web report

전체적인 구조

클롤링

css-selector에 익숙해서 크롬브라우저 개발자 도구에서 css-selector copy로 가져온 뒤 변형이 필요하면 변형을 했습니다. 그리고 한경페이지가 더보기를 누를 때마다 동적으로 페이지가 생성되는 데 이것이 셀레니움 코드와 동기화가 되지 않아 오류가 났었습니다. 찾아보니 selenium이 `WebDriverWait` 으로 이런 문제를 해결 할 수 있도록 지원하는 것을 알았고 해결했습니다.

예외처리

예외처리를 부족하게 했습니다. 우선 검색해본 selenium에서 많이들 발생하는 Exception과 제가 경험한 TimeoutException만 예외발생시 읽은 냉용까지 기록하도록 구현했습니다. bs4에서 파싱하다가 오류가 난 것을 어떻게 처리할 지 고민을 했습니다. bs4가 따로 세분화 되게 예외들을 나눈 것이 아니여서 어려움이 있었고 bs4에 대한 예외처리는 하지 않았습니다.

성능

생각보다 30일치 뉴스들이 엄청 많았고 selenium이 느려서 초기에 데이터를 다가져오는 데 한시간이 넘게 걸렸습니다. 성능이 중요한 것은 아니겠지만 너무 답답한 너머지손을 봤습니다. 우선 개별page들에 request를 날리는 것을 selenium에서 selenium의 의존 패키지중 하나인 urllib3로 구현했지만 이부분에서 가장 많은 병목이 있었습니

다. 하나의 스레드에서 응답이 느린 뉴스페이지로 요청을 날린 것이니 당연한 것 같습니다. 응답 시간이 길면 비동기로 처리해보는 게 어떨까 싶어서 비동기로 구현을 바꾸어줬더니 시간이 확실히 줄어들었습니다. 다만 개별 기사를 파싱하는 시간을 줄어들지 않았습니다.
그리고 selenium에 브라우저 탭을 띄우지 않는 옵션과 이미지 로딩을 방지하는 옵션으로 성능을 좀더 올려봤습니다.