

Report

김준호

1번 과제: 간단한 감상

R과 pandas의 comparison에 대해 자세히 알 수 있어서 좋았다.

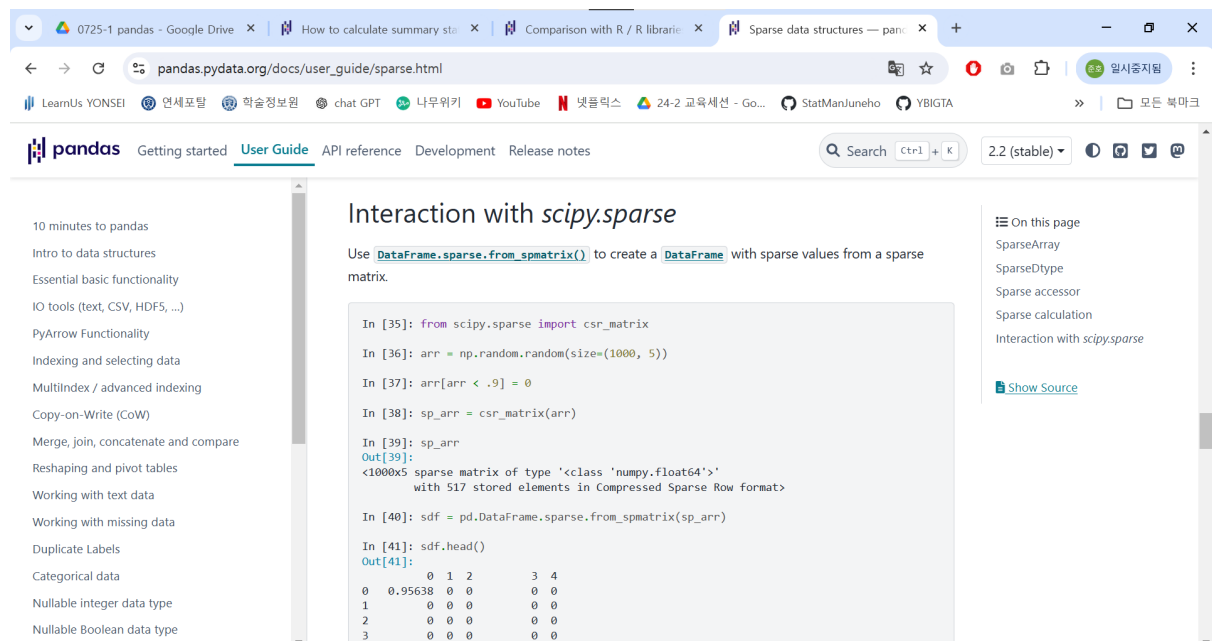
This screenshot shows the pandas documentation page for 'Comparison with R / R libraries'. The page is titled 'Transforming' and lists several R functions alongside their pandas equivalents. The R functions are: `select(df, col_one = col1)`, `rename(df, col_one = col1)`, and `mutate(df, c=a-b)`. The pandas equivalents are: `df.rename(columns={'col1': 'col_one'})['col_one']`, `df.rename(columns={'col1': 'col_one'})`, and `df.assign(c=df['a']-df['b'])`. The page also includes a sidebar with navigation links and a search bar.

This screenshot shows the pandas documentation page for 'plyr'. The page is titled 'plyr' and explains that it is an R library for the split-apply-combine strategy for data analysis. The functions revolve around three data structures in R: `a` for arrays, `l` for lists, and `d` for data.frame. The table below shows how these data structures could be mapped in Python.

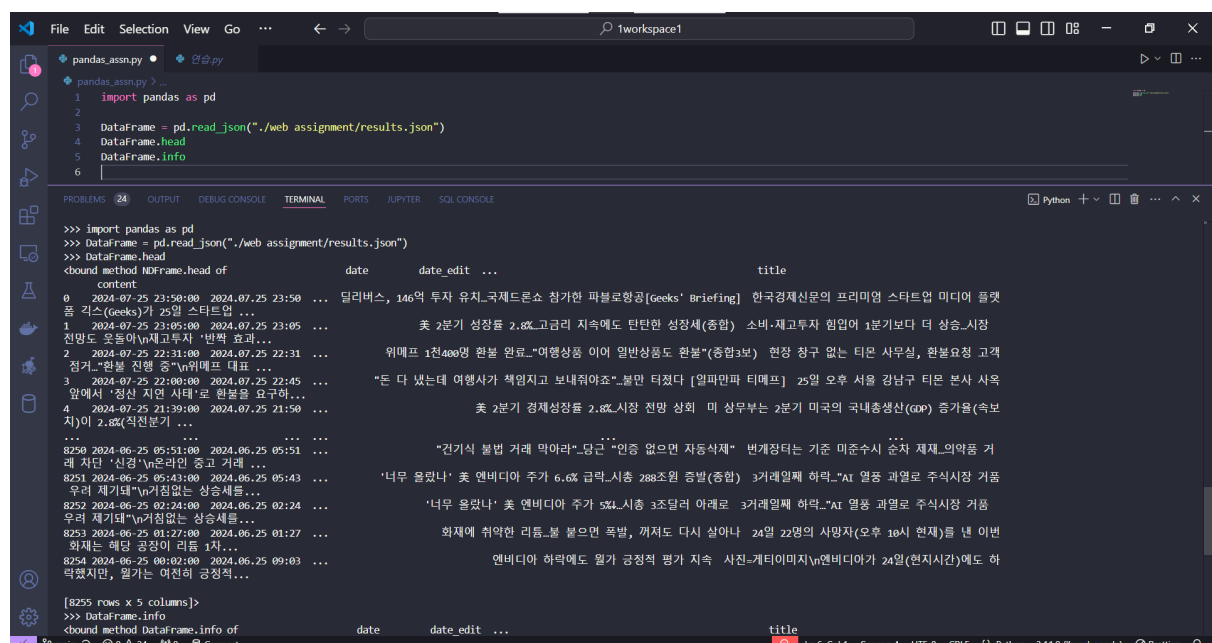
R	Python
array	list
lists	dictionary or list of objects
data.frame	dataframe

