

|  |
| --- |
| 중간 보고서 |
| 미디어 통합 리뷰 어플  Vol.2 |

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

* 미디어는 연령, 성별, 자라온 환경 등 많은 요인에 따라 각기 다른 후기를 남기며 개개인마다 보는 관점이 다르다. 이에 일반적인 제품 리뷰와 다르게 자신들의 생각, 가지고 있는 신념, 자라온 문화 환경, 평소 자주 사용하는 어투 등이 더 여실 없이 리뷰에 드러난다.
* 미디어를 접할 수 있는 환경에도 많은 변화가 생겼다. 과거에는 현재 개봉한 영화들은 주로 영화관에서, 지난 영화들은 DVD, 블루레이와 같은 소장본 혹은 다시 보기 서비스를, 책의 경우, 신간이 나온다면 서점이나 도서관을 이용했다면 최근에는 스트리밍 플랫폼이라는 새로운 서비스가 음악이나 단순 동영상이 아니라 영화, TV, 서적까지 분야에 있어 확장되면서 미디어계의 새로운 패러다임으로 떠오르기 시작했다. 영화와 TV라는 미디어는 “넷플릭스”를 선두로 현재까지 많은 스트리밍 플랫폼이 생겨났는데 단순히 미디어 스트리밍뿐만이 아닌 자체 오리지널 시리즈를 제작하기도 하고 기존 영화 개봉은 영화관이라는 틀을 바꾸어 플랫폼 최초 공개까지 영화 산업의 전반적인 흐름을 바꾸고 있다고 해도 과언이 아니다. 이런 스트리밍 서비스는 출판업에까지 그 분야를 더욱 확장하여 기존 서점이나 도서관에서 실제 종이로 된 책을 인터넷으로 더욱 쉽게 가볍게 접할 수 있는 E-북 시장에 이어 이젠 “밀리의 서재”와 같이 손 안의 도서관, 서적계의 “넷플릭스”와 같은 느낌의 플랫폼까지 만들어냈다. 단순히 소설뿐 아니라 신문, 잡지와 같은 출판물도 “넷플릭스”같이 구독하여 볼 수 있는 애플의 “뉴스플러스”같은 서비스도 나오게 되었다.
* 이렇게 과거 문화 소비와는 많은 변화를 거친 현대에 미디어를 접하는 환경도 이에 대한 리뷰를 남기는 방식이나 플랫폼도 사용자에 따라 특유의 특성을 보이고 저마다 다른 의견을 보이기 때문에 해당 미디어에 대한 리뷰를 단 한 곳에서만 보고 이 미디어가 어떤 미디어인지 판단하는 것은 불가능해졌다.
* 많은 젊은 층의 사람들은 소셜미디어를 통해 자신의 리뷰를 남긴다. 그리고 소셜미디어 또한 10대, 20대, 30대, 혹은 성별에 따라서 선호하는 소셜미디어가 다르다. 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.
* 소셜미디어의 특성에 따라서도 리뷰를 남기는 방식의 차이가 있다. 인스타그램이 키워드 중심의 간결한 표현이 많다면 트위터는 일상생활 속 자신이 사용하는 어투 그대로를 적어놓는 경우가 많다. 블로그나 리뷰는 상대적으로 또박또박한 말로 적어 놓기도 한다.
* 최근 미디어 리뷰 사이트에서 빈번하게 일어나는 별점 테러와 같은 행위와 광고성 포스트들로 인해 진정한 리뷰를 찾기 힘들어진 지금, 소셜미디어를 통한 사실적인 리뷰를 파악하고 현재 미디어 트렌드에 대해 파악해보고자 한다.

1. 목적

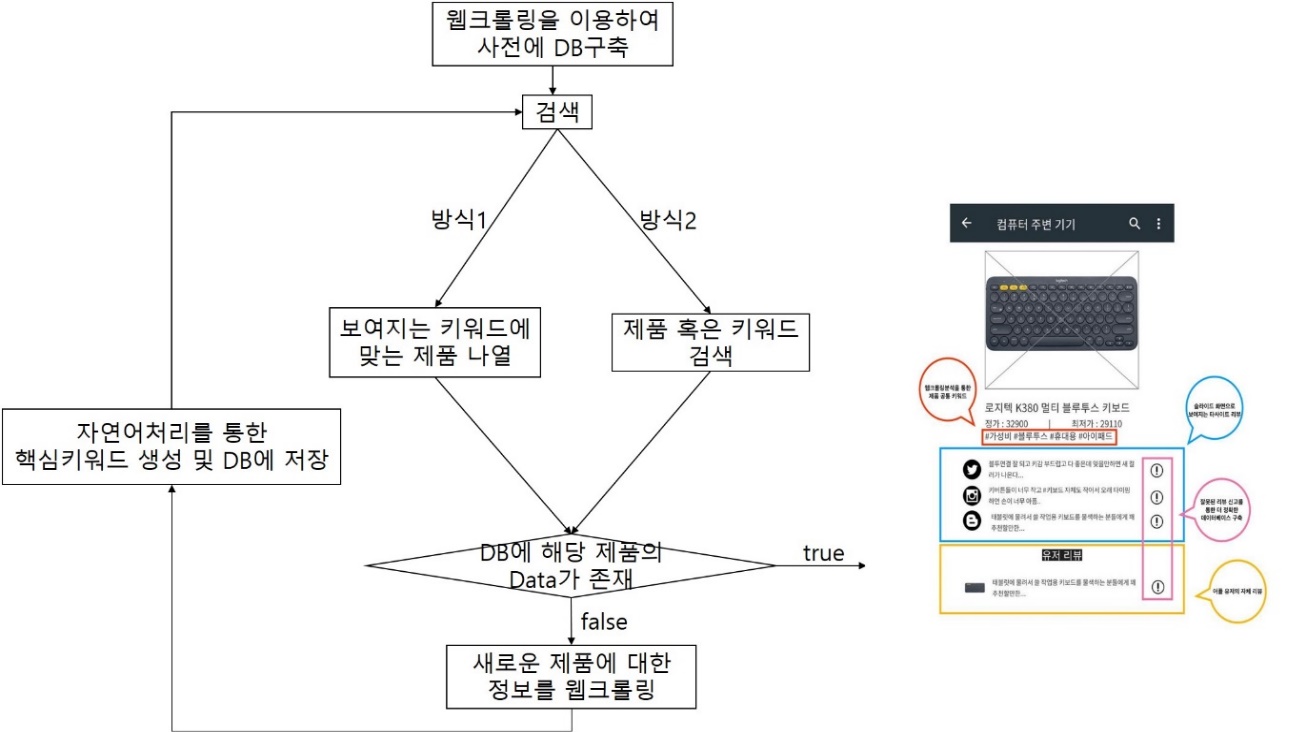
* 정보의 신뢰성

기존의 리뷰 사이트뿐 아니라 여러 소셜미디어에 있는 관련 리뷰를 분석합니다. 여기서 관련 중요 키워드를 중심으로 데이터베이스를 구축하여 이렇게 만들어진 데이터베이스를 통해 현재 미디어 트렌드를 파악해보고 이와 관련되어 긍정적인 반응을 얻고 있는 작품을 추천할 수 있습니다. 전반적인 작품에 대한 반응을 살펴보고 이에 대해 긍정적 리뷰와 부정적 리뷰를 모두 보여줌으로써 이러한 리뷰를 통해 사용자는 해당 미디어에 대한 정보를 사전 습득할 수 있고 이와 비슷한 평가를 받은 미디어가 어떤 것이 있는지 확인하며 미디어의 추천으로 이어질 수 있고 또는 반대되는 평가를 받은 미디어 확인을 통해 미디어 선택에 도움을 줄 수 있다.

* 접근의 용이성

이를 어플리케이션(혹은 웹사이트)로 구현하여 사용자가 언제 어디서든 쉽게 접근하고 사용할 수 있도록 합니다.

1. 설계내용
2. Flow Chart



1. 개념 설계

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

- DB를 구축하기 위해 해당 웹사이트 접속 (\*리뷰 사이트, 블로그, 트위터, 인스타그램)

- 작품 기본 정보와 사용자 및 구매자의 리뷰가 담긴 웹사이트 코드 분석

- BS4를 이용해 페이지 데이터 호출

- 작품의 기본 정보 tag를 찾아 추출

- 각 사이트와 페이지별로 링크를 재귀적으로 검색하여 데이터를 추출

**2) 크롤링된 정보를 이용한 DB구축(Mysql or sqlite)**

**-** MySQL 서버에 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

- 오피니언 마이닝(감성 분석)을 이용하여 리뷰의 긍부정을 평가하고 어떤 평가가 많은지 분석한다. 최근 1-2달간의 리뷰를 토대로 현재의 트렌드를 파악한다.

- 긍부정 리뷰 평가 과정에서 각 단어의 속성 값을 추출하여 판별을 하는데 이 속성 값이 작품을 선정함에 있어 사용자의 기준이 되므로 이 키워드를 중요 키워드로 설정하고 관련 데이터 현황을 조사, 분석한다.

