

|  |
| --- |
| 중간 보고서 |
| 통합 리뷰를 기반으로 한 제품 안내 어플  Vol.1 |

클립아트이(가) 표시된 사진

자동 생성된 설명

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 제출일 | 2020. 01, 24 |  | 전공 | 컴퓨터공학과 |
| 과목 | 졸업작품 프로젝트 |  | 학번 | 2015722084  2015722083 |
| 담당교수 | 이기훈 |  | 이름 | 한승주  김성종 |

**목 차**

|  |  |
| --- | --- |
|  |  |
| Ⅰ 개요 |  |
| 1. 배경 및 필요성 |  |
| 2. 목적 |  |
| 3. 설계 내용 |  |
| 1. Flow Chart |  |
| 1. 개념 설계 |  |
|  |  |
| Ⅱ 과제 수행 |  |
| 1. 수행 일정 |  |
| 2. 웹크롤링 |  |
| 3. DB구축 |  |
| 4. 단어 사용 빈도 추출 |  |
| 5. UI 제작 |  |
|  |  |
| Ⅲ 과제 평가 |  |
| 1. 개선방안 |  |
| 2. 기대효과 |  |
| 1. 기업적 측면 |  |
| 1. 사용자 측면 |  |
|  |  |
| **< 참고문헌 >** |  |
|  |  |

1. **개요**
2. 배경 및 필요성

* 서로 다른 플랫폼과 웹사이트는 사용하는 연령대나, 나이 대 등이 달라 제품의 정보를 단 하나의 사이트만 보고는 신뢰성 있는 정보를 통한 구매가 불가능하다.
* 실제 국내 페이스북 유저 비율은 남성이 여성 대비 14% 많고, 인스타그램은 여성이 남성 대비 4% 많은 비율을 가지고 있다. 또한 페이스북은 연령대가 고른 반면, 인스타그램은 20~30대 비율이 상대적으로 높다.
* 페이스북, 인스타그램, 트위터와 같은 소셜미디어에 자주 노출되는 광고와 함께 실제 사용자가 올리는 리뷰, 그리고 제품 구매처인 쇼핑몰(G마켓, 옥션 등), 제품 가격 비교 사이트(다나와, 네이버 비교 쇼핑 등)은 여러 존재하지만, 이렇게 많은 웹사이트와 플랫폼에 퍼져 있는 사용자의 구매 후기를 통합적으로 파악할 수 있는 곳은 없다.

1. 목적

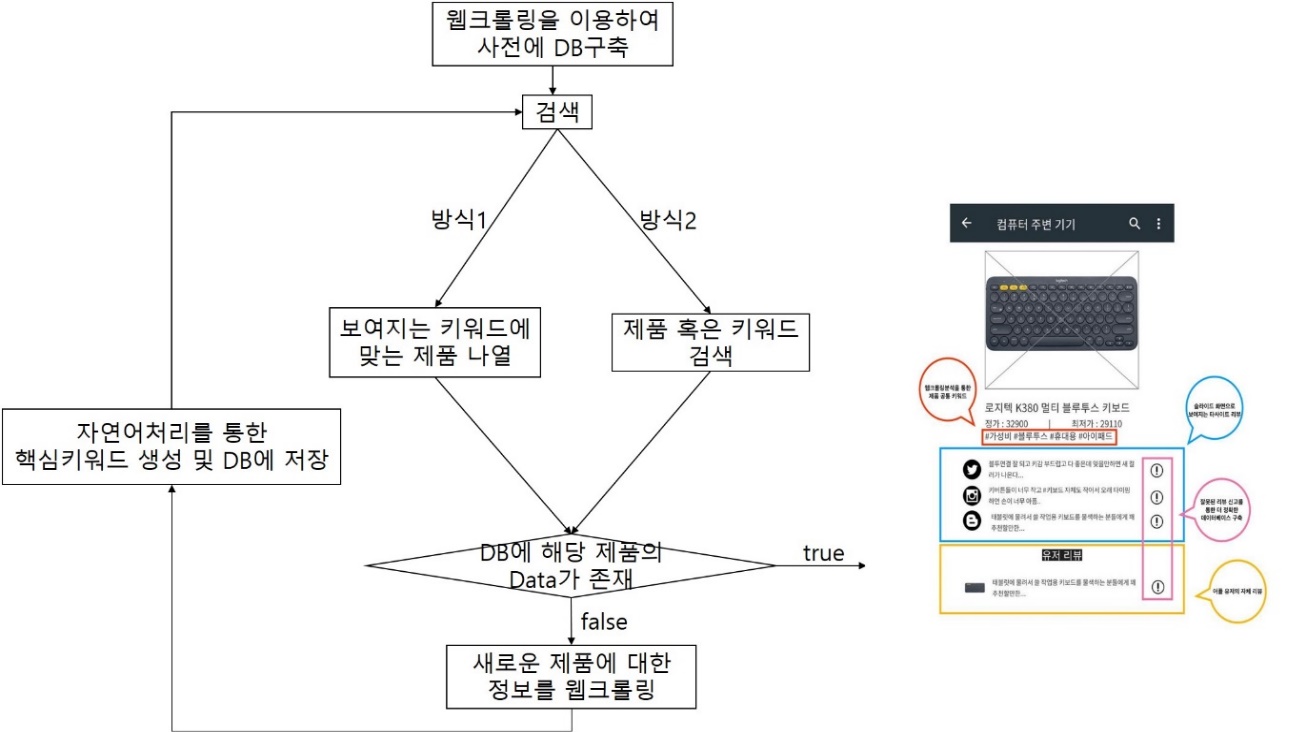
* 정보의 신뢰성

여러 플랫폼에 퍼져 있는 제품의 리뷰를 중요 키워드를 중심으로 데이터베이스를 구축하면 이렇게 만들어진 데이터베이스를 통해 키워드만으로 사용자가 원하는 조건의 제품을 소개하고 찾아볼 수 있다. 긍정적 리뷰와 부정적 리뷰를 다 같이 한번에 보여줌으로써 제품의 장점뿐만 아니라 단점을 파악하고 같은 제품군을 비교 구매하고 싶어 하는 사용자에게 편의성을 제공할 수 있다.

* 접근의 용이성

이런 결과를 어플리케이션(혹은 웹사이트)으로 추가적인 구현을 하면 사용자가 언제 어디서든 쉽게 접근하고 사용할 수 있다.

