웹사이트 크롤링 프로젝트

소개

- 네이버 증권의 시가총액 페이지의 코스피, 코스닥 데이터를 크롤링하는 함수입니다.
- 데이터를 크롤링 후 csv파일로 데이터를 저장합니다.
- 최종적으로 크롤링 소요 시간과 파일 저장 성공 여부를 출력합니다.

입력

- 공지사항에 기재된 네이버 증권 URL과 원하는 페이지의 시작번호와 끝 번호를 입력 인자로 받습니다.
  - 첫번째 인자 `URL`: 네이버 증권 시가총액 페이지 "https://finance.naver.com/sise/sise_market_sum.naver?&page=1"
  - 두번째 인자 `PAGE_FROM` : 크롤링 할 페이지 시작 번호
  - 세번째 인자 `PAGE_TO` : 크롤링 할 페이지 마지막 번호

출력

- 크롤링한 네이버 증권 시가총액 페이지의 KOSPI, KOSDAQ 데이터를 각각 csv파일로 저장합니다.
- csv 파일명은 현재 날짜와 시간을 포함하며, 형식은 2024년 5월 12일 8시 5분일 경우 `202405120805_KOSPI.csv`, `202405120805_KODOQ.csv` 이 됩니다.
- 파일 저장후 크롤링 초 단위의 소요 시간과 파일 저장 여부를 출력합니다.

출력 예)

```
Crawling and saving completed successfully in 0.687723 seconds.
Success: Crawling and saving completed.
```

기타

처음에는 공지에 기재된 사이트 중 ***Reddit*** 으로 선정하여 포스팅 작성자, 제목, Upvote, 내용, 댓글을 크롤링하는 크롤러를 만들고자 하였으나 완성하지 못했습니다. 완성된 네이버 증권 크롤러와 함께 레딧 크롤러를 첨부합니다.

미완성 원인

1. 정적 콘텐츠만 가져올 수 있는 BeautifulSoup
    a. 클라이언트 사이드로 렌더링되는 부분(동적으로 로드되는 콘텐츠들)이 있기 때문에, 정적 웹페이지의 데이터만을 가져오는 BeautifulSoup만으로 필요한 데이터를 다 가져올 수 없었음
2. 중복되며 복잡한 클래스 이름
    a. *tailwind* 와 같이 클래스명으로 css를 조절하는 css library를 사용한 페이지이기 때문에 중복되는 클래스명, 장황한 클래스명들이 많아 클래스명만으로 필요한 요소를 가져와야하는 경우에 어려움이 있었음
3. 기타 이슈
    a. 비디오 콘텐츠 URL, 댓글 등 클라이언트 사이드로 렌더되는 요소들을 Selenium을 이용하여 가져오려고 시도 했으나 요소가 잡히지 않는 이슈, 또 댓글은 전체 댓글이 아닌 상위 몇개 댓글만 감지되는 이슈가 있었음

계획했던 코드 구조는 아래와 같습니다.

1. 상위 for문에서 순차적으로 모든 2023 best 포스팅의 작성자, 제목, 포스팅 url을 긁어온다.
2. 하위 for문에서 각 포스팅url로 페이지를 이동해 내용과 댓글을 가져온다.