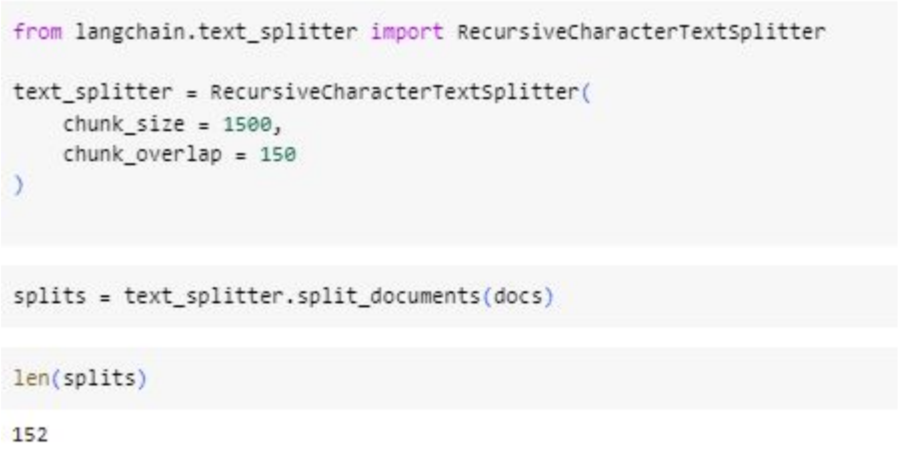
```
import os
import openai
import sys
sys.path.append('../..')

from dotenv import load_dotenv, find_dotenv
_ = load_dotenv(find_dotenv()) # read local .env file

openai.api_key  = os.environ['OPENAI_API_KEY']
```

```
from langchain_community.document_loaders import PyPDFLoader
```



```
loaders = [
    # Duplicate documents on purpose - messy data
    PyPDFLoader(
      "/content/MachineLearning-Lecture01 (1).pdf"),
    PyPDFLoader(
      "/content/MachineLearning-Lecture02.pdf"),
    PyPDFLoader(
      "/content/MachineLearning-Lecture03.pdf")
]
docs = []
for loader in loaders:
    docs.extend(loader.load())
```

SPLITTING



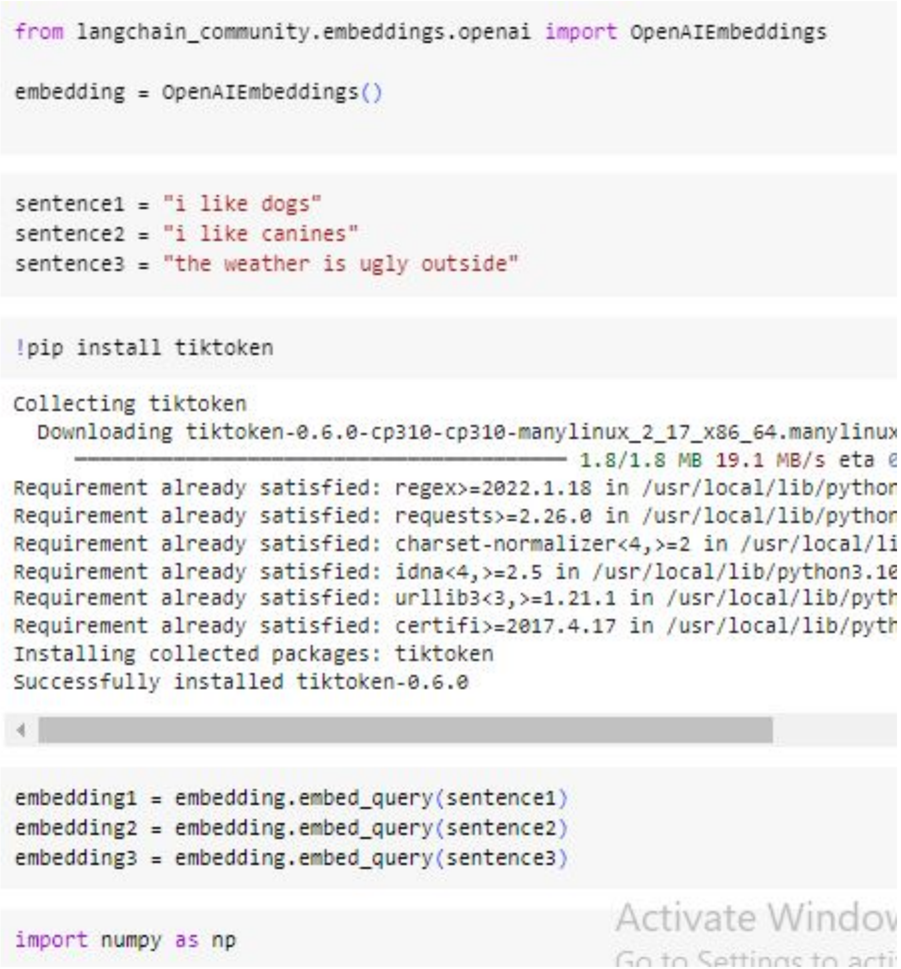
```
from langchain.text_splitter import RecursiveCharacterTextSplitter

text_splitter = RecursiveCharacterTextSplitter(
    chunk_size = 1500,
    chunk_overlap = 150
)
```

```
splits = text_splitter.split_documents(docs)
```

```
len(splits)
```

```
152
```

