
- 출처: LangChain 공식 문서 또는 해당 교재명
- 원본 URL: https://smith.langchain.com/hub/teddynote/summary-stuff-documents

---

CSV

- **Comma-Separated Values (CSV)**
  - **쉼표** 로 **값** 을 구분하는 구분된 **텍스트 파일**
  - 파일의 각 줄 = 데이터 레코드
  - 각 레코드 = 쉼표로 구분된 하나 이상의 필드로 구성

CSVLoader

- **CSV 데이터를 문서당 한 행씩 로드**

```
# API KEY를 환경변수로 관리하기 위한 설정 파일
import os
from dotenv import load_dotenv

# API KEY 정보로드
load_dotenv()               # true
```

- 사전에 `VS Code` **터미널에 설치**

```
pip install langchain-community
```

```
from langchain_community.document_loaders.csv_loader import CSVLoader

# CSV 로더 생성
loader = CSVLoader(file_path="../06_Document_Loader/data/titanic.csv")

# 데이터 로드
docs = loader.load()

print(len(docs))                        # 891
print(docs[0].metadata)                 # {'source': '../06_Document_Loader/data/titanic.csv', 'row': 0}
```

CSV 파싱 및 로딩 커스터마이징

- **csv module** 문서를 참조하여 지원되는 csv args에 대한 자세한 정보를 확인하기

```
# 컬럼정보:
# PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked

# CSV 파일 경로
loader = CSVLoader(
    file_path="../06_Document_Loader/data/titanic.csv",
    csv_args={
        "delimiter": ",",                                   # 구분자
        "quotechar": '"',                                   # 인용 부호 문자
        "fieldnames": [
            "Passenger ID",
            "Survival (1: Survived, 0: Died)",
            "Passenger Class",
            "Name",
```

```
            "Sex",
            "Age",
            "Number of Siblings/Spouses Aboard",
            "Number of Parents/Children Aboard",
            "Ticket Number",
            "Fare",
            "Cabin",
            "Port of Embarkation",
        ],                                                  # 필드 이름
    },
)

# 데이터 로드
docs = loader.load()

# 데이터 출력
print(docs[1].page_content)
```

- 셀 출력

```
Passenger ID: 1
Survival (1: Survived, 0: Died): 0
Passenger Class: 3
Name: Braund, Mr. Owen Harris
Sex: male
Age: 22
Number of Siblings/Spouses Aboard: 1
Number of Parents/Children Aboard: 0
Ticket Number: A/5 21171
Fare: 7.25
Cabin:
Port of Embarkation: S
```

- `source_column` **인자를 사용** → 각 행에서 생성된 문서의 출처 지정하기
  - 그렇지 않으면 `모든 문서의 출처` = `file_path`**가 사용됨**
  - `CSV` 파일에서 로드된 문서를 **출처를 사용해 질문에 답하는 체인에 사용할 때 유용**

```
loader = CSVLoader(                                         # CSV 로더 설정
    file_path="../06_Document_Loader/data/titanic.csv",     # 1_파일 경로 지정
    source_column="PassengerId"                             # 2_소스 컬럼 지정
)

# 데이터 로드
docs = loader.load()

# 데이터 출력
print(docs[1])
```

- 셀 출력

```
page_content='PassengerId: 2
Survived: 1
Pclass: 1
Name: Cumings, Mrs. John Bradley (Florence Briggs Thayer)
Sex: female
Age: 38
SibSp: 1
Parch: 0
Ticket: PC 17599
Fare: 71.2833
Cabin: C85
Embarked: C' metadata={'source': '2', 'row': 1}
```

- `CSV Loader` vs `source_column`

`CSV_Loader`	`source_column`
Passenger ID: 1	page_content='PassengerId: 2

CSV_Loader	source_column
Survival (1: Survived, 0: Died): 0	Survived: 1
Passenger Class: 3	Pclass: 1
Name: Braund, Mr. Owen Harris	Name: Cumings, Mrs. John Bradley (Florence Briggs Thayer)
Sex: male	Sex: female
Age: 22	Age: 38
Number of Siblings/Spouses Aboard: 1	SibSp: 1
Number of Parents/Children Aboard: 0	Parch: 0
Ticket Number: A/5 21171	Ticket: PC 17599
Fare: 7.25	Fare: 71.2833
Cabin:	Cabin: C85
Port of Embarkation: S	Embarked: C' metadata={'source': '2', 'row': 1}