- 이 과정에서 마이닝의 정확도를 높이는 작업이 필요하다. 이 부분은 현재 나와있는 다양한 모델 기반 접근법을 이용한 감성사전을 이용하여 데이터의 정확도를 끌어올릴 계획이다. 각 감성 사전마다 어떤 단어나 조사를 제거하고 비속어, 신조어, 오타의 처리를 어떻게 하냐에 따라 정확도가 달라진다. 현 소셜미디어의 특성상 표준어를 사용하지 않고 발음을 있는 그대로 적어 두거나 비속어, 신조어, 약어 혹은 “이 장면 정말 사이다였다.”와 같이 원래 가지고 있는 역할과 다르게 빗대어 많이 사용되고 있는 언어들이 있으므로 이 부분에서 더 효과적인 판별법과 사전이 어떤 것인지 정확도를 높이는 데 중점을 두고자 한다. 표준어만을 이용한 감성 분석 API 및 툴들은 많이 존재하지만 비표준어는 경우의 수가 굉장히 다양하기 때문에 이를 처리하는 API나 사전 구축에 대한 조사는 더 필요하다.

- 참고 분석 예정 : word2vec, sentiwordnet, [www.openhanquel.com](http://www.openhanquel.com), KTS

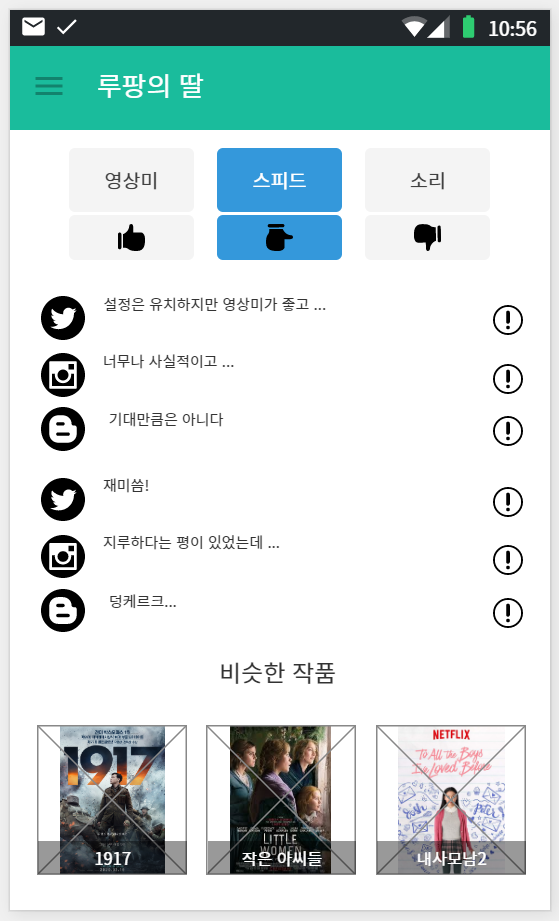
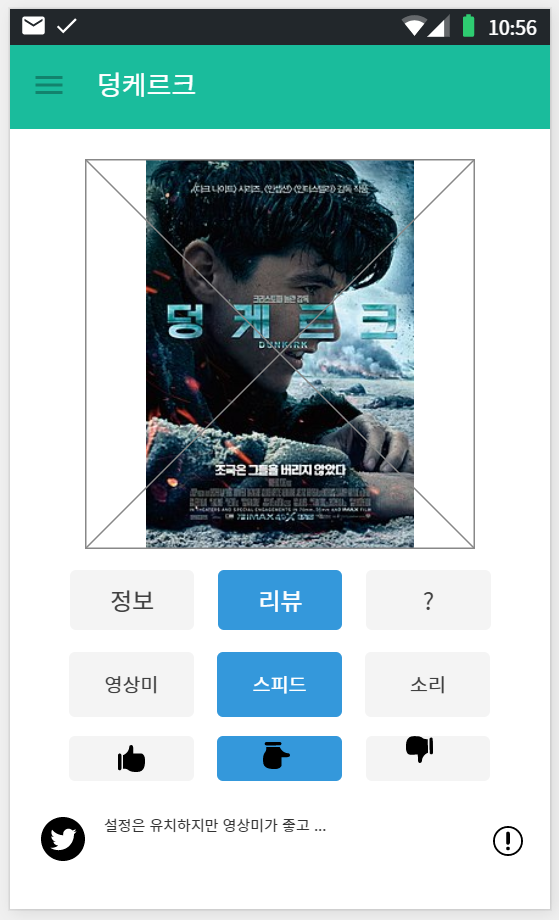
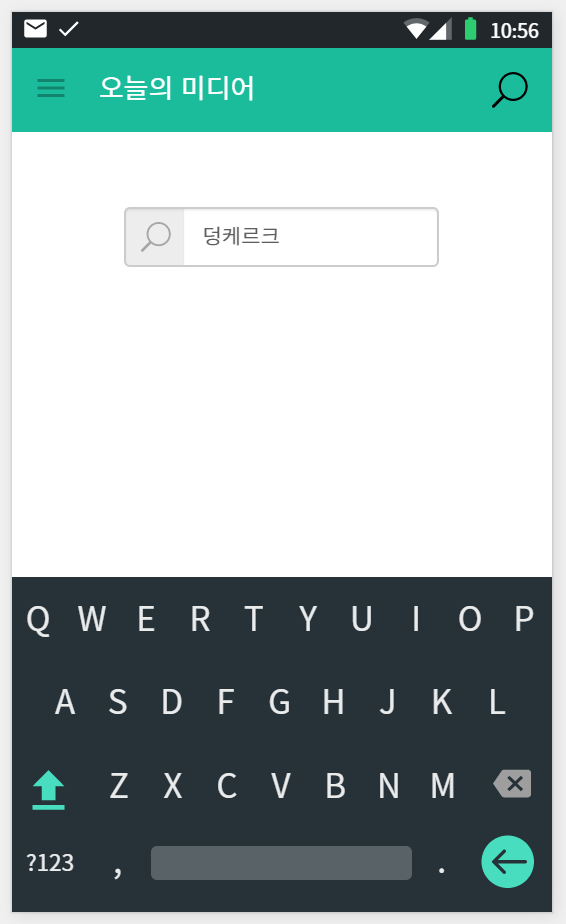
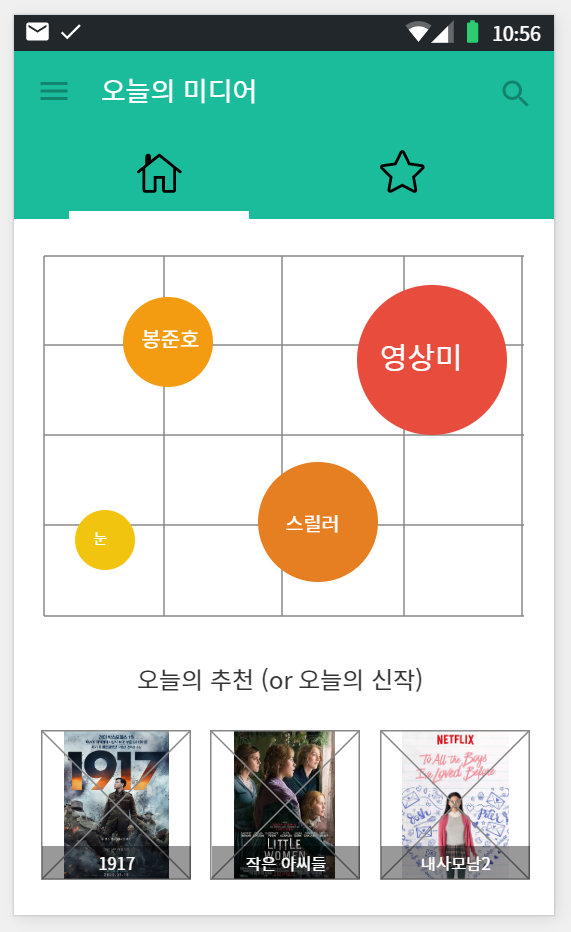
- 음소 단위 분할 조합을 통해 뜻을 파악하는 Trigram-Signature를 사용하면 오타가 있는 “조ㅎ아”, “실ㅎ어”와 같은 문맥을 파악하는 데 유용할 것이라고 생각한다. 이를 이용한 API가 있는지 혹은 이와 관련된 사전을 제작하여 사용하는지 더 조사할 계획이다.