1. 설계내용
2. Flow Chart



1. 개념 설계

**1) 플랫폼의 웹사이트 코드 분석 및 크롤링(Beautiful soup)**

- DB를 구축하기 위해 해당 웹사이트 접속

- 제품 정보와 사용자 및 구매자의 리뷰가 담긴 웹사이트 코드 분석

- BS4를 이용해 페이지 데이터 호출

- 제품의 기본 정보(이름, 가격, 리뷰 등) tag를 찾아 추출

- 각 사이트와 페이지별로 링크를 재귀적으로 검색하여 데이터를 추출

**2) 크롤링된 정보를 이용한 DB구축(Mysql)**

**-** MySQL서버에 접속하여 데이터베이스 생성

- cursor를 추출하여 execute 메서드로 SQL을 실행, 테이블 생성

- Execute 메서드에 데이터를 계속 확장

**3) Kkma, Hannanum을 이용한 KoNLP(키워드 생성)**

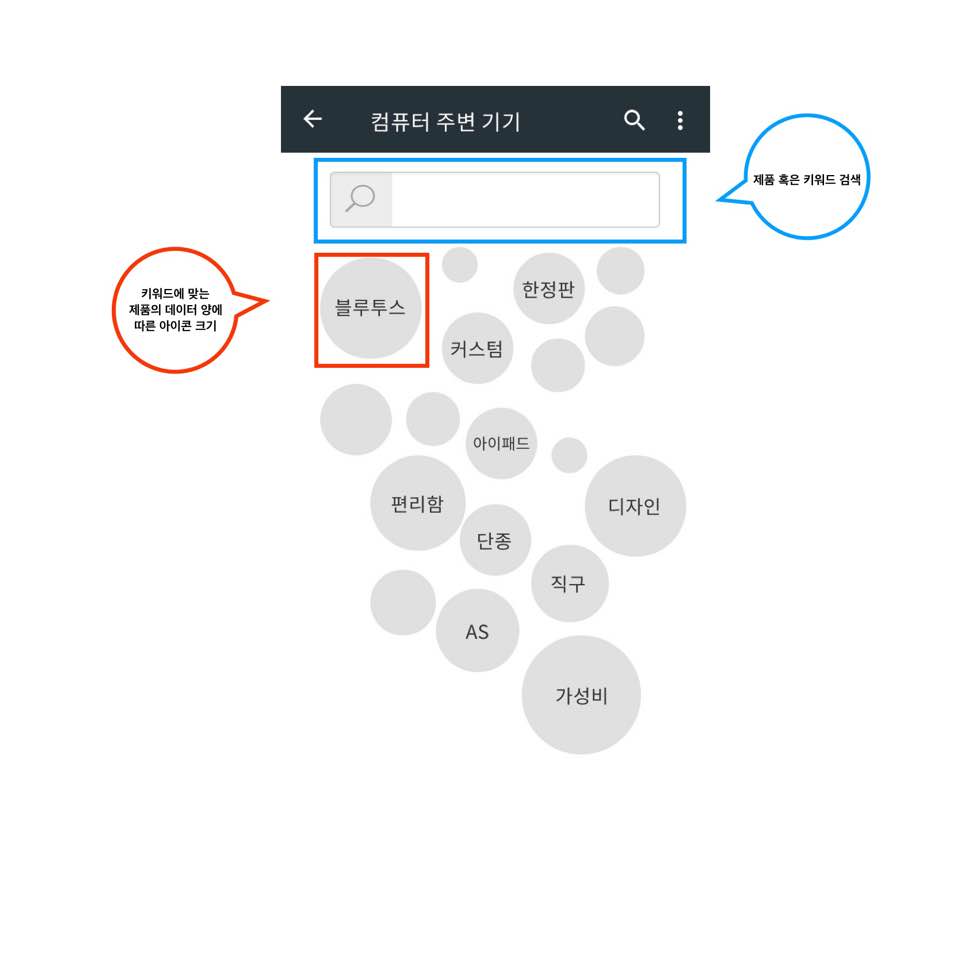
- Kkma나 Hannanum 모듈을 이용하여, 해당 모듈에 맞추어 입력된 문자열에서 키워드로 표현할 품사 추출

- 가장 빈도수가 높은 단어(키워드로 설정할 단어)를 DB에 저장

**4) Android UI제작**

**-** 자신이 구매하고자 하는 제품의 리뷰를 보기 위한 제품의 검색창과 위에서 나타낸 키워드 중 전체 제품에서 가장 많은 비중을 차지하는 몇 개의 키워드를 다음과 같이 제품 검색창 아래에 사용자가 보기 편하도록 UI로 구현

- 검색 시, 제품 검색 및 키워드(조건)를 검색할 수 있게 구현

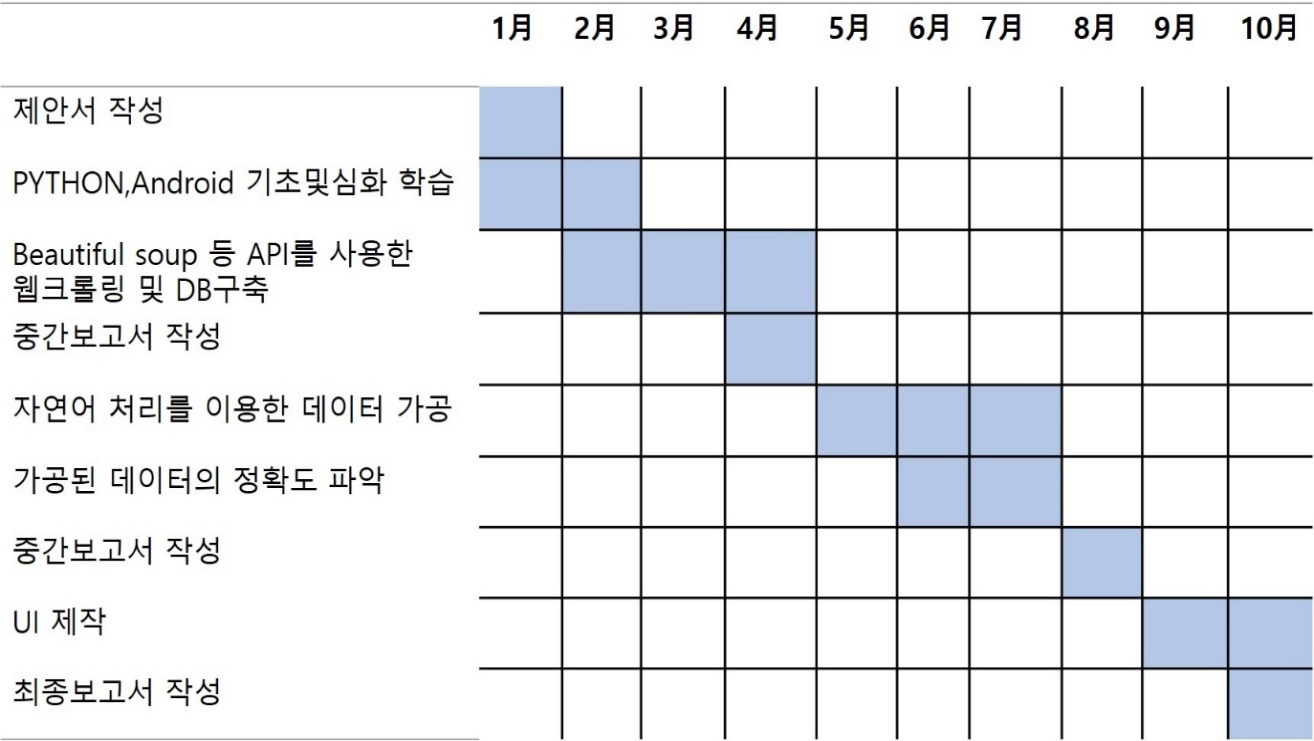


- 제품 검색 시, 제품의 가격 정보와 리뷰의 중요 키워드를 노출

- 실제 리뷰를 사이트 별로 노출

- 어플 자체 리뷰를 남겨 사용자가 어플을 더 다양하게 사용하도록 유도

1. **과제 수행**
2. 수행 일정



1. 웹 크롤링
2. 트위터

트위터의 정식 API인 tweepy가 존재는 하나 단점이 최근 7일간의 트윗만을 가져올 수 있기 때문에, 리뷰분석을 하는데에 있어서 한계가 존재한다. 그래서 사용한 API는 getoldtweet3(<https://github.com/Jefferson-Henrique/GetOldTweets-python)이며>, 해당 API는 기간을 설정하여 트윗을 수집할 수 있다. 기간은 2019년 1년동안 ‘로지텍’이라는 검색어가 들어간 트윗을 수집해보았다.

가져온 트윗에서 사용할 정보는 유저, 트윗 개시 날짜, 트윗 링크, 내용이며 해당 내용을 csv 파일로 저장한다.

