

중간 보고서

데이터 마이닝과 분석을 통한 웹소설 종합 인포 사이트

Vol. 9



제출일	2020. 05. 01	전공	컴퓨터공학과
과목	졸업작품 프로젝트	학번	2015722084
			2015722083
			20172020672
담당교수	이기훈	이름	한승주
			김성종
			조예슬

목 차

I 배경 및 필요성

1. 시장 성장

2. 문제 정의

ㄱ. 양산화

ㄴ. 다양화

3. 설계 내용

ㄱ. Flow Chart

ㄴ. 개념 설계

II 과제 수행

1. 수행 일정

2. Scraping

ㄱ. 플랫폼

ㄴ. SNS

ㄷ. 커뮤니티

3. Data Processing

ㄱ. 트렌드 분석

ㄴ. 감성 분석

ㄷ. Aspect 분석

4. DB Construction

5. UI Development

III 과제 평가

1. 개선방안

2. 기대효과

ㄱ. 기업적 측면

ㄴ. 사용자 측면

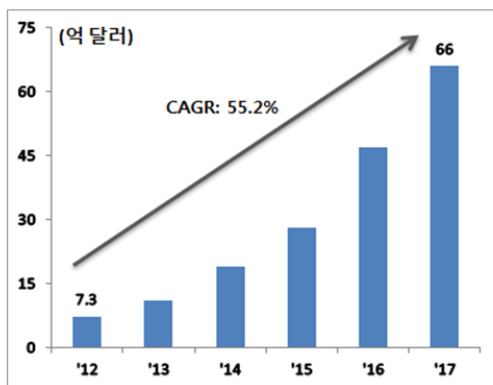
< 참고문헌 >

I. 배경 및 필요성

1. 시장 성장

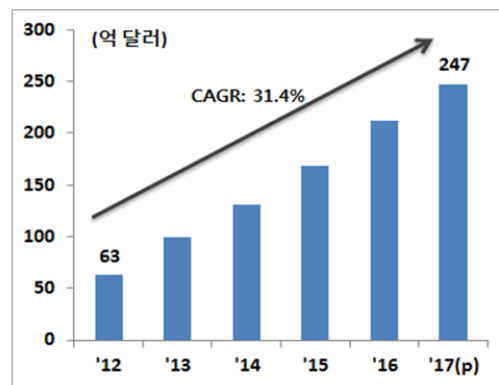
현대에 이르며 미디어 시장에 많은 변화가 이뤄졌다. 미디어의 다양화뿐만 아니라 미디어 플랫폼에도 변화가 이루어지며 현대인들이 더 쉽고 빠르게 그리고 편리하게 미디어에 접근할 수 있도록 환경이 만들어졌다. 음악에 있어 CD 앨범과 같은 아날로그식의 접근 방법이 음원 다운로드와 같은 디지털화를 넘어 스트리밍 플랫폼으로 진화하였고 영화 TV와 같은 영상 매체들은 TV, 영화관과 같은 제한적 접근이 “넷플릭스”와 같은 플랫폼이 생성되며 통합 스트리밍 플랫폼의 패러다임이 열리게 되었다고 할 수 있다.

< 전 세계 음악 스트리밍시장 규모 >



자료 : 국제음반산업협회(IFPI).

< 전 세계 OTT 서비스시장 규모 >



자료 : PwC(2017), ITU(2017), 정보통신진흥원(2018) 재인용.

그림 1. OTT 서비스 시장 규모 (출처 : 현대경제연구원)¹

출판업계 또한 마찬가지다. 책, 신문과 같은 인쇄물은 어느새 디지털화되어 신문은 웹으로 책은 이북과 오디오북 등 다양한 형태로 변화되어 출판업 시장이 많은 다양화를 이루어내고 있다.

그중 우리가 오늘 주목할 것은 웹소설시장이다. 웹소설의 전신은 과거 인터넷 소설과 그 전 PC 통신에서 퍼지던 소설부터 시작되었다고 할 수 있다. 스마트폰의 보급이 활성화되고 인터넷에 대한 접근이 쉬워지며 모바일 환경에서 쉽게 볼 수 있는 웹소설이라는 새로운 시장이 설립되었다. 기존의 인터넷 소설을 연재하던 사이트부터 카카오페이지, 네이버 등 덩치가 큰 기업들이 뛰어들며 시장은 더욱 커지고 있다.

¹ 콘텐츠 스트리밍 산업의 성장동력화가 시급하다. (현대경제연구원, 2019.02)

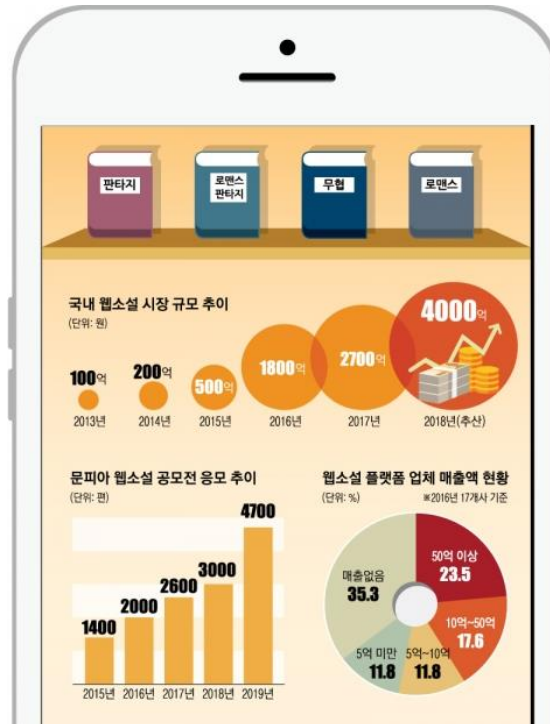


그림 2. 국내 웹소설 시장 규모 추이 (출처 : 서울신문, 2019.05)²

웹소설의 시장이 커질 수 있게 된 것에 원 소스 멀티 유즈, 즉 OSMU가 큰 역할을 한 것처럼 보인다. 2018년 한국 콘텐츠 진흥원의 조사에 따르면 국내 웹소설 시장은 2013년 100억 원 규모에서 2018년 4000억 원까지 40배 성장했다. 이러한 배경에는 웹소설을 바탕으로 제작된 드라마의 성공이 있다. 또한, 주변국인 일본은 남녀노소 상관없이 서브컬처 문화를 수용하고 이를 즐기고 있어 이러한 콘텐츠 시장이 이미 발달한 상태이고, 중국도 연간 2조 원의 시장 규모를 가지고 있을 만큼 웹툰 및 웹소설 시장이 큰 편이다. 최근에는 중국, 일본 외에도 세계 각국의 작품들을 수입 및 수출하고 있고 새로이 제작되는 웹툰이나 드라마, 영화 심지어는 게임까지도 웹소설을 기반으로 한 것이 굉장히 많다.

이렇게 커진 웹소설 시장을 방증하듯 독점 연재도 하며 인기도 많은 플랫폼이 약 6곳, 그 외에도 여러 작품의 연재 서비스를 제공하는 곳까지 합하면 그 수는 10곳을 훌쩍 넘는다. 이렇게 커진 플랫폼들은 저마다 신인 작가를 육성하기 위한 공모전을 열기도 하고 독점 연재작 계약도 진행하며 독자 유입을 위한 노력을 하고 있다.

² '하루 5분' SNS 하듯 쓰윽~ 4000억 시장 펼친 웹소설 (서울신문, 2019.05)

2. 문제 정의

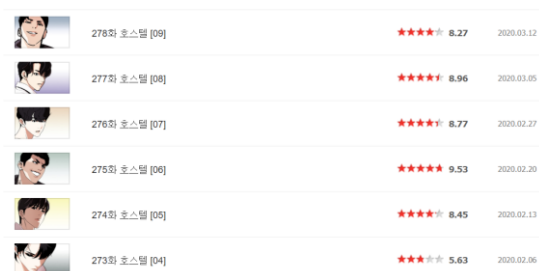
ㄱ. 양산화

웹소설의 인기와 트렌드에 따른 양산화에 따라 수적으로 선택지는 많아졌지만 작품의 질이 받쳐주지 않고 있다. 그렇기 때문에 소비자는 이를 직접 다양한 형태의 내리고 그 평가를 근거로 소비한다. 플랫폼 내부의 댓글과 별점, SNS 나 각종 커뮤니티에 자신의 리뷰를 남김으로써, 직접 칭찬이나 불만을 터트리기도 하여 작품에 대한 평가가 여러 곳에 퍼지며 이에 따라 작품에 직접적인 타격을 주고 있다. 예를 들어 웹툰 [외모지상주의]는 초반 9점대 이상의 좋은 평으로 인기를 얻은 작품이다. 하지만, 시간이 지날수록 줄거리에 지루함을 느낀 독자는 댓글과 별점에 이를 표하며 현재는 7점대까지 떨어졌다.



88화	불법 도도 [06]	★★★★★	9.87	2018.07.21
87화	불법 도도 [05]	★★★★★	9.89	2018.07.14
86화	불법 도도 [04]	★★★★★	9.89	2018.07.07
85화	불법 도도 [03]	★★★★★	9.88	2018.06.30
84화	불법 도도 [02]	★★★★★	9.85	2018.06.23
83화	불법 도도 [01]	★★★★★	9.82	2018.06.16

그림 3. 좋은 평을 받고 있는 작품 초반기 (출처 : 네이버 만화 외모지상주의)



278화	호스텔 [09]	★★★★☆	8.27	2020.03.12
277화	호스텔 [08]	★★★★☆	8.96	2020.03.05
276화	호스텔 [07]	★★★★☆	8.77	2020.02.27
275화	호스텔 [06]	★★★★★	9.53	2020.02.20
274화	호스텔 [05]	★★★★☆	8.45	2020.02.13
273화	호스텔 [04]	★★★☆☆	5.63	2020.02.06

그림 4. 최근 평이 떨어지고 있는 작품 (출처 : 네이버 만화 외모지상주의)

ㄴ. 다양화

웹소설을 제공하는 플랫폼뿐만 아니라 SNS 나 커뮤니티에도 리뷰를 많이 남기는데 사이트들마다 서로 다른 특성을 보이기 때문에 이용하는 나이대나 성별 등이 다르다 보니 리뷰를 남기는 방식에도 차이가 있다. 따라서, 이러한 리뷰들을 사용자에게 한 번에 보여주고, 특성에 맞춰서 서로 다른 분석 결과를 추가적으로 제공하여 작품 판별에 대한 지표를 제공할 필요성이 있다.

(*³ 예를 들어, 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.)

3. 설계내용

ㄱ. Flow Chart

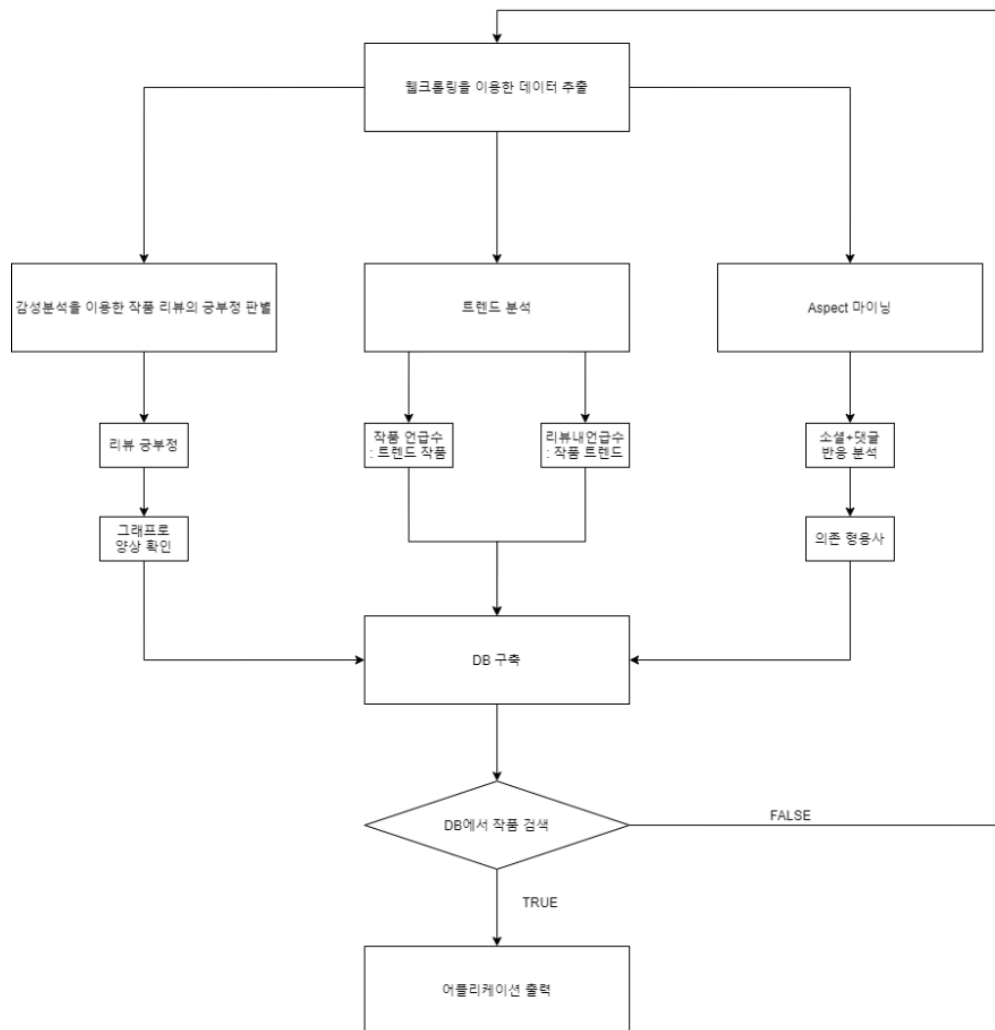


그림 5. 작품 설계

³ 출처 Facebook Internal Data (2016.09) / Instagram Internal Data (2016.09)

ㄴ. 개념 설계

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

- DB 를 구축하기 위해 해당 웹사이트 접속 (*리뷰 사이트, 블로그, 트위터, 인스타그램)
- 작품 기본 정보와 리뷰가 담긴 웹사이트 코드 분석
- BS4를 이용해 페이지 데이터 호출
- 작품의 기본 정보 tag 를 찾아 추출
- 각 사이트와 페이지별로 링크를 재귀적으로 검색하여 데이터를 추출
- Scraping 할 사이트 선정

플랫폼	비고
조아라 (프리미엄)	자유로운 연재 가능 정식 연재작 선정기준 필요
문피아 (유료 웹소설)	자유로운 연재 가능 정식 연재작 선정기준 필요
카카오페이지	리뷰 크롤링 불가
리디북스	
네이버 시리즈	화수별 리뷰가 전체 리뷰에 포함
네이버 웹소설	네이버 단독 연재작

표 1

SNS 및 커뮤니티	비고
네이버 블로그	리뷰 중심
티스토리	리뷰 중심
트위터	독백형 추천작으로 언급이 多
인스타그램	해시태그 이용한 검색만 가능 리뷰 중심
디씨인사이드	독백, 리뷰, 추천 多
인스티즈	추천, 독백, 리뷰 多

표 2

2. Kkma, Hannanum 을 이용한 KoNLP(키워드 생성)

- Kkma 나 Hannanum 모듈을 이용하여, 해당 모듈에 맞추어 입력된 문자열에서 키워드로 표현할 품사 추출하여 가장 빈도수가 높은 단어(키워드로 설정할 단어)를 DB 에 저장
- 오피니언 마이닝(감성 분석)을 이용하여 리뷰의 긍부정을 판단하고 전체적인 긍부정에 대한 평가를 진행한다. 긍부정 평가에 대한 점수를 기간에 대한 그래프로 표현하여 시간이 지남에 따라 작품의 평가 변화 양상을 확인
- 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 하나에 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 언어들 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 API 나 사전 구축에 대한 조사는 더 필요함.
- 참고 분석 예정 : word2vec, sentiwordnet, www.openhanquel.com, KTS
- 음소 단위 분할 조합을 통해 뜻을 파악하는 Trigram-Signature 를 사용하면 오타가 있는 "조ㅎ아", "실ㅎ어"와 같은 문맥을 파악하는 데 유용할 것이라고 생각한다. 이를 이용한 API 가 있는지 혹은 이와 관련된 사전을 제작하여 사용하는지 더 조사할 계획이다.
- word2vec 을 이용하여 Aspect 마이닝 분석을 한다. 이때 주인공, 작품, 분위기, 스토리 등 작품의 다양한 요소에 대해 의존적인 연결 형용사나 동사를 뽑으면서 작품 내 트렌드를 파악한다.
- matlab 을 이용한 데이터 분석을 통해 소셜미디어와 커뮤니티에서 작품의 언급수를 파악해보고 작품의 화제성을 판단한다.
- 위와 같은 방법으로 작품 내 명사들의 언급수를 통해 어떤 키워드가 화제가 되고 있는지 분석해본다.