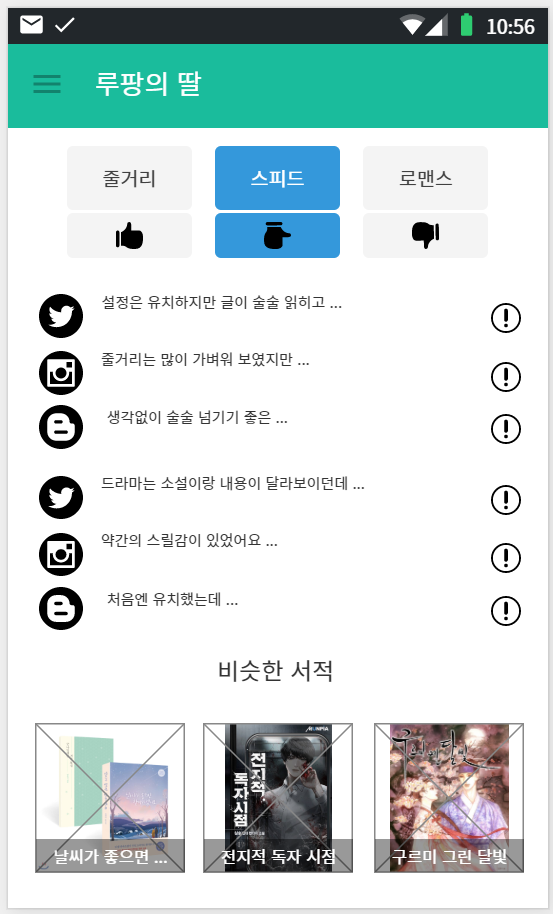
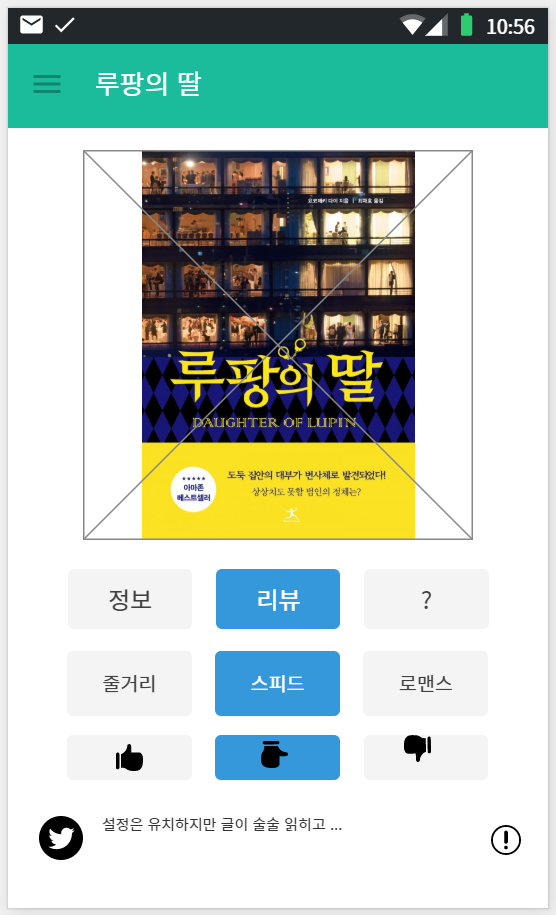
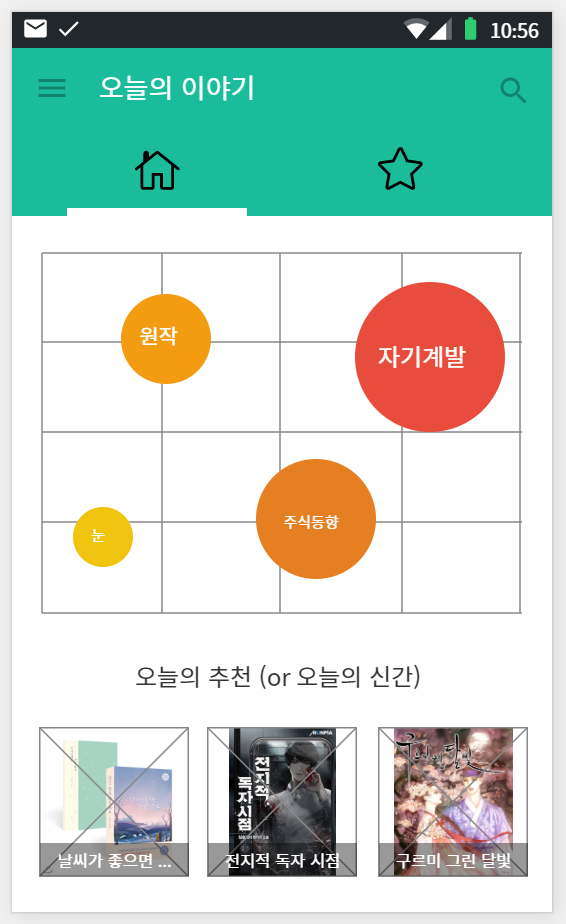
**4) Android UI제작**

**-** 자신이 보고 싶은 작품의 리뷰를 보기 위한 제품의 검색창을 사용자가 보기 편하도록 UI로 구현

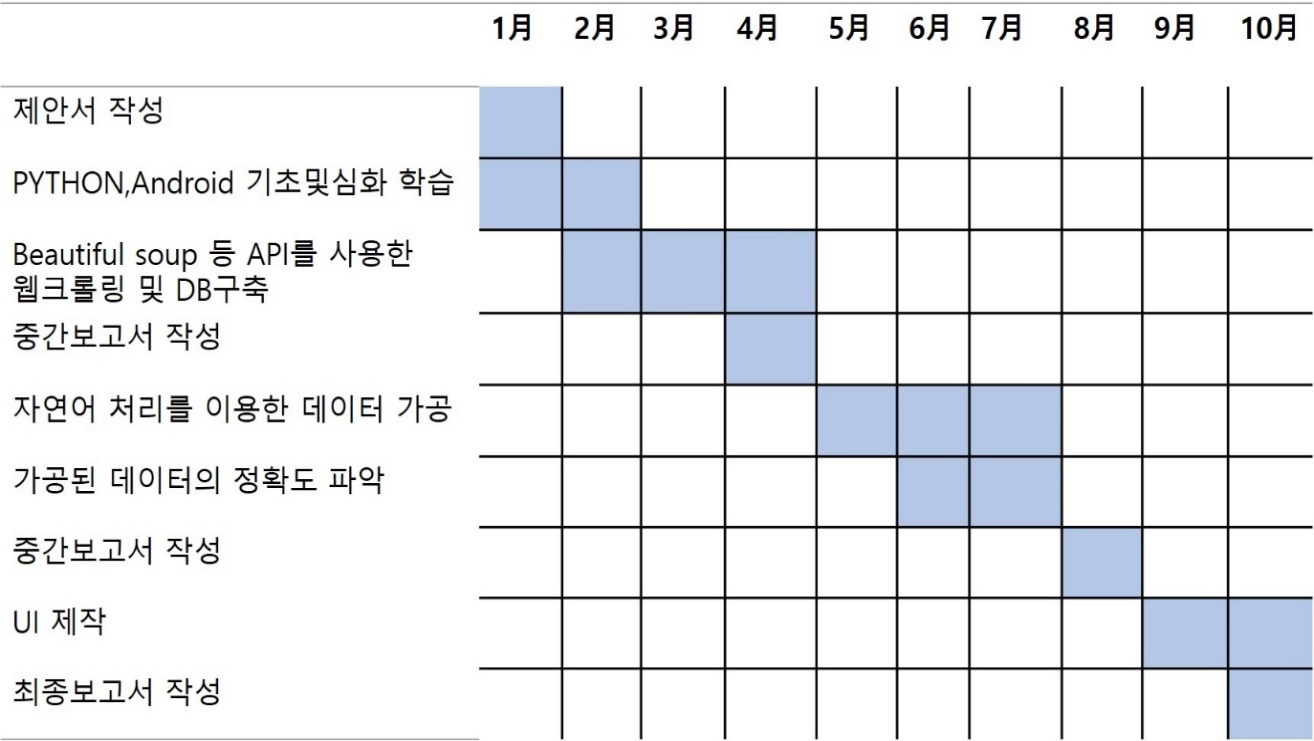
- 검색 결과로 작품의 기본 정보와 함께 리뷰를 보여준다. 리뷰의 경우 공통적으로 많이 언급이 되는 속성(엔터티, 특징)을 보여주고 이에 대한 리뷰를 선택 시 행당 리뷰를 사이트 별로 보여준다.

**5) 예시 결과**





1. **과제 수행**
2. 수행 일정



1. 웹 크롤링
2. 트위터

Vol1. 트위터의 정식 API인 tweepy가 존재는 하나 단점이 최근 7일간의 트윗만을 가져올 수 있기 때문에, 리뷰 분석을 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 1년 동안 ‘로지텍’이라는 검색어가 들어간 트윗을 수집해보았다.