CREATE AN INDEX



```
from langchain_community.embeddings.openai import OpenAIEmbeddings

embedding = OpenAIEmbeddings()
```

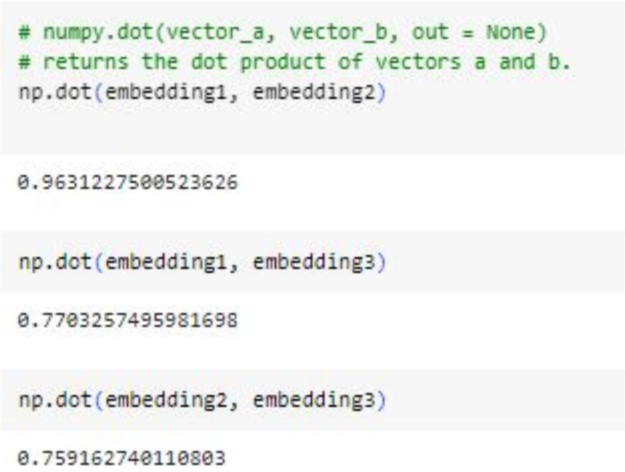
```
sentence1 = "i like dogs"
sentence2 = "i like canines"
sentence3 = "the weather is ugly outside"
```

```
!pip install tiktoken
```

```
Collecting tiktoken
  Downloading tiktoken-0.6.0-cp310-cp310-manylinux_2_17_x86_64.manylinux
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 19.1 MB/s eta 0
Requirement already satisfied: regex>=2022.1.18 in /usr/local/lib/python
Requirement already satisfied: requests>=2.26.0 in /usr/local/lib/python
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/li
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/pyth
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/pyth
Installing collected packages: tiktoken
Successfully installed tiktoken-0.6.0
```

```
embedding1 = embedding.embed_query(sentence1)
embedding2 = embedding.embed_query(sentence2)
embedding3 = embedding.embed_query(sentence3)
```

```
import numpy as np
```



```
# numpy.dot(vector_a, vector_b, out = None)
# returns the dot product of vectors a and b.
np.dot(embedding1, embedding2)
```

```
0.9631227500523626
```

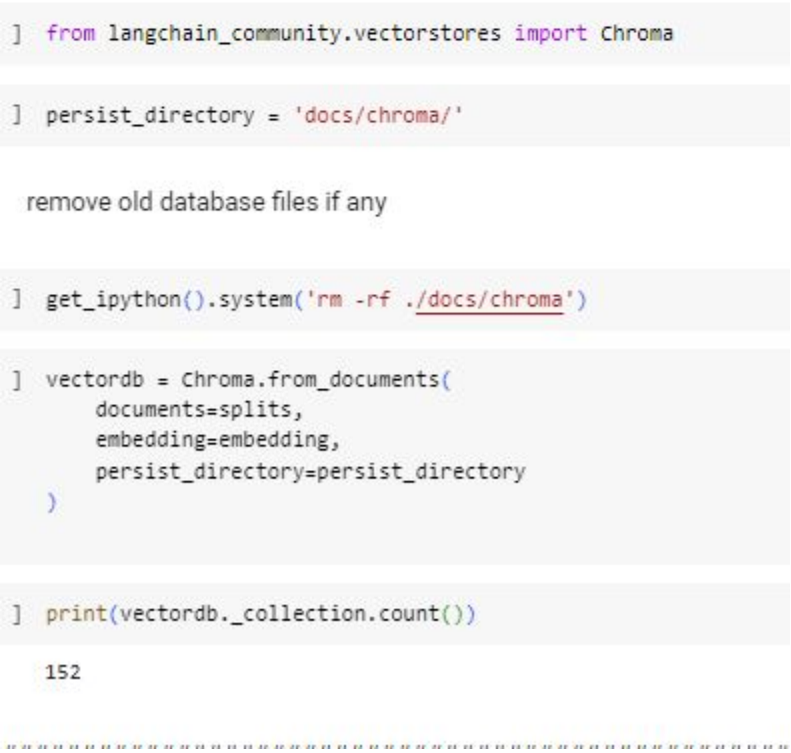
```
np.dot(embedding1, embedding3)
```

```
0.7703257495981698
```

```
np.dot(embedding2, embedding3)
```

```
0.759162740110803
```