3. Scraping 된 정보를 이용한 DB 구축(MySql or sqlite)

- MySQL 서버에 접속하여 데이터베이스 생성
- cursor 를 추출하여 execute 메서드로 SQL 을 실행, 테이블 생성
- Execute 메서드를 이용해 데이터를 지속적 확장

4. 웹사이트 UI 제작

자신이 보고 싶은 작품의 리뷰를 보기 위한 제품의 검색창을 사용자가 보기 편하도록 UI 로 구현한다.

결과물을 보여줄 플랫폼으로 안드로이드 어플과 웹사이트를 고민하였으나 사용자의 접근 등을 고려하여 최종적으로 웹사이트로 선정하였다.

검색 결과로 작품의 기본 정보와 함께 리뷰를 보여준다. 리뷰의 경우 긍정적 평가에 대한 그래프를 제공하여 시간이 지남에 따라 작품 평가 양상을 확인 가능하도록 하며 또한 작품 내 트렌드는 어떤 것인지 보여준다. 메인화면에는 현재 많이 화제 되고 있는 작품 Top 10을 선정하여 보여준다.

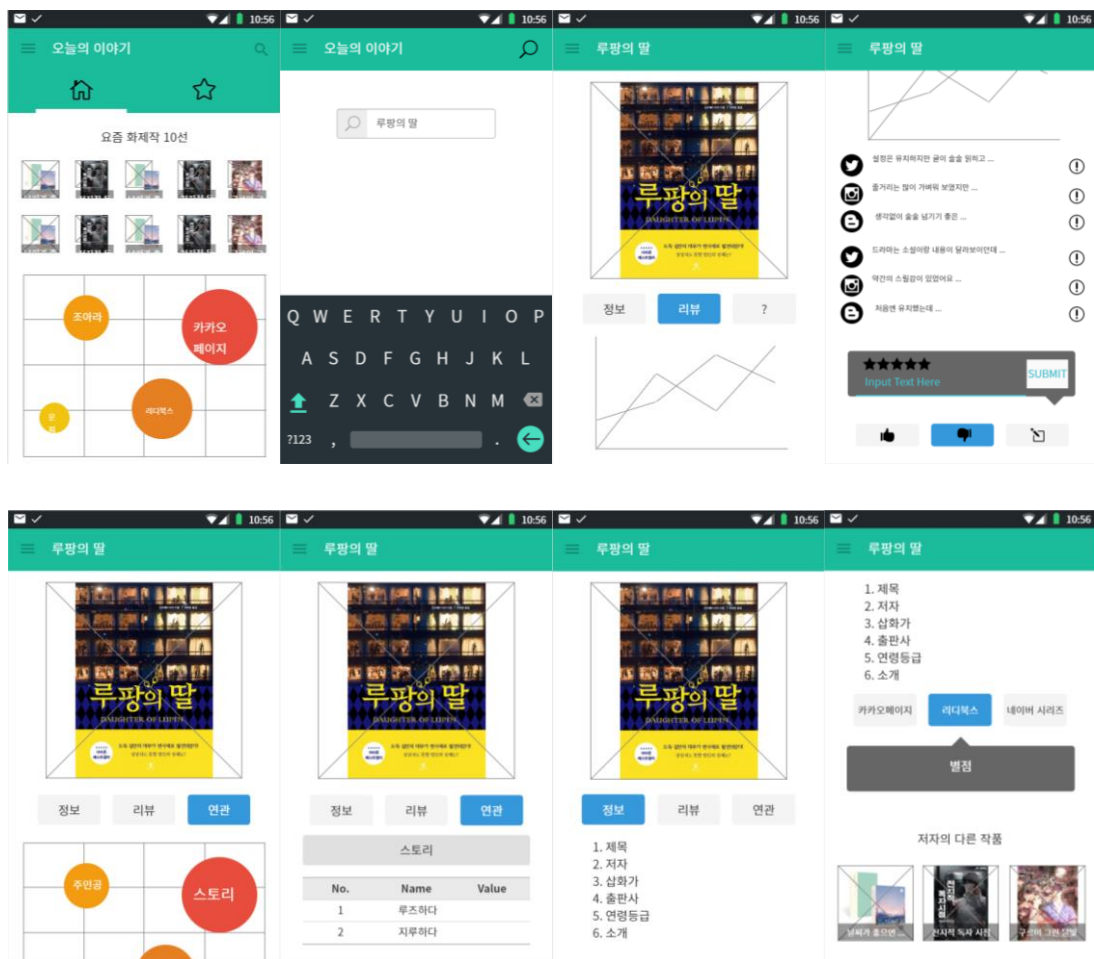


그림 6. 결과 예시

II. 과제 수행

1. 수행 일정

데이터 마이닝과 분석을 통한 웹소설 종합 인포 사이트

프로젝트 이름	데이터 마이닝과 분석을 통한 웹소설 종합 인포 사이트	회사명	DAWA 한승주 김성종 조예순
프로젝트 관리자	한승주 김성종 조예순	날짜	20년 5월 29일

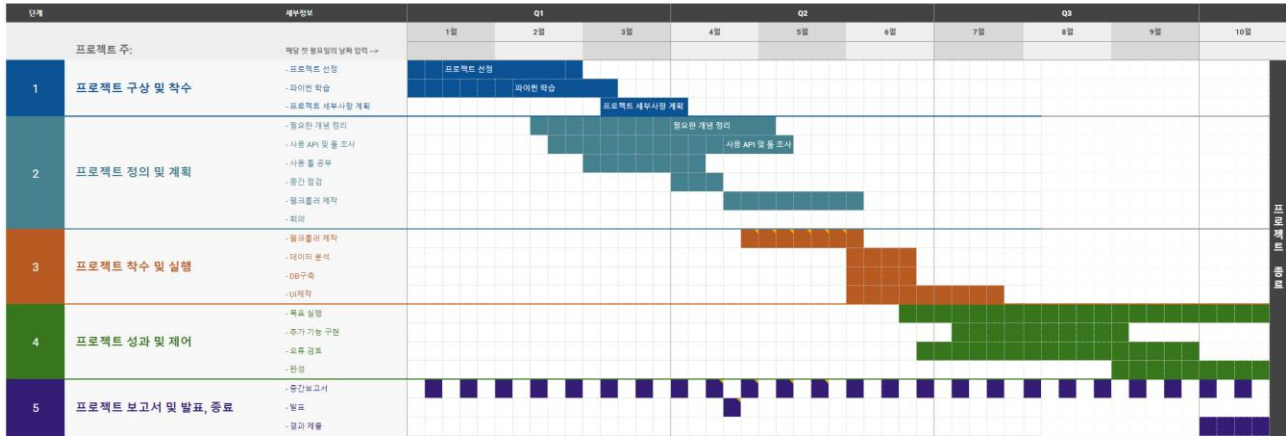


그림 7. 수행 일정 진행상황

2. Scraping

ㄱ. 플랫폼

Request 와 Selenium 의 경우 스크레이핑에 걸리는 시간이 확연하게 차이가 나므로, Request 를 활용하였다. 단, 카카오페이지의 작품 정보를 가져오는 부분은 페이지가 Java-Script 로 작성되어 동적 웹으로 구성되어 있기 때문에 Selenium 을 활용하였다.

1) 카카오페이지

```

import bs4
import requests
import urllib.request
from selenium import webdriver
import bs4
import requests
from urllib import parse
import time
import re

print("작품이름 -> ")
search_name=input()
tmp = search_name
search_name_nospace=search_name.replace(" ", "")

url = "https://page.kakao.com/search?word=" + parse.quote(tmp)

r=requests.get(url)
c=r.content
bs_obj = bs4.BeautifulSoup(c, "html.parser")

boxes = bs_obj.findAll("div", {"class": "css-c09e5i"})

for box in boxes:
    is_novel=box.find("div", {"class": "css-vurnku"}).text
    if("소설" in is_novel):
        parent=box.parent.parent
        title=parent.find("div", {"class": "text-ellipsis css-11602e0"}).text.replace(" ", "")
        if("단행본" in title):
            continue
        elif(search_name_nospace not in title):
            continue
        else:
            main=parent.parent.parent
            monopoly=True
    else:
        continue

title=main.find("div", {"class": "text-ellipsis css-11602e0"}).text

if '[' in title:
    print("작품이름 : " + title.split('[')[0])
    print("독점여부 : " + title.split('[')[1][:-1])
else:
    print("작품이름 : " + title)
    print("독점여부 : 미독점")

print("감상인원 : " + main.find('div', {'class', 'css-zlhhis'}).text)

pages=set()

```

그림 8. 카카오 작품 기본 정보 크롤링 - 1

```

title=main.find("div", {"class": "text-ellipsis css-11602e0"}).text

if '[' in title:
    print("작품이름 : " + title.split('[')[0])
    print("독점여부 : " + title.split('[')[1][:-1])
else:
    print("작품이름 : " + title)
    print("독점여부 : 미독점")

print("감상인원 : " + main.find('div',{'class','css-z1hhis'}).text)

pages=set()

for link in main.findAll("a", href = re.compile('^(/home)(?!:.)+$')):
    if 'href' in link.attrs: # 위에서 왔든 link에 href 속성이 있는지 확인
        if link.attrs['href'] not in pages: # 새로운 페이지인지 확인
            newPage = link.attrs['href']

#기다무소설 or 소설
#[단행본] X
#[악터최태수 or 악터 최태수 포함]

#작품의 링크
url="https://page.kakao.com" + newPage
driver2 = webdriver.Chrome("C:/chromedriver.exe")
driver2.get(url)

driver2.find_element_by_xpath('//*[@id="root"]/div[3]/div/div/div[1]/div[2]/div[2]/div[2]/button[1]').click()

title=driver2.find_element_by_xpath('//*[@id="root"]/div[3]/div/div/div[1]/div[2]/div[1]/h2')
print("제목 : " + title.text)

update_days = driver2.find_element_by_xpath('//*[@id="root"]/div[3]/div/div/div[1]/div[2]/div[2]/div[1]/p[1]')
print("연재일 : " + update_days.text)

writer=driver2.find_element_by_xpath('//*[@id="root"]/div[3]/div/div/div[1]/div[2]/div[2]/div[1]/p[2]')
print("작가 : " + writer.text)

genre=driver2.find_element_by_xpath('/html/body/div[2]/div/div/div/div[2]/div/div/table/tbody/tr[1]/td[2]/div[2]/div[2]')
print("장르 : " + genre.text)

age=driver2.find_element_by_xpath('/html/body/div[2]/div/div/div/div[2]/div/div/table/tbody/tr[1]/td[2]/div[4]/div[2]')
print("연령등급 : " + age.text)

publisher=driver2.find_element_by_xpath('/html/body/div[2]/div/div/div/div[2]/div/div/table/tbody/tr[1]/td[2]/div[3]/div[2]')
print("출판사 : " + publisher.text)

time.sleep(1)
driver2.quit()

```

그림 9. 카카오 작품 기본 정보 크롤링 - 2

작품이름 ->
 악터 최태수
 작품이름 : 악터 최태수
 독점여부 : 미독점
 감상인원 : 190.5만명
 제목 : 악터 최태수
 연재일 : 연재 중
 작가 : 조석호
 장르 : 기다무소설현판
 연령등급 : 전체미용가
 출판사 : 시나브로

그림 10. 카카오 작품 기본 정보 크롤링 - 추출 결과

2) 네이버 시리즈 (+ 네이버웹소설)

```
#!/usr/bin/env python
# coding: utf-8

# 네이버 웹소설

import requests
from bs4 import BeautifulSoup

req = requests.get('https://novel.naver.com/webnovel/list.nhn?novelId=699567')
source = req.content
soup = BeautifulSoup(source, 'html.parser')
#print(soup)

# 제목
container = soup.find('h2')
con = container.text
print("제목: " + con)

f_wri_ill = soup.find('p', {'class': 'writer'})
author = f_wri_ill.find('a', {'class': 'NPI=a:writer'})
aut = author.text
illustrator = f_wri_ill.find('a', {'class': 'NPI=a:illustrator'})
ill = illustrator.text
print("글: " + aut)
print("그림: " + ill)
```

그림 11. 네이버 웹소설 작품 기본 정보 크롤링 - 1

```
f_stargrade = soup.find('p', {'class': 'grade_area'})
ff_stargrade = f_stargrade.find('em')
stargrade = ff_stargrade.text
print("별점: " + stargrade)

info = soup.find('p', {'class': 'info_book'})
f_like = info.find('span', {'id': 'concernCount'})
like = f_like.text

f_publish = info.find('span', {'class': 'publish'})
publish = f_publish.text

f_genre = info.find('span', {'class': 'genre'})
genre = f_genre.text

print("관심: " + like)
print("연재일: " + publish)
print("장르: " + genre)

list_count = soup.find('span', {'class': 'total'})
count1 = list_count.text.replace("(", " ")
count = count1.replace(")", " ")
print("연재 화수: " + count)

f_dsc = soup.find('p', {'class': 'dsc'})
dsc = f_dsc.text
print("소개: " + dsc)
```

그림 12. 네이버 웹소설 작품 기본 정보 크롤링 - 2

```

##### 검색 #####
search = "https://novel.naver.com/search.nhn?keyword="
search_for = input("검색: ")
print("\n-----\n")
web_search = search + search_for.replace(" ", "+")

req = requests.get(web_search)
source = req.text
soup = BeautifulSoup(source, 'html.parser')

link = soup.find('ul', {'class', 'list_type2 v3'})
href = link.find('a').attrs['href']

novel_url = "https://novel.naver.com" + href

req = requests.get(novel_url)
source = req.text
soup = BeautifulSoup(source, 'html.parser')

##### 제목 #####
container = soup.find('h2')
con = container.text
print("제목: " + con)

##### 글, 그림 #####
f_wri_ill = soup.find('p', {'class', 'writer'})
author = f_wri_ill.find('a', {'class', 'NPI=a:writer'})
aut = author.text
illustrator = f_wri_ill.find('a', {'class', 'NPI=a:illustrator'})
ill = illustrator.text
print("글: " + aut)
print("그림: " + ill)

##### 별점 #####
f_stargrade = soup.find('p', {'class', 'grade_area'})
ff_stargrade = f_stargrade.find('em')
stargrade = ff_stargrade.text
print("별점: " + stargrade)

##### 관심, 연재일, 장르 #####
info = soup.find('p', {'class', 'info_book'})
f_like = info.find('span', {'id': 'concernCount'})
like = f_like.text

```

그림 13. 네이버 웹소설 작품 기본 정보 크롤링 - 3

제목: 검은 늑대가 나를 부르면
글: 임혜
插圖: 홍복
별점: 9.96
관심: 23,039
연재일: 화, 금 연재
장르: 로맨스
연재 화수: 8
소개: 첫 번째 삶은 남편의 손에 죽임을 당했고, 두 번째 삶은 가족을 몰살한 남편 앞에서 자살했다. 하지만 그것은 마지막이 아닌 또 다른 시작이었다. 과거로 돌아가 세 번째 삶을 살게 된 연우. 그녀 앞에 나타난 검은 늑대 휘타. 가족과 제 목숨을 지키고 싶었던 연우는 휘타에게 자신을 맡기게 되는데……. 사랑하는 사람을 위해 영혼마저 팔아버린 그들의 가슴 시리도록 아픈 사랑 이야기가 펼쳐집니다.