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

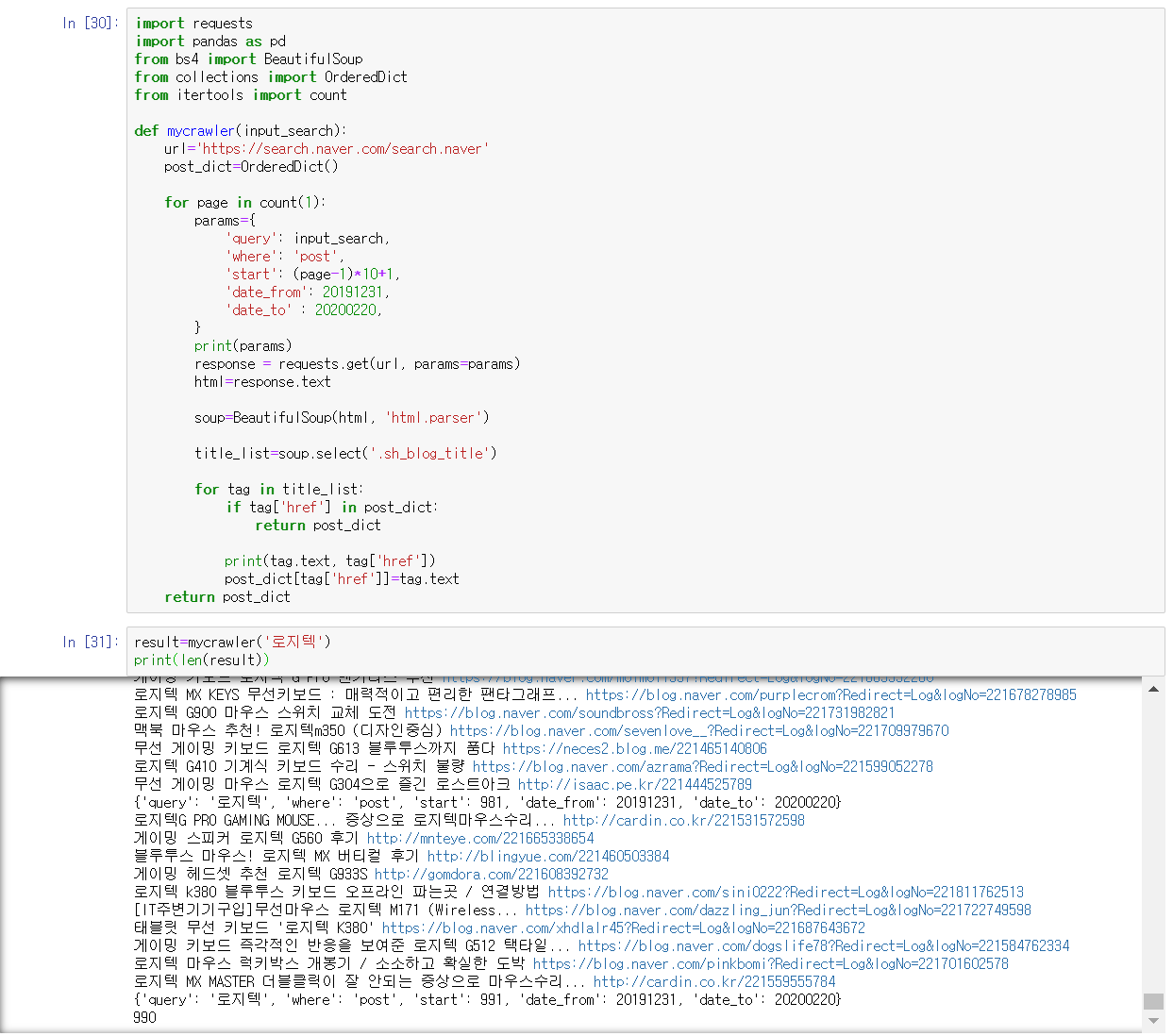
저장한 파일을 확인하면



이러한 결과를 얻을 수 있다.

1. 네이버

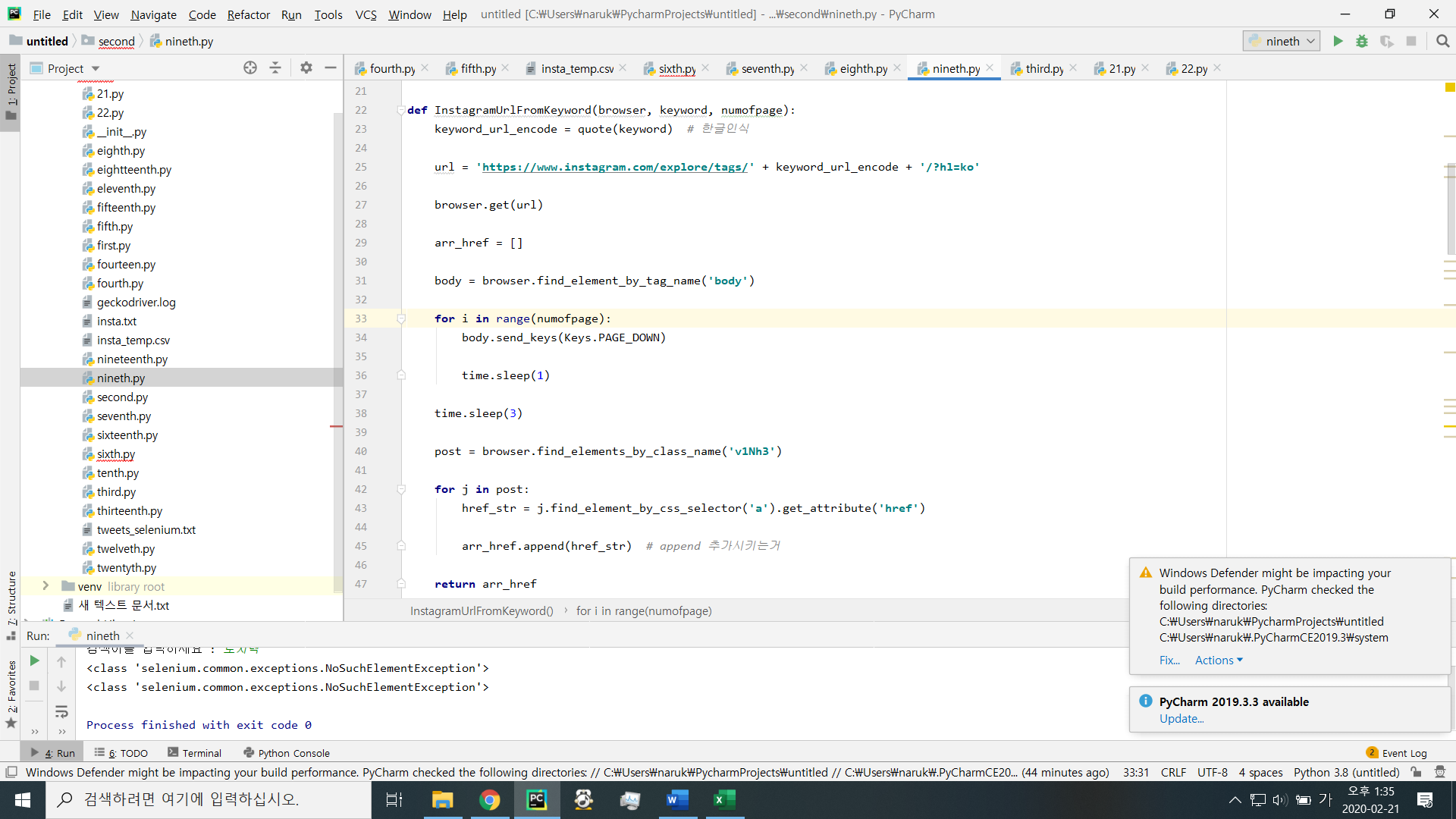
1차시: 실제로 많은 사용자들이 많이 보는 리뷰는 네이버일 것이다. 따라서 네이버 블로그의 검색어에 따른 결과 크롤링을 해보면 제목과 블로그 링크를 가져올 수 있다. 무한정으로 많은 양의 검색 결과를 가져올 수 없는데 이는 좋은 검색 결과를 위해 네이버가 1000건의 검색 결과만을 보여주고 있기 때문이다. 하지만 이것만으로도 충분히 키워드를 추출한다던지, 제품 한 개를 분석함에 있어서 부족함이 보이지는 않는다.