STORE



```python
from langchain_community.vectorstores import Chroma
```

```python
persist_directory = 'docs/chroma/'
```

remove old database files if any

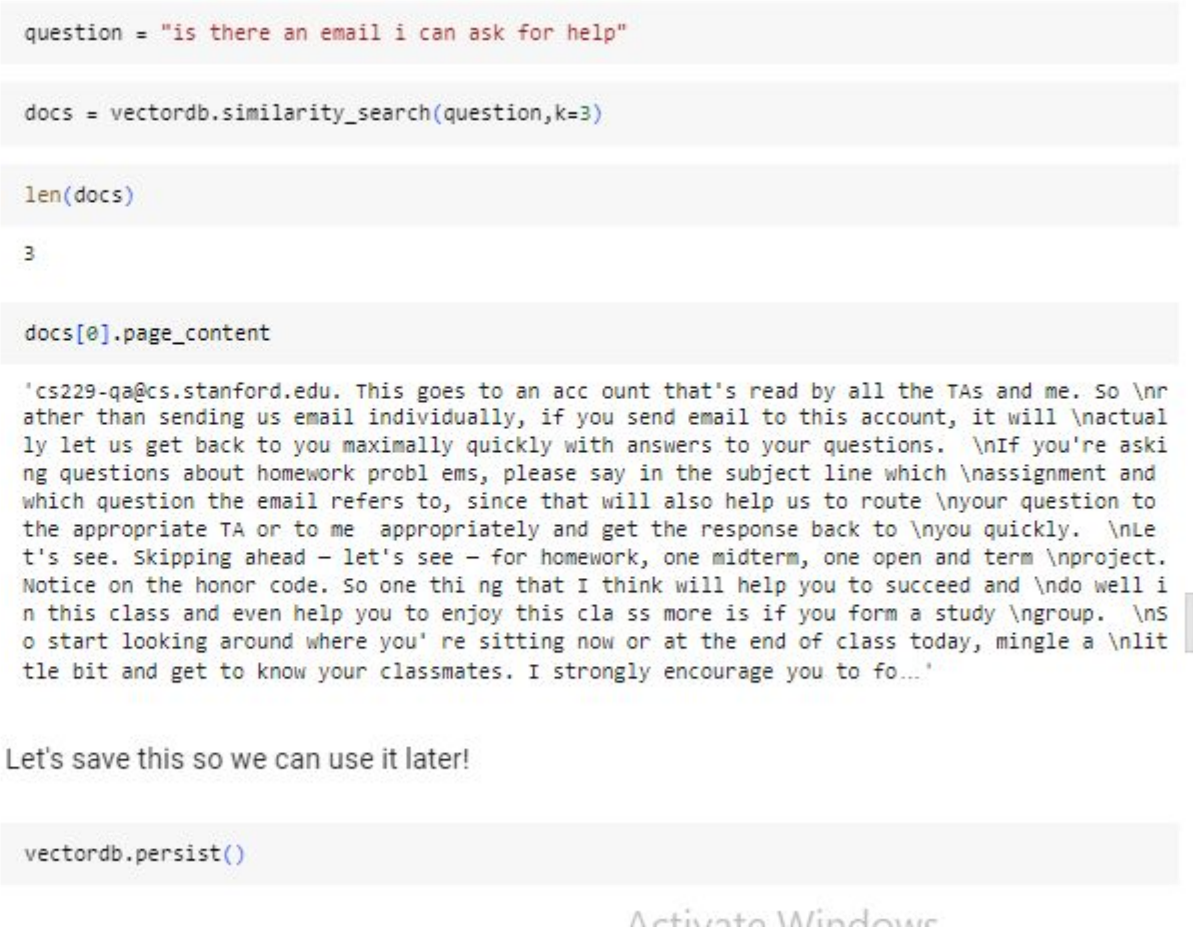
```python
get_ipython().system('rm -rf ./docs/chroma')
```

```python
vectordb = Chroma.from_documents(
    documents=splits,
    embedding=embedding,
    persist_directory=persist_directory
)
```

```python
print(vectordb._collection.count())
```

```
152
```

