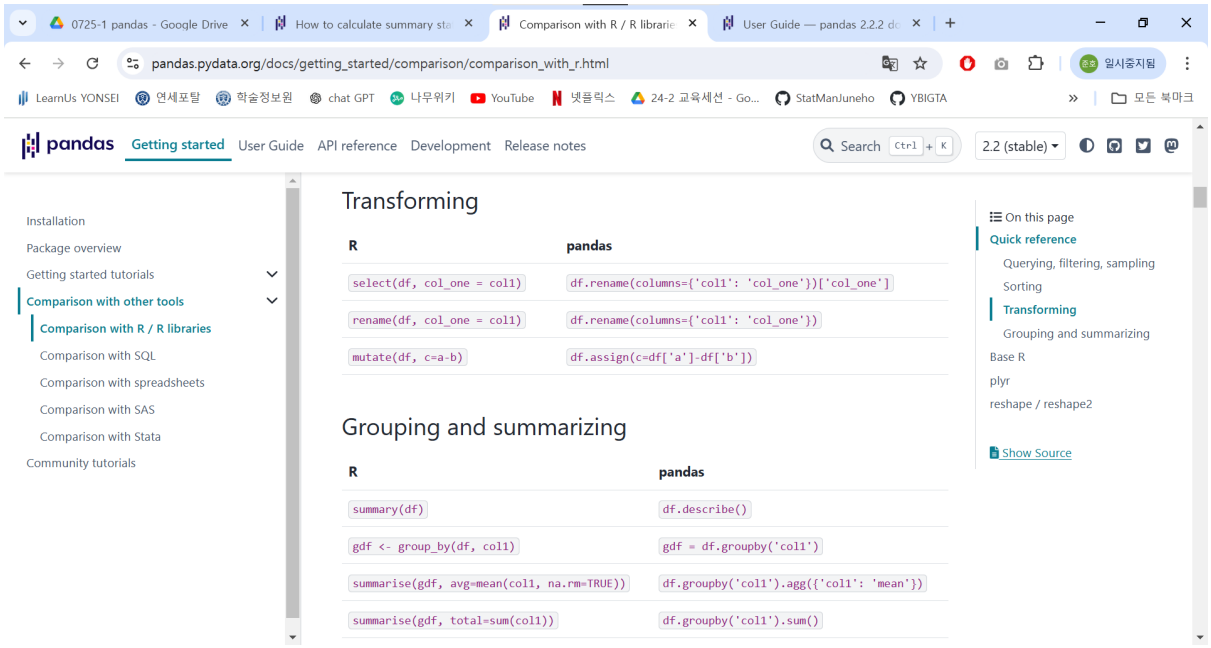
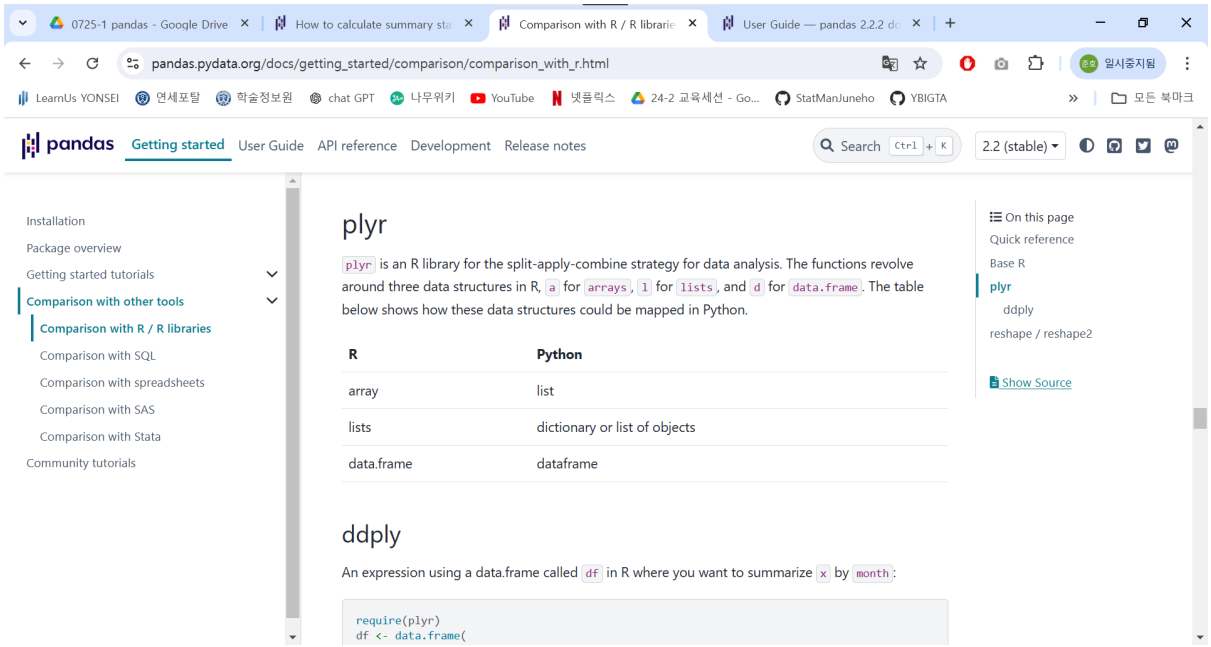
Report