그림 14. 네이버 웹소설 작품 기본 정보 크롤링 - 추출 결과

검색: 검은 늑대가 나를 부르면

제목: 검은 늑대가 나를 부르면

글: 임해

그림: 홍복

별점: 9.96

관심: 23,025

연재일: 화, 금

장르: 로맨

연재 화수: 8

소개: 첫 번째 삶은 남편의 손에 죽임을 당했고, 두 번째 삶은 가족을 몰살한 남편 앞에서 자살했다. 하지만 그것은 마지막이 아닌 또 다른 시작이었다. 과거로 돌아가 세 번째 삶을 살게 된 연우. 그녀 앞에 나타난 검은 늑대 휘타. 가족과 제 목숨을 지키고 싶었던 연우는 휘타에게 자신을 맡기게 되는데... 사랑하는 사람을 위해 영혼마저 팔아버린 그들의 가슴 시리도록 아픈 사랑 이야기가 펼쳐집니다.

Process finished with exit code 0

그림 15. 네이버 웹소설 작품 기본 정보 크롤링 - 추출 결과2

```
import requests
from bs4 import BeautifulSoup

req = requests.get('https://series.naver.com/novel/detail.nhn?productNo=32000031')
source = req.text
soup = BeautifulSoup(source, 'html.parser')

f_container = soup.find('div', {'class', 'end_head'})
container = f_container.find('h2')
con = container.text.split(" ")[0]
print("제목: " + con)
count = container.find('em')
count1 = count.text.replace("- 출", "")
count2 = count1.replace("화/", "")
if '미완결' in count2:
    count3 = count2.replace("미완결", "")
else:
    count3 = count2.replace("완결", "")
print("화수: " + count3)
```

그림 16. 네이버 웹소설 작품 기본 정보 크롤링 - 시리즈 1

```
box = soup.find(id='container')
info_list = box.find('li', {'class', 'info_list'})
li = info_list.find('li')
print("저자: " + li[0].find('a').text)
if "그림" in li[1].find('span'):
    n = 1
else:
    n = 0
if n == 1:
    print("그림: " + li[1].find('a').text)
print("장르: " + li[1+n].find('a').text)
print("출판사: " + li[2+n].find('a').text)
print("등급: " + li[3+n].text.replace("등급 ", ""))
update = li[4+n].text.replace("업데이트 ", "")
if n == 1:
    print("최근 업데이트: ", update.replace("(미완결)", ""))
else:
    print("최근 업데이트: ", update.replace("(미완결)", ""))
if n == 1:
    print("완결여부: 완결")
else:
    print("완결여부: 미완결")

f_scorearea = soup.find('div', {'class', 'score_area'})
ff_scorearea = f_scorearea.find('em')
scorearea = ff_scorearea.text
print("별점: " + scorearea)

f_dsc = soup.find('div', {'class', '_synopsis'})
print("소개: " + f_dsc.text)
```

그림 17. 네이버 웹소설 작품 기본 정보 크롤링 - 시리즈 2

```
##### 검색 #####
search = "https://series.naver.com/search/search.nhn?t=all&fs=novel&q="
search_for = input("검색: ")
print("\n-----\n")
web_search = search + search_for.replace(" ", "+")

req = requests.get(web_search)
source = req.text
soup = BeautifulSoup(source, 'html.parser')

link = soup.find('ul', {'class', 'lst_list'})
href = link.find('a').attrs['href']

novel_url = "https://series.naver.com/" + href

req = requests.get(novel_url)
source = req.text
soup = BeautifulSoup(source, 'html.parser')
```

그림 18. 네이버 웹소설 작품 기본 정보 크롤링 - 시리즈 3

```
제목: 입술이 너무해
화수: 106
저자: 갯너
그림: RM
장르: 로맨스
출판사: 와이앰블록스
등급: 전체 이용가
최근 업데이트: 2018.09.21.
완결여부: 완결
별점: 9.8
소개: "키스하면 변해?" 남자가 되는 병에 걸린 지 7년, 그 남자와의 하룻밤은 서연을 다시 여자로 만들었다. 머리카락은 길어지고, 가슴은 볼록해지고, 입술은 더욱더 새빨갳게 피어났다. "예쁘네요." "네?" "입술 예쁘네." 설령 대폭발, 심층 주의, 치명적으로 썩어 한 작진남의 줄 떨어지는 막강 들이음이 시작했다. 운명에 얽힌 세계 최고 달달한 씬씬과 더 달달한 끈적끈적 연애! 예측불허 입술 로맨스.
```

그림 19. 네이버 웹소설 작품 기본 정보 크롤링 - 추출 결과

```
검색: 검은 사슬
-----

제목: 검은 사슬
화수: 86
저자: 유세라
장르: 로판
출판사: n.fic
등급: 12세 이용가
최근 업데이트: 2020.05.15.
완결여부: 미완결
별점: 9.0
죽겠다 했는데, 이상하게 죽어지지가 않..
```

```
Process finished with exit code 0
```

그림 20. 네이버 웹소설 작품 기본 정보 크롤링 - 추출 결과 2

3) 조아라

```
import requests
from bs4 import BeautifulSoup

req=requests.get('http://www.joara.com/premium_new/book_intro.html?book_code=1360670')
source=req.text
soup=BeautifulSoup(source, 'html.parser')

# 제목
title_list=soup.select("#premium_content > div.latest > div.best_list_list_wrap > div.box sty01 > div > div > div.txt_c sty01 > str
title=title_list[0].text

# 저자
author_list=soup.select("#premium_content > div.latest > div.best_list_list_wrap > div.box sty01 > div > div > div.txt_c sty01 > st
author=author_list[0].text

# 총 연재화수 - 점수
book_count_list=soup.select("#premium_content > div.latest > div.best_list_list_wrap > div.box sty01 > div > div > div.txt_c sty01
book_count=book_count_list[0].text
book_count=book_count.replace("만", "")
book_count=int(book_count)

# 완결 여부
complete_list=soup.select("#premium_content > div.latest > div.best_list_list_wrap > div.box sty01 > div > div > div.txt_c sty01 >
if complete_list[0].text=="마지막 연재일":
    complete='미완결'
else:
    complete='완결'

# 출간일
published_date_list=soup.find("table")
published_date_list=published_date_list.find_all("tr")
published_date_list=published_date_list[book_count]
published_date_list=published_date_list.find_all("a")
published_date=published_date_list[2].text

# 최근 업데이트일
recent_update_list=soup.select("#premium_content > div.latest > div.best_list_list_wrap > div.box sty01 > div > div > div.txt_c sty
update=recent_update_list[0].text
update=update[:10]+"..."

# 표지 사진 url
img=soup.find("div")
img=soup.find(class_="img_s")
img=img.find("img")
img_src=img.get("src")
```

그림 21. 조아라 작품 기본 정보 크롤링 - 1

```
# 장르
category_list=soup.select("#premium_content > div.latest > div.best_list_list_wrap > div.box sty01 > div > div > div.txt_c sty01 >
category=category_list[0].text
category=category.split(' ')[0].replace("[", "")
category=category.replace("]", "")

# 조회수 - 점수
count_list=soup.select("#premium_content > div.latest > div.best_list_list_wrap > div.box sty01 > div > div > div.txt_c sty01 > div
views=count_list[0].text
views=views.replace(", ", "")
views=int(views)

# 추천수 - 점수
recommend=count_list[2].text
recommend=recommend.replace(", ", "")
recommend=int(recommend)

# 관심도 - 점수
preference=count_list[4].text
preference=preference.replace(", ", "")
preference=int(preference)

# 소개
introduction_list=soup.select("#premium_content > div.latest > div.best_list_list_wrap > div.t_cont_v")
introduction=introduction_list[0].text
introduction=introduction.replace("<br>", "")
introduction=introduction.replace("<hr>", "")
introduction=introduction.replace("<hr>", "")

##### 결과 출력 #####

print("제목: " + title)
print("장르: " + category)
print("저자: " + author)
print("표지 url: " + img_src)
print("총 연재화수: ", book_count)
print("최근 업데이트일: " + update)
print("출간일: " + published_date)
print("완결 여부: " + complete)
print("조회수: ", views)
print("추천수: ", recommend)
print("관심도: ", like)
print("소개: " + introduction)
```

그림 22. 조아라 작품 기본 정보 크롤링 - 2

제목: 무한서고의 계약자!
 장르: 판타지
 저자: 준술
 표지 url: http://cf.joara.com/literature_file/20190418_121420.jpg_thumb.png
 총 연재 화수: 220
 최근 업데이트일: 2019.06.04.
 출간일: 19/04/03
 완결 여부: 완결
 조회수: 613218
 추천수: 3965
 관심도: 724
 소개:
 스릴북을 포함한 모든 서적들이 기록되어 있는 무한서고.
 나는 그런 무한서고를 열람할 수 있는 권능을 얻게 되는데

그림 23. 조아라 작품 기본 정보 크롤링 - 추출 결과

```

joara_base = "http://www.joara.com/search/search.html?sl_search=book&sl_keyword="
joara_work = input("작품 검색 : ")
print("\n-----\n")
joara_search = joara_base + joara_work.replace(" ", "+")

req = requests.get(joara_search)
source = req.text
soup = BeautifulSoup(source, 'html.parser')

url_list = soup.select("#content > div.view > div.layout > div.series")
check = 0
for i in url_list:
    # 프리미엄 카테고리 검색
    if (i.find("h3").text.find("프리미엄") >= 0):
        joara_list = i.find("div")
        joara_list = joara_list.find("a")
        joara_list = joara_list.get("href")
        joara_workID = joara_list.split('=')[1]

        # url 링크 추출
        joara_url = "http://www.joara.com/premium_new/book_intro.html?book_code=" + joara_workID
        get_info(joara_url)
        check = 1
        break;

if check == 0:
    print("검색 결과가 없습니다.")
else:
    print("\n-----\n검색 결과 출력 완료")
  
```

그림 24. 조아라 작품 기본 정보 크롤링 - 3

작품 검색 : 소설 속 엑스트라

제목: 소설 속 엑스트라
장르: 판타지
저자: [지강송]
표지 url: http://cf.joara.com/literature_file/20181228_161523.jpg_thumb.png
총 연재화수: 432
최근 업데이트일: 2020.05.15.
출간일: 18/08/02
완결 여부: 미완결
조회수: 1205484
추천수: 14652
관심도: 4060
소개:
소설은 하나의 세계와 수십억의 등장인물이 존재한다.
하지만 히로인이나 조력자 같은
'비중' 있는 조역이라면 몰라도
그 외의 모두에게 이름이 있을 리는 없다.

"훈동아 너는 몇 위야?"

나는 나를 모른다. 이름이 왜 훈동인지도 모르겠다.

이 세상은 내가 쓴 소설.
그러나 나는 내가 단 한 번도 쓰지 않은 인물이 되어 있다.

요원사관학교에 입학했다는 것 말고는 평범하기 그지없는,
소설 속 그 누구와도 접점이 없는,
소설의 지면 그 어디에도 이름이 적히지 않을 그런 인물.

그러니까, 나는 소설 속 엑스트라가 되었다.
.....아니, 소설 속 먼지가 되었다.

[소설 속 엑스트라]

검색 결과 출력 완료

그림 25. 조아라 작품 기본 정보 크롤링 - 추출 결과

4) 문피아

```
import bs4
import requests

r=requests.get("https://novel.munpia.com/11799")
c=r.content
html = bs4.BeautifulSoup(c, "html.parser")

box=html.find('div',{'class','dd detail-box'})
h2=box.find('h2')
monopoly=h2.find('span')
if(monopoly==None):
    print("독점여부 : 미독점")
    title=h2.text.replace("\n","")
    print("제목 : " + title)
    #print("제목 : " + title,end='')
else:
    print("독점여부 : " + monopoly.text)
    title=h2.text.split("독점")[1]
    print("제목 : " + title,end='')

p=box.find('p',{'class','meta-path'})
genre=p.find('strong').text
print(genre)
if '.' in genre:
    print("장르 : " + genre.split('.')[0] + ", " + genre.split('.')[1])
else:
    print("장르 : " + genre)

days = dict.fromkeys(["mon", "tue", "wed", "thu", "fri", "sat", "sun"], False)
novel_period=box.find('div',{'class','novel-period'})
period=""
if novel_period!=None:
    if(novel_period.find('span',{'class','pluszone-mon'})):
        days.update(dict.fromkeys(["mon"], True))
        period=period+"월,"
    if(novel_period.find('span',{'class','pluszone-tue'})):
        days.update(dict.fromkeys(["tue"], True))
        period=period+"화,"
    if(novel_period.find('span',{'class','pluszone-wed'})):
        days.update(dict.fromkeys(["wed"], True))
        period=period+"수,"
    if(novel_period.find('span',{'class','pluszone-thu'})):
        days.update(dict.fromkeys(["thu"], True))
        period=period+"목,"
    if(novel_period.find('span',{'class','pluszone-fri'})):
        days.update(dict.fromkeys(["fri"], True))
        period=period+"금,"
    if(novel_period.find('span',{'class','pluszone-sat'})):
        days.update(dict.fromkeys(["sat"], True))
        period=period+"토,"
    if(novel_period.find('span',{'class','pluszone-sun'})):
        days.update(dict.fromkeys(["sun"], True))
        period=period+"일,"
```

그림 26. 문피아 작품 기본 정보 크롤링 - 1


```

# 별점 점수
star_rate_list=soup.select("#page_detail > div.detail_wrap > div.detail_body_wrap > section > article.detail_header.trackable > div
star_rate=star_rate_list[0].text

# 별점 참여자
star_rate_count_list=soup.select("#page_detail > div.detail_wrap > div.detail_body_wrap > section > article.detail_header.trackable
star_rate_count=star_rate_count_list[0].text

# 관심도
preference_list=soup.select("#page_detail > div.detail_wrap > div.detail_body_wrap > section > article.detail_header.trackable > di
preference=preference_list[0].text

# 책소개
introduction_list=soup.select("#introduce_book")
introduction=introduction_list[2].text
introduction=introduction.replace("<[연재] "+title+"> ", "")
introduction=introduction.replace("¶", "")
introduction=introduction.replace("책 소개", "")
introduction=introduction.replace("펼쳐보기 ", "")

##### 결과 출력 #####

print("제목: " + title)
print("저자: " + author)
print("출판사: " + publisher)
print("표지 url: " + img_src)
print("총 연재화수: ", book_count)
print("완결 여부: " + complete)
print("출간일: " + published_date)
print("최근 업데이트일: " + update)
print("장르: " + category)
print("별점: " + star_rate)
print("별점 참여자: " + star_rate_count)
print("관심도: " + preference)
print("소개: " + introduction)

```

그림 30. 리디북스 작품 기본 정보 크롤링 - 2

```

제목: 백작가의 말나기가 되었다
저자: 유령한
출판사: 도서출판 청어람
표지 url: //img.ridicdn.net/cover/875125819/xxlarge
총 연재화수: 568
완결 여부: 미완결
출간일: 2018.10.02
최근 업데이트일: 2020.04.24
장르: 휴전 판타지
별점: 4.8점
별점 참여자: 3,900명
관심도: 0
소개:
눈을 떠보니 소설 속이었다.
그것도 말나기로 유명한 백작가 도련님 몸으로.

하지만,
그렇다고 말나기가 될 순 없잖아?