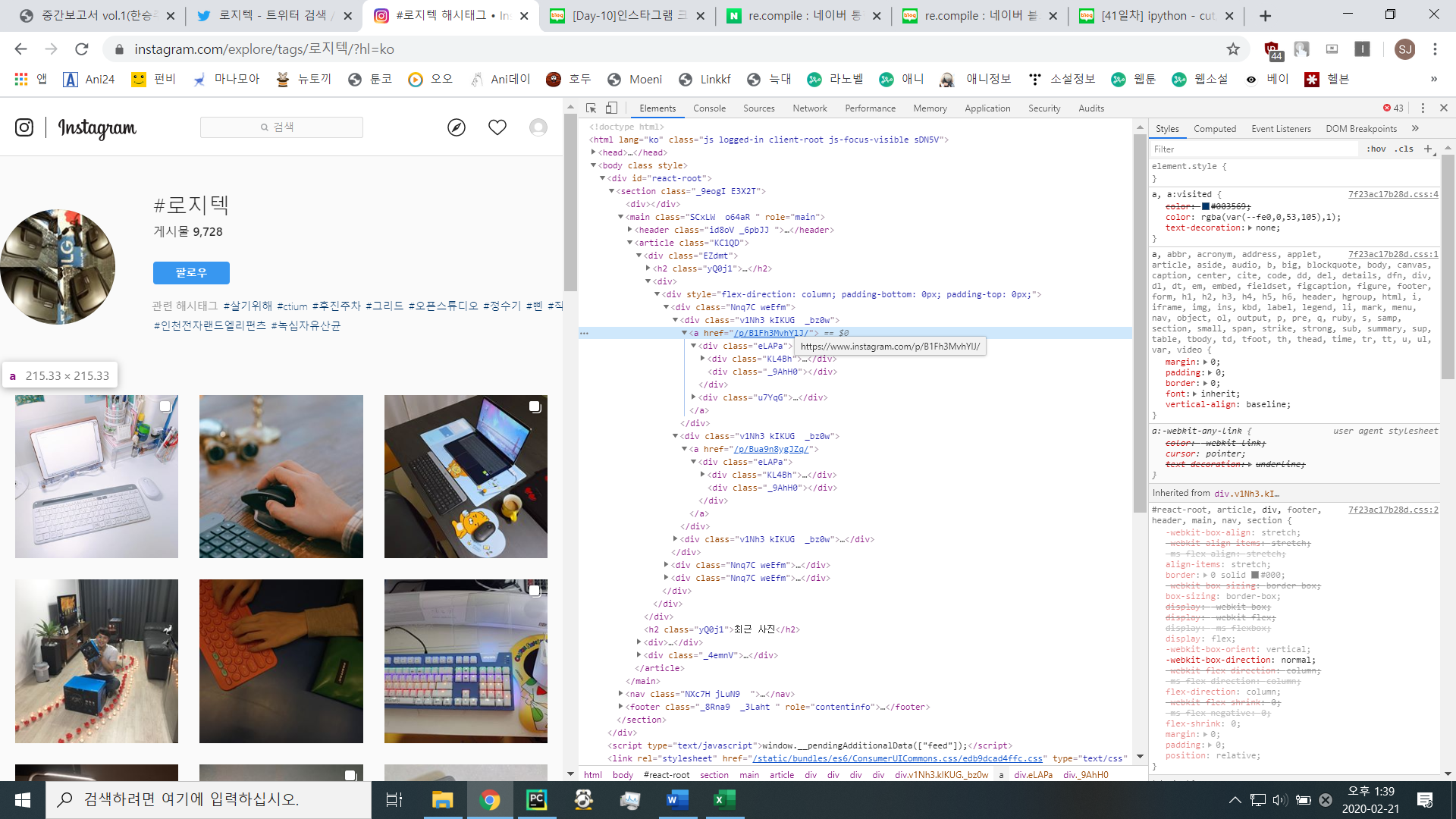
위는 글에서 글의 제목과 링크 주소를 가져온 스크래핑의 결과인데, 해당 링크를 타서 본문을 가져올 수 있지만, 해당 기능은 다음인 인스타그램에서 구현해보고자 한다.

1. 인스타그램

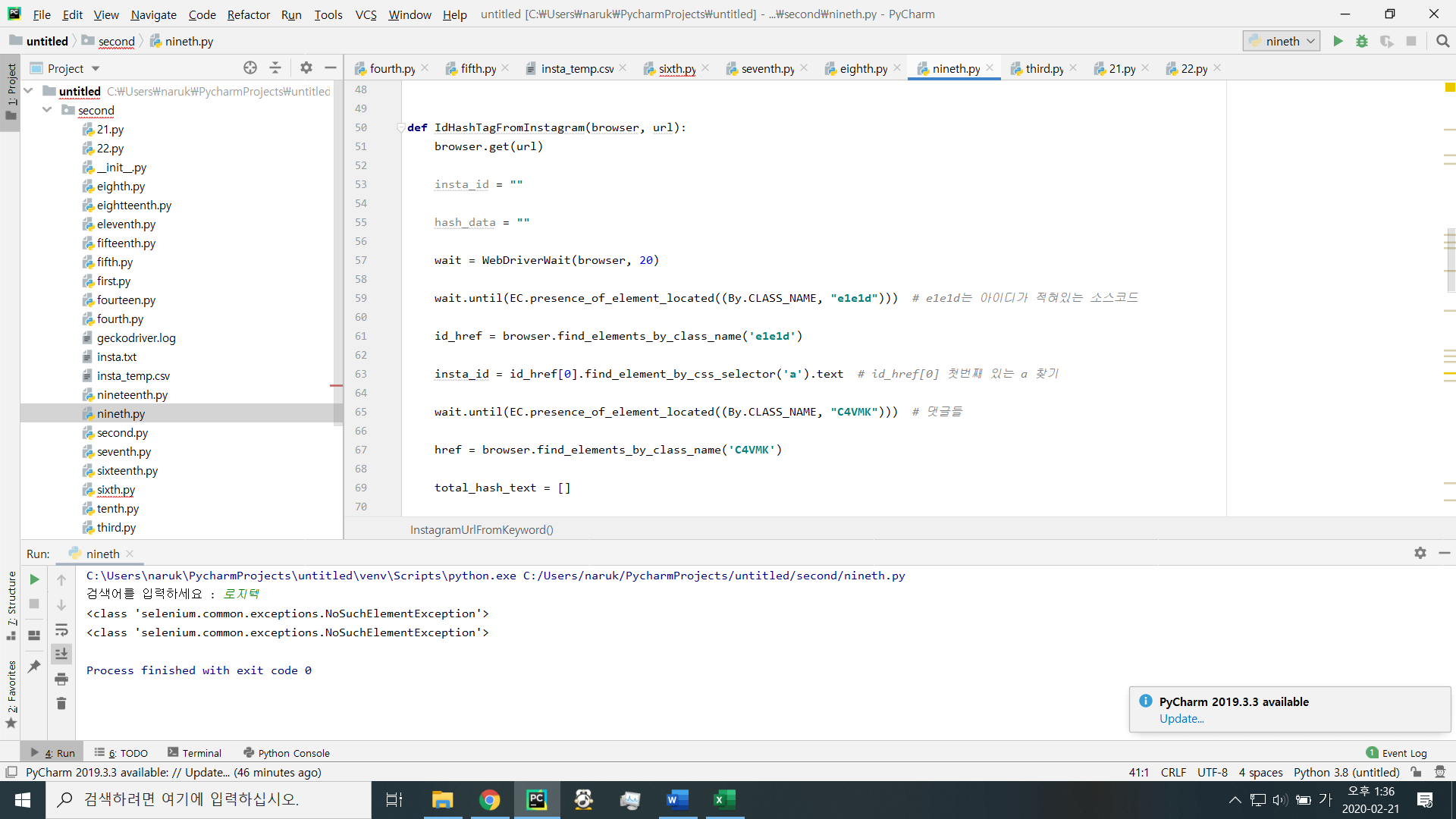
Vol1. 네이버 블로그가 긴 리뷰, 트위터가 짧은 리뷰라고 하면 인스타그램은 해쉬태그라는 자신만의 키워드를 표현하는 기능이 있지만, 실제로 분석함에 있어서는 해쉬태그를 따로 볼 것이 아니라 전체적인 리뷰로 보고 트위터와 유사한 짧은 리뷰라 판단하여 전체적인 분석을 하고자 한다.