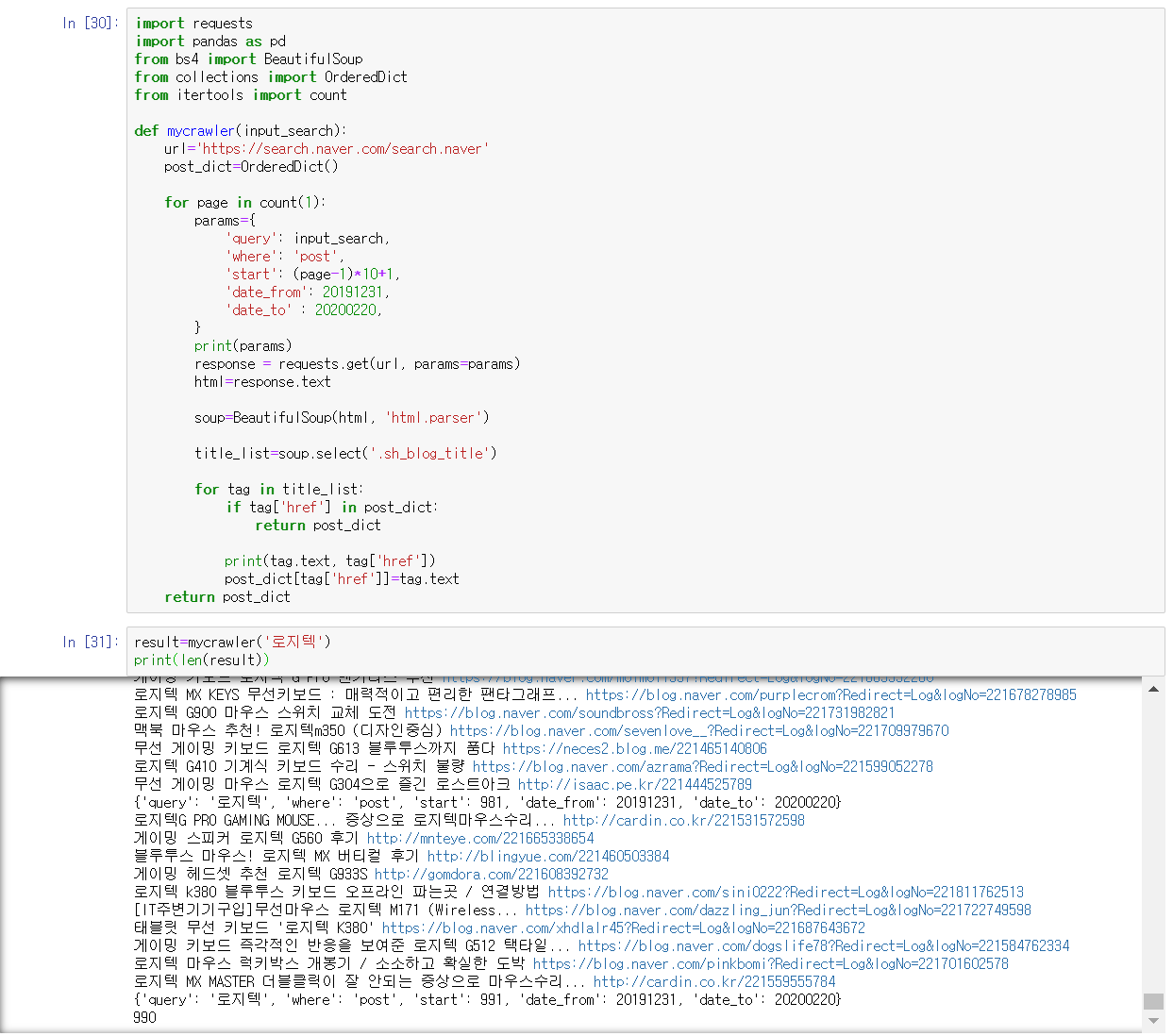
저장한 파일을 확인하면



이러한 결과를 얻을 수 있다.

1. 네이버

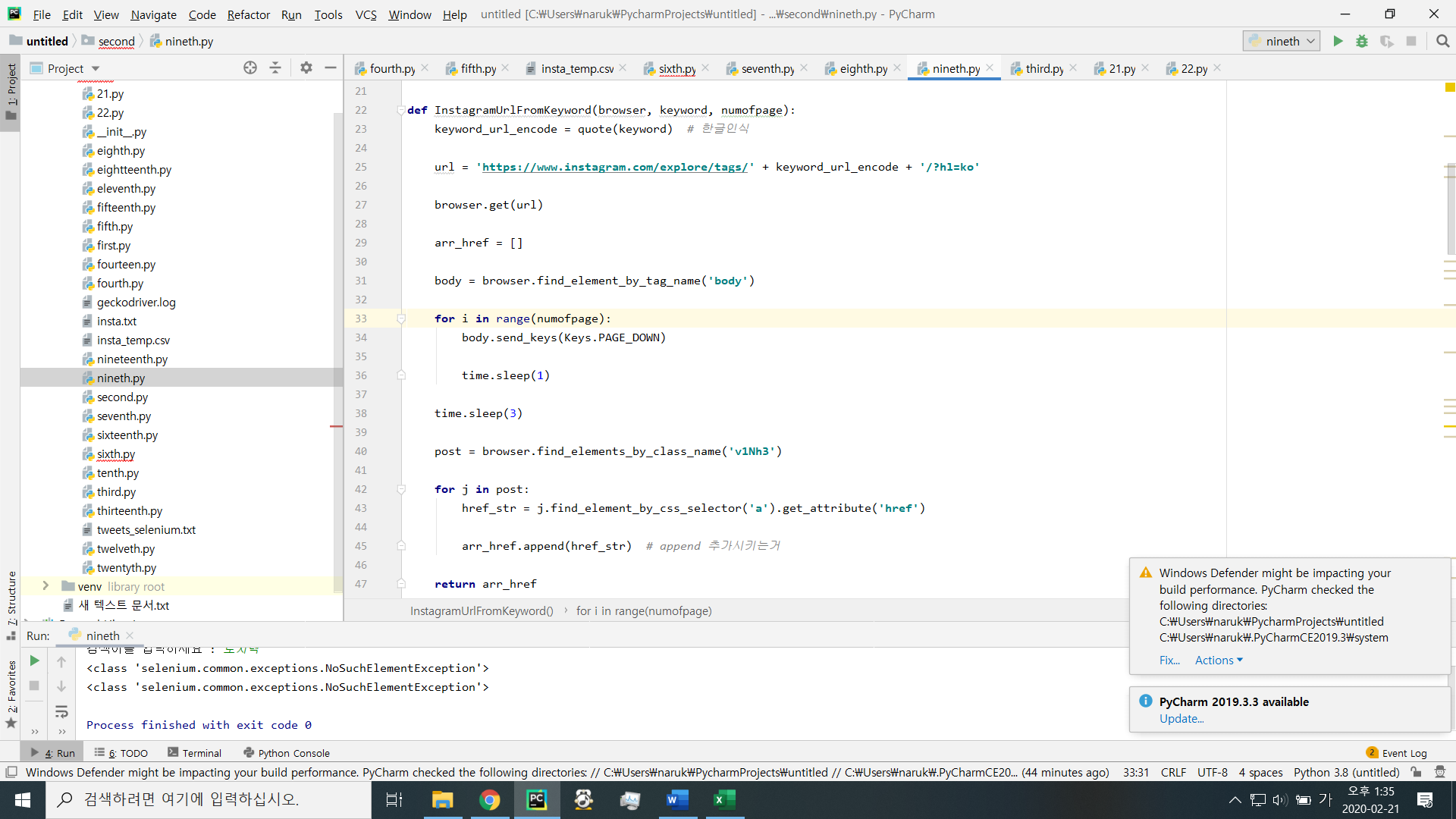
실제로 많은 사용자들이 많이 보는 리뷰는 네이버일 것이다. 따라서 네이버 블로그의 검색어에 따른 결과 크롤링을 해보면 제목과 블로그 링크를 가져올 수 있다. 무한정으로 많은 양의 검색결과를 가져올 수 없는데 이는 좋은 검색결과를 위해 네이버가 1000건의 검색결과만을 보여주고 있기 때문이다. 하지만 이것만으로도 충분히 키워드를 추출한다던지, 제품 한 개를 분석함에 있어서 부족함이 보이지는 않는다.