```

그림 31. 리디북스 작품 기본 정보 크롤링 - 추출 결과

```

import requests
from bs4 import BeautifulSoup

def get_info(url, book_id):
    print(url)
    print("리디북스 ID no. " + book_id)
    print("\n-----\n")
    req = requests.get(ridi_url)
    source = req.text
    soup = BeautifulSoup(source, 'html.parser')

```

그림 32. 리디북스 작품 기본 정보 크롤링 - 3

```

ridi_base = "https://ridibooks.com/search?q="
ridi_work = input("작품 검색 : ")
print("\n-----\n")
ridi_search = ridi_base + ridi_work.replace(" ", "+")

req = requests.get(ridi_search)
source = req.text
soup = BeautifulSoup(source, 'html.parser')

url_list = soup.select(
    "#page_search_result > div.result_list_wrapper > article > div.book_macro_wrapper.js_book_macro_wrapper > div")
check = 0
for i in url_list:
    if (i.text.find("연재") > 0):
        ridi_bookID = i.find(class_="book_thumbnail_wrapper")
        ridi_bookID = ridi_bookID.get("data-book_id_for_tracking")
        ridi_url = "https://ridibooks.com/books/" + ridi_bookID
        get_info(ridi_url, ridi_bookID)
        check = 1
        # print(book_id)
        break;

if check == 0:
    print("검색 결과가 없습니다.")
else:
    print("\n-----\n검색 결과 출력 완료")

```

그림 33. 리디북스 작품 기본 정보 크롤링 - 4

작품 검색 : 내게 북중하세요

<https://ridibooks.com/books/3049006830>
리디북스 ID no. 3049006830

제목: 내게 북중하세요
저자: 권우
출판사: 북컴즈
표지 url: //img.ridicdn.net/cover/3049008246/oxlarge
총 연재화수: 122
연결 여부: 미완료
출간일: 2020.02.03
최근 업데이트일: 2020.05.08
장르: 판타지물
분량: 4.6권
별첨 한페이지: 1,515명
관심도: 0
소개: 황세자에게 일방적인 파혼을 당하고,
자속 차 온 여탈길에서 자유를 만끽하던 그림.

[안녕.]

그것, 아니, 그를 깨워 버렸다.

[나는 니타니엘.]

그가 말했다.

[여기 사람들은 나를 ‘종말’이라 부르더구나.]

겨울의 왕 같은 아름다운 남자가, 권태롭고 오만하게 미소 지었다.

“나에게 해 줘요.”

니타니엘이 손을 뻗었다. 커리에가 그것을 뿌리쳤으나, 뺨가 도드라진 흰 손은 오히려 더 느리고 부드럽게, 커리에의 귀와 뺨 근처를 머무르렀다.

[그런 생각은 하지 않는 게 좋을걸. 뭐든 지내고 싶지 않다면.]

대답 대신, 커리에의 이가 니타니엘의 손가락을 깨물었다.
그는 천천히 고개를 숙여, 커리에와 이마를 맞았다. 코끝의 푸른 눈은 커리에의 보라색 눈동자가 불안에 흔들릴수록 더 활활에 휘하는 것 같았다.

[웃은 일어서 벗도록.]

니타니엘이 엉망이 된 자신의 쇼츠자락을 내려다보며 사납게 미소지었다.

[또 허튼것하면 목줄을 채울 줄 알아.]

그림 34. 리디북스 작품 기본 정보 크롤링 - 추출 결과

ㄴ. SNS

1) 트위터

트위터의 특징은 불특정 다수가 독백을 하는 경향이 있다. 불특정 다수의 독백을 통해 작품의 언급수는 자연스레 현재 작품의 화제성을 대변해줄 수 있다. 이를 통해 트렌드 작품을 파악하기 좋을 것이라고 판단했다.

트위터 크롤링을 위한 API 는 여러 가지가 존재한다. 트위터의 정식 API 는 키 신청을 하면 이용이 가능하지만 최근 7일간의 트윗만 확인이 가능하다. 이외에도 현재 많이 사용되고 있는 사설 API GetOldTweets3⁴, twitterscraper⁵ 등 여러 가지가 있는데 이를 이용하여 이전 정보를 가져오기 위해서는 사설 API 를 이후 계속되는 최신 정보를 가져오기 위해서는 Tweepy 를 사용할 예정이다.

사용한 API getoldtweet3은 기간 설정이 가능하다. 2020년 5월 1일부터 4일까지, 약 4일 간 조건에 맞게, 즉, 해당 키워드를 포함해 작성된 트윗들을 수집하였다.

```
import GetOldTweets3 as got
import datetime
import time
from random import uniform
from tqdm import tqdm notebook
import pandas as pd

# 트윗 수집 기간
days_range = []

starting_date = input("트윗 수집 시작일 [ex. 양식 : 2010-07-23] : ")
ending_date = input("트윗 수집 마지막일 [ex. 양식 : 2015-07-23] : ")

start = datetime.datetime.strptime(starting_date, "%Y-%m-%d")
end = datetime.datetime.strptime(ending_date, "%Y-%m-%d")
date_generated = [start + datetime.timedelta(days=x) for x in range(0, (end - start).days)]

for date in date_generated:
    days_range.append(date.strftime("%Y-%m-%d"))

print("=== 설정된 트윗 수집 기간은 {} 에서 {} 까지 입니다 ===".format(days_range[0], days_range[-1]))
print("=== 총 {}일 간의 데이터 수집 ===\n".format(len(days_range)))

# 수집 기간 맞추기
start_date = days_range[0]
end_date = (datetime.datetime.strptime(days_range[-1], "%Y-%m-%d") + datetime.timedelta(days=1)).strftime("%Y-%m-%d")

# 작품 검색
search_key = input("작품 검색 #")
print()

# 트윗 수집 기준 정의
tweetCriteria = got.manager.TweetCriteria().setQuerySearch(search_key).setSince(start_date).setUntil(
    end_date).setMaxTweets(-1)
```

그림 35. 트위터 작품 검색 내용 추출

⁴ <https://github.com/Jefferson-Henrique/GetOldTweets-python>

⁵ <https://github.com/taspinar/twitterscraper>

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

```
# 트윗 수집
print("트윗 수집 시작")
start_time = time.time()

tweet = got.manager.TweetManager.getTweets(tweetCriteria)

print("트윗 수집 완료 [{0:0.2f} Minutes]".format((time.time() - start_time) / 60))
print("=== 수집 트윗 총 {}개 ===".format(len(tweet)))

# initialize
tweet_list = []

for index in tqdm_notebook(tweet):
    # 데이터 목록
    username = index.username
    link = index.permalink
    content = index.text
    tweet_date = index.date.strftime("%Y-%m-%d")

    info_list = [tweet_date, username, content, link]
    tweet_list.append(info_list)

    time.sleep(uniform(1, 2))

tw_df = pd.DataFrame(tweet_list, columns=["date", "user_name", "text", "link"])

# csv 파일 만들기
tw_df.to_csv("{}_tw.csv".format(search_key), index=False, encoding='utf-8')
print("저장 완료.\n")

## In[59]:

## 파일 확인하기
df_tweet = pd.read_csv("{}전지적독자시점_tw.csv".format(search_key))
df_tweet
```

그림 36. 트위터 작품 검색 내용 추출 내용 저장


```

=== 총 4일 간의 데이터 수집 ===

작품 검색 #전지적독자시점

트윗 수집 시작
트윗 수집 완료 [0.18 Minutes]
=== 수집 트윗 총 120개 ===

HBox(children=(IntProgress(value=0, max=120), HTML(value='')))

저장 완료.

In [59]: # 파일 확인하기
df_tweet = pd.read_csv("전지적독자시점_tw.csv".format(search_key))
df_tweet

Out [59]:

```

	date	user_name	text	link
0	2020-05-04	autumn_teatime	沈清秋你没事吧? #전지적_독자_시점 #백작가의_망나니가_되었다 #人渣反派自救系统	https://twitter.com/autumn_teatime/status/1257...
1	2020-05-04	R_LA_	[전지적 독자 시점] 한수영_리아님 @kangela1117 어린 독자_R_LA_ 갓...	https://twitter.com/R_LA_/status/125732622727...
2	2020-05-04	ripi32478	오소마츠상 사이파즈 헬프튜 러브엔프로듀서 병드림 음악사 페이트 패그오 패스나 fat...	https://twitter.com/ripi32478/status/125732364...
3	2020-05-04	kangela1117	전지적 독자 시점 - 5부 한수영 - 리아 (@kangela1117) 어린 독자 ~...	https://twitter.com/kangela1117/status/1257322...
4	2020-05-04	cooing_o91o	유중혁씨 찾아요!! 중혁이 오면 알티 이변? 김티 뭐라도 올릴테니 상위 트윗 알티부...	https://twitter.com/cooing_o91o/status/1257321...
...
115	2020-05-01	silversphere_	전지적 독자 시점 108화 이 부분이 앞쪽 회차들 중에선 제일 웃겼던 기억ㅠ #도겔쓰다	https://twitter.com/silversphere_/status/12560...
116	2020-05-01	silversphere_	전지적 독자 시점 107화 #도겔쓰다	https://twitter.com/silversphere_/status/12560...
117	2020-05-01	silversphere_	전지적 독자 시점 97화 제일 좋아하는 탑3 안에 드는 대사들이예요 특히 서슬! 유...	https://twitter.com/silversphere_/status/12560...
118	2020-05-01	silversphere_	전지적 독자 시점 76, 88화 #도겔쓰다 #전독시	https://twitter.com/silversphere_/status/12560...
119	2020-05-01	wtf5cx	전지적 독자 시점 단행본 헬통 외전 양장본 설정집 비나이다 비나이다 공식 먹방 5월...	https://twitter.com/wtf5cx/status/125602080460...

120 rows × 4 columns

그림 37. 트위터 크롤링 결과

그림 36을 통해 저장한 파일을 확인하면 그림 37과 같은 결과를 얻을 수 있다.

2) 인스타그램

인스타그램은 트위터와 같은 소셜미디어보다는 작품에 대한 얘기를 많이 하지만 네이버나 티스토리보다는 단순한 리뷰를 중심으로 많이 얘기한다. 이를 토대로 트렌드 지표를 알아보기 위한 언급수에 분석 결과를 쓸 것이다.

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.

```

def InstagramUrlFromKeyword(browser, keyword, numofpage):
    keyword_url_encode = quote(keyword) # 한글인식

    url = 'https://www.instagram.com/explore/tags/' + keyword_url_encode + '/?hl=ko'

    browser.get(url)

    arr_href = []

    body = browser.find_element_by_tag_name('body')

    for i in range(numofpage):
        body.send_keys(Keys.PAGE_DOWN)

        time.sleep(1)

    time.sleep(3)

    post = browser.find_elements_by_class_name('v1Nh3')

    for j in post:
        href_str = j.find_element_by_css_selector('a').get_attribute('href')

        arr_href.append(href_str) # append 추가시키는거

    return arr_href

```

그림 38. 인스타그램 크롤링

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

```

▼<div class="v1Nh3 kIKUG _bz0w">
  ▼<a href="/p/B1Fh3MvhY1J/"> == $0
    ▼<div class="eLAPa"> https://www.instagram.com/p/B1Fh3MvhY1J/

```

그림 39. 인스타그램 크롤링

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.

```

def IdHashTagFromInstagram(browser, url):
    browser.get(url)

    insta_id = ""

    hash_data = ""

    wait = WebDriverWait(browser, 20)

    wait.until(EC.presence_of_element_located((By.CLASS_NAME, "e1e1d"))) # e1e1d는 아이디가 적혀있는 소스코드

    id_href = browser.find_elements_by_class_name('e1e1d')

    insta_id = id_href[0].find_element_by_css_selector('a').text # id_href[0] 첫번째 있는 a 찾기

    wait.until(EC.presence_of_element_located((By.CLASS_NAME, "c4VMK"))) # 댓글들

    href = browser.find_elements_by_class_name('c4VMK')

    total_hash_text = []

```

그림 40. 인스타그램 크롤링 - 해시태그

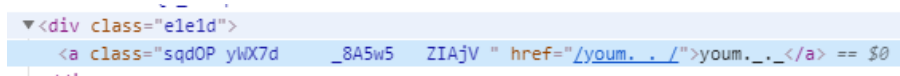


그림 41. 인스타그램 크롤링 - ID

아이디를 찾는 부분으로, e1e1d 클래스 영역에 href로 text만 뽑아내면 해당 부분이 인스타그램 id이므로 이를 추출한다.

```

for i in range(0, len(href)): # 댓글 가져와서 하나씩 끝까지 보는 거 len 몇개 개수

    hash_text = href[i].find_element_by_css_selector('span').text

    total_hash_text.append(hash_text)

    image_src = ''

try:

    image_temp = browser.find_element_by_class_name('KL4Bh').find_element_by_css_selector('img') # 이미지 찾기

    image_list = image_temp.get_attribute('srcset') # srcset이란 속성을 가지고 있는 애를 가져와라

    temp = image_list.split(',') # ,로 구분해서 temp로 가져와라

    for i in temp:

        if '1080w' in i: # 사진의 많은 url중에서 1080w있는 문자열 찾기

            image_src = i.split(' ')[0] # url 1080w이 있는 링크에서 1080w를 떼고 공백 앞의 정보를 가져오기

except:

    image_src = '' # 동영상이면(이미지가 아니면) 빈칸으로 뒤라

    pass

return insta_id, image_src, total_hash_text

```

그림 42. 인스타그램 크롤링 내용 추출

그림 2-27의 코드는 for 문을 이용하여 참조할 reference 가 있는 동안 해당 ref 의 본문을 축적하는 부분과 본문에 덧붙인 이미지의 src 를 찾는 부분으로 구성이 되어있다.

```
▼<span class title="수정됨"> == $0
  "저한테 로지텍 핑크 k380이 있지만"
  <br>
  "늘 화이트색상의 키보드를 가지고 싶다는 생각이 마음 저 구석에 있는데, 드디어 저도 화이트 키보드와 마우스를 하나 더 가지게 되었어요!"
  <br>
  <br>
  "제 책상과도 아주 찰떡인 k580과 m350"
  <br>
```

그림 43. 인스타그램 크롤링 - 내용

본문의 부분으로 댓글 또한 위와 같이 span 의 영역에 글이 작성되어 있다.

```
▼<div class="KL4Bh" style="padding-bottom: 100%;">
  
</div>
```

그림 44. 인스타그램 크롤링 - 댓글

Img_src 또한 KL4Bh 의 영역에 img 를 find 하여 찾을 수 있다.

```
browser = webdriver.Chrome('C:\chromedriver.exe')

keyword = input("검색어를 입력하세요 : ")

num_of_pages = 2

arr = InstagramUrlFromKeyword(browser, keyword, num_of_pages)

insta_df = pd.DataFrame(columns=['Insta ID', 'Image Src', 'Content'])

for url in arr:
```

그림 45. 인스타그램 크롤링 - 이미지

```
try:
    insta_id, image_src, hash_data = IdHashTagFromInstagram(browser, url)

    char = re.compile('[^0-9a-zㄱ-ㅣ-가-힣!#?]*') # 재정의
    """
    정규식을 두 번 이상 사용하면, 모듈의 match, search 함수는 효율적이지 않다.
    매번 match 혹은 search를 수행할 때마다, 정규식을 분석해서 처리하기 때문이다.
    효과적인 처리 방법은 정규식을 내부 표현식으로 일단 변환하고, 그것을 계속 활용하는 것이다.
    compile 함수가 정규식을 내부 표현식으로 변환하여 정규식 객체를 리턴한다.
    """
    hash_data_str = ""
```

그림 46. 인스타그램 크롤링

```
for data in hash_data:
    hash_data_str = hash_data_str + data

hash_data_str = char.sub("", hash_data_str) # ""를 hash_data_str으로 바꿔주기

dic_insta = {"Insta ID": insta_id, "Image Src": image_src, "Content": hash_data_str}

temp_df = pd.DataFrame(dic_insta, index=[0]) # index=0은 dic을 temp로 바꾸는데 여려가 나지 않도록 하는 것

insta_df = insta_df.append(temp_df, ignore_index=True)

except:

    print(sys.exc_info()[0])

    pass

insta_df.to_csv('insta_temp.csv', mode='w', encoding='euc-kr')
```

그림 47. 인스타그램 클롤링

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

	포스트 URL	유저ID	내용	작성 날짜
0	https://www.instagram.com/p/B-00H9PHaS-/	diana.pontin	리셋팅 레이디. 이셀라 예반스 낙서사실 읽은지는 좀 돼서 가들가들하기 때문에 외모모...	2020-04-11
1	https://www.instagram.com/p/B7ivmF3Jr6a/	choinuri17	리셋팅 레이디. 이셀라 예반스 낙서사실 읽은지는 좀 돼서 가들가들하기 때문에 외모모...	2020-01-20
2	https://www.instagram.com/p/BxvfvoFIYk/	gillllip	. "이번에도 저와 결혼해 주시겠습니까?"... 이번이 두 번째로 최악인 청혼이예요. ...	2019-06-16
3	https://www.instagram.com/p/B8BazVend_w/	raaahleen	2020.01. [완독] 리셋팅레이디회귀를 토판 회귀도 토판도 새롭게 해석한 소설...	2020-02-01
4	https://www.instagram.com/p/B8ekVDkHyMx/	ilikehouse1	#리셋팅레이디 #리디북스 #로맨스판타지소설 #차서진 #회귀를 #피레를 #완결 #완결...	2020-02-13
5	https://www.instagram.com/p/Bj8_8GIHh_y/	happy_hji	#리셋팅레이디등장인물들마저도 깨알같이 취져 시은경. 밀바닥 인생에서 돈 되는 일이면...	2018-06-13
6	https://www.instagram.com/p/CAIvqHlgSO6/	u_u9oo	한줄 감상 : 불합륵을연보다 매운 로맨스윙러.... 『리셋팅 레이디』, 차서진...	2020-05-14
7	https://www.instagram.com/p/B73jRjhnQwX/	roommr_0202	캐런은 117세 생일을 맞이하여 살인마가 되기로 결심했다 #리셋팅레이디	2020-01-29

그림 48. 인스타그램 크롤링 통한 작품 관련 게시물 크롤링

ㄷ. 커뮤니티

1) 네이버 블로그

네이버 블로그와 티스토리는 기타 소셜미디어 (트위터, 인스타그램)과 다르게 세세한 리뷰 중심의 글이 돋보인다. 상대적으로 더 정확한 연관어를 찾을 수 있을 것 같다고 예상되기에 Aspect 마이닝에 쓰일 계획이다.