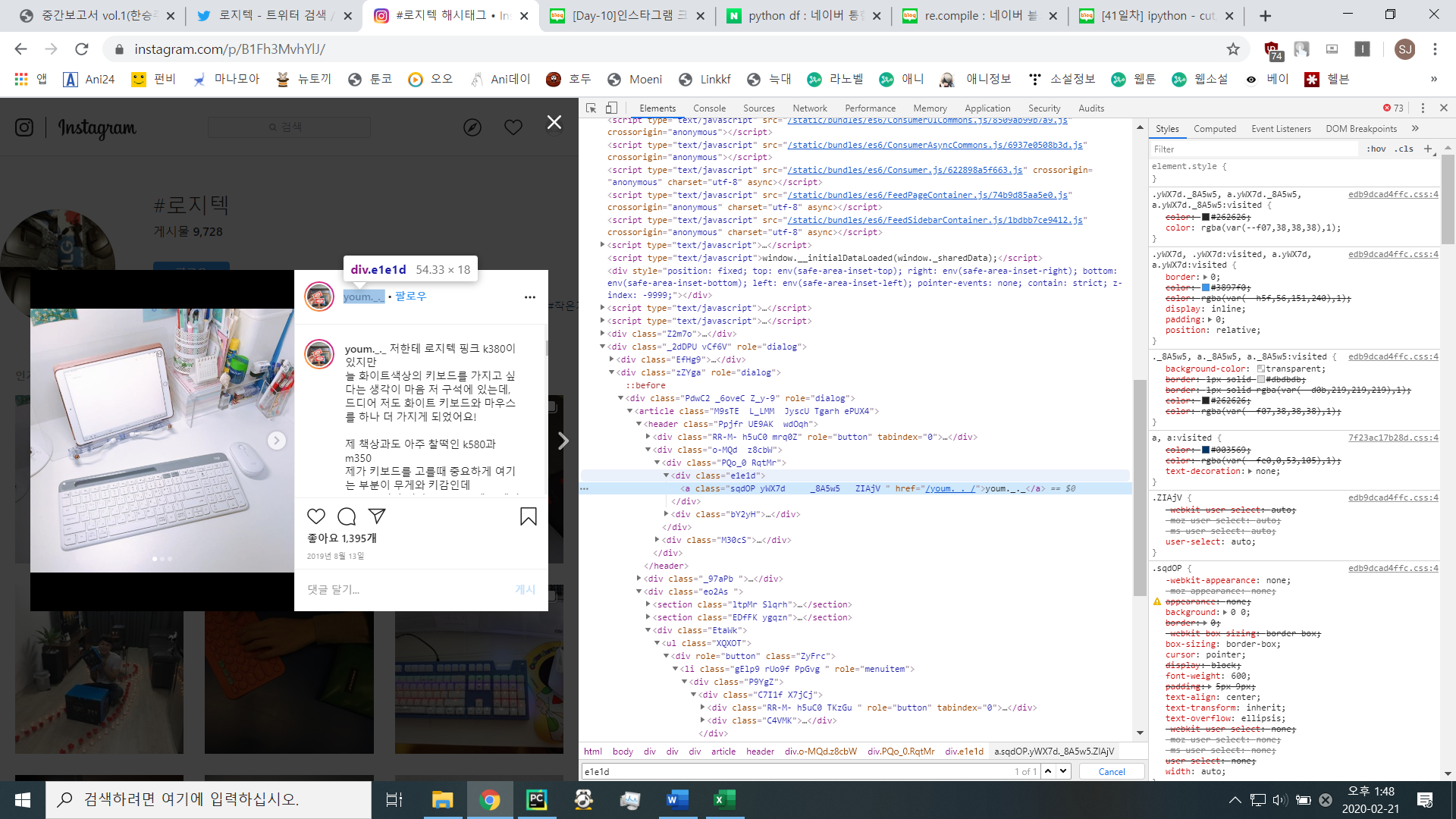


인스타그램에서 어떤 키워드로 검색어를 하면 해당 키워드인 한글이나 영어는 컴퓨터가 해석하기 난해하므로, 인코딩을 통해 해당 검색어에 맞추어 주소변환이 가능하다. 또한,

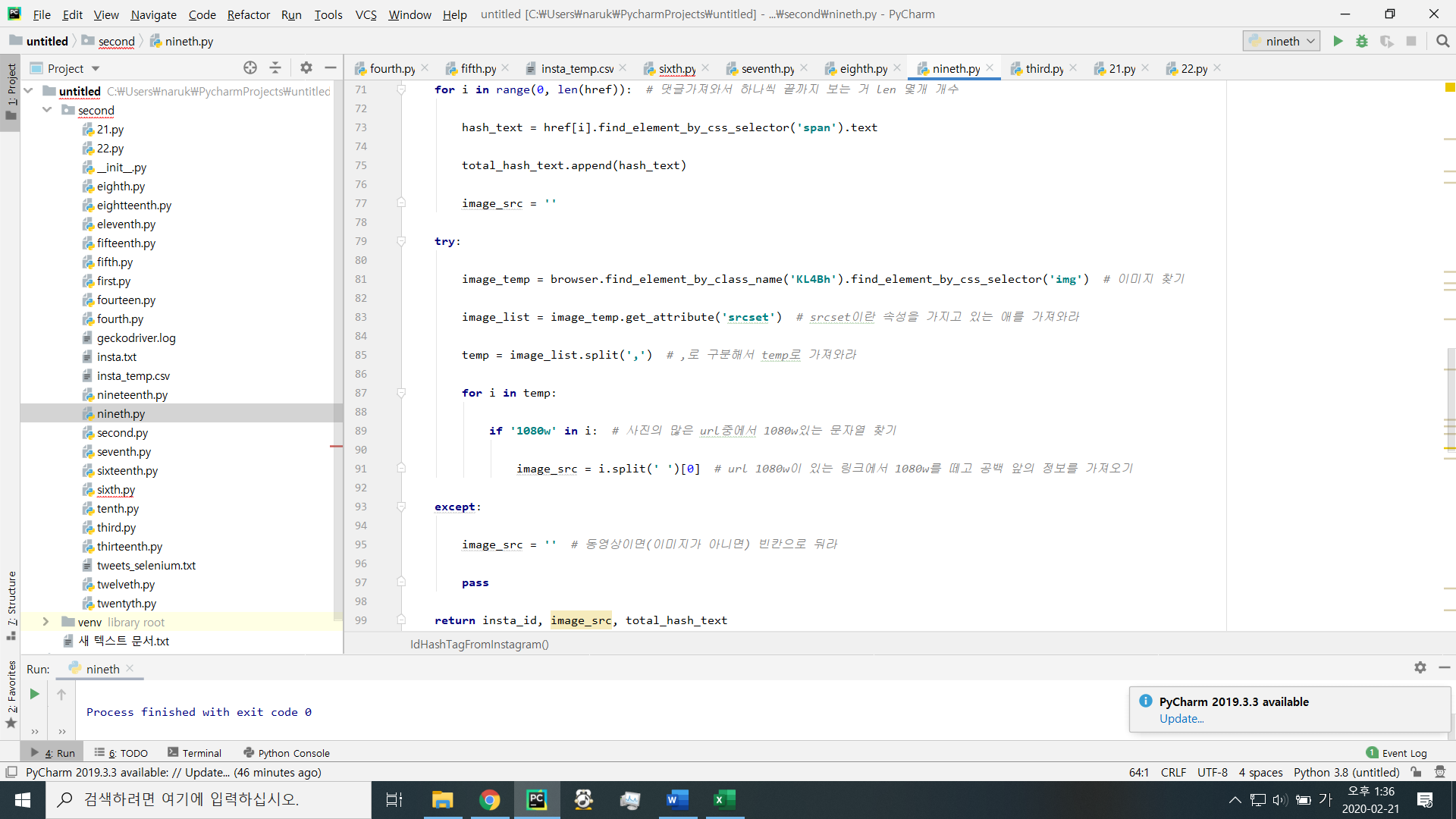


여러 글의 본문을 찾기 위해서 해당 함수를 사용하게 되는데, 인스타그램 웹의 소스코드에는 v1Nh3 class 밑에 href를 통해 모든 글의 주소를 가져올 수 있어서 해당 글의 ref를 가져오기 위한 정의이다.

