

数据加载，分割与向量化入库

1.数据load

根据实验文档说明，使用 `langchain.document_loaders` 中的 `CSVLoader` 来加载的数据

```
1 from langchain_community.document_loaders import CSVLoader
2
3 # Load CSV data
4 file_path = "law_data_3k.csv"
5 loader = CSVLoader(file_path=file_path)
6 data = loader.load()
```

load之后观察data的数据情况

```
Total number of records: 3013
Processed 1000 records.
Processed 2000 records.
Processed 3000 records.
```

共计3013组数据，而观察law_data_3k.csv实际共有近6000行，这是因为load操作把2400余条问答的问题和回答两项记在一起了，：

```
仍不能正常使用的；，（三）发动机、变速器、动力蓄电池、行驶驱动电机、转向系统、制动系统、悬架系统、传动系统、污染控制装置、车身的同一主要零部件因其质量问题累计更换2次，仍不能正常使用的；，（四）因质量问题累计修理时间超过30日，或者因同一质量问题累计修理超过4次的。；发动机、变速器、动力蓄电池、行驶驱动电机的更换次数与其主要零部件的更换次数不重复计算。；需要根据车辆识别代号（VIN）等定制的防盗系统、全车主线束等特殊零部件和动力蓄电池的运输时间，以及外出救援路途所占用的时间，不计入本条第一款第（四）项规定的修理时间。
data: 夫妻一方欠下的赌债是个人债务还是夫妻
根据《婚姻法》规定，离婚时，原为夫妻共同生活所欠的债务，应当共同偿还。共同财产不足清偿的，或财产归各自所有的，由双方协议清偿；协议不成时，由人民法院判决。共同生活所欠的债务即夫妻共同债务，是夫妻为了共同生活或履行抚养、赡养义务所欠的债务。而一方赌博所欠的债务，由于该债务未用于夫妻共同生活和家庭生活，属于一方个人不合理的开支，不属于夫妻共同债务的范围，因而应由债务人个人自行承担，另一方不承担偿还责任。
data: 男方提出离婚，一年后要求返还彩礼，女方必须退吗
当事人请求返还按照习俗给付的彩礼的，如果查明属于以下情形，人民法院应当予以支持：（一）双方未办理结婚登记手续的；（二）双方办理结婚登记手续但确未共同生活的；（三）婚前给付并导致给付人生活困难的。适用前款第（二）、（三）项的规定，应当以双方离婚为条件。已登记结婚，尚未共同生活，一方或双方受赠的礼金、礼物应认定为夫妻共同财产，具体处理时应考虑财产来源、数量等情况合理分割。各自出资购置、各自使用的财物，原则上归各自所有。借婚姻关系索取的财物，离婚时，如结婚时间不长，或者因索要财物造成对方生活困难的，可酌情返还。对取得财物的性质是索取还是赠与难以认定的，可按赠与处理。不予返还彩礼的情形1、不符合最高人民法院《关于适用〈中华人民共和国民事诉讼法〉若干问题的解释》（二）》第十条规定情形，一方请求返还彩礼的，不予支持。另外对该条中的“婚前给付并导致给付人生活困难的”的情形，应做限制性的解释。该情形是指给付彩礼的一方婚前举债给付、婚后无经济来源偿还债务的，或者是婚前用家庭财产给付、婚后无固定经济来源、依靠自己的力量无法维持最基本的生活水平。确定“生活困难”需根据给付彩礼的数额、给付人的生活来源、当地生活
```

如上图，一个data块里包括了问题和解答，而在csv里他们属于不同的行。

2.数据分割

数据分割参考 `langchain.text_splitter` 中的 `CharacterTextSplitter` ,其中分割参数有下面这些选项：

```
1 # Initialize text splitter
2 text_splitter = CharacterTextSplitter(
3     separator="\n\n",          # Use paragraph separator
4     chunk_size=1000,          # Set chunk size
5     chunk_overlap=100,        # Set overlap between chunks to 100 characters
6     length_function=len,      # Use len to calculate text length
7     is_separator_regex=False,
8 )
```

测试发现当分割符选择 `\n\n` 时，数据块分割后数量不变，说明文本里并没有 `\n\n` ,这个不能起到好的分割作用，而选择 `\n` ,的情况下块的数量会有所增长。

关于块大小的问题，选择不同 `chunk size` 的块进行测试，所得分块数量如下：

chunk = 1000

```
Total number of records: 3013
Processed 1000 records.
Processed 2000 records.
Processed 3000 records.
Total number of chunks: 3261
```

chunk = 500

```
Total number of records: 3013
Processed 1000 records.
Processed 2000 records.
Processed 3000 records.
Total number of chunks: 4254
```

chunk = 300

```
Total number of records: 3013
Processed 1000 records.
Processed 2000 records.
Processed 3000 records.
Total number of chunks: 4254
```

chunk = 200

```
Total number of records: 3013
Processed 1000 records.
Processed 2000 records.
Processed 3000 records.
Total number of chunks: 4921
```

chunk = 150

```
Total number of records: 3013
Processed 1000 records.
Processed 2000 records.
Processed 3000 records.
Total number of chunks: 5092
```

分块100时会有提示分块太小，根据 csv 数据分布，个人觉得选则 200 块大小比较合适。

3. 向量化处理与入库

选取实验文档里的 m3e-base 嵌入模型构建向量，把这个模型部署在本地后封装在

MyEmbeddingFunction 里作为参数传给 chroma，chroma 将 document 里的数据向量化后存放在建立的数据库里：

```

1
2 class MyEmbeddingFunction(EmbeddingFunction):
3     def embed_documents(self, texts: Documents) -> Embeddings:
4         # 使用 tqdm 显示嵌入进度条
5         embeddings = [
6             model.encode(text, convert_to_numpy=True) for text in
7             tqdm(texts, desc="Embedding Texts", unit="text")
8         ]
9         # 返回嵌入结果
10        return [embedding.tolist() for embedding in embeddings]
11
12 # Load pre-trained SentenceTransformer model
13 model = SentenceTransformer('D:/2024 fall/web/lab/实验三/m3e-base')

```

```

1 try:
2     # Chroma expects an embedding function, passing MyEmbeddingFunction as
the embedding function
3     persist_directory = "./chroma_data_200"
4     db = Chroma.from_documents(
5         documents,
6         MyEmbeddingFunction(),
7         persist_directory=persist_directory
8     )
9     print("Chroma database saved")
10 except Exception as e:
11     print(f"Error saving Chroma database: {e}")
12

```

最后运行结果

[illegible]

chunk= 200 的运行结果以及源代码在这里。(附上chunk=300 的结果, 以便后续选择)

链接: <https://rec.ustc.edu.cn/share/e1899490-d0ce-11ef-ab47-e9e1edc2baf9>

密码: 1958