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 추출한다면, 제품 한 개를 분석함에 있어서 부족함이 보이지는 않는다.

그림 49와 같이 네이버 검색창 검색 결과를 통해 나온 글들의 url 을 받아온다.

```
import bs4
import requests
from urllib import parse
import re

session = requests.Session()
header = {
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/81.0.4044.138 Safari/537.36"
}

print("작품이름 -> ")
search_name = input()
tmp = search_name + "리뷰"

url = "https://search.naver.com/search.naver?date_from=&date_option=0&date_to=&dup_remove=1&nso=&post_blogurl=&post_blogurl_without=&query=" + parse.quote(
    tmp) + "&sm=tab_pge&srchby=all&st=sim&where=post&start=1"

r = requests.get(url)
c = r.content
html = bs4.BeautifulSoup(c, "html.parser")

main = html.find('div', {'class', 'blog section _blogBase _prs_blg'})

total = main.find('span').text.split("/ ")[1].replace("건", "").replace(",", "")
print(int(total))
```

그림 49. 네이버 블로그 크롤링 - 검색

그림 49와 같이 검색을 통해 글의 제목과 작성자, 작성 일시, 게시물의 내용을 스크레이핑한다.

```

for li in li_s:
    url = li.find('a').attrs['href'].replace("://", "://m.")
    req = session.get(url, headers=header)
    html = bs4.BeautifulSoup(req.text, "html.parser")
    title_area = html.find('div', {'class', 'post_tit_area'})
    if title_area != None:
        title = title_area.find('h3', {'class', 'tit_h3'})
        print("제목 : " + title.text)

        author = title_area.find('strong', {'class', 'ell'})
        print("작성자 : " + author.text)

        date = title_area.find('p', {'class', 'se_date'})
        print("작성일시 : " + date.text)

        post_area = html.find('div', {'class', 'post_ct'})
        p = post_area.find('p')
        if p != None:
            posts = post_area.findAll('p')
        else:
            posts = post_area.findAll('span')

        print("\n-----\n" + "게시글\n" + "-----\n")
        for post in posts:
            print(post.text)
        print("\n-----\n")

    else:
        title = html.find('span', {'class', 'se-fs- se-ff-'})
        if title == None:
            title = html.find('div', {'class', 'se_textView'})
        print("제목 : " + title.text)

        info_area = html.find('div', {'class', 'blog_authorArea'})

        author = info_area.find('strong', {'class', 'ell'})
        print("작성자 : " + author.text)

        date = info_area.find('p', {'class', 'blog_date'})
        print("작성일시 : " + date.text)

        print("\n-----\n" + "게시글\n" + "-----\n")

        post_area = html.find('div', {'class', 'se-main-container'})
        if post_area != None:
            posts = post_area.findAll('p')
            for post in posts:
                print(post.text)
        else:
            post_area = html.findAll('div', {'class', 'se_component_wrap'})[1]
            posts = post_area.findAll('span')
            for post in posts:
                print(post.text)

        print("\n-----\n")

```

그림 50. 네이버 블로그 크롤링 – 게시글

다음과 같은 결과가 출력되는 것을 확인할 수 있다.

```
작품이름 ->
닥터최태수
140
제목 : [리뷰] 닥터 최태수
작성자 : RM MC
작성일시 : 2020. 3. 4. 8:19

-----
게시글

[닥터 최태수 리뷰] 닥터 최태수 읽으며 문득문득 드는 생각이극한 상황과 제한된 인력으로 살리기 위해어떻게 해야하나 하는 게임을 보
는 것 같다.게임 보고 있지만 게이머가 펼쳐는 완벽하고 절제된 전개에 넋을 놓게 되어 그냥 빨려들어 물입의 경지가 되고갑자기 정전
이 되거나 광고가 나오는 것 같이란 편이 끝난다. ^^

-----

제목 : [현대판타지] 웹소설 닥터 최태수리뷰
작성자 : 퐁
작성일시 : 2019. 4. 14. 17:03
```

그림 51. 네이버 블로그 크롤링 - 추출 결과

2) 티스토리(다음 블로그)

네이버 블로그와 마찬가지로 검색어 결과로 나온 글들의 링크를 뽑고 그 링크들의 제목, 작성자, 작성일시를 출력한다. 티스토리나 다음 블로그는 해당 항목 각각의 class 명이 다를 뿐, 비슷한 포맷으로 갖춰져 있다.

```
main = html.find('div', {'class', 'coll_cont'})

total = html.find('span', {'class', 'txt_info'}).text.split("/ ")[1].replace("건", "").replace(", ", "")
print(int(total))

ul = main.find('ul')
li_s = ul.findAll('li')
for li in li_s:
    url = li.find('a').attrs['href'].replace("://", "://m.")
    r = session.get(url, headers=header)
    html = bs4.BeautifulSoup(r.content, "html.parser")

    title_area = html.find('div', {'class', 'view_head'})

    title = title_area.find('h3', {'class', 'tit_view'})
    print("제목 : " + title.text.strip())

    info_area = html.find('div', {'class', 'info_writer'})

    author = info_area.find('span', {'class', 'txt_writer'})
    print("작성자 : " + author.text.strip())

    date = title_area.find('time', {'class', 'txt_time'})
    print("작성일시 : " + date.text)

    post_area = html.find('div', {'class', 'small'})

    posts = post_area.findAll('p')
    print("\n-----\n" + "게시글\n" + "-----\n")
    for post in posts:
        print(post.text)
    print("\n-----\n")
```

그림 52. 티스토리 크롤링


```

total = html.find('span', {'class', 'txt_info'}).text.split("/ ")[1].replace("건", "").replace(",", "")
print(int(total))
ul = main.find('ul')
li_s = ul.findAll('li')
for li in li_s:
    url = li.find('a').attrs['href'].replace(".com/", ".com/m/")
    r = session.get(url, headers=header)
    html = bs4.BeautifulSoup(r.content, "html.parser")

    title_area = html.find('div', {'class', 'blogview_tit'})

    title = title_area.find('h2', {'class', 'tit_blogview'})
    print("제목 : " + title.text)

    author = title_area.find('span', {'class', 'txt_by'})
    print("작성자 : " + author.text)

    info_area = html.find('div', {'class', 'blogview_info'})
    date = info_area.find('time', {'class', 'txt_date'})
    print("작성일시 : " + date.text)

    post_area = html.find('div', {'class', 'blogview_content'})

    posts = post_area.findAll('p')
    print("\n-----\n" + "게시글\n" + "-----\n")
    for post in posts:
        print(post.text)
    print("\n-----\n")

```

그림 53. 티스토리 크롤링

그림 52, 다음 블로그의 실행 결과는 다음(그림 54)과 같다.

```

작품이름 ->
재혼 황후
301
제목 : 【역사/로맨스】 연록혼 재현 - 한수영
작성자 : 원고래
작성일시 : 2008.01.18 01:58

-----
게시글
-----

▶원 제 - 연록혼 재현
▶시 리 즈 - 1_5[완]
▶작 가 - 한수영
▶발 행 일 - 2007년 08월
▶발 행 처 - 마야

```

그림 54. 다음 블로그 크롤링 - 추출 결과

그림 53, 티스토리의 실행 결과는 다음(그림 55)과 같다.

작품이름 -> 닥터최태수 15 제목 : [리뷰] 닥터 최태수[조석호_현대 판타지_완결] 작성자 : 잔말행이 작성일시 : 2019. 4. 8. 00:55
----- 게시글 -----
<닥터 최태수>
저자 : 조석호 출판사 : 마미더스스토리 장르 : 현대판타지, 성장물, 전문직, 환상체험 완결 유무 : 완결(3,236화) 책 소개
“그래, 환자를……, 무서워해야 돼.

그림 55. 티스토리 크롤링 - 추출 결과

3) 디시인사이드

디시인사이드는 갤러리 전체를 통틀어 검색이 가능하기도, 메이저 및 마이너 갤러리별로 검색이 가능하기도 하다. 갤러리 전체를 통틀어 검색을 할 경우 최대 120 페이지(한 페이지 당 25 개의 게시물을 출력한다.)까지밖에 검색이 되지 않는다.

디시인사이드의 경우 리뷰 성향을 띄기보다는 언급 회수에 따른 관심도 및 인지도 지표를 파악하기 좋은 수단이라고 판단, 언급된 회수를 카운트업 해 반영하기로 결정했다.

다음은 디시인사이드 장르소설 갤러리의 게시글을 크롤링하기 위한 코드이다. Base_url 을 디시인사이드로 설정, 각 갤러리가 가지고 있는 아이디를 받아 params 값으로 입력을 해 연결하는 방식이다.

```
BASE_URL = "https://gall.dcinside.com/mgallery/board/lists?id=genrenovel"

params = {
    'id': 'genrenovel',}

headers = {
    'User-Agent': "Mozilla/5.0 (Intel Mac OS X 10_13_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/81.0.4044.138 Safari/537.36"}

resp = requests.get(BASE_URL, params=params, headers=headers)

soup = BeautifulSoup(resp.content, 'html.parser')

contents = soup.find('tbody').findAll('tr')
```

그림 56. 디씨인사이드 크롤링

장르소설 갤러리의 첫 페이지, 작성된 15 개의 게시글의 제목, 글쓴이(닉네임 혹은 ip), 날짜, 조회수, 추천수, 내용을 출력한다.

```

for i in contents:
    print('-' * 15)

    title_tag = i.find('a')
    title = title_tag.text
    print("제목: ", title)

    writer_tag = i.find('td', class_='gall_writer ub-writer').find('span', class_='nickname')
    if writer_tag is not None:
        writer = writer_tag.text
        print("글쓴이: ", writer)

    else:
        print("글쓴이: ", "없음")

    ip_tag = i.find('td', class_='gall_writer ub-writer').find('span', class_='ip')
    if ip_tag is not None:
        ip = ip_tag.text
        print("ip: ", ip)

    date_tag = i.find('td', class_='gall_date')
    date_dict = date_tag.attrs

    if len(date_dict) == 2:
        print("날짜: ", date_dict['title'])

    else:
        print("날짜: ", date_tag.text)
        pass

    views_tag = i.find('td', class_='gall_count')
    views = views_tag.text
    print("조회수: ", views)

    recommend_tag = i.find('td', class_='gall_recommend')
    recommend = recommend_tag.text
    print("추천수: ", recommend)

    link = i.find('td', class_='gall_tit ub-word')
    href = link.find('a').attrs['href']
    if href=="javascript:;":
        continue
    content_url = "https://gall.dcinside.com/" + href

```

그림 57. 디씨인사이드 크롤링

출력 결과는 다음과 같다. 공지글과 공지글이 아닌 것을 구분하는 방법이 고안되어야 한다.

제목: 예의를 좀 아는 그랜젤 마스터 연예인은?
글쓴이: 없음
날짜: 20.05.12
조회수: -
추천수: -

제목: 장궤 통합 공지 0.8
글쓴이: o o
날짜: 2020-02-01 21:20:41
조회수: 16590
추천수: 40

<https://gall.dcinside.com/mgallery/board/view/?id=genrenovel&no=457640&page=1>
[일반] 장궤 통합 공지 0.8

차단 및 삭제 대상글목 분홍소설 제목 말 안 하고 튀기소설과 전혀 무관한 억박이나 잡담 (막줄에 소설이야기 알아두는 것도 포함)참돔
머그로타 갈 올 퍼오기특히 특정 소설 얘기 없는 처1,2년 , TS 빨굴은 댓글로 호응해 주는 것도 차단함 (글쓴놈 7일 차단, 댓글 단 놈 3일
차단)특정 소설 얘기 없는 백합,보림총 글정치색 섞인 글(차단)낙시 뇌질늑연급미련 소설 없냐? 등등의 제목으로 소설과 관련없는 짤 올리
려고 쓰는 글 (차단)소설미랑 관련 없는 씹덕 만화(3일 차단)--- 당분간 소설 관련 없는 빨굴들은 최소 1일 차단 하겠음 ----- 당분간 연
독률,구매수 언급하는 글들 최소 1일 차단 하겠음 ---작가 홍보는 15화 이상 쓴 글만해당 글 제외 작가 티 내지 마셈차단 단어,아이피http
s://gall.dcinside.com/mgallery/board/view/?id=genrenovel&no=698098

제목: 장마궤 신문고
글쓴이: Qwer1234m
날짜: 2020-04-25 20:24:45
조회수: 4111
추천수: 6

<https://gall.dcinside.com/mgallery/board/view/?id=genrenovel&no=708312&page=1>
[일반] 장마궤 신문고앱에서 작성

원장 호출됩니다 댓글달 때 링크 뒤에 글자 붙이지 말아주십시오 기본적으로 새벽반이긴 한데 새벽 아닐 때 사용해도 상관 없습니다 용
은 예) <https://m.dcinside.com/board/genrenovel/708312> 삭제 좀 들린 예) <https://m.dcinside.com/board/genrenovel/708312삭제> 좀

그림 58. 디씨인사이드 크롤링 - 추출 결과

3. Data Processing

추출 데이터의 분석은 각 서비스 플랫폼의 리뷰와 댓글 커뮤니티 및 소셜 데이터를 통해 이뤄진다.

1. 트렌드 분석 : 작품의 언급이 커뮤니티와 소셜을 통해 얼마나 되었는지를 통해 알아본다.
 - A. 작품 자체가 얼마나 언급이 되었는지 카운트를 통해 작품의 화제성을 판단한다.
주기는 1달 간격으로 설정하고 추이를 살펴본다.
 - B. 작품 관련 글에서 어떤 이야기가 많이 언급되고 있는지 텍스트 분석을 통해 단어를 추출해
내고 마찬가지로 1달 간격으로 어떤 것이 작품 내에서 화제가 되고 있는지 살펴본다.
 2. 감성 분석 : 작품 댓글의 긍부정도를 판단하여 작품에 대해 1달 간격의 긍정적인 반응과 부정적
인 반응을 살펴보고 작품에 대한 반응 양상을 살펴본다.
 3. Aspect 분석 : 작품 사이트 속 리뷰와 네이버, 티스토리 등 리뷰에 대한 게시글이 많은 플랫폼
을 가지고 작품 내 주인공, 분위기 , 스토리 등과 연결되어 자주 나오는 반응을 분석한다.
- 분석 과정은 김성중, 한승주를 주축으로 진행한다.

