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

---

`Excel`

- **`UnstructuredExcelLoader`** = `Microsoft Excel 파일`을 `로드`하는 데 사용
  - `.xlsx`, `.xls` 파일 모두에서 작동 → 페이지 내용은 `Excel`파일의 원시 텍스트가 됨
  - **`elements`** 모드: 문서 메타데이터의 **`text_as_html`** 키 아래에서 `Excel` 파일의 `HTML` 표현으로 제공됨

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

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

- 사전에 `VS Code` 터미널에 설치할 것

```
pip install -qU langchain-community unstructured openpyxl
pip install msoffcrypto-tool                                    # msoffcrypto 패키지
```

```
from langchain_community.document_loaders import UnstructuredExcelLoader

# UnstructuredExcelLoader 생성
loader = UnstructuredExcelLoader(
    "../06_Document_Loader/data/titanic.xlsx",              # 경로 설정
    mode="elements"                                         # elements 모드
    )

# 문서 로드
docs = loader.load()

# 문서 길이 출력
print(len(docs))                    # 1 (셀 출력: 4.6s)
```

- 1개의 문서가 로드되었음을 확인

- **`page_content`** = 각 행의 `데이터`가 저장
- **`metadata`** 의 **`text_as_html`** = 각 행의 데이터를 `HTML 형식`으로 저장

```
# page_content 출력해보기

print(docs[0].page_content[:200])
```

- 셀 출력

```
PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Cabin Embarked 1 0 3 Braund, Mr. Owen Harris mal
```

```
# metadata 의 text_as_html 출력

print(docs[0].metadata["text_as_html"][:1000])
```

- 셀 출력

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

- 마크다운 테이블 형식으로 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`

- `Excel` 파일을 로드하는 **`read_excel()` 기능을 사용 → `DataFrame` 으로 만든 뒤, 로드**
  - (`CSV` 파일과 같은 방법)

```
import pandas as pd

# Excel 파일 읽기
df = pd.read_excel("../06_Document_Loader/data/titanic.xlsx")
```

```
from langchain_community.document_loaders import DataFrameLoader

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

# 문서 로드
docs = loader.load()

# 데이터 출력
print(docs[0].page_content)

print("\n", "="*50, "\n")

# 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
```

- *next: 워드(`Microsoft Word`)*