`UnstructuredCSVLoader`

- `UnstructuredCSVLoader` 사용 → **`테이블 로드`할 수도 있음**
- 장점: **`elements`** 모드에서 사용 → **`메타데이터`에서 `테이블`의 `HTML 표현`이 제공된다는 것**
- 사전에 `VS Code` 터미널에 설치할 것

```
pip install unstructured
```

```
from langchain_community.document_loaders.csv_loader import UnstructuredCSVLoader

# 비구조화 CSV 로더 인스턴스 생성
loader = UnstructuredCSVLoader(
    file_path="../06_Document_Loader/data/titanic.csv",         # 파일 경로
    mode="elements")                                            # elements 모드 설정

# 문서 로드
docs = loader.load()

# 첫 번째 문서의 HTML 텍스트 메타데이터 출력
print(docs[0].metadata["text_as_html"][:1000])
```

- 셀 출력 (13.8s)

```
<table><tr><td>PassengerId</td><td>Survived</td><td>Pclass</td><td>Name</td><td>Sex</td><td>Age</td><td>SibSp</td>
```

- 마크다운 테이블로 변환

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	S	
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925	S	
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S

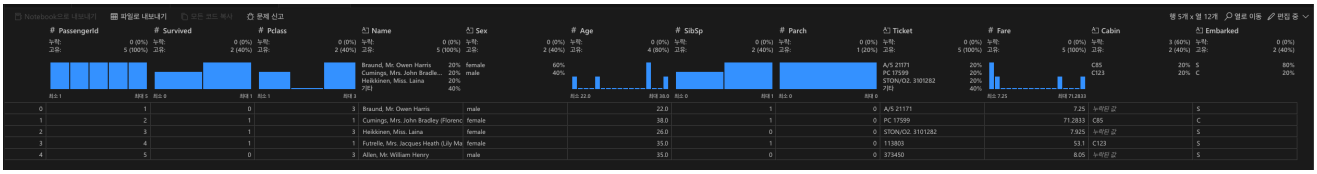
`DataFrameLoader`

- **`Pandas`**
  - **`Python`** **프로그래밍 언어를 위한 오픈 소스 데이터 분석 및 조작 도구**
  - 데이터 과학, 머신러닝, 그리고 다양한 분야의 데이터 작업에 널리 사용

```
import pandas as pd

# CSV 파일 읽기
df = pd.read_csv("../06_Document_Loader/data/titanic.csv")

# 첫 5개행 조회해보기
df.head()
```

- 첫 5개 행 조회 결과
  - 

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25	누락된 값	S
1	2	1	1	Cumings, Mrs. John Bradley (Florenc	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925	누락된 값	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily Mz	female	35.0	1	0	113803	53.1	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05	누락된 값	S

```
# 데이터 프레임 로더 임포트
from langchain_community.document_loaders import DataFrameLoader

# 데이터 프레임 로더 설정, 페이지 내용 컬럼 지정
# df=데이터 프레임 객체, page_content_column=페이지 내용 컬럼 이름
loader = DataFrameLoader(df, page_content_column="Name")

# 문서 로드
docs = loader.load()

# 첫 번째 문서의 페이지 내용 출력
print(docs[0].page_content)

# 첫 번째 문서의 메타데이터 출력
print(docs[0].metadata)
```

- 셀 출력 (0.1s)

```
Braund, Mr. Owen Harris
{'PassengerId': 1, 'Survived': 0, 'Pclass': 3, 'Sex': 'male', 'Age': 22.0, 'SibSp': 1, 'Parch': 0, 'Ticket': 'A/5
```

```
# 큰 테이블에 대한 지연 로딩, 전체 테이블을 메모리에 로드하지 않음

for row in loader.lazy_load():
    print(row)
    break                                   # 첫 행만 출력
```

- 셀 출력

```
page_content='Braund, Mr. Owen Harris' metadata={'PassengerId': 1, 'Survived': 0, 'Pclass': 3, 'Sex': 'male', 'Ag
```

- *next: 엑셀(`Excel`)*