SIMILARITY SEARCH



```
question = "is there an email i can ask for help"
```

```
docs = vectordb.similarity_search(question,k=3)
```

```
len(docs)
```

```
3
```

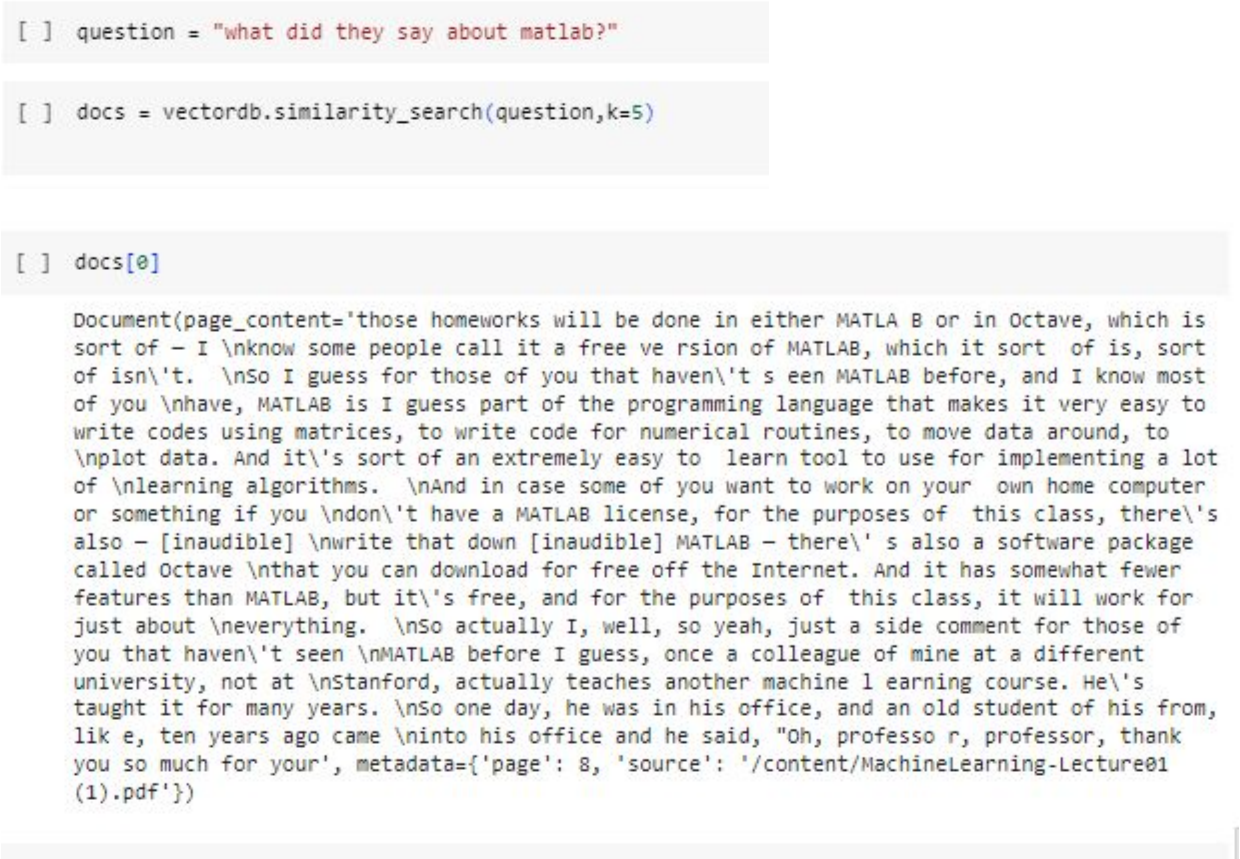
```
docs[0].page_content
```

```
'cs229-qa@cs.stanford.edu. This goes to an acc ount that's read by all the TAs and me. So \nr
ather than sending us email individually, if you send email to this account, it will \nactual
ly let us get back to you maximally quickly with answers to your questions.  \nIf you're aski
ng questions about homework probl ems, please say in the subject line which \nassignment and
which question the email refers to, since that will also help us to route \nyour question to
the appropriate TA or to me  appropriately and get the response back to \nyou quickly.  \nLe
t's see. Skipping ahead - let's see - for homework, one midterm, one open and term \nproject.
Notice on the honor code. So one thi ng that I think will help you to succeed and \ndo well i
n this class and even help you to enjoy this cla ss more is if you form a study \ngroup.  \nS
o start looking around where you' re sitting now or at the end of class today, mingle a \nlit
tle bit and get to know your classmates. I strongly encourage you to fo...'
```

