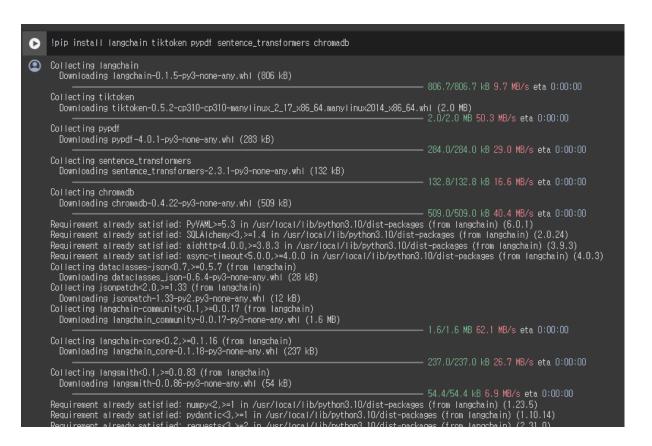
<Vetor Database>

https://colab.research.google.com/drive/1me5iRJCmrlpyUgByrRLZOjC3q1AVXmwo?usp=sharing

- 일반적인 데이터베이스는 글자가 똑같지 않으면 검색이 안 되는데
- 벡터 데이터베이스는 오탈자가 있어도 됨 (벡터 기반이기 때문!!)
- 그래서 나름 유용함!

필요한 거 싹 설치하자



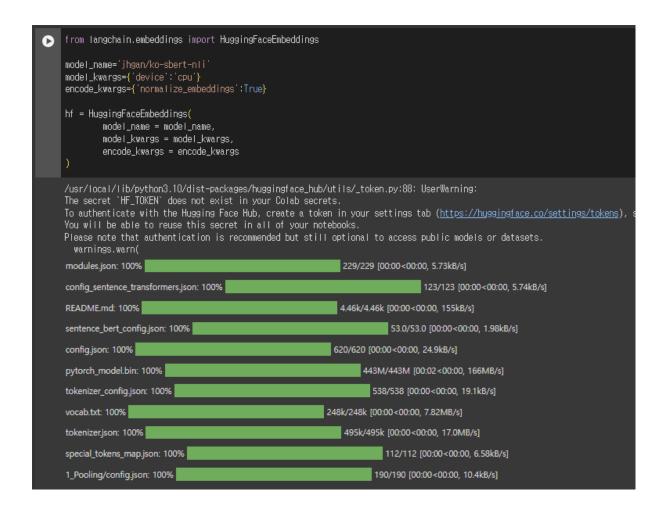
- chromadb가 벡터 데이터베이스임
- 나중에 실무에선 이런 거 하나하나가 컨테이너로 되어있을 것!
- ⇒ 특정한 작업에 대한 컨테이너 숫자 탄력적으로 관리 가능. 도커 컨테이너...

- 토큰 단위로 자르는 splitter로 자름
- 알아서 잘라주는 recursive splitter, 토큰 500개 단위로 자름

```
from langchain.document_loaders import PyPDFLoader

loader = PyPDFLoader('/content/[정책브리프 2021-04] 탄소중립 대응을 위한 정부 정책과 동향 (1).pdf')
pages = loader.load_and_split()
```

• pdf 파일 업로드 & 페이지 단위로 쪼개기



- Hugging Face에 있는 임베딩 모델 중에서 가져오겠다~
- 거대 언어 모델 말고도 작은 언어 모델에서도 사용
- microsoft에서 만든 small language model: "PHI"
 성능이 좋아지면 소프트웨어에도 탑재될 수도 있음

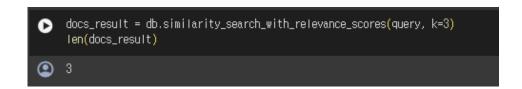
```
[ ] # GIOIE의 Split
docs = text_splitter.split_documents(pages)
# Vector Store 저장
from langchain.vectorstores import Chroma
db = Chroma.from_documents(docs, hf)

[ ] query = '탄소중립'
docs_result = db.similarity_search(query)

② print(docs_result[0].page_content)
print('-' * 100)
print(docs_result[0].metadata['page'])

② 010]노비즈 정책브리프
이 탄소중립과 대용 필요성
탄소중립은 개인・회사・단체 등에서 배출되는 탄소와 흡수되는 탄소량을 같게 해 탄소 '순배출이 제 로(0)' 가 되게 하는 것으로, '넷·제료(Net-Zero)' 라 부르기도 함
- 인간의 활동에 의한 존실까소 * 배출을 최대한 줄이고, 남은 존실까스는 흡수, 제거하며 '순배출이 제로
(0)' 가 되게 하는 것으로, '넷·제료(Net-Zero)' 라 부르기도 함
- 인간의 활동에 의한 존실까소 * 배출을 해변한 줄이고, 남은 존실까스는 흡수, 제거하며 '순배출이 제로
(0)' 가 되는 개념이며, 기후변화에 대용하기 위한 글로벌 추세로 자리 잡고 있음
*온실가스란 '지구 대기를 오염시켜 온실 관리를 일으키는 가스'를 통들이 이르는 말로, 이산화탄소 (602), 메탄(어4), 아산화질소 (N20), 수
소불하탄소 (바Cs), 과불화탄소 (바FCs), 목불화할 (SF6) 등은 배출 규제 및 거래의 대상이 되고 있음
대용 필요성
·지구 온난화로 인한 폭염, 폭설, 태풍, 산별 등 이상기후 현상이 세계 곳곳에서 발생되고 있는 가
윤대 기후변화 문제에 대한 국제사회의 공동 대용 필요성이 대두될
```