ㄱ. 전처리 과정(불용어 제거, 어근 동일화, N-gram)

```

In [3]: import nltk
        nltk.download('stopwords')

[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\maruk\AppData\Local\nltk_data...
[nltk_data] Unzipping corpora\stopwords.zip.

Out[3]: True

In [5]: words_Korean=["초식","연호","민국","대이들","시작","놀이","교통광","교통사고","복히","자동차"]
        stopwords=["가디","놀이","나리","것","복히"]
        [i for i in words_Korean if i not in stopwords]

Out[5]: ['초식', '연호', '민국', '대이들', '시작', '교통광', '교통사고', '자동차']

In [7]: from nltk.corpus import stopwords
        words_English=["apple","banana","chief","roberts",',','.', 'president','you','.']
        print([w for w in words_English if not w in stopwords.words('english')])

['apple', 'banana', 'chief', 'roberts', ',', '.', 'president', '.']

```

그림 59. NLTK API 사용 예시

Nltk api 의 stopwords 를 이용하여 불용어를 제거한 결과입니다. 이 외에도 re.compile 과 정규식의 조합을 이용하여 특수문자나 특정 조합(메일)등을 지우는 처리를 해보았습니다.

```

In [1]: ##### 동인화 #####
        from nltk.stem import PorterStemmer
        from nltk.tokenize import word_tokenize

In [4]: import nltk

In [7]: nltk.download('punkt')

[nltk_data] Downloading package punkt to
[nltk_data] C:\Users\maruk\AppData\Local\nltk_data...
[nltk_data] Unzipping tokenizers\punkt.zip.

Out[7]: True

In [9]: ps_stemmer=PorterStemmer()
        new_text="It is important to be immersed while you are pythoning with python. All pythoners have pythoned poorly at least once."
        words=word_tokenize(new_text)
        for w in words:
            print(ps_stemmer.stem(w),end=" ")

It is import to be immers while you are python with python . all python have python poorli at least onc .

In [10]: from nltk.stem.lancaster import LancasterStemmer
        LS_stemmer=LancasterStemmer()

In [11]: for w in words:
            print(LS_stemmer.stem(w),end=" ")

It is import to be immers whil you ar python with python . al python hav python poor at least ont .

In [15]: from nltk.stem.revep import RegexpStemmer
        RS_stemmer=RegexpStemmer('python')#python이라는 글자 제거
        for w in words:
            print(RS_stemmer.stem(w),end=" ")

It is important to be immersed while you are ing with . All ers have ed poorly at least once .

```

그림 60. NLTK API 사용 예시

해당 실행문은 영어의 경우 시제에 따라 같은 동사임에도 형태가 달라지는 경우가 있는데, 이때 달라지게 하는 원인을 제거하는 것을 실행해 보았습니다. 한국어에도 이러한 형태가 있는데 형태소 분석 뒤에 가능할 것이라고 판단합니다.

```

In [2]: ##### n-gram #####
        from nltk.tokenize import word_tokenize
        from nltk import ngrams
        sentence="Chief Justice Roberts, President Clinton, President Obama, President Carter, the citizen of Americans and people of the w
        gram=ngrams(sentence,3)
        for gram in gram:
            print(gram,end=" ")

Chief, President, Obama, Carter, the, citizen, of, Americans, and, people, of, the, world, thank, you,

```

그림 61. NLTK API 사용 예시 – n-gram

그림 2-37

n-gram 을 이용하여 '대통령 트럼프'와 같이 한 단어로 취급해야 할 필요가 있는 단어를 표기해주는 전처리 과정을 실행해 보았습니다.

ㄴ. KoNLPy 한국어 분석 中 Hannanum

```
In [24]: import pandas as pd
import nltk
from nltk.stem.porter import PorterStemmer
from nltk.tokenize import RegexpTokenizer
from nltk.corpus import stopwords

In [30]: from konlpy.tag import Hannanum
hannanum=Hannanum()

In [31]: temp = []
for i in range(len(lines)):
    temp.append(hannanum.nouns(lines[i]))#명사만추출

In [32]: # 중첩 리스트(개념을 알 것) 하나의 리스트로 변환하는 함수
def flatten(l):
    flatList = []
    for elem in l:
        if type(elem) == list:
            for e in elem:
                flatList.append(e)
        else:
            flatList.append(elem)
    return flatList

word_list=flatten(temp)
# 두글자 이상인 단어만 추출
word_list=pd.Series([x for x in word_list if len(x)>1])
word_list.value_counts().head(10)

Out [32]: 대통령    29
국민    19
대한민국    9
우리    8
여러분    7
역사    6
국민들    6
나라    6
대통령의    5
세상    5
dtype: int64
```

그림 62. KoNLPy - Hannanum

대통령 담화문 중, 명사만을 추출하여 빈도를 추출하고자 한 실행문입니다. 담화문의 경우 문단별 구성에 따라서 리스트가 중첩이 되므로 이를 하나로 만들어줄 필요가 있습니다. 따라서 def 문을 통해 함수를 정의하여 실행하였습니다.

```

In [37]: from wordcloud import WordCloud
         from collections import Counter

In [38]: font_path = 'C:\Users\hnrak\Desktop\잡아라! 텍스트마인팅\NanumBarunGothic.ttf'

In [39]: wordcloud = WordCloud(
         font_path = font_path,
         width = 800,
         height = 800,
         background_color="white"
         )

In [40]: count = Counter(stopped_tokens2)
         wordcloud = wordcloud.generate_from_frequencies(count)

In [41]: def __array__(self):
         """Convert to numpy array.
         Returns
         -----
         image : nd-array size (width, height, 3)
         Word cloud image as numpy matrix.
         """
         return self.to_array()

         def to_array(self):
         """Convert to numpy array.
         Returns
         -----
         image : nd-array size (width, height, 3)
         Word cloud image as numpy matrix.
         """
         return np.array(self.to_image())
         array = wordcloud.to_array()

In [43]: count = Counter(word_list)
         wordcloud = wordcloud.generate_from_frequencies(count)
         array = wordcloud.to_array()

In [44]: get_ipython().run_line_magic('matplotlib', 'inline')
         import matplotlib.pyplot as plt

         fig = plt.figure(figsize=(10, 10))
         plt.imshow(array, interpolation="bilinear")
         plt.show()
         fig.savefig('wordcloud.png')

```

그림 63. 자연어 처리를 통한 결과 예시

자연어 처리를 한 결과만을 이용하여 UI 를 구성하고자 했는데, 시각적인 처리가 있다면 UI 가 좀 더 다양화할 수 있지 않을까 하여 수행해보았습니다. 이 외에도 matplotlib 를 이용하여 그래프 등의 시각적인 자료를 제작해보았는데, 이러한 시각적인 자료가 트렌드를 파악할 때 사용될 수 있으리라 판단합니다.

ㄷ. 감성분석(감성사전이용)

```

In [1]: #사전 기반 감성 분석
import pandas as pd
import glob
from afinn import Afinn
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from nltk.tokenize import RegexpTokenizer
import numpy as np
import matplotlib.pyplot as plt

In [2]: pos_review=(glob.glob("C:\Users\naruk\Desktop\잡아라!엑스트마이닝\데이터\train\pos*.txt"))[20]
#1008 리뷰의 긍정(pos) 데이터셋 중 20번째 데이터의 경로를 받아옴
f = open(pos_review, 'r')
lines1 = f.readlines()[0]
#해당 문장을 받아옴
f.close()

In [3]: afinn = Afinn()
afinn.score(lines1)#감성 점수 산출(긍정 이니까 하이스코어)

Out[3]: 8.0

In [4]: #부정
neg_review=(glob.glob("C:\Users\naruk\Desktop\잡아라!엑스트마이닝\데이터\train\neg*.txt"))[20]
f = open(neg_review, 'r')
lines2 = f.readlines()[0]
f.close()
afinn.score(lines2)

Out[4]: -4.0

```

그림 64. 감성사전 사용 예시

영화사이트 IDMB 의 리뷰 자료를 이용하였고, 2500개의 data set 을 가지고 있는 Afinn 을 이용하여 불러온 리뷰의 긍부정 척도를 계산하는 것을 실행해 보았습니다.

저희는 한국어 감성사전을 이용하여 이러한 척도를 계산하는 방법도 있지만 한국어를 영어로 번역하여 해당 사전을 활용하는 방안도 고려 중에 있습니다.

```

In [7]: #Regex
NRC=pd.read_csv("C:\Users\naruk\Desktop\잡아라!엑스트마이닝\데이터\train\nrc.txt",engine="python",header=None,sep="\t")
#감성사전 오픈

NRC=NRC[(NRC != 0).all()]
NRC=NRC.reset_index(drop=True)
#감성 어휘 감성표현이 유의미한 것만 추출

tokenizer = RegexpTokenizer('[^\s]+')
stop_words = stopwords.words('english')
#불용어 처리

p_stemmer = PorterStemmer()

raw = lines1.lower()
tokens = tokenizer.tokenize(raw)
stopped_tokens = [i for i in tokens if not i in stop_words]

#감성 텍스트 처리

match_words = [x for x in stopped_tokens if x in list(NRC[0])]
emotion=[]
for i in match_words:
    temp=list(NRC.iloc[np.where(NRC[0] == i)[0],1])
    for j in temp:
        emotion.append(j)
#감성사전과 텍스트의 감성어들의 매핑

sentiment_result1=pd.Series(emotion).value_counts()
#감성표현 출력결과

sentiment_result1

Out[7]: positive      8
        trust        7
        negative     5
        joy          4
        anticipation  4
        sadness      3
        fear         3
        anger        2
        surprise     2

```

그림 65. 감성사전 사용 예시

자연어 처리를 하기 전에 불용어 처리를 하여 단어별 감정을 매핑하여 감정이 몇 번 매핑되었나를 확인하는 실행문입니다.

ㄹ. 감성분석(지도 기계학습기반 감성 분석)

```
In [2]: import pandas as pd
import glob
from afinn import Affin
from nltk.corpus import stopwords
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer

In [3]: pos_review=(glob.glob("C:\Users\naruk\Desktop\잡아라! 텍스트 마이닝 데이터\aclImdb\train\pos\*.txt"))
# 긍정, 부정 텍스트 읽어오기
lines_pos=[]
for i in pos_review:
    try:
        f = open(i, 'r')
        temp = f.readlines()[0]
        lines_pos.append(temp)
        f.close()
    except Exception as e:
        continue

len(lines_pos)

Out[3]: 12490

In [4]: neg_review=(glob.glob("C:\Users\naruk\Desktop\잡아라! 텍스트 마이닝 데이터\aclImdb\train\neg\*.txt"))
lines_neg=[]
for i in neg_review:
    try:
        f = open(i, 'r')
        temp = f.readlines()[0]
        lines_neg.append(temp)
        f.close()
    except Exception as e:
        continue

len(lines_neg)

Out[4]: 12489

In [5]: total_text=lines_pos+lines_neg
len(total_text)

Out[5]: 24979
```

그림 66. 감성사전 사용 예시

그림 2-45의 IDMB 데이터셋을 불러오는 과정입니다.

```
In [7]: x = np.array(["pos", "neg"])
class_index=np.repeat(x, [len(lines_pos), len(lines_neg)], axis=0)
# 긍정, 부정 클래스 라벨링
stop_words = stopwords.words('english')

vect = TfidfVectorizer(stop_words=stop_words).fit(total_text)
X_train_vectorized = vect.transform(total_text)
# TF-IDF가중치를 준 후에 문서-단어 매트릭스로 바꾸어줌

In [8]: from sklearn.linear_model import LogisticRegression,SGDClassifier
model = LogisticRegression()
model.fit(X_train_vectorized, class_index)
#로지스틱 회귀모형을 세움

C:\Users\naruk\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarning: Default solver will be changed to
'lbfgs' in 0.22. Specify a solver to silence this warning.
FutureWarning)

Out[8]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
intercept_scaling=1, l1_ratio=None, max_iter=100,
multi_class='warn', n_jobs=None, penalty='l2',
random_state=None, solver='warn', tol=0.0001, verbose=0,
warm_start=False)

In [17]: pos_review_test=(glob.glob("C:\Users\naruk\Desktop\잡아라! 텍스트 마이닝 데이터\aclImdb\test\pos\*.txt"))[10]
test=[]
f = open(pos_review_test, 'r')
test.append(f.readlines()[0])
f.close()

predictions = model.predict(vect.transform(test))
predictions

Out[17]: array(['pos'], dtype='<U3')

In [18]: neg_review_test=(glob.glob("C:\Users\naruk\Desktop\잡아라! 텍스트 마이닝 데이터\aclImdb\test\neg\*.txt"))[20]
test2=[]
f = open(neg_review_test, 'r')
test2.append(f.readlines()[0])
f.close()
predictions = model.predict(vect.transform(test2))
predictions

Out[18]: array(['neg'], dtype='<U3')
```

그림 67. 감성사전 사용 예시

여러 모델 중 로지스틱 회귀분석 모델을 사용한 결과로 불러온 리뷰의 감정을 나타낸 결과입니다.

위와 동일하게 서포트벡터머신, 의사결정 나무 모형 등 지도 기계학습 기반의 여러 모델을 동일한 데이터에 대해 수행해보았습니다.

```
In [19]: #의사결정나무모형으로 위와 동일한 실험
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier()
clf.fit(X_train_vectorized, class_Index)
predictions = clf.predict(vect.transform(test))
predictions

Out[19]: array(['neg'], dtype='<U3')
```

```
In [20]: predictions = clf.predict(vect.transform(test2))
predictions

Out[20]: array(['neg'], dtype='<U3')
```

```
In [15]: #서포트벡터머신-리나오래결립
from sklearn.svm import SVC
clf = SVC() #SVC(gamma='scale') 이면식으로 변경가능
clf.fit(X_train_vectorized, class_Index)
predictions = clf.predict(vect.transform(test))
predictions

C:\Users\maruk\AppData\Local\anaconda3\lib\site-packages\sklearn\svm\base.py:193: FutureWarning: The default value of gamma will change from 'auto' to 'scale' in version 0.22 to account better for unscaled features. Set gamma explicitly to 'auto' or 'scale' to avoid this warning.
"avoid this warning.", FutureWarning)

Out[15]: array(['pos'], dtype='<U3')
```

```
In [16]: predictions = clf.predict(vect.transform(test2))
predictions

Out[16]: array(['pos'], dtype='<U3')
```

그림 68. 로지스틱 회귀 분석 모델 - 긍정 결과

그 결과로 로지스틱 회귀분석의 경우 원하는 대로 긍정, 부정의 결과를 가져왔고 수행 속도도 해당 데이터에 한해서는 다른 두 모델보다 속도도 빨랐습니다. 하지만, 나머지 두 개의 모델은 수행 속도도 굉장히 느리고 원하는 결과를 도출해내지도 않았습니다. 이를 통해서 모델별로 적용해서 실제 데이터의 예측률을 파악하는 게 중요할 것이라고 생각하였으며, 저희가 수행하고자 할 프로젝트에서 사용할 수 있는 api 나 모델들이 여러 가지 있으므로 이를 적용하여 최선의 값을 도출해낼 필요가 있다고 판단하여 추후에 작업해보고자 합니다.