Let's save this so we can use it later!

```
vectordb.persist()
```

EDGE CASE



```
[ ]  question = "what did they say about matlab?"
```

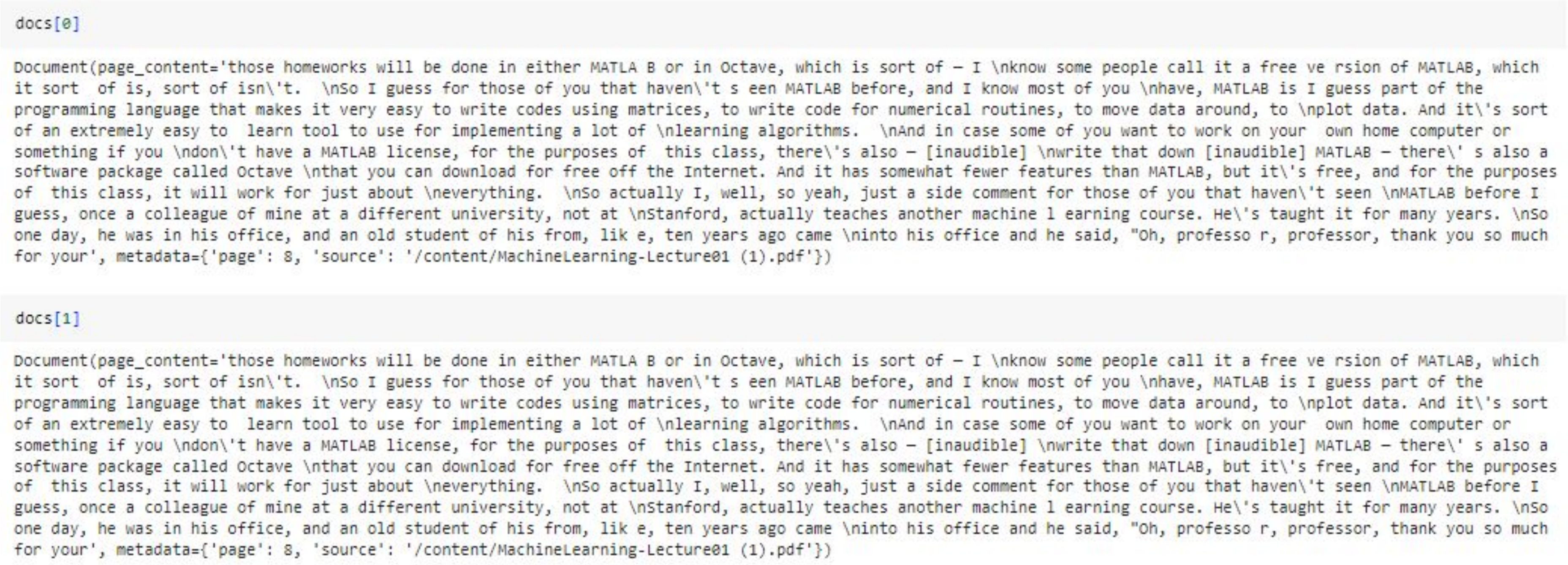
```
[ ]  docs = vectordb.similarity_search(question,k=5)
```

```
[ ]  docs[0]
```

```
Document(page_content='those homeworks will be done in either MATLA B or in Octave, which is
sort of — I \nknow some people call it a free ve rsion of MATLAB, which it sort  of is, sort
of isn\'t.  \nSo I guess for those of you that haven\'t s een MATLAB before, and I know most
of you \nhave, MATLAB is I guess part of the programming language that makes it very easy to
write codes using matrices, to write code for numerical routines, to move data around, to
\nplot data. And it\'s sort of an extremely easy to  learn tool to use for implementing a lot
of \nlearning algorithms.  \nAnd in case some of you want to work on your  own home computer
or something if you \ndon\'t have a MATLAB license, for the purposes of  this class, there\'s
also — [inaudible] \nwrite that down [inaudible] MATLAB — there\' s also a software package
called Octave \nthat you can download for free off the Internet. And it has somewhat fewer
features than MATLAB, but it\'s free, and for the purposes of  this class, it will work for
just about \neverything.  \nSo actually I, well, so yeah, just a side comment for those of
you that haven\'t seen \nMATLAB before I guess, once a colleague of mine at a different
university, not at \nStanford, actually teaches another machine l earning course. He\'s
taught it for many years. \nSo one day, he was in his office, and an old student of his from,
lik e, ten years ago came \ninto his office and he said, "Oh, professo r, professor, thank
you so much for your', metadata={'page': 8, 'source': '/content/MachineLearning-Lecture01
(1).pdf'})
```