김준호

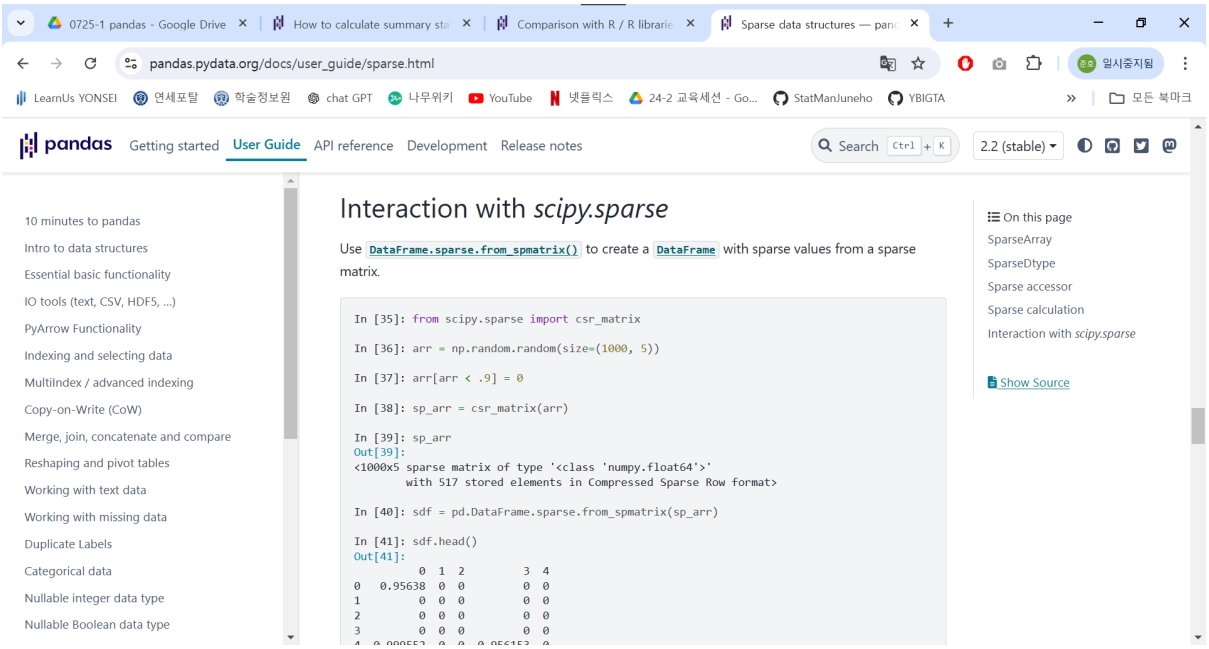
1번 과제: 간단한 감상

R과 pandas의 comparison에 대해 자세히 알 수 있어서 좋았다.

