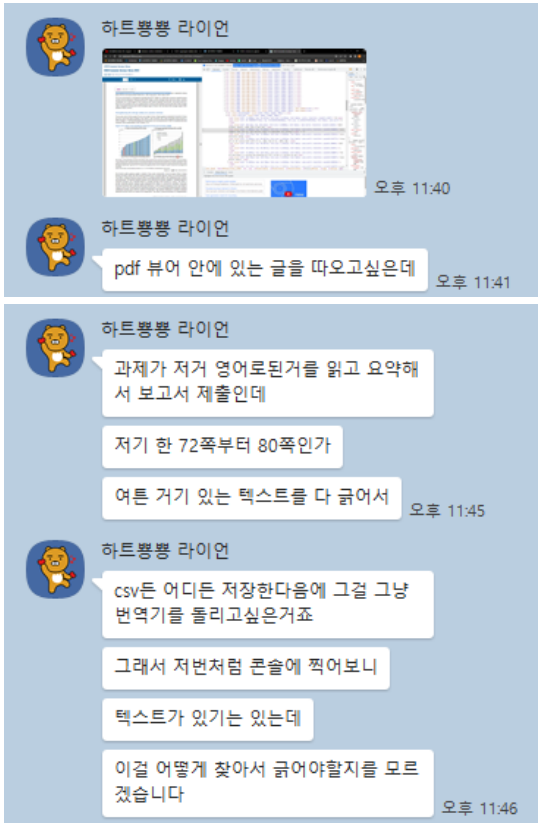
지난 학기 대학 동아리에서 파이썬 멘토링을 진행했었는데, 멘티분께 흥미로운 질문을 받았다.



PDF 뷰어로 보이는 텍스트를 추출해서 저장한 다음 번역기를 돌리고 싶은데 추출하는 방법을 모르시는 상황.
흥미가 생겨서 도와드렸다.
텍스트를 추출하고자 하는 페이지는 아래 링크해두었다.
https://read.oecd-ilibrary.org/economics/oecd-economic-surveys-korea-2022_20bf3d6e-en

40분 정도 삽질을 했다.
삽질 과정은 아래와 같다.

bs4, requests를 이용해서 긁어보자.
↓
get 요청을 하면 403 에러가 난다.
header를 추가해 보자.
↓
에러는 나지 않는데 원하는 페이지가 오지 않는다.
동적 페이지라서 그런 것 같으니 selenium을 써보자.
↓
될 것 같긴 한데 쉽게 안 된다.
그냥 개발자 도구 콘솔 창에 자바스크립트로 작성해서 손쉽게 해결하자.

지금부터는 자바스크립트로 상단 링크된 복사 안 되는 페이지의 문서 텍스트를 추출하는 방법에 대해 적어보려고 한다.
나는 크롬 브라우저를 이용했다.

1.
상단 링크에 들어가서 개발자 도구 실행 후 콘솔 창에 다음과 같이 입력한다.
iframe 내 document html에 한 번 포커싱이 되고 나서야 아래 코드가 정상 동작한다.
bs4, requests, selenium으로 페이지 안의 페이지를 가져오려면 기존과는 다른 방식을 써야겠다는 사실을 직감했다.
참고로 NodeList의 길이는 페이지 수와 일치한다.

```
pages = document.querySelectorAll('.page')

// 포커싱 전
// NodeList []length: 0[[Prototype]]: NodeList
// 포커싱 후
// NodeList(132) [div.page, div.page, div.page, div.page, div.page, div.page, div.page,
```

2.
i 번째 인덱스에 접근하여 텍스트만 저장한다.
i에 본인이 원하는 페이지 수 - 1을 입력해 주면 된다.
이것도 콘솔 창에서 실행하면 좀 까다롭다.
보고 있는 페이지에 따라 fragment가 바뀌는데, 이에 따라 innerText가 생겼다 말았다 한다.
queue에 innerText를 저장하다가 10개를 초과하면 가장 처음 인덱스 요소를 pop 시키고 지금 본 페이지를 저장하는 듯한 느낌.
반복문을 써도 되지만, 중간에 끊기는 건 싫으니 하나씩 저장하기로 했다.
보고 있는 페이지 innerText는 확실히 있는 것 같으니 한 페이지씩 내려가며 text에 저장해 주면 된다.

```
text = document.pages[i].innerText
// text += document.pages[i].innerText
```

3.
웹에서 Javascript 만으로 텍스트 파일을 생성할 수 있는 코드가 있다.
크롬에서 작동하는 코드를 찾아 text를 txt 파일로 저장해 준다.
윈도우 10을 사용하고 있는 나의 경우 text.txt 파일이 다운로드 폴더에 저장되었다.

```
// 코드 출처: http://www.gisdeveloper.co.kr/?p=5564
function saveToFile_Chrome(fileName, content) {
    var blob = new Blob([content], { type: 'text/plain' });
    objURL = window.URL.createObjectURL(blob);

    // 이전에 생성된 메모리 해제
    if (window.__Xr_objURL_forCreatingFile__) {
        window.URL.revokeObjectURL(window.__Xr_objURL_forCreatingFile__);
    }
    window.__Xr_objURL_forCreatingFile__ = objURL;
    var a = document.createElement('a');
    a.download = fileName;
    a.href = objURL;
    a.click();
}

saveToFile_Chrome('text.txt', text)
```

감사하게도 선물까지 받았다.
단순한 코드지만, 확실히 해당 사이트에서 텍스트만 추출해서 번역하고 싶을 때는 유용할 듯.
해당 사이트에서 구글 번역을 하면 문서 상으로 텍스트가 번역은 되는데, 화면 상에 반영이 안 된다.

다음번에는 python으로 iframe 크롤링/스크래핑 하는 방법에 대해 알아보아야겠다.