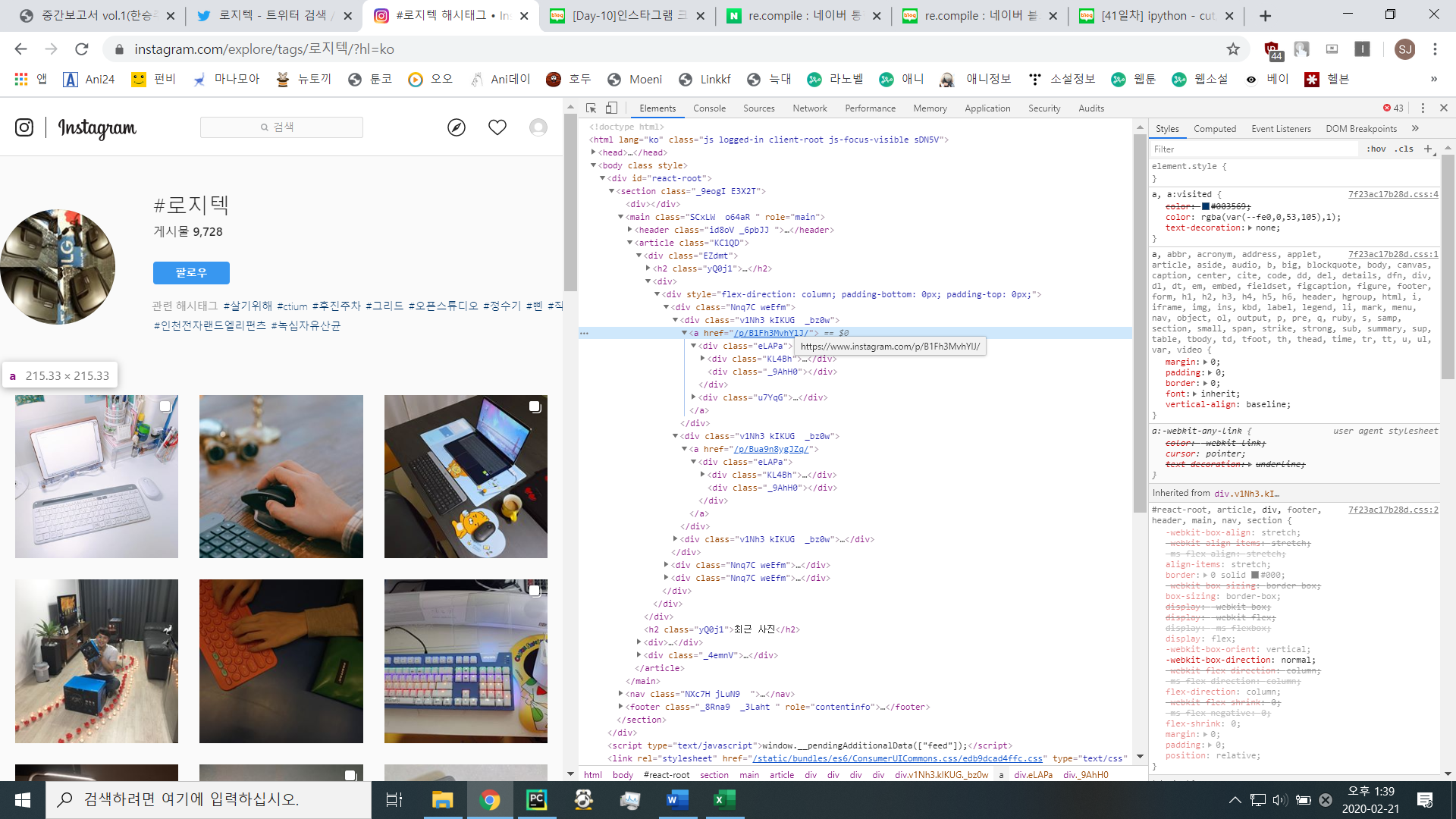
위는 글에서 글의 제목과 링크 주소를 가져온 스크래핑의 결과인데, 해당 링크를 타서 본문을 가져올 수 있지만, 해당 기능은 다음인 인스타그램에서 구현해보고자 한다.

1. 인스타그램

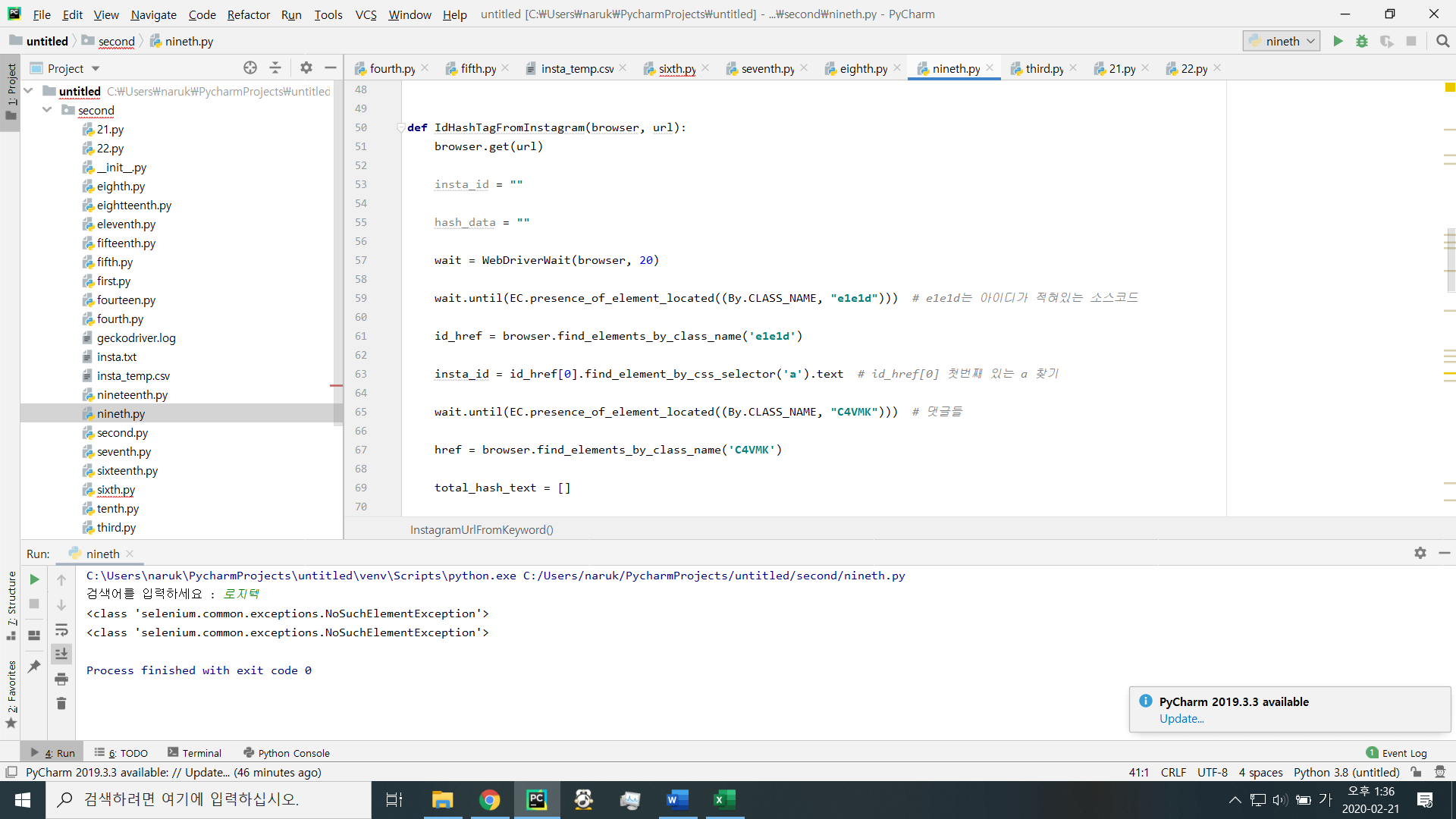
네이버 블로그가 긴 리뷰, 트위터가 짧은 리뷰라고 하면 인스타그램은 해쉬태그라는 자신만의 키워드를 표현하는 기능이 있지만, 실제로 분석함에 있어서는 해쉬태그를 따로 볼 것이 아니라 전체적인 리뷰로 보고 트위터와 유사한 짧은 리뷰라 판단하여 전체적인 분석을 하고자 한다.