□. Word2vec을 이용한 단어 임베딩 中 단어 유사도 판단

추후에 Aspect Analysis 를 할 때 단어 유사도 판단의 과정이 포함될 것이라 생각하여 Word2vec 을 이용하여 단어 임베딩을 실시해보았습니다. 그중에서 이번에는 단어 유사도를 판단하여, 특정 단어와 유사한 단어가 무엇이 있는지 확인해보았으며, 좌표로서 나타낼 수 있음을 확인하였습니다. 이때, 네이버 영화 말뭉치 training 이 완료된 Set 를 이용하여 Word2vec 과정을 수행하였으며, Konlpy 를 이용하여 한글에 대한 분석을 실시해 보았습니다. 특히, konlpy 사용 중에 norm 이나 stem 을 이용하여 오타나 형태소의 원형을 이용하였습니다.

```
In [5]: import codecs
        #konlpy 0.5.0 버전 이후부터 이틀이 Twitter에서 Okt로 바뀌었다.
        from konlpy.tag import Okt
        from gensim.models import word2vec
        from konlpy.utils import pprint

In [6]: def read_data(filename):
        with codecs.open(filename, encoding='utf-8', mode='r') as f:
            data = [line.split('\t') for line in f.read().splitlines()]
            data = data[1:] # header 제외
        return data

In [7]: #파일 위치. 본인의 파일경로로 변경필요
        ratings_train = read_data('ratings_train.txt')
        #konlpy 중에서 트위썬 형태소분석기 사용 (1)
        tw_tagger = Okt()

In [8]: # 토큰나이즈(의미단어결합) 함수. 트위터 형태소 분석기 사용 (2)
        # 형태소 / 품사 형태로 리스트화
        def tokens(doc):
            return ['/'.join(t) for t in tw_tagger.pos(doc, norm=True, stem=True)]
        #norm 기능을 이용해 오타를 정정(ex. 사롱해를 사랑해로), stem을 이용해 원형으로 반환(ex. 입니다를 이다로)

In [9]: # 파일중에서 영화 리뷰 데이터만 달기
        docs = []
        for row in ratings_train:
            docs.append(row[1])

        data = [tokens(d) for d in docs]

In [11]: # [TRAIN] word2vec 으로 모델 생성 (3)
        w2v_model = word2vec.Word2Vec(data)

        # init_sims 명령어로 필요없는 메모리 반환
        w2v_model.init_sims(replace=True)

        # [TEST] 가장 유사한 단어 출력 (4)
        pprint(w2v_model.wv.most_similar(positive=tokens(u'남자 여배우'),
            negative=tokens(u'배우'), topn=1))

[('여자/Noun', 0.8150866031646729)]
```

그림 69. Word2vec 단어 유사도 판단 사용예시

```
In [14]: pprint(w2v_model.wv.most_similar(positive=tokens(u'주인공'),topn=10))
```

```
[('기선/Noun', 0.7544881105422974),
 ('얼굴/Noun', 0.6906263828277588),
 ('행동/Noun', 0.6544679307937622),
 ('남자/Noun', 0.6500486636161804),
 ('여자/Noun', 0.6532753705978394),
 ('악역/Noun', 0.6524346470832825),
 ('등장인물/Noun', 0.6311010122299194),
 ('성격/Noun', 0.630095899105072),
 ('캐릭터/Noun', 0.6252544522285461),
 ('악당/Noun', 0.6231570839881897)]
```

```
In [16]: from sklearn.manifold import TSNE
from matplotlib import font_manager as fm
from matplotlib import rc
movie_tsne = TSNE(n_components=2)
```

```
Out[16]: TSNE(angle=0.5, early_exaggeration=12.0, init='random', learning_rate=200.0,
method='barnes_hut', metric='euclidean', min_grad_norm=1e-07,
n_components=2, n_iter=1000, n_iter_without_progress=300, perplexity=30.0,
random_state=None, verbose=0)
```

```
In [18]: movie_vocab = w2v_model.wv.vocab
movie_similarity = w2v_model[movie_vocab]
movie_similarity
```

```
C:\Users\naruk\Anaconda3\lib\site-packages\ipykernel_launcher.py:2: DeprecationWarning: Call to deprecated `__getitem__` (Method will be removed in 4.0.0, use self.wv.__getitem__() instead).
```

```
Out[18]: array([[ 0.01583514, -0.01967263,  0.01102952, ..., -0.06493731,
-0.11728221, -0.10047553],
[ 0.02801962, -0.07917922, -0.02239081, ..., -0.11629011,
-0.03628932,  0.07144445],
[ 0.2009029 , -0.05290857,  0.02421443, ..., -0.1569652 ,
 0.07382585, -0.17336367],
...,
[ 0.10578212,  0.0587996 , -0.027095 , ...,  0.02845743,
-0.18845902, -0.03497209],
[ 0.00156862,  0.11541089, -0.0601042 , ...,  0.02754588,
-0.17417371,  0.06905237],
[ 0.00297765, -0.00140684,  0.05425932, ..., -0.0171057 ,
-0.16980827,  0.06580625]], dtype=float32)
```

그림 70. Word2vec 단어 유사도 판단 사용 예시

자료를 나타내 주는 코드를 수행하였습니다. 단어의 유사도 이므로 특정 단어와 유사한 단어를 보여주는 것인데, Aspect Analysis 의 경우 문법적인 연관성을 판단해야 하므로 임베딩에서 다른 과정을 추가적으로 요구함을 확인하였습니다.

```
In [20]: import pandas as pd
movie_transform_similarity = movie_tsne.fit_transform(movie_similarity)
movie_df = pd.DataFrame(movie_transform_similarity, index=movie_vocab, columns=['x', 'y'])
movie_df[0:10]
```

Out [20] :

	x	y
아/Exclamation	22.172894	-57.955566
더빙/Noun	6.981616	45.611614
../Punctuation	18.876213	-60.003292
진짜/Noun	17.515112	-63.976002
짜증나다/Adjective	36.130859	53.686295
목소리/Noun	-56.858418	-29.822529
홀/Noun	23.910069	-57.069225
.../Punctuation	18.831528	-59.998028
포스터/Noun	46.868126	26.451733
보고/Noun	50.604397	-23.869593

그림 71. Word2vec 단어 유사도 판단 사용예시

작업하고자 하는 타겟이 리뷰이다 보니, 'ㅋㅋㅋ'나 '꿀잼'과 같은 은어들에 대해서 처리하는 과정이 필요하리라 판단하였는데 Okt 의 경우 이러한 은어들도 한국어 조사 등으로 처리가 되고 있음을 따로 확인하였습니다. 그 외의 분석기에 대해서도 확인이 필요할 듯합니다. 분석기 별로 수행 속도가 다를 것을 확인하였고, 분석을 하여 보여주는 결과의 종류도 각기 다르기 때문입니다.

- 트위터에서 마이닝한 트윗 정보로 유사도 판단하기.

```
In [1]: #https://jeongweokie.github.io/2019/08/10/190810-twitter-data-crawling/
# GetOldTweet3 사용 준비
try:
    import GetOldTweets3 as got
except:
    !pip install GetOldTweets3
    import GetOldTweets3 as got

In [2]: # datetime을 사용해 가지를 범위를 정의
# 예제 : 2020-01-01 ~ 2020-04-01

import datetime

days_range = []

start = datetime.datetime.strptime("2020-01-01", "%Y-%m-%d")
end = datetime.datetime.strptime("2020-04-01", "%Y-%m-%d")
date_generated = [start + datetime.timedelta(days=x) for x in range(0, (end-start).days)]

for date in date_generated:
    days_range.append(date.strftime("%Y-%m-%d"))

print("=== 설정된 트윗 수집 기간은 {} 에서 {} 까지 입니다 ===".format(days_range[0], days_range[-1]))
print("=== 총 {} 일 간의 데이터 수집 ===".format(len(days_range)))

=== 설정된 트윗 수집 기간은 2020-01-01 에서 2020-03-31 까지 입니다 ===
=== 총 91일 간의 데이터 수집 ===

In [4]: # 특정 검색어가 포함된 트윗 검색하기 (query search)
# 검색어 : 기묘한이야기

import time

# 수집 기간 맞추기
start_date = days_range[0]
end_date = (datetime.datetime.strptime(days_range[-1], "%Y-%m-%d")
            + datetime.timedelta(days=1)).strftime("%Y-%m-%d") # setUntil() 끝을 포함하지 않으므로, day + 1

# 트윗 수집 기준 정의
tweetCriteria = got.manager.TweetCriteria().setQuerySearch('기묘한이야기')\
    .setSince(start_date)\
    .setUntil(end_date)\
    .setMaxTweets(-1)

# 수집 with GetOldTweet3
print("Collecting data start.. from {} to {}".format(days_range[0], days_range[-1]))
start_time = time.time()

tweet = got.manager.TweetManager.getTweets(tweetCriteria)

print("Collecting data end.. [0:0.2f] Minutes".format((time.time() - start_time)/60))
print("=== Total num of tweets is {} ===".format(len(tweet)))

Collecting data start.. from 2020-01-01 to 2020-03-31
Collecting data end.. 21.67 Minutes
=== Total num of tweets is 7758 ===

In [5]: # GetOldTweet3 에서 제공하는 기본 변수
# 유저 아이디, 트윗 링크, 트윗 내용, 날짜, 리트윗 수, 관심글 수
# 원하는 변수 골라서 저장하기

from random import uniform
from tqdm import tqdm_notebook

# initialize
tweet_list = []

for index in tqdm_notebook(tweet):

    # 메타데이터 목록
    username = index.username
    link = index.permalink
    content = index.text
    tweet_date = index.date.strftime("%Y-%m-%d")
    tweet_time = index.date.strftime("%H:%M:%S")

    info_list = [tweet_date, tweet_time, username, content, link]
    tweet_list.append(info_list)

# 휴식
time.sleep(uniform(1,2))

HBox(children=(IntProgress(value=0, max=7758), HTML(value='')))
```

그림 72. 트위터에서 정보 추출

예시로 2020년 1월 1일부터 3월 31까지 "기묘한이야기"를 검색한 결과를 마이닝하였습니다.

```
In [6]: # csv 파일로 결과 저장 - pandas
# 파일 저장하기

import pandas as pd

#twitter_df = pd.DataFrame(tweet_list,
#                           # columns = ["date", "time", "user_name", "text", "link", "retweet_counts", "favorite_counts",
#                           # "user_created", "user_tweets", "user_followings", "user_followers"])

twitter_df = pd.DataFrame(tweet_list,
                           columns = ["date", "time", "user_name", "text", "link"])

# csv 파일 만들기
twitter_df.to_csv("sample_twitter_data_{to_}.csv".format(days_range[0], days_range[-1]), index=False)
print("=== {} tweets are successfully saved ===".format(len(tweet_list)))

=== 7758 tweets are successfully saved ===

In [7]: #생성 파일 확인
# 파일 확인하기

df_tweet = pd.read_csv('sample_twitter_data_{to_}.csv'.format(days_range[0], days_range[-1]))
df_tweet.head(10) # 위에서 10개만 출력

Out[7]:
```

	date	time	user_name	text	link
0	2020-03-31	23:52:03		요즘 킬링시리즈가 유행하던데 난 넷을 잘, 많이 안봐서 킬링, 기묘한이야기, 내가 ...	https://twitter.com/status/124...
1	2020-03-31	21:07:31		혁 기묘한 이야기 보러구 했는데 모던 패밀리 재밌어요! 액들이 너무 귀엽습미 다...	https://twitter.com/B.../2450954049...
2	2020-03-31	20:58:52		기묘한 이야기 재밌고 다 좋은데 시즌 3는 특히 엄청 그로테스크하고 고여하니 주의하...	https://twitter.com/iz.../12450932...
3	2020-03-31	20:57:28		꿈 속 인물한테 꿈에서 꿈 곧 거 꿈꿨다고 말했는데 자기도 똑같은 꿈 꿔다고 하더...	https://twitter.com/hy.../124508287...
4	2020-03-31	20:08:00		논아 기묘한이야기 심령어를 보고자єм	https://twitter.com/91.../2450804244...
5	2020-03-31	19:13:12		헬모야 기묘한이야기 시즌3 촬영예행연출알았는데 진짜예 출시됐네	https://twitter.com/f.../124506663...
6	2020-03-31	19:03:14		헬 넷을 보세요?카 산타클라리타다이어트 / 엘리트들 / 나를 자버린 스파이 / ...	https://twitter.com/m.../1245064127...
7	2020-03-31	18:22:17		넷플릭스 추천해주시용 무서운거 백고 ... 저 겁 완전 많아서 기묘한이야기도 못보...	https://twitter.com/.../2450538196...

그림 73. 트위터 정보 추출 결과

크롤링한 내용을 csv 파일로 저장하여 위와 같이 저장하고 저장한 파일에서 내용과 관련된 부분만 추출하여 konlpy와 word2vec을 이용한 분석을 진행해보았습니다.

```
In [24]: from konlpy.tag import Twitter
twitter = Okt()

word_dic={}

for i in df["text"]:
    mallist=twitter.pos(i)
    for word in mallist:
        if word[1]=="Noun" or word[1]=="Verb" or word[1]=="Adjective":
            if not(word[0]in word_dic):
                word_dic[word[0]]=0
            word_dic[word[0]]+=1
print(word_dic)
```

```
{'같은데': 48, '있지': 9, '나': 519, '어제': 68, '학교': 28, '귀신': 30, '봤다': 71, '년': 9, '도': 119, '헛소리': 4, '어떻고': 1, '그래': 21, '일정': 2, '팔로워': 2, '도달': 2, '시': 24, '개장': 4, '확정': 3, '됩니다': 13, '너': 102, '그': 339, '소문': 7, '들은': 5, '적': 23, '있어': 36, '아가': 5, '발음': 4, '플래': 2, '일': 15, '모어': 7, '걸스': 26, '임': 118, '낮': 73, '오케이': 80, '부류클린': 68, '나인': 141, '글리': 10, '앨리스': 1, '실뉴플': 1, '있는데': 63, '다스틴': 21, '엄마': 64, '둘': 51, '나와서': 20, '괴리감': 1, '들어요': 2, '몬다': 41, '나탈': 1, '리아': 1, '캐플': 7, '나오는': 85, '뒤': 33, '실제': 14, '사귀게': 1, '영': 33, '취향': 76, '비슷하면': 1, '아메리칸': 14, '반달': 1, '리즈': 6, '보워': 21, '좋아하면': 31, '반': 16, '봐': 222, '다': 207, '플리': 12, '다전': 11, '종아': 220, '집': 265, '오뉴': 80, '불': 81, '오중': 3, '자리': 3, '재밌음': 51, '보고싶는데': 25, '진도': 6, '나갈': 1, '클루': 15, '리스': 17, '웨': 209, '레이스': 170, '내일': 33, '만나': 4, '보살수': 3, '보신': 28, '분': 100, '있나용': 1, '다음': 61, '불파': 78, '말파': 3, '고민': 33, '중': 297, '봤더니': 12, '가발': 2, '바느질': 1, '끝났서': 1, '용': 6, '와': 93, '옌': 196, '레오': 3, '색': 5, '국내': 3, '발자': 1, '알아서': 8, '이런': 66, '별짓': 1, '하계': 29, '만드': 1, '헛말': 1, '해아하는데': 5, '자아': 5, '해서': 113, '일단': 81, '글라다니면': 1, '인형': 24, '머리': 175, '씩씩': 1, '불': 149, '꼭같다': 1, '게이': 10, '튼이다': 1, '보시나요': 10, '금': 1, '이제': 156, '이집트': 1, '피라미드': 1, '세울': 1, '있게': 6, '되었습니': 4, '비장한': 1, '표정': 11, '진심': 42, '열': 25, '동네': 4, '부': 31, '다봤다': 1, '악': 16, '보고싶다': 32, '보려고': 31, '달': 49, '무료': 17, '하는건데': 2, '제': 176, '끝내고': 8, '온다': 4, '이': 211, '북': 6, '불': 1, '아시': 10, '전': 131, '수박': 2, '결': 1, '할기로': 1, '파는': 7, '면이': 16, '스킨스': 11, '같은것도': 4, '그냥': 63, '힐': 2, '분위기': 45, '홀당': 2, '아러고': 5, '좋아함': 31, '성격': 9, '대사': 39, '관계도': 1, '이런건': 4, '알바': 3, '무지개': 1, '전구': 4, '배송': 5, '와라': 1, '느낌': 108, '내면': 2, '방': 32, '차': 14, '박해있을거라고': 1, '데이': 8, '결제': 49, '했다': 1, '봐라': 16, '안보': 28, '아님': 48, '기묘한이야기': 1, '존나': 111, '좋아한다고': 2, '오빠': 22, '나가고': 4, '나가지': 1, '배우': 58, '주하면': 8, '마지막': 7, '한국': 54, '요괴': 27, '대백': 10, '프로젝트': 9, '후원자': 8, '님': 33, '공유': 19, '님': 182, '결': 89, '쓰요': 1, '재밌음': 1, '바': 51, '속': 16, '다박습': 1, '게': 64, '위앞요': 1, '많이': 1, '아닌것': 3, '같': 33, '달와지만': 1, '머가': 10, '다
```

