**크롤러 작동 방식**

1. URL 불러오기

* 엑셀 파일에서 A.지역 이름 + B.카페 이름 + C.상호명으로 검색

1. Webdriver 실행 – 현재 설치 안돼 있을 시 설치
2. 리뷰 전체 보기 버튼 element 찾기 -> 클릭
3. 가장 최근에 달린 순서대로 정렬

* **최신 순으로 정렬 시**

장점: 카페의 더 빠른 정보들을 접할 수 있음

단점: 리뷰의 질보다 작성 순에 초점을 맞춰 아예 내용이 없는 것도 많이 크롤링됨

**관련성 순으로 정렬 시**

장점: 장문의 디테일한 부분 내용이 많이 담겨있는 리뷰일수록 위에 있어서 좀 더 실속 있는 내용들로 크롤링이 가능함

단점: 3,4년 전 리뷰도 포함되어있음

Q. 최신 순으로 정렬과 관련성 순으로 정렬 중 어떤 것 사용이 좋을까요?

1. 리뷰들의 내용을 불러올 수 있게 아래로 스크롤 하는 동작을 19번 반복(한 번 내릴 때마다 리뷰 10개씩 로드됨 기본 리뷰 10개 + 스크롤로 190개 추가)

* 현재 카페 1곳에서 200개의 리뷰를 크롤링하는데 50~60초 가량 소요됨(headless 사용 안한 기준, 200개보다 리뷰가 적으면 그만큼 빨리 끝남)

스크롤을 내리고 로딩을 위해 time.sleep 명령어로 쉬어가는 것 때문에 현재 크롤링 하는 과정에서 가장 많은 시간을 소요시키고 있음.

좀 더 효율적인 방법이 있는지 알아볼 계획입니다.

1. 별점과 리뷰, 작성 기간의 element를 불러오고 for문을 이용해 .text 명령어를 이용해 텍스트화하고 append하고 pandas를 이용해 dataframe화해 csv 파일로 저장함

해결된 부분: 장문의 리뷰 더 보기를 클릭해야 모든 내용을 확인할 수 있음.

* 크롤링 과정에 if문으로 ’자세히’ 라는 이름의 버튼이 있으면 클릭하게 선택