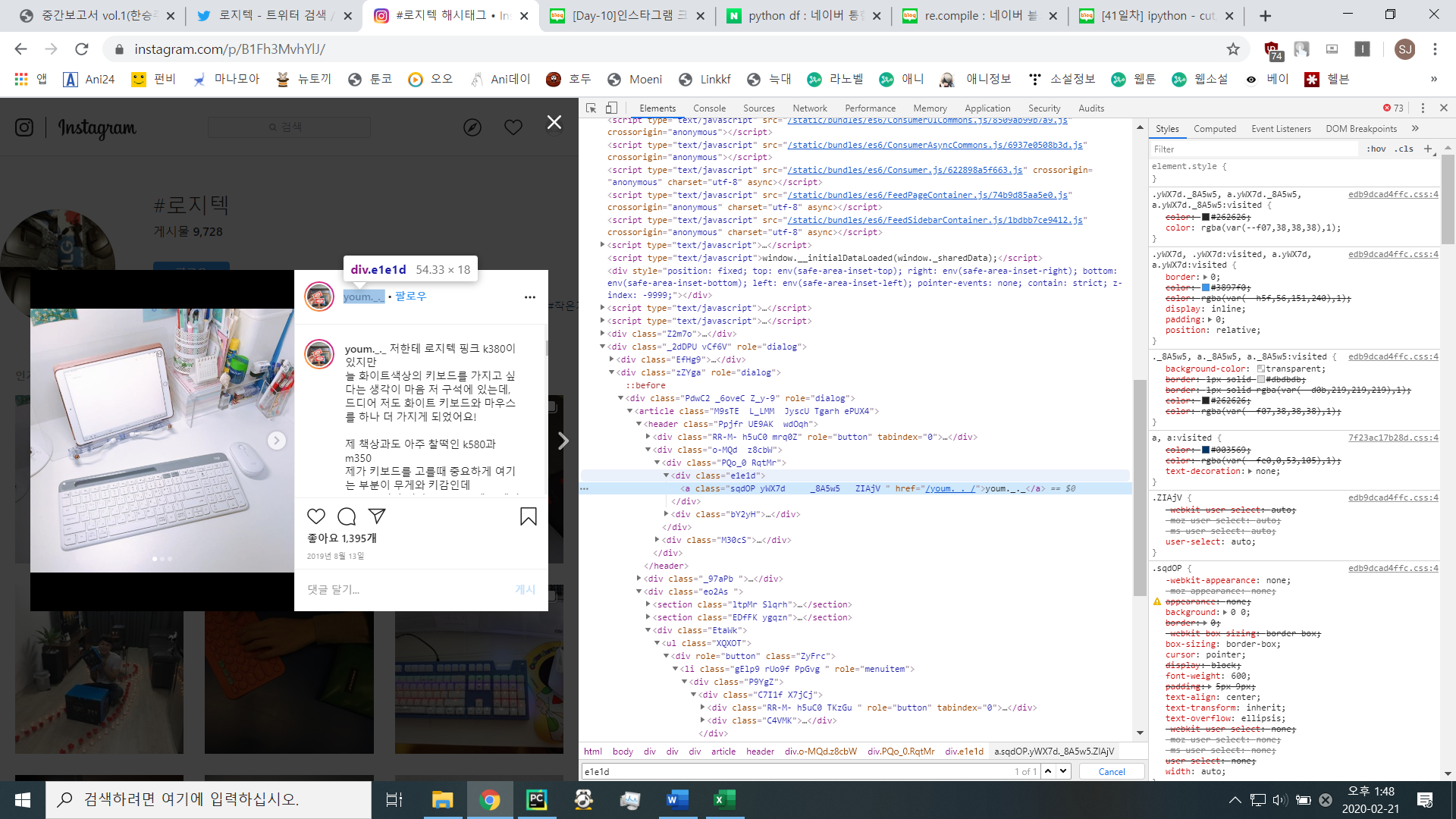


인스타그램에서 어떤 키워드로 검색어를 하면 해당 키워드인 한글이나 영어는 컴퓨터가 해석하기 난해하므로, 인코딩을 통해 해당 검색어에 맞추어 주소변환이 가능하다. 또한,