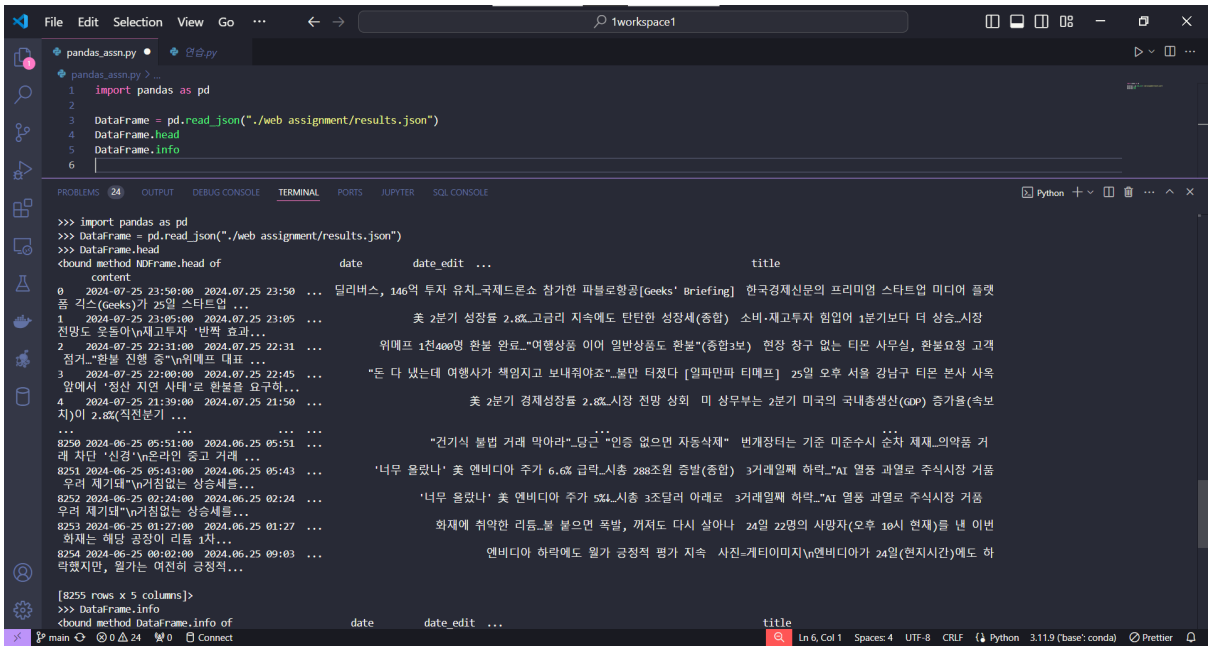




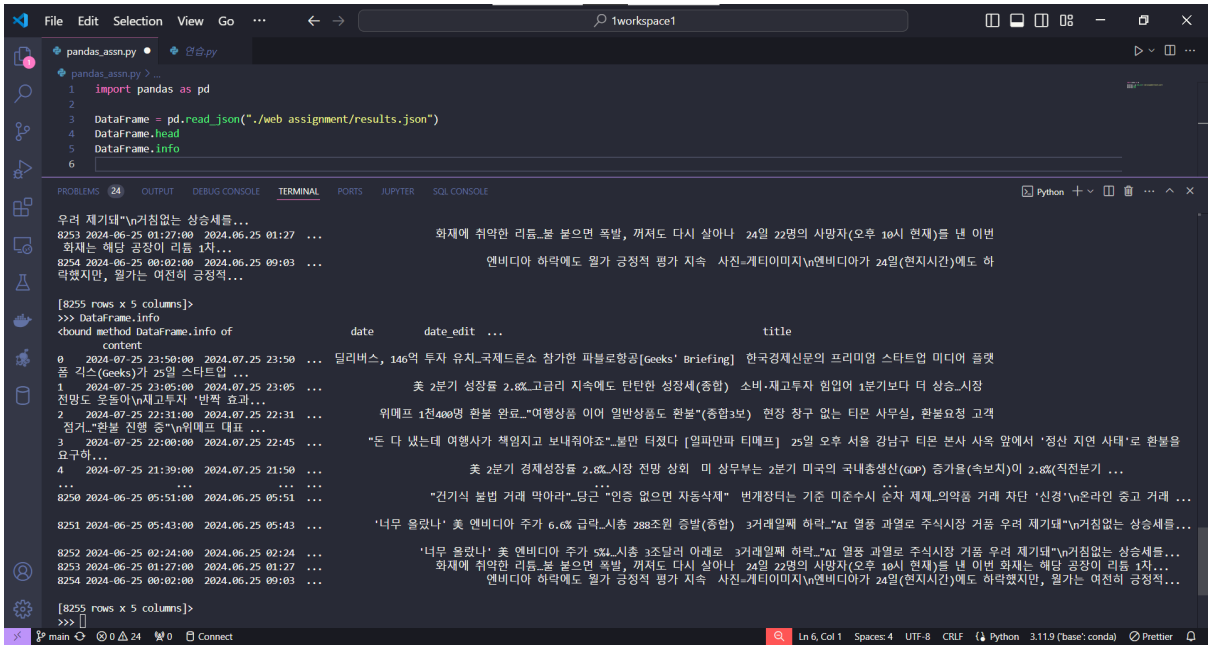
또한 평소 수업이든 연구든 Sparse한 Dataset을 이용해야 할 일이 많은데, Pandas를 통해 Sparse 한 Dataset을 만드는 방법에 대해서 배울 수 있어 좋았다.



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



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



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.