주제분석 1주차 패키지 '크롤링'

1. 설치

A. 주어진 자료를 활용하여 크롬 드라이버를 설치하세요

B. Selenium과 Beautiful Soup 라이브러리를 다운 받으세요

C. Pandas를 설치하세요

2. 크롤링 초기 작업

A. 네이버 영화를 크롤링할 계획입니다. 각 영화의 URL이 어떻게 구성되어 있고

어떻게 접근하면 좋을지 적어보세요!

B. 다양한 영화들을 검색해보면서, 영화들이 공통된 구조를 가지고 있는지 확인하세요

3. 크롤링 실행!

A. 오늘 크롤링할 데이터는 네이버 영화의 '현재 상영영화'들입니다. 이 영화들의 url를 추출해보세요
(hint : 현재 이 페이지는 java로 구성되어 있어, html을 한 번에 읽기 어렵습니다. Beautiful soup(정적)과 selenium(동적) 중 어떤 것을 써야 할지 생각해보세요!)

B. A에서 추출한 url를 이용하여 영화의 데이터를 추출해보세요. 추출해야 할 데이터는

'영화 한국 이름', '영화 영어 이름', '관람객 평점', '관람객 평점 참여 수', '네티즌 평점', '네티즌 평점 참여 수', '장르', '국가', '개봉시간', '감독', '출연 배우', '상영 등급' 입니다.

C. B에서 추출한 데이터를 데이터프레임 형태로 저장하세요!

(Hint: dictionary와 list를 이용해서 데이터를 저장하면 데이터프레임으로 깔끔하게 만들 수 있습니다.)

D. 저장된 데이터를 실제로 엑셀을 이용하여, 오류가 있는지 확인해보세요. 오류가 있다면 왜 오류가 있었는지 간단하게 피드백하시면 됩니다.
(5~10% 정도의 오류만 나오게 코드를 구성하시면 됩니다)

(첨부된 파일과 비슷하게 나오면 괜찮습니다)

* 이번 패키지는 크롤링을 해보면서 PYTHON과 익숙해지는 것에 의의가 있습니다 :)

* 최대한 크롤링이 돌아가는 for loop에서 전처리가 알아서 이루어지게 구현해주세요 :)

* 정답이 없는 문제입니다. 코드 구성과 결과물로 채점하도록 하겠습니다!