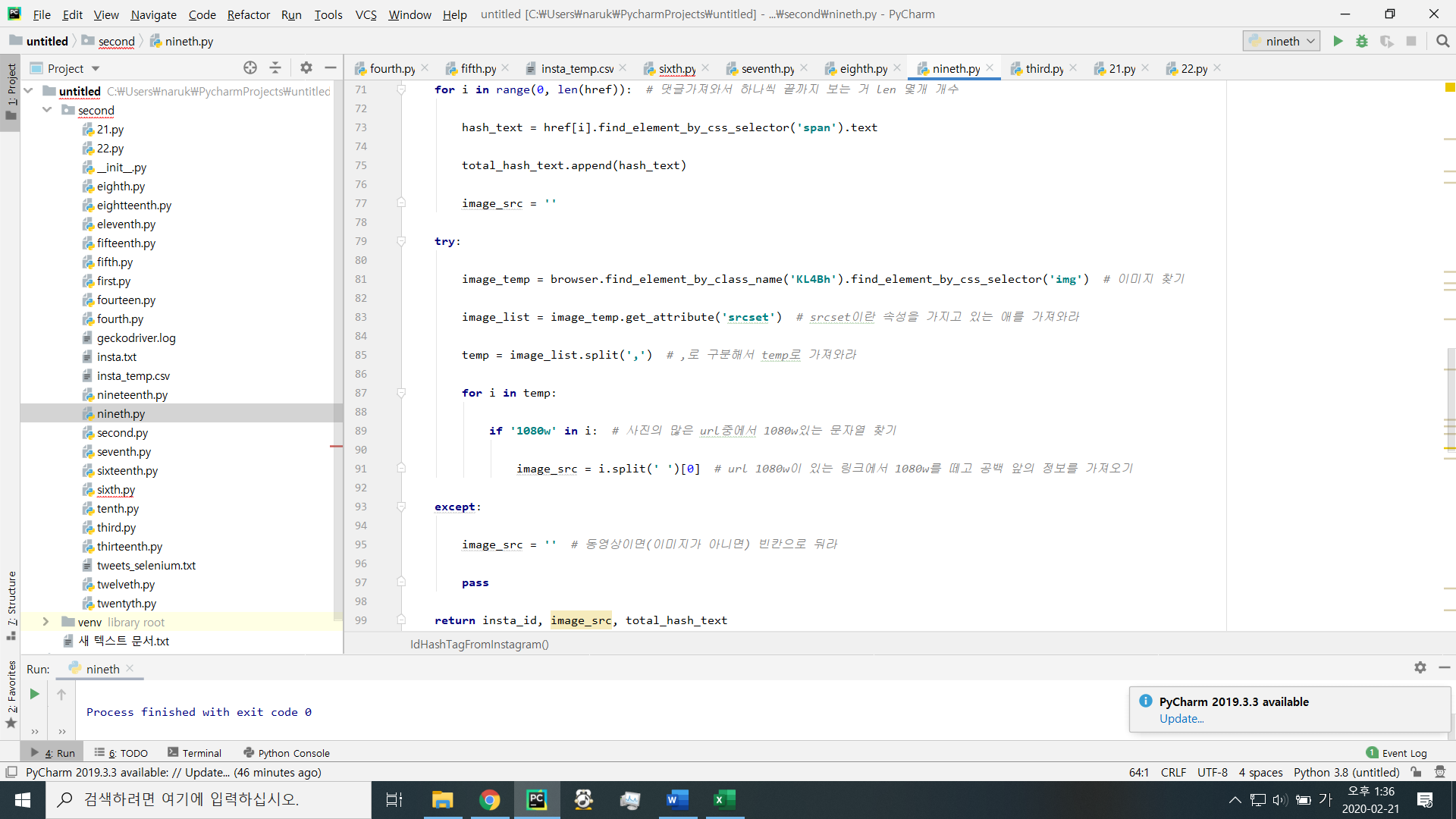


여러 글의 본문을 찾기 위해서 해당 함수를 사용하게 되는데, 인스타그램 웹의 소스코드에는 v1Nh3 class 밑에 href를 통해 모든 글의 주소를 가져올 수 있어서 해당 글의 ref를 가져오기 위한 정의이다.