The page also includes a sidebar with navigation links and a search bar.

또한 평소 수업이든 연구든 Sparse한 Dataset을 이용해야 할 일이 많은데, Pandas를 통해 Sparse한 Dataset을 만드는 방법에 대해서 배울 수 있어 좋았다.



2번 과제: 캡처본 (DataFrame.head, DataFrame.info 실행결과)



pd.read.json 함수를 이용해 웹 과제의 환경 기사들을 불러왔다.

```
File Edit Selection View Go ... 1workspace1
pandas_asn.py 2 @.py
pandas_asn.py > ...
1 import pandas as pd
2
3 DataFrame = pd.read_json("../web assignment/results.json")
4 DataFrame.head()
5 DataFrame.info
6

PROBLEMS 24 OUTPUT DEBUG CONSOLE TERMINAL PORTS JUPYTER SQL CONSOLE Python + -
우려 제기돼 "거침없는 상승세들..." 화제에 취약한 리튬... 불 붙으면 폭발, 꺼져도 다시 살아나 24일 22명의 사망자(오후 10시 현재)를 낸 이번
8253 2024-06-25 01:27:00 2024.06.25 01:27 ... 화제는 해당 공장이 리튬 1차...
8254 2024-06-25 00:02:00 2024.06.25 00:03 ... 엔비디아 하락에도 월가 긍정적 평가 지속 사진-게티이미지\n엔비디아가 24일(현지시간)에도 하
[8255 rows x 5 columns]>
>>> DataFrame.info
<bound method DataFrame.info of
content
0 2024-07-25 23:50:00 2024.07.25 23:50 ... 딜리버스, 146억 투자 유치-국제드론쇼 참가한 파블로랑공[Geeks' Briefing] 한국경제신문의 프리미엄 스타트업 미디어 플랫폼
꿈 키스(Geeks)가 25일 스타트업 ...
1 2024-07-25 23:05:00 2024.07.25 23:05 ... 美 2분기 성장률 2.8%고금리 지속에도 탄탄한 성장세(종합) 소비-재고투자 힘입어 1분기보다 더 상승-시장
전망도 우호적\n재고투자 '반짝 효과...'
2 2024-07-25 22:31:00 2024.07.25 22:31 ... 위메프 1천400명 환불 완료-"여행상품 이어 일반상품도 환불"(종합3보) 현장 창구 없는 티몬 사무실, 환불요청 고객
점거-"환불 진행 중"\n위메프 대표 ...
3 2024-07-25 22:00:00 2024.07.25 22:45 ... "돈 다 있는데 여행사가 책임지고 보내줘야죠"...불만 터졌다 [일파만파 티메프] 25일 오후 서울 강남구 티몬 본사 사옥 앞에서 '정신 지원 사태'로 환불을
요구하...
4 2024-07-25 21:39:00 2024.07.25 21:50 ... 美 2분기 경제성장률 2.8%시장 전망 상회 미 상무부는 2분기 미국의 국내총생산(GDP) 증가율(속보치)이 2.8%(직전분기 ...
...
8250 2024-06-25 05:51:00 2024.06.25 05:51 ... "건기식 불법 거래 막아라" 당근 "인증 없으면 자동삭제" 변경장터는 기존 미준수시 순차 체재-의약품 거래 차단 '신경'\n온라인 중고 거래 ...
8251 2024-06-25 05:43:00 2024.06.25 05:43 ... '너무 올랐나' 美 엔비디아 추가 6.6% 급락-시장 288조원 증발(종합) 3거래일째 하락-"AI 열풍 과열로 주식시장 거품 우려 제기돼"\n거침없는 상승세들...
8252 2024-06-25 02:24:00 2024.06.25 02:24 ... '너무 올랐나' 美 엔비디아 추가 5.4% 시총 3조달러 아래로 3거래일째 하락-"AI 열풍 과열로 주식시장 거품 우려 제기돼"\n거침없는 상승세들...
8253 2024-06-25 01:27:00 2024.06.25 01:27 ... 화제에 취약한 리튬... 불 붙으면 폭발, 꺼져도 다시 살아나 24일 22명의 사망자(오후 10시 현재)를 낸 이번 화제는 해당 공장이 리튬 1차...
8254 2024-06-25 00:02:00 2024.06.25 00:03 ... 엔비디아 하락에도 월가 긍정적 평가 지속 사진-게티이미지\n엔비디아가 24일(현지시간)에도 하락했지만, 월가는 여전히 긍정적...
[8255 rows x 5 columns]>
>>> []
main 0 24 0 Connect Ln 6, Col 1 Spaces: 4 UTF-8 CRLF Python 3.11.9 (base: conda) Prettier
```

