대본

안녕하십니까, 크롤링 프로젝트를 진행한 2조 김연주, 장예서, 홍지우, 이윤경, 백채원 입니다.

크롤링 과정을 먼저 설명드리겠습니다.

크롤링을 위해 필요한 라이브러리들을 import해줍니다.

크롬 웹드라이버 설정을 하고 / 크롤링 조건에 맞게 아우터 중 자켓에서, 리뷰 많은 순대로 링크를 통해 해당 사이트로 이동합니다.

이 브라우저에서 html 소스를 가져온 후 / 뷰티플숲을 이용해 파싱을 해줍니다.

클래스 이름을 이용해 상품 정보가 포함된 영역을 선택합니다. (지그재그_프로덕트_에리어)

그리고 상품 정보와 리뷰를 저장할 result, result_review 리스트를 각각 생성합니다.

for문을 통해 상품 정보에서 상위 30개의 상품을 수집합니다.

상품 정보 수집 내용입니다.

인덱싱을 통해 현재 상품을 a에 저장하고 / a에서 리뷰 평점, 리뷰 개수의 텍스트를 각각 리뷰, 리뷰_넘 변수에 저장합니다.

리뷰 개수는 숫자 정보만 사용합니다.(*리뷰_넘이 숫자(형태)인지 판단하고 숫자가 아닌 부분 빼고 다시 합치는 코드가 있다. (1,234)를 1234로 저장하는 코드.*)

상품 제목 텍스트를 타이틀 변수에 저장하고 / 썸네일 이미지의 주소를 ‘썸네일’ 변수에 저장합니다.

가격과 할인율도 같은 방법으로 가져오는데 할인 정보가 표시되지 않는 경우에는 ‘할인 정보 없음’이라는 텍스트를 넣었습니다.

그리고 [상품 제목, 가격, 할인율, 리뷰 평점, 리뷰 개수, 썸네일] 순서대로 result 리스트에 추가합니다.

상품 리뷰 수집 내용입니다.

각 상품의 상세페이지로 이동할 수 있는 위치를 찾아서 클릭해줍니다.

상품 상세페이지로 이동한 후, 상세페이지의 링크 내용을 가져와서, 처음 했던 것처럼 html 소스를 가져오고 뷰티플숲으로 파싱합니다.

리뷰 창에 있는 리뷰들을 지그재그_리뷰 변수와 지그재그_리뷰_텍스트 변수에 저장했습니다. ()

여기서 for문으로 5번씩 반복하면서, 지그재그_리뷰변수에서 리뷰어 이름, 리뷰 날짜 텍스트 정보를 각각 리뷰_네임, 리뷰_데이트 변수에 저장했습니다.

다만 리뷰글은 리뷰 창의 정보에서 크롤링하면 긴 내용이 더보기…로 잘리기 때문에 지그재그_리뷰_텍스트 변수 위치에서 내용을 가져와 리뷰_텍스트 변수에 저장했습니다.

이렇게 [리뷰어, 리뷰 날짜, 리뷰 텍스트] 순서대로 result_review 리스트에 추가해줍니다.

그리고 다시 처음 상품 목록 페이지로 이동하면서 html 소스를 가져오고, 뷰티플숲 파싱을 합니다.

이렇게 30번 반복하면서, 상품 목록 페이지와 각 상품 페이지를 이동하며, 상품 정보와 리뷰 정보를 가져오게 됩니다.

모든 과정이 끝나면 result, result_review 정보에 열 이름을 추가해서 csv 형태로 저장합니다.

(시연?)

```
browser = webdriver.Chrome("chromedriver.exe")
```

크롤링 결과물은 아래와 같습니다.

상품 정보는 이렇게 조건에 맞게 30개 들어 있는 것을 확인하실 수 있고,

리뷰도 30개 상품에 대해 리뷰 5개씩, 총 150개가 들어갑니다.

전체 코드 내용입니다.