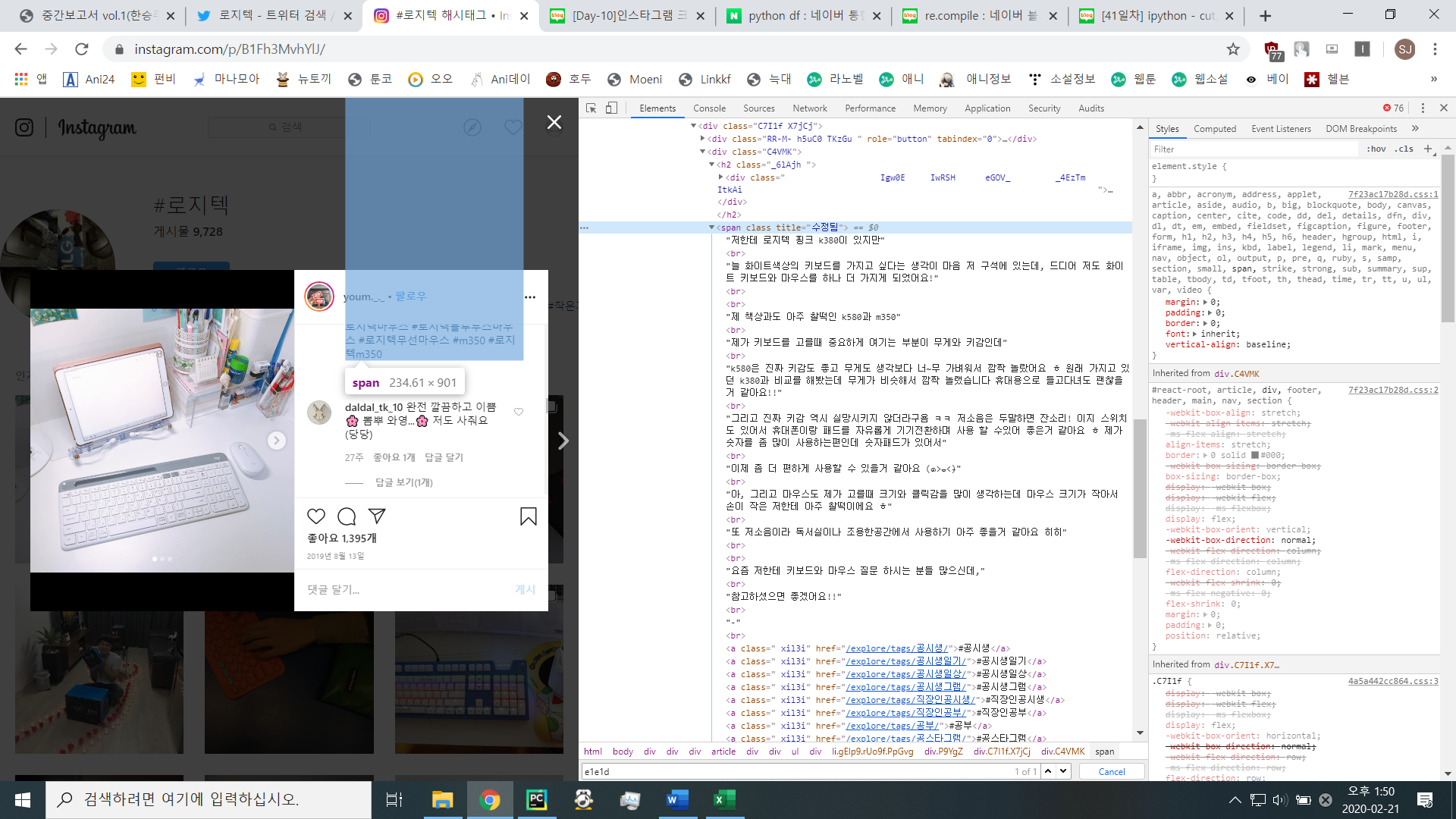




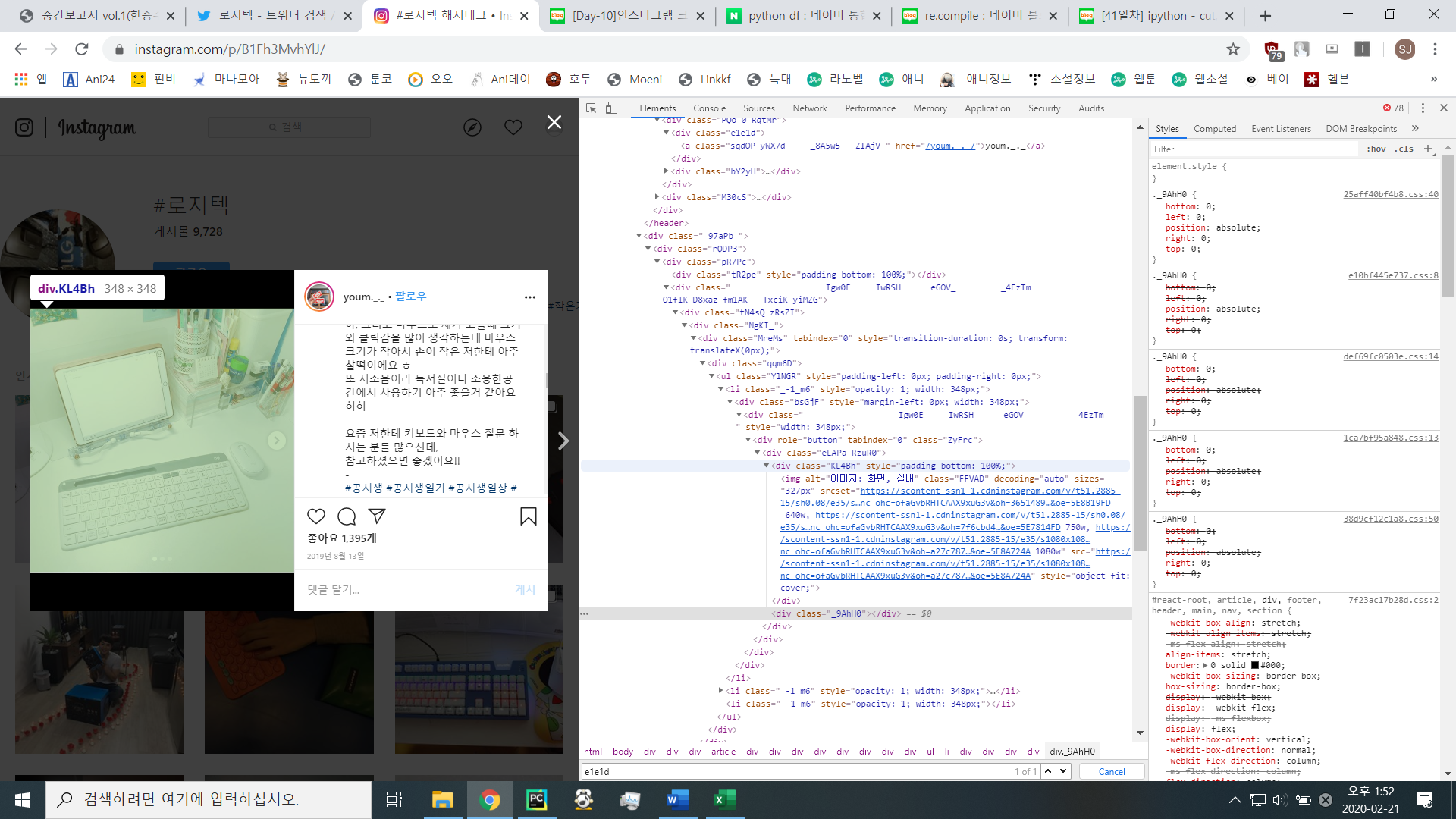
아이디를 찾는 부분으로, eleld 클래스 영역에 href로 text만 뽑아내면 해당 부분이 인스타그램 id이므로 이를 추출한다.



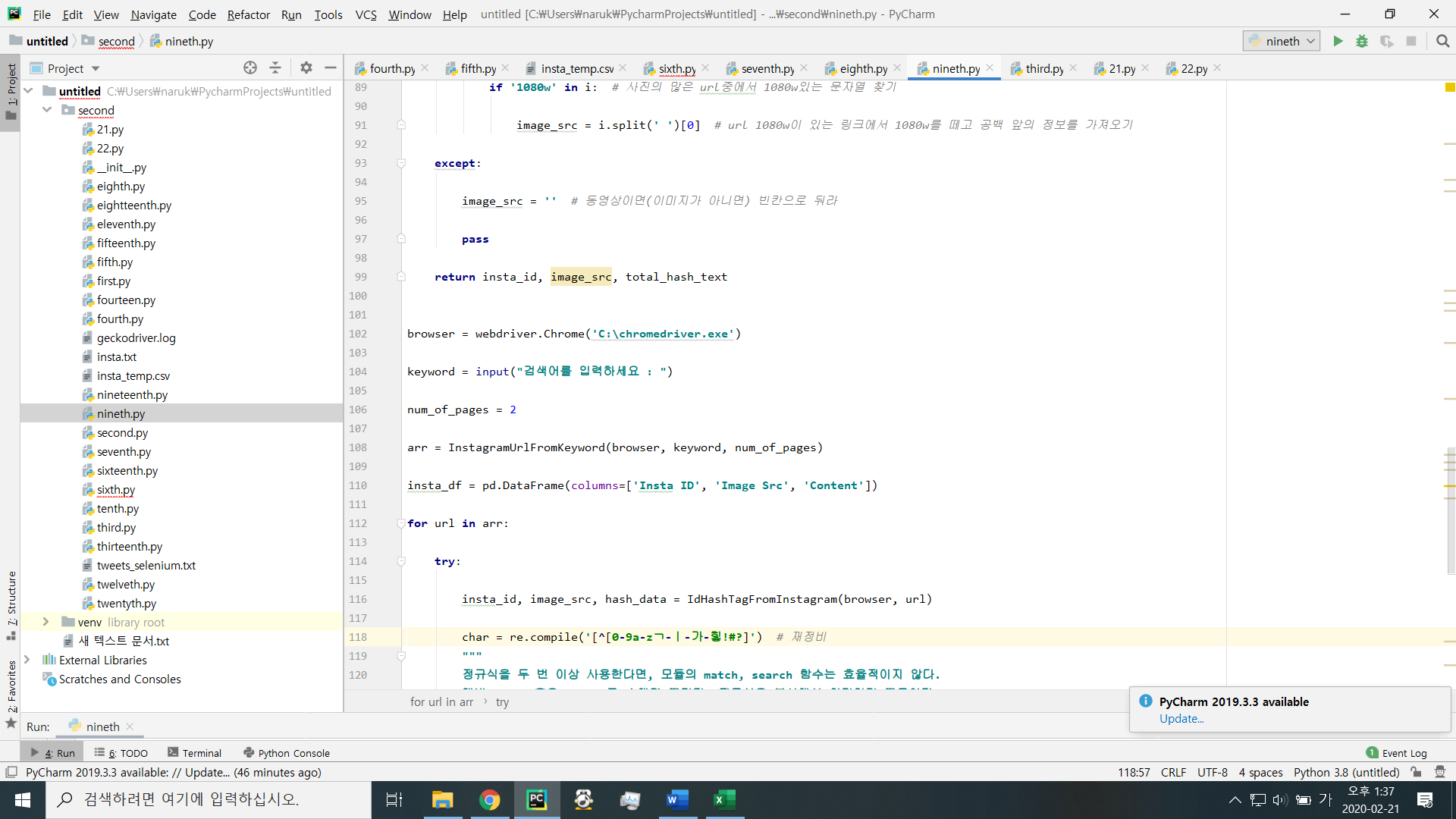
해당 코드는 for문을 이용하여 참조할 reference가 있는 동안 해당 ref의 본문을 축적하는 부분과 본문에 덧붙인 이미지의 src를 찾는 부분으로 구성이 되어있다.