3번 과제: 조사 내용

1. **IO는 입력(Input)과 출력(Output)을 의미한다.** pandas에서는 다양한 형태의 데이터를 읽어들이고(입력) 그 결과를 다양한 형태로 저장(출력)할 수 있는 기능을 지원한다. 예를 들어, 파일 형태의 데이터를 읽어 DataFrame으로 변환하거나, DataFrame의 데이터를 파일 형태로 저장하는 것이 이에 해당한다.
2. **각각의 데이터 포맷에 대한 설명:**
 - **Pickle:** 파이썬의 객체 직렬화를 위해 사용되는 포맷이다. 객체의 상태를 그대로 파일에 저장하여 나중에 불러와서 사용할 수 있게 해준다. 이는 파이썬 전용이며, 데이터 복구나 파이썬 객체를 임시 저장할 때 유용하다.
 - **CSV (Comma-Separated Values):** 각 데이터 필드가 쉼표로 구분된 텍스트 파일 포맷이다. 대부분의 데이터베이스와 프로그램에서 널리 지원되며, 테이블 형태의 데이터를 저장하고 공유하는 데 많이 사용된다.
 - **TSV (Tab-Separated Values):** 각 데이터 필드가 탭으로 구분된 텍스트 파일 포맷으로, CSV와 유사하나 구분자만 다르다. 역시 테이블 형태의 데이터를 간단하게 저장하고자 할 때 사용된다.
 - **JSON (JavaScript Object Notation):** 데이터를 키-값 쌍으로 저장하는 포맷이다. 웹 데이터 교환에 매우 흔하게 사용되며, 구조가 유연하여 다양한 시스템 간의 통

신에 적합하다.

- **HTML (HyperText Markup Language):** 웹 페이지의 구조를 기술하기 위해 사용되는 마크업 언어이다. pandas에서는 테이블 형태의 데이터를 HTML 파일로 저장하거나, HTML 내의 테이블 데이터를 추출할 때 사용할 수 있다.
- **XML (eXtensible Markup Language):** HTML과 유사하나 태그의 목적이 자유롭게 정의될 수 있어 데이터의 저장과 전송에 널리 사용된다. 데이터 구조가 복잡한 문서나 메타데이터를 표현하는 데 적합하다.
- **Parquet:** 컬럼 기반의 데이터 저장 포맷으로, 대용량의 구조화된 데이터를 효율적으로 압축 및 저장할 수 있다. 분석을 위해 대량의 데이터를 빠르게 읽을 수 있도록 설계되었다.
- **YAML (YAML Ain't Markup Language):** 구성 파일 등을 작성하기 위해 사용되는 데이터 직렬화 포맷이다. JSON과 비슷하지만 가독성이 더 좋고, 중첩된 데이터 구조를 표현하는데 효과적이다.
- **TOML (Tom's Obvious, Minimal Language):** 주로 소프트웨어 프로젝트의 설정 파일을 작성하는 데 사용되며, YAML이나 JSON에 비해 더 간결하고 명확하다는 것이 특징이다.

3. Pickle의 직렬화와 역직렬화에 대한 설명:

- ****직렬화(serialization)****는 파이썬 객체(예: 리스트, 딕셔너리, DataFrame 등)를 바이트 스트림(파일이나 바이트 배열 등)으로 변환하는 과정이다. 이렇게 저장된 데이터는 파일 등에 저장되어 영구히 보관할 수 있다.
- ****역직렬화(deserialization)****는 저장된 바이트 스트림을 다시 파이썬 객체로 복원하는 과정이다. 이를 통해 프로그램이 종료되어도 이전에 사용하던 데이터를 복원할 수 있다.

pandas의 다양한 데이터 포맷 지원은 데이터 분석 시 여러 소스에서 데이터를 수집하고, 다양한 목적으로 데이터를 내보낼 때 유연성을 제공한다.