Edge Case 1 - Failure modes: Diversity

Notice that we're getting duplicate chunks (because of the duplicate`MachineLearning-Lecture01.pdf` in the index). Semantic search fetches all similar documents, but does not enforce diversity. `docs[0]` and `docs[1]` are indentical.



```
docs[0]
```

```
Document(page_content='those homeworks will be done in either MATLA B or in Octave, which is sort of — I \nknow some people call it a free ve rsion of MATLAB, which
it sort  of is, sort of isn\'t.  \nSo I guess for those of you that haven\'t s een MATLAB before, and I know most of you \nhave, MATLAB is I guess part of the
programming language that makes it very easy to write codes using matrices, to write code for numerical routines, to move data around, to \nplot data. And it\'s sort
of an extremely easy to  learn tool to use for implementing a lot of \nlearning algorithms.  \nAnd in case some of you want to work on your  own home computer or
something if you \ndon\'t have a MATLAB license, for the purposes of  this class, there\'s also — [inaudible] \nwrite that down [inaudible] MATLAB — there\' s also a
software package called Octave \nthat you can download for free off the Internet. And it has somewhat fewer features than MATLAB, but it\'s free, and for the purposes
of  this class, it will work for just about \neverything.  \nSo actually I, well, so yeah, just a side comment for those of you that haven\'t seen \nMATLAB before I
guess, once a colleague of mine at a different university, not at \nStanford, actually teaches another machine l earning course. He\'s taught it for many years. \nSo
one day, he was in his office, and an old student of his from, lik e, ten years ago came \ninto his office and he said, "Oh, professo r, professor, thank you so much
for your', metadata={'page': 8, 'source': '/content/MachineLearning-Lecture01 (1).pdf'})
```

```
docs[1]
```