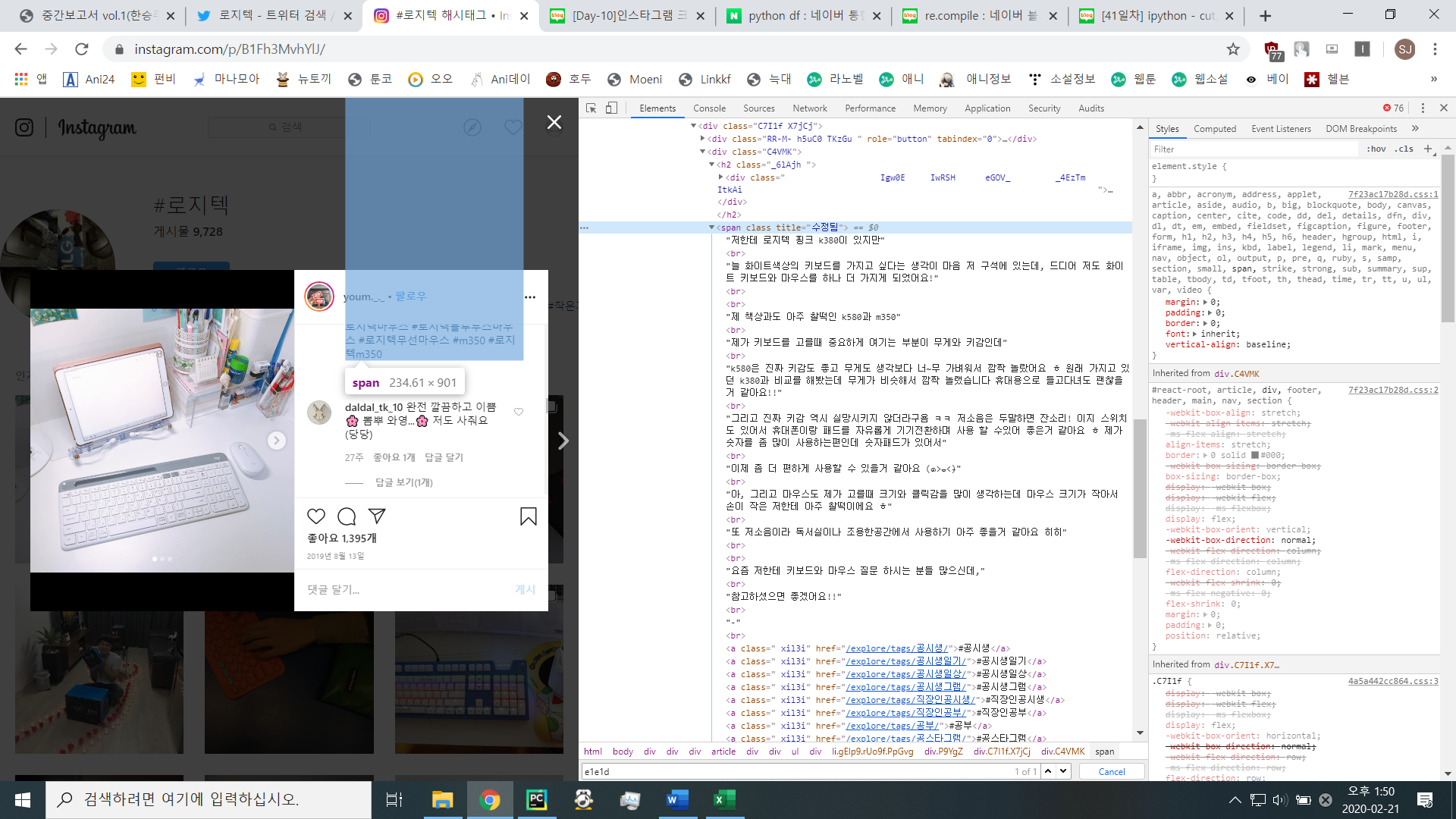




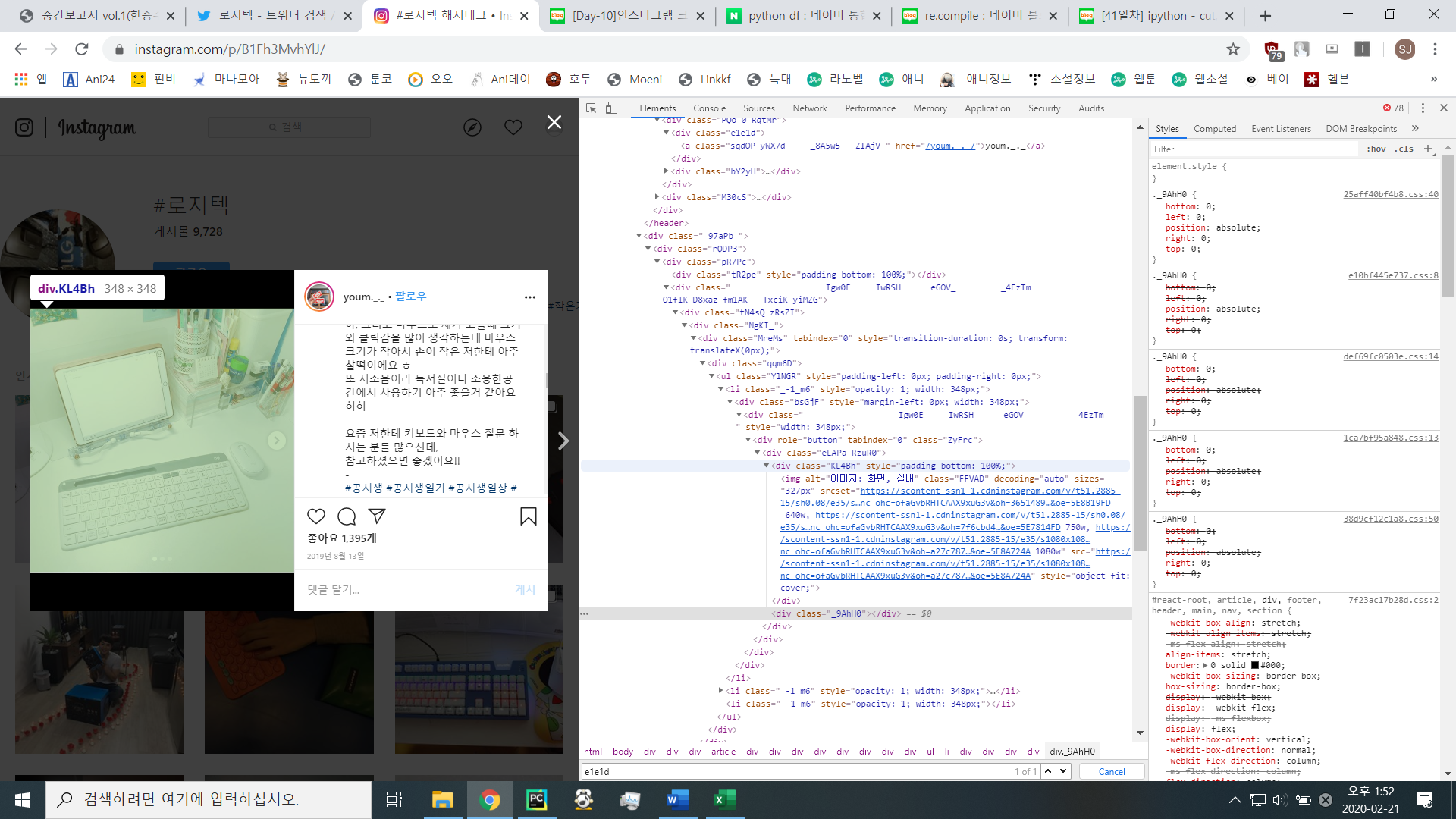
아이디를 찾는 부분으로, eleld 클래스 영역에 href로 text만 뽑아내면 해당 부분이 인스타그램 id이므로 이를 추출한다.



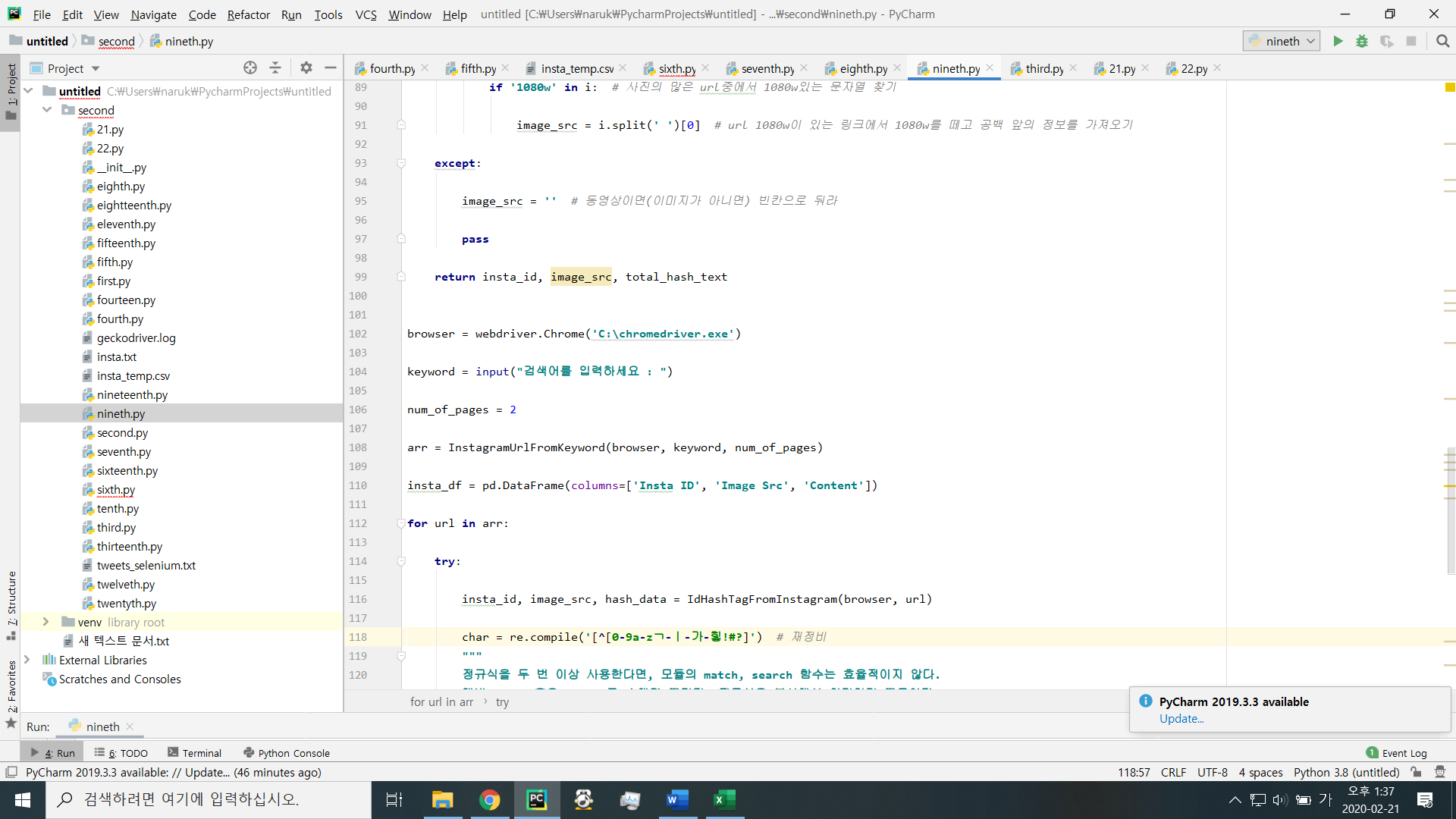
해당 코드는 for문을 이용하여 참조할 reference가 있는 동안 해당 ref의 본문을 축적하는 부분과 본문에 덧붙인 이미지의 src를 찾는 부분으로 구성이 되어있다.