- 아까 설치한 chromadb를 langchain에서 랩핑해놓은 애가 Chroma
- 그냥 집어넣으면 안 되고 벡터 기반으로 임베딩해서 집어넣어야 함
- 모든 DB는 파일에 있지만 얘는 메모리에 있음. 그래서 껐다 키면 날라감. 대신 속도는 엄청 빠름.
- 암튼... 지금 순식간에 파일을 임베딩해서 DB를 생성하는 작업을 함. 위에 있는 코드가...
- 정보 세 가지
 - 1. 페이지 컨텐츠: 갖고 온 데이터의 진짜 내용
 - 2. 백 개의 -
 - 3. 페이지 0번의 메타 데이터. 실제론 54페이지의 내용임
- 그 다음 할 거: 이걸 싹 긁어다가 GPT한테 묻기
- 그럼 GPT는 이 내용을 기반으로 해서 나한테 말을 만들어서 답할 것
- "RAG" 증폭시켜서 답한다~

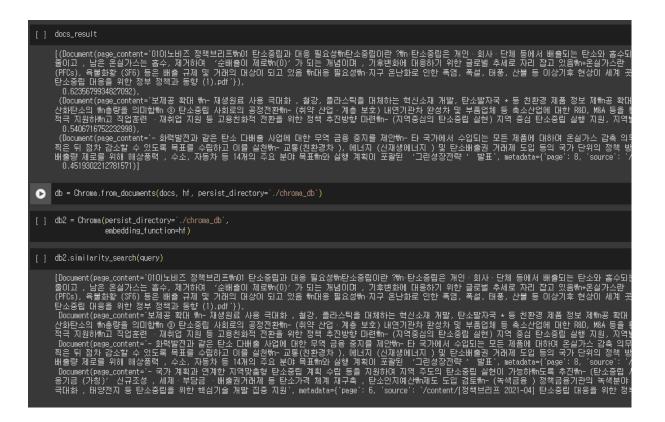


- 유사도가 높은 걸 세 개까지 뽑아준다
- 여기서 k=3은 k-means 알고리즘에서의 k와 비슷한 느낌임

• 정확도를 높이기 위해서 k=3를 갖고 세 개를 같이 뭉쳐서 대답하도록 할 수 있음



- 임베딩해서 DB를 만들면서 chroma_db라는 파일로 저장해줌
- 그래서 이후부턴 파일만 읽어들이면 실행되도록 함
- 이걸 클라우드로 한다면 이 파일을 클라우드의 스토리지로 넣어두면 됨! (AWS EC2 같은)



• 결과가 이렇게 나옴

<FAISS>

페이스북이 만든 벡터 DB. 근데 별로임 이런 게 있다~ 정도만 봐두기



<Retirivel 그래서 이걸 어떻게 연결하느냐~> 검색해온 결과를 토대로 말을 만들어주는 걸 볼 것

```
| Display | Disp
```

```
os.environ['OPENAI_API_KEY'] = 'bb6860ae08e147eeae418efdce560e4e'
os.environ['AZURE_OPENAI_ENDPOINT'] = 'https://helloai-openai-025.openai.azure.com/'
os.environ['OPENAI_API_TYPE'] = 'azure'
os.environ['OPENAI_API_VERSION'] = '2023-05-15'

[] # Data의 loading
import openai
import tiktoken

from langchain.document_loaders import PyPOFLoader
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.text_models import AzureChardOpenAl
from langchain.vectorstores import Chroma
from langchain.chains import RetrievalQA

loader = PyPOFLoader('/content/[정책브리프 2021-04] 탄소중립 대응을 위한 정부 정책과 동향 (1).pdf')
pages = loader.load_and_split()
```