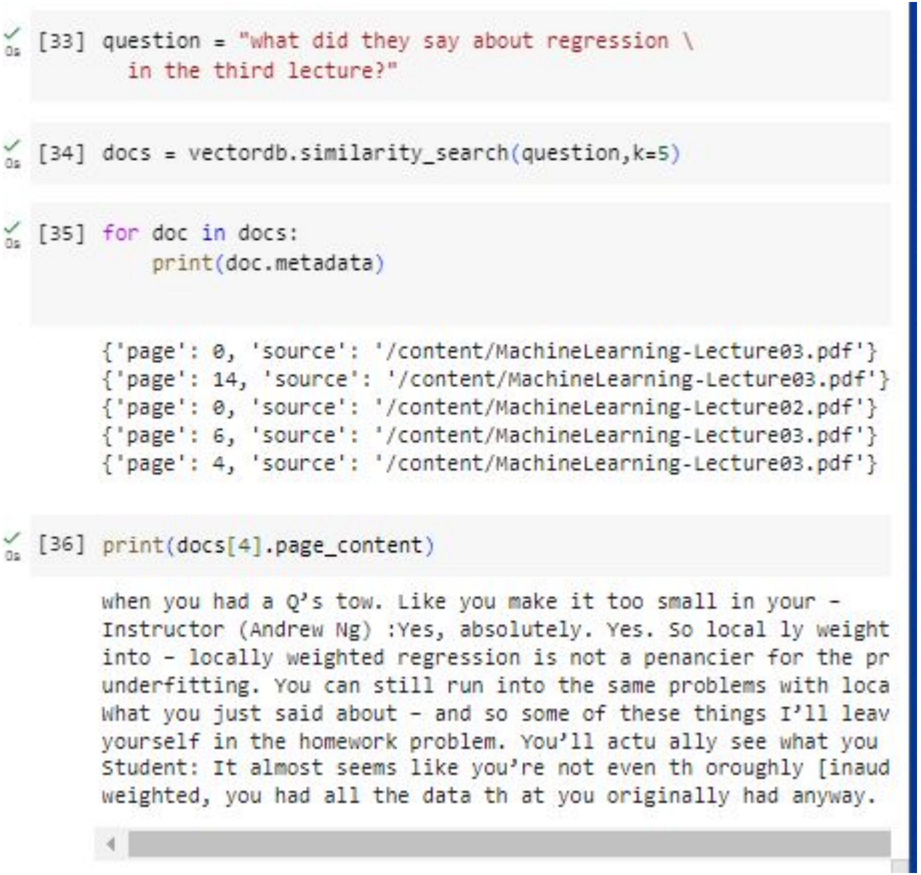
```
Document(page_content='those homeworks will be done in either MATLA B or in Octave, which is sort of — I \nknow some people call it a free ve rsion of MATLAB, which
it sort  of is, sort of isn\'t.  \nSo I guess for those of you that haven\'t s een MATLAB before, and I know most of you \nhave, MATLAB is I guess part of the
programming language that makes it very easy to write codes using matrices, to write code for numerical routines, to move data around, to \nplot data. And it\'s sort
of an extremely easy to  learn tool to use for implementing a lot of \nlearning algorithms.  \nAnd in case some of you want to work on your  own home computer or
something if you \ndon\'t have a MATLAB license, for the purposes of  this class, there\'s also — [inaudible] \nwrite that down [inaudible] MATLAB — there\' s also a
software package called Octave \nthat you can download for free off the Internet. And it has somewhat fewer features than MATLAB, but it\'s free, and for the purposes
of  this class, it will work for just about \neverything.  \nSo actually I, well, so yeah, just a side comment for those of you that haven\'t seen \nMATLAB before I
guess, once a colleague of mine at a different university, not at \nStanford, actually teaches another machine l earning course. He\'s taught it for many years. \nSo
one day, he was in his office, and an old student of his from, lik e, ten years ago came \ninto his office and he said, "Oh, professo r, professor, thank you so much
for your', metadata={'page': 8, 'source': '/content/MachineLearning-Lecture01 (1).pdf'})
```

```
##########################################################
```

Edge Case 2 - Failure modes: Specifity

We can see a new failure mode.

The question below asks a question about the third lecture, but includes results from other lectures as well.



```
[33] question = "what did they say about regression \
       in the third lecture?"
```

```
[34] docs = vectordb.similarity_search(question,k=5)
```

```
[35] for doc in docs:
         print(doc.metadata)
```

```
{'page': 0, 'source': '/content/MachineLearning-Lecture03.pdf'}
{'page': 14, 'source': '/content/MachineLearning-Lecture03.pdf'}
{'page': 0, 'source': '/content/MachineLearning-Lecture02.pdf'}
{'page': 6, 'source': '/content/MachineLearning-Lecture03.pdf'}
{'page': 4, 'source': '/content/MachineLearning-Lecture03.pdf'}
```

```
[36] print(docs[4].page_content)
```

```
when you had a Q's tow. Like you make it too small in your -
Instructor (Andrew Ng) :Yes, absolutely. Yes. So local ly weight
into - locally weighted regression is not a penancier for the pr
underfitting. You can still run into the same problems with loca
What you just said about - and so some of these things I'll leav
yourself in the homework problem. You'll actu ally see what you
Student: It almost seems like you're not even th oroughly [inaud
weighted, you had all the data th at you originally had anyway.
```