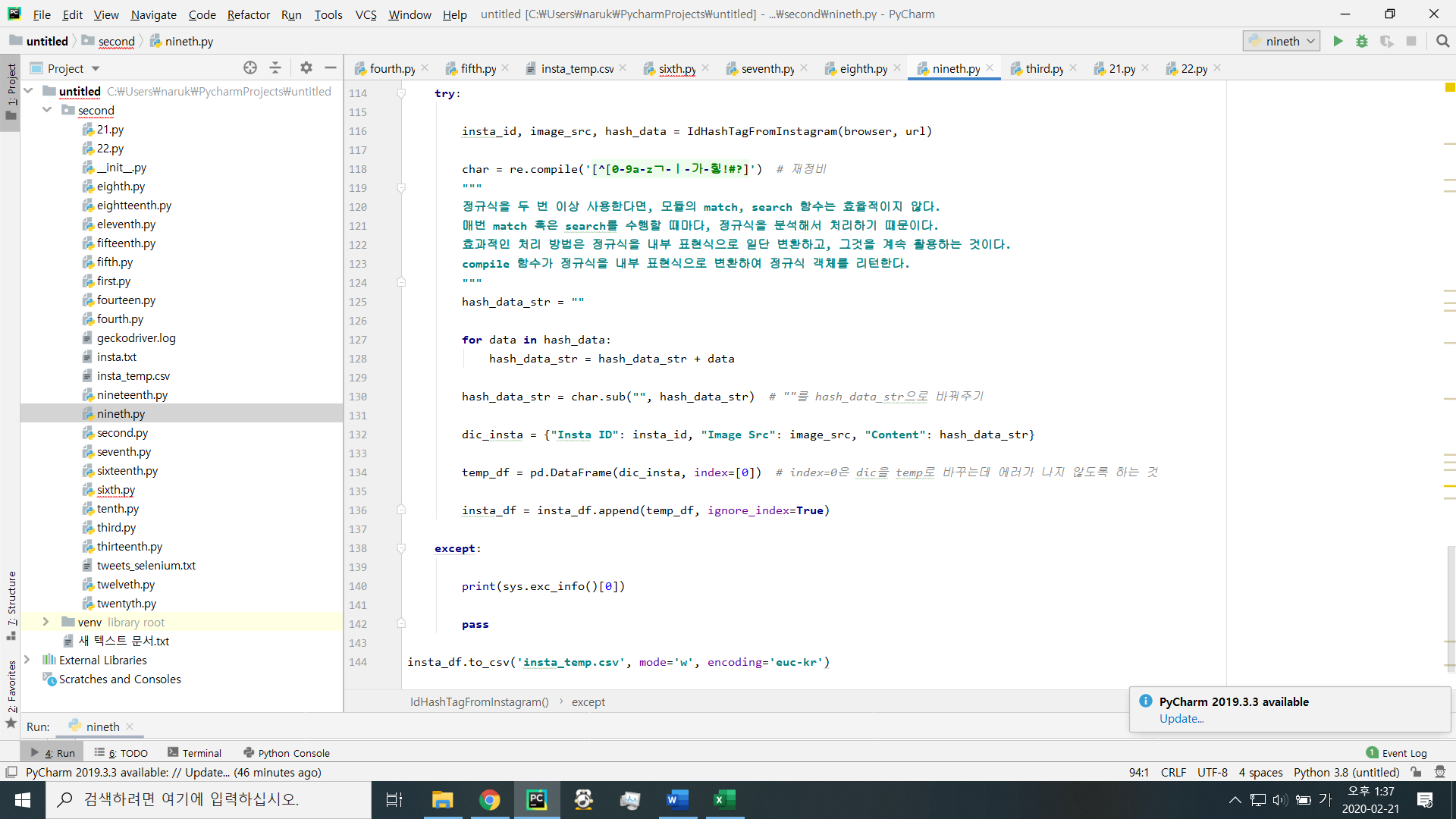


본문의 부분으로 댓글 또한 위와 같이 span의 영역에 글이 작성되어 있다.



Img\_src 또한 KL4Bh의 영역에 img를 find하여 찾을 수 있다.





실제로 동작하는 부분은 검색어를 입력받아서 위의 두 가지 함수를 활용하여 인스타그램의 주소를 가져와서 해당 주소의 해쉬태그를 데이터 프레임에 쌓고 이를 csv로 저장하는 부분이다.



라는 결과를 얻을 수 있다. 여기서, 활용한 것은 저번 2주 차 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 자연어 처리를 실습하는 것을 계획하고 있으며, 어떠한 DB를 사용할지까지 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. KoNLPy 한국어 분석

KoNLPy는 한국어 분석을 위한 정보 처리기로 형태소 분석에 용이하다.



Vol2. 분석 대상 변경으로 크롤링 대상은 리뷰 사이트와 현대 많은 사람들이 쉽게 이용하고 후기를 남기는 블로그, 트위터, 인스타그램을 대상으로 분석한다. 한국어의 특성상 문맥파악, 새롭게 나타나는 신조어 구어와 문어체가 많이 달라 구어를 그대로 사용할 경우 이를 최대한 정확하게 사용자의 뜻과 맞게 분석하는 방법을 찾아야할 것이다.

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 오래 걸리는 것을 확인했는데, html을 파싱하는 데에 있어서 다소 문제가 있어서 selenium을 사용했던 것인데 이 파싱하는 부분을 바꿔서 코드를 수정해보고자 한다. 또한, 딕셔너리로 return을 받을 때 나중에 데이터를 사용할 때는 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 오른쪽 클릭이 안되어서 html 소스를 못 보는 줄 알았는데, F12를 사용해서 할 수 있는 것을 확인하고 추가로 해보고자 한다. 특히, 이 경우 위에서처럼 제품 정보를 얻으면서 링크를 구할 수 있고, 그 링크로부터 리뷰를 뽑아내는 것까지 해보고자 한다.
* 자연어 처리 konlpy 설치하는 데에 있어서 pip upgrade 등의 시작부터 문제가 있는 듯하여 해결하고자 한다.

1. 기대효과
2. 기업적 측면

- 즉각적인 피드백이 필요한 문화 산업의 경우 통합적으로 리뷰를 확인 가능함으로서 앞으로의 홍보, 제작, 투자 방향 선택에 도움이 되는 지표가 될 것이다.

1. 사용자 측면

- 별점 테러와 같이 실제 작품에 대한 후기가 아닌 평가 반영으로 실제 작품의 후기를 원하는 사용자에게 더욱 사실적인 후기를 각기 다른 플랫폼에서 검색해 볼 필요 없이 한 곳에서 확인이 가능할 것이다.

- 리뷰에서 자주 언급된 단어를 통해 중요 키워드를 산출해내기 때문에 선호하는 양상의 작품을 기호에 맞춰 선택하기 쉽다.

- 비슷한 다른 작품을 추천받을 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/>
* 2010, 오피니언 마이닝 기술을 이용한 효율적 상품평 검색 기법, 윤홍준, 김한준, 장재영
* 2011, 한글 텍스트의 오피니언 분류 자동화 기법, 김진옥, 이선숙, 용환승
* 2013, 상품평가 텍스트에 암시된 사용자 관점추출, 장경록, 이강욱, 맹성현
* 2013, 텍스트 마이닝을 이용한 2012년 한국대선 관련 트위치 분석, 배정환, 손지은, 송민
* 2014, 한글 감성어 사전 api구축 및 자연어 처리의 활용, 안정국, 김희웅
* 2015, 한글 음소단위 trigram-signature 기반의 오피니언 마이닝, 장두수, 김도연, 최용석
* 2016, 소셜네트워크서비스에 활용할 비표준어 한글처리 방법연구, 이종화 레환수, 이현규
* 2017, <https://www.samsungsds.com/global/ko/support/insights/1195888_2284.html>, 윤병운 교수
* 2017, 한국어 비정형 데이터 처리를 위한 효율적인 오피니언 마이닝 기법, 남기훈
* 2020, A study on Sentiment Analysis with Multivariate ratings in Online Reviews, 임소현