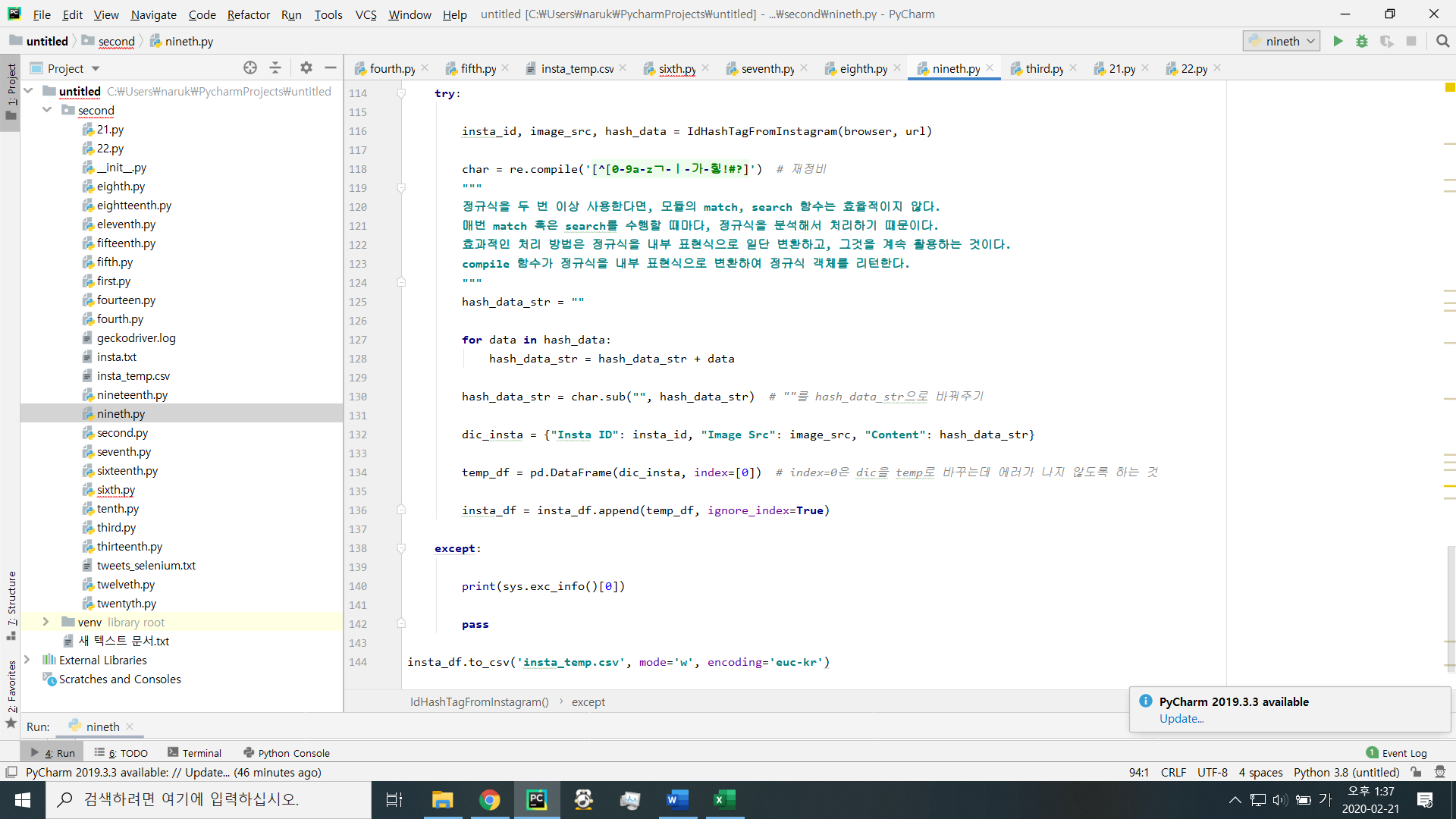


본문의 부분으로 댓글또한 위와같이 span의 영역에 글이 작성되어 있다.



Img\_src또한 KL4Bh의 영역에 img를 find하여 찾을 수 있다.





실제로 동작하는 부분은 검색어를 입력받아서 위의 두가지 함수를 활용하여 인스타그램의 주소를 가져와서 해당 주소의 해쉬태그를 데이터프레임에 쌓고 이를 csv로 저장하는 부분이다.



라는 결과를 얻을 수 있다. 여기서, 활용한 것은 저번 2주차 동안의 목표였던 링크를 타서 본문을 가져오는 것과, 검색어를 입력하여 해당 검색어에 맞는 리뷰를 가져오는 것을 사용해보았으며, 추가적으로 csv파일로 저장하는 것까지 구현해보았다.

해당 주차를 크롤링하면서 생각한 부분은 페이스북은 리뷰를 분석하기에는 자료가 너무 적어서 의미가 없다는 점과 크롤링된 결과에서 필요한 키워드를 뽑아내는 것을 연구해야 할 것 같다. 따라서 자연어처리를 실습하는 것을 계획하고 있으며, 어떠한 DB를 사용할 지까지 고민해보고자 한다. 또한, 최근 1년 사이의 리뷰를 확인 기간을 정하고자 하는데 이는 리뉴얼된 제품 즉, 최신의 정보도 업데이트가 가능할 것이라고 판단하여 해당 기간으로 리뷰를 분석하고자 한다. 해당 기간 설정이 필요한 이유는 리뉴얼된 제품의 경우, 과거의 제품과는 다른 제품력을 보여주기 때문이다.

1. DB 구축

(진행도가 있을 때 추가 작성예정)

1. 단어 사용 빈도 추출

(진행도가 있을 때 추가 작성예정)

1. UI 제작

(진행도가 있을 때 추가 작성예정)

1. **과제 평가**
2. 개선방안

* Request를 사용했을 때보다 selenium을 사용했을 때 동작에 있어 시간이 다소 오래걸리는 것을 확인했는데, html을 파싱하는데에 있어서 다소 문제가 있어서 selenium을 사용했던 것인데 이 파싱하는 부분을 바꿔서 코드를 수정해보고자 한다. 또한, 딕셔너리로 return을 받을 때 나중에 데이터를 사용할때는 반점과 괄호로 구분이 되어서 크게 상관은 없지만, 직접적인 view에 있어서는 다소 불편함이 있어서 중간중간 ‘\n’가 삽입이 가능한지 확인해보고자 한다.
* 이번 같은 경우에는 상품정보는 외국 몰에서, 리뷰는 영화로 대체를 했는데, 국내 쇼핑몰의 경우 오른쪽클릭이 안되어서 html소스를 못보는줄 알았는데, F12를 사용해서 할 수 있는 것을 확인하고 추가로 해보고자 한다. 특히, 이 경우 위에서처럼 제품정보를 얻으면서 링크를 구할 수 있고, 그 링크로부터 리뷰를 뽑아내는 것까지 해보고자 한다.
* 자연어처리 konlpy설치하는데에 있어서 pip upgrade 등의 시작부터 문제가 있는 듯하여 해결하고자 한다..

1. 기대효과
2. 기업적 측면

- 제품에 대한 긍정적인 리뷰를 많이 보이는 제품의 경우, 구매자로 하여금 다양한 웹사이트의 공통적인 리뷰를 확인가능함으로써 리뷰의 신뢰성이 높아져 구매력이 증가할 것이다.

- 해당 제품의 실제 이용고객이 제품을 사용하고 실제로 좋은 제품이라면 회사의 신뢰성이 증가하여 차기 출시될 제품의 신규 구매력도 증가할 것이다.

1. 사용자 측면

- 제품에 대한 리뷰를 알아보기 위해 번거롭게 여러 플랫폼을 일일이 찾아 검색하지 않아도 한번에 파악할 수 있다.

- 리뷰에서 자주 언급된 단어를 통해 중요 키워드를 산출해내기 때문에 신뢰가 가는 제품인지 쉽게 파악할 수 있다.

- 비슷한 조건의 다른 좋은 선택지가 없는지 한 눈에 알아볼 수 있다.

참고문헌

* 파이썬을 활용한 클로러 개발과 스크레이핑 입문, 2019, 카토 카츠야, 요코야마 유우키, 위키북스
* 파이썬 데이터 수집 자동화 한방에 끝내기 한입에 웹크롤링, 2018, 김경록, 서영덕, 비제이퍼블릭
* 파이썬을 이용한 웹크롤링과 스크레이핑, 2018, 카토 코타, 위키북스
* 파이썬을 이용한 머신러닝, 딥러닝 실전 개발 입문, 2019, 쿠지라 히코우즈쿠에, 위키북스
* Web Scraping with Python, 2019, 라이언미첼, 한빛미디어
* https://www.crummy.com/software/BeautifulSoup/bs4/doc.ko/