RAG_mission

소유자	민 민영
태그	

복습한다고 생각하고 수업시간에 사용한 모델들 위주로 사용함.

```
#사용한 모델
OpenAIEmbeddings(model='text-embedding-3-large')
llm = ChatOpenAI(
    model="gpt-4o-mini",
```

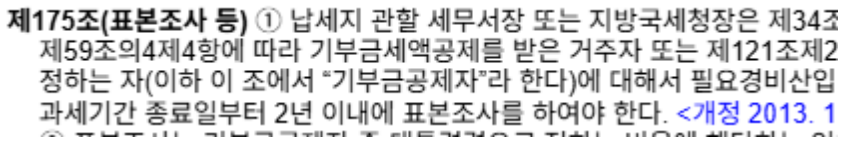
#데이터로드

docx 는 txt 가 아니라서 TextLoader아니라 Docx2txtLo
그래도 안되서
%pip install docx2txtader 설치

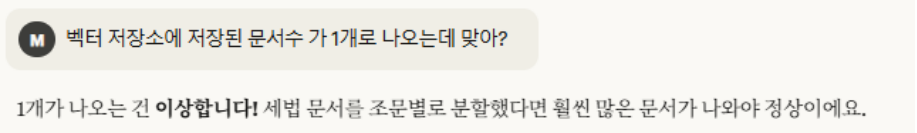
#청킹

^제\d+조
로 했는데 문서 계속 1개라고.,.

그러다가 조항의 내용이
^제\d+조
로 시작하는 것 발견해서 클로드에 질의


제175조(표본조사 등) ① 납세지 관할 세무서장 또는 지방국세청장은 제34조
제59조의4제4항에 따라 기부금세액공제를 받은 거주자 또는 제121조제2
정하는 자(이하 이 조에서 "기부금공제자"라 한다)에 대해서 필요경비산입
과세기간 종료일부터 2년 이내에 표본조사를 하여야 한다. <개정 2013. 1

볼드만 걸러지게 하라고 알려줘서 했는데도 계속 1개문서



여러 패턴으로 테스트 해보라고 알려줘서 해보니 3번째 패턴이 205로 가장 근접한 값으로 출력됨

```
# 여러 패턴 테스트
patterns = [
    r"\*\*제\d+조\([^)]+\)\*\*",  # 현재 사용하는 패턴
    r"\*\*제\d+조",               # 간단한 패턴
    r"제\d+조\([^)]+\)",          # 볼드 없는 패턴
    r"제\d+조"                    # 가장 간단한 패턴
]
```

chunk_size 조절
100 -> 500까지 단계적으로 늘려가며 확인한 결과 = 480이 192개로 나와서 이걸로 함.

```
...
Created a chunk of size 704, which is longer than the specified 480
Created a chunk of size 483, which is longer than the specified 480
Created a chunk of size 616, which is longer than the specified 480
Created a chunk of size 1680, which is longer than the specified 480
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
한국어 문서 수:  192
```

=> 10조, 11조와 같이 내용이 적은 조항들은 하나로 묶임.

> 나중에 보니 같은 조항이 반복되는 경우도 있어서 chunk_size를 더 적게 했어도 됐을 것 같음.

<query>
과세기간은 어떻게 되나요?
과세 기간 -> 띄어쓰기에 따라 다름..?
0.5454 / 0.4869 : 문서상 붙여서 적혀있으나 띄어쓰기한 질의가 더 유사도 높게 나옴

- 유사도 출력값이 0.5가 넘어서 score_threshold를 0.5로 하면 결과값이 안나와서 0.1로 여러 개 테스트 해본 후 0.3으로 함.

```
tax_retriever = chroma_db.as_retriever(
    search_type = "similarity_score_threshold",
    search_kwargs = {
        "k" : 3,
        "score_threshold" : 0.3,    #0.4 이상으로 하면 결과가 안나옴.
    },
```

```
)
```

#퓨샷 사용

모델에서 자체 데이터로 답변해서 계속 답이 틀림
클로드에 질의하여 퓨샷 동적으로 사용하도록 수정함.
그리고 퓨샷 더 추가

히스토리 추가하니까 다시 위의 문제가 반복됨.
인줄 알았으나 같은 질문 반복하니 답이 달라지는 문제였음.

그래서 히스토리 일단 지우고 다시 확인

- 그리고 같은 질문 반복할때마다 답변이 다름. (퓨샷 넣은것만 정확하게 나오는거 같기도 함니다..)
- 억 단위가 넘어가면 계산을 아래 범주와 해당 범주 두 가지 모두 계산해서 결과를 도출하거나(이렇게하면 오히려 결과값은 맞음) 다른 범주의 계산식을 적용하거나 어쨋든 틀리게 계산함...

모델을 바꾸려고보니 시간이 부족해서 일단 여기까지....입니다🫠