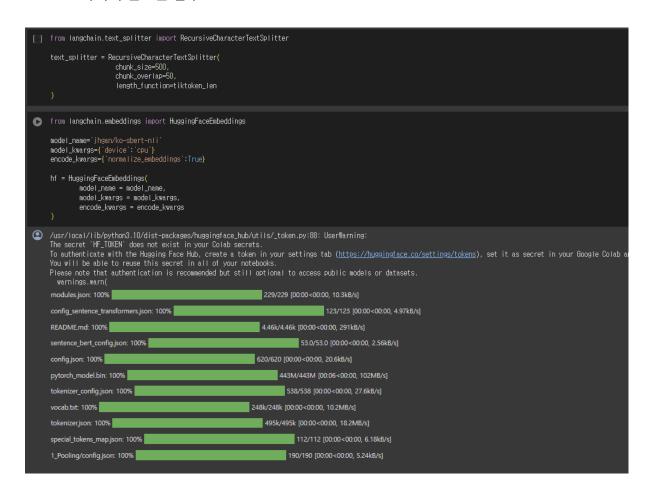
- key & endpoint 값 바꿔주고 pdf 파일 업로드
- 필요한 패키지 업로드

- RetrievalQA를 통해 묶어 놓으면 질문을 던졌을 때 다 연결되어있는 채로 답함
- 경로 복사 & 로딩 & 스플릿

```
[ ] import tiktoken
    tokenizer = tiktoken.get_encoding('c|100k_base')

    def tiktoken_len(text):
        tokens = tokenizer.encode(text)
        return len(tokens)
```

• 토크나이저 만드는 함수



- splitter를 가져오고 아까 사용했던 HugginFace 임베딩 모델 가져옴
- 나중에 코딩할 땐 무료라.. 오류 생길 수 있음.

- ESG 평가... 탄소 얼마나 사용하는지 다 따짐. 준수 X→ 펀드도 안 들어옴
- 답을 만드는 모델을 좋은 걸 쓰면 또 더 좋으니까... 우리는 GPT 3.5 turbo를 사용할 것
- chain_type: 결과가 나오면 묶어줌. 오늘은 가장 기본적인 방법을 쓸 것
- retriever: db는 아까 위에서 만든 chromadb... chromadb의 유사도 검색을 할 것
 k=3 유사도 3개... fetch_k=10: 내가 질문을 던지면 chromadb에서 10개를 뽑아내고, 그 중에 가장 괜찮을 걸 뽑겠다는 거
- return_source_documents=True: 어느 pdf 파일의 몇 번째 페이지... 이런 정보 돌려줌
- // 암튼 이렇게 gna chain을 연결함. chain한테 일을 시키면 거기 묶여있는 애들까지 다~

[] result = qa(query) result

{'query': '탄소중립에 중소기업이 필요한 이유는?',
'result': '중소기업의 적극적인 참여가 필요한 이유는 탄소중립을 실현하기 위해 국가 경제에서 중소기업이 차지하는 비중이 높기 때문입니다. 중소기업의 참여 있는 기회를 얻을 수 있습니다.',
'source_document(pase_content='대한 대응 필요성을 느끼지 못한다는 결과를 보여주고 있음\ 항호 중소기업을 대상으로 도입될 정부 정책과 제 불가능합\ "\- (따라서 탄소중립에 대한 중소기업의 적극적 참여가 필요하며 , 중소기업의 참여를 유인할 수 있는 다양한 \ 청적 전책과 제도가 마련될 예정임\ "\- '성장할 수 있는 동력을 마련할 수 있을 것임', metadata='pase': 11, 'source': '/content/(장책브리프 2021-04) 탄소중립 대응을 위한 정부 정책과 동향 (1), 'Bournent(pase_content='숙화 추진\ "\- 산업 분야에서는 재생에너지 , 수소, 에너지 IT 등 3대 에너지 신산업 육성에 집중\ "\- (고탄소 산업구조 혁신) 철강, 석업 등 다배출 중소기업을 대상으로 공정진단을 통해 맞춤형 공정개선 실비보급을 지원\ "\하는 등 산업 전과정 (연.원료→공정→제품→소바.자원 순환)에서의 저단 거점별 수소 충전소 구축\ "\- (도시·국토 저탄소화) 신규 건축물 제로에너지 건축 의무화 , 국토 계획 수립 시 생태자원 활용한 탄소중\ "하는 기능 강화 \ "\ ② 신 Document(pase_content='010)보다 조책보리프��이를 다용 필요성\ "한자 한소중립이라 '\ "*\ "한소 본소중립은 개인 '화사 '단체 등에서 배출되는 탄소와 흡수되는 중심 이고 , 남은 온실가스는 흡수, 제거하여 '소배출이 제로\ "(0)' 가 되는 개념이며 , 기흥변화에 대응하기 위한 글로벌 추세로 자리 접고 있음\ "\- 는원 중입 대응을 위한 정부 정책과 동향 (1),pdf'})}

- qa 체인한테 질문을 던지고 결과를 가져옴
- 결과 잘 나옴~
- 그냥 하면 조각 데이터만 나왔을 것. 근데 여긴 gpt가 말을 만들어서 답해줌
- 소스 document... page 60에서 가져왔다고까지 알려줌

// 끝...