그림 74. KoNLPy를 통한 분석과 Word2vec을 활용한 유사도 판정

그림 74과 같이 명사, 동사, 형용사 같은 리뷰와 작품과 연관 지어 중요한 품사의 단어들만 다시 추출하였고 그 단어가 마이닝한 결과 내에서 얼마큼의 횟수만큼 사용되었는지 파악하였습니다.

```
In [26]: for word, count in keys[:20]:
        print("{} [1]".format(word, count))
```

```
이야기 7966
기묘한 7753
시즌 942
전체 607
나 519
거 511
넷플릭스 465
보고 462
추천 459
저 418
넷플 369
것 352
그 339
드라마 338
내 313
중 297
때 265
집 265
안 259
곳 238
```

그림 75. 최다빈출 단어 추출

위와 같이 리뷰에서 가장 많은 쓰인 단어를 추출할 수 있었고 이를 토대로 다음 과정에서는 유용한 단어에 대한 추출방법이 과제가 될 것 같습니다.

```
In [31]: results = []
        for i in dff["text"]:
            mallist = twitter_pos(i, norm=True, stem=True)
            # stem=True -> 어근으로 출력하라는 의미
            # ex) "그려요" -> "그린다"
            #print(mallist)
            r = []
            for word in mallist:
                if not word[0] in ["Josa", "Punctuation", "Foreign", "Suffix", "Eomi"]:
                    r.append(word[0])
            # 결과에서 제외 할 품사 임력하기
            print(r)
            rl = (" ".join(r)).strip() #공백제거
            results.append(rl)
        print(results)
```

```
['기묘하다', '이야기', '저는', '1', '2', '3', '운저대로', '재미', '올려올라', '올려', '저는', '4', '올라', '기', '내려', '보가']
['하이큐', '회지', '하산', '기묘하다', '이야기', '하산', 'In', 'our', 'time', '하산', '한지붕', '세남자', '양도', '판매', '구함',
'나', '임', '멘션', '부탁드립니다']
['그냥', '보다', '거', '계속', '보구', '요즘', '기묘하다', '이야기', '볼', '까말다', '생각', '중', '이다', '닷', '췌언님', '뽕',
'보다']
['기묘하다', '이야기']
['발레', '영상', '유튜브', '보다', '저', '찾다', '보다', '바', '갸다', '기묘하다', '이야기', '저', '좋다', '거의', '보다', '같다',
'에', 'ㅋㅋ', 'ㅠ', '재롭다', '나오다', 'ㅠㅠ', '노래', '저', '방탄', '소년단', '런', '들다', 'ㅋㅋ', '모', 'ㅋㅋ']
['버즈', '오브', '플레이', '할리퀸', '에브리', '그냥', '할리퀸', '다', '함', '클로켓', '한국판', '기묘하다', '이야기', '일드', '갑
룩뽕', '다', '이상', '노뽕', 'ㅠ']
['기묘하다', '이야기', '하나', '영기다', '있다', '엿떡', '오다']
['앗', 'ㅋㅋ', '저', '그때', '그때', '다르다', '요즘', '꽤', '발레', '영상', '클립', '몇개', '주', '구', '장창', '들리다', '보
구', '드라마', '기묘하다', '이야기', '시즌', '4', '나오다', '같다', '재랑', '중이', 'at', 'It', '영화', '요즘', '재랑하다', '없
다', 'ㅋㅋ', '유님', '요즘', '뽕', '들다', 'at', 'It']
['기묘하다', '이야기', '미치다', '마감', '하다', '되다', '더', '과', '이', '십', 'ㅈㅇ', 'ㅇ', 'ㅇㅇ']
['기묘하다', '이야기', '한번', '보다', '보다']
['마블', '제시카', '존스', 'DC', '타이탄', '기묘하다', '이야기', '한니발']
['기묘하다', '이야기', '4', '언제', '나오다', 'ㅠ', '내', '기', '달리', '있다']
['하이큐', '회지', '양도', '방다', '레스큐일', '원더풀', '데이즈', '기묘하다', '이야기', '마법사', '기록', 'oh', 'kids', '디엠',
'좋다', 'ㅠ']
```

그림 76. 어근 변경

다음은 형용사와 동사를 어근으로 바꿔주는 작업을 하였습니다.

```

In [33]: from gensim.models import word2vec
data = word2vec.LineSentence(data_file)
print(data)
model = word2vec.Word2Vec(data, size=100, window=10, hs=1, min_count=2, sg=1)
# CBOW, Skip-gram(0)
model.init_sims(replace=True) #필요없는 메모리는 unload
model.save("news.model")
print("ok")

<gensim.models.word2vec.LineSentence object at 0x00000178B8BC6488>
ok

In [36]: model = word2vec.Word2Vec.load("news.model")
print(model.similarity("기묘하다", "네티플릭스"))
print(model.similarity("기묘하다", "밀리"))
#두 단어의 유사도 -> 1에 근접할 수록 서로 상관관계가 있다.
print(model.most_similar("기묘하다"))

print(model.most_similar(positive=["기묘하다"]))
print(model.most_similar(positive=["기묘하다", "이아기"], negative=["정그럽다"], topn=5))

0.5386249
0.32553965
[('개굴', 0.6190359592437744), ('역시', 0.6100665926933289), ('월', 0.6065816283226013), ('기정', 0.6044149398803711), ('이따', 0.6040357947349548), ('알다', 0.6020412445068359), ('이러다가', 0.6007715463638306), ('거랑', 0.6000221967697144), ('도리', 0.5988008975982666), ('혹', 0.5978043079376221)]
[('개굴', 0.6190359592437744), ('역시', 0.6100665926933289), ('월', 0.6065816283226013), ('기정', 0.6044149398803711), ('이따', 0.6040357947349548), ('알다', 0.6020412445068359), ('이러다가', 0.6007715463638306), ('거랑', 0.6000221967697144), ('도리', 0.5988008975982666), ('혹', 0.5978043079376221)]
[('개굴', 0.4821982681751251), ('월', 0.46736279129981995), ('좋아하다', 0.45916640758514404), ('추강', 0.44668009877204895), ('렌즈', 0.4345983564853668)]

```

그림 77. 어근 변경 후 유사성 판단

이전의 작업을 토대로 단어 간의 유사성을 판단해보았습니다.

이번 과정에서는 추출한 데이터를 konlpy 및 word2vec 과 같은 분석에 유용한 툴과 연결시키는 작업을 해보았고 이 결과 유용한 데이터의 선별 및 축약어와 같은 비속어에 대한 판별 또한 다음 과제가 될 것으로 보입니다.

4. DB Construction

추출한 작품 정보를 DB에 저장하고 리뷰와 댓글 및 커뮤니티와 SNS 글의 경우 많은 데이터로 인해 분석 후 결과를 DB에 저장한다.

DB 구축 과정은 한승주, 조예슬을 주축으로 진행한다.

ㄱ. Logical-ERD 구성

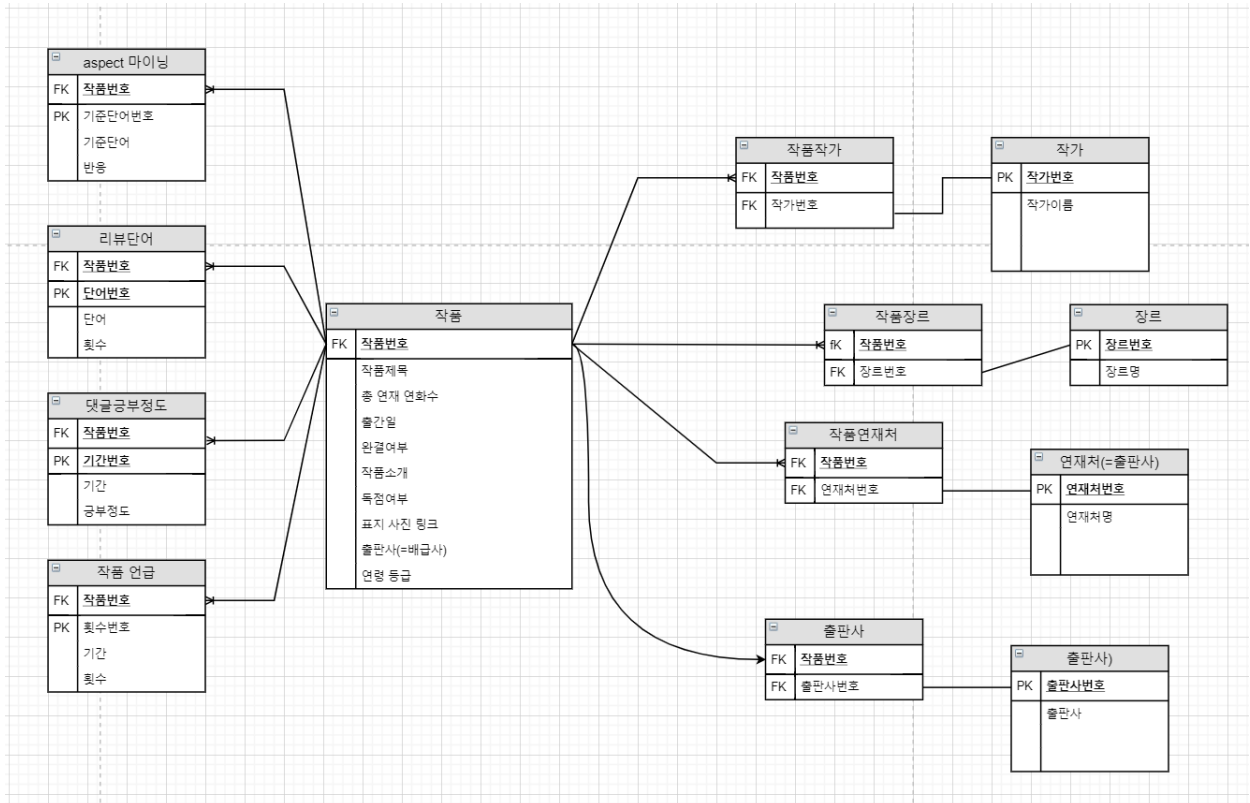


그림 78. DB ERD

기본적인 작품 정보로 작품 제목, 작가, 출판사, 연재사, 완결 여부, 현재 총 작품 화수, 출간일, 작품 소개, 장르, 독점 여부, 완결 여부가 들어가게 됩니다. 이 중 작가, 장르와 연재처, 출판사는 단일 테이블로 만들어 각 장르나 작가, 연재처 혹은 출판사 별 작품도 살펴볼 수 있도록 구성할 예정이다. 가장 중요한 리뷰의 경우 크게 4가지로 나누어 소셜 속 작품의 언급도, 작품내 어떤 주제가 화제되고 있는지 작품의 반응을 살펴보기 위한 댓글 공부정도와 작품의 aspect 분석 정보를 넣을 테이블 4개를 구성하였다.

*이는 추후 제작과정에서 변경될 수 있으며 계속 진행되는 회의를 통해 상세 내역을 수정해 나가고 있습니다.

L. DB구축 및 분석결과 축적

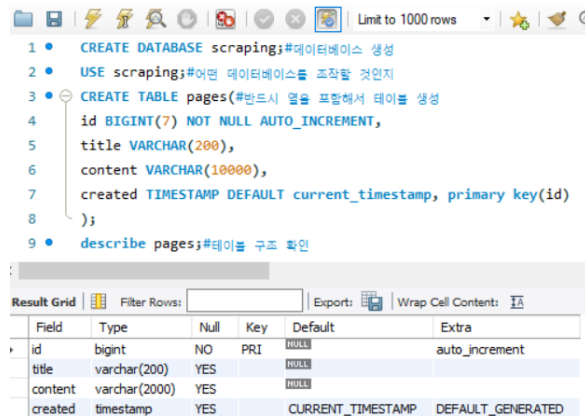


그림 79. 테이블 구축

여러 개의 테이블이 아닌 한 개의 테이블을 이용하여, 제목과 내용을 저장해 보기 위한 테이블을 구성하였습니다.

```

22 #####문자열을 utf8mb4에서 unicode_ci로 바꿔주는 코드#####
23 • alter database scraping character set=utf8mb4 collate=utf8mb4_unicode_ci;
24 • use scraping;
25
26 • alter table pages convert to character set utf8mb4 collate utf8mb4_unicode_ci;
27 • alter table pages change title title varchar(200) character set utf8mb4 collate utf8mb4_uni
28 • alter table pages change content content varchar(200) character set utf8mb4 collate utf8mb4
29 #####
30 • alter table pages modify column content varchar(2000);

```

그림 80. 테이블 열 정보 수정

이후, 데이터를 저장하는 과정에서 하나의 열에 내용을 포함시키지 못하는 경우가 많아 alter 기능을 이용하여 데이터 열의 정보를 수정하였습니다.

```

In [1]: pip install PyMySQL

Requirement already satisfied: PyMySQL in c:\users\mnaruk\anaconda3\lib\site-packages (0.9.3)
Note: you may need to restart the kernel to use updated packages.

In [3]: import pymysql
conn=pymysql.connect(host='127.0.0.1',user='root',passwd='3721',db='mysql')

In [6]: cur=conn.cursor()
cur.execute("USE scraping")
cur.execute("SELECT * from pages WHERE id=2")

Out[6]: 1

In [7]: print(cur.fetchone()) #마지막에 실행한 쿼리 결과 출력
(2, 'A new title', 'Some new content', datetime.datetime(2020, 4, 2, 7, 28, 17))

In [8]: cur.close()
conn.close()

```

그림 81. DB 구축

pyMySQL 을 이용하여 cur 과 conn 이라는 변수를 활용하여 execute 문을 통한 MySQL 명령어를 실행해 보았습니다.

```

insert into pages(title,content)#id는 자동증가, timestamp는 현재시간 자동 저장
values(
    "Test page title",
    "This is some test page content. It can be up to 10,000 characters long."
);
select * from pages where id=2;#id가 2인 행이없으므로 none return
select * from pages where title like "%test%";
select id,title from pages where content like "%page content%";
#delete를 실행하기전에는 select를 먼저 실행하는 것이 좋다.;
select * from pages where id=1;
delete from pages where id=1;
update pages set title="A new title", content="Some new content" where id=2;

```

그림 82. DML 명령어 활용

크롤링한 결과를 DB에 저장해 보기 앞서서 DML 명령어를 기본적으로 활용해보았습니다.

```

In [5]: from urllib.request import urlopen
        from bs4 import BeautifulSoup
        import datetime
        import random
        import pymysql
        import re

        conn = pymysql.connect(host='127.0.0.1', user='root', passwd='3721', db='mysql', charset='utf8')
        cur = conn.cursor()
        cur.execute('USE scraping')

        random.seed(datetime.datetime.now())

        def store(title, content):
            cur.execute('INSERT INTO pages (title, content) VALUES ("%s", "%s")', (title, content))
            cur.connection.commit()

        def getLinks(articleUrl):
            html = urlopen('http://en.wikipedia.org'+articleUrl)
            bs = BeautifulSoup(html, 'html.parser')
            title = bs.find('h1').get_text()
            content = bs.find('div', {'id': 'mw-content-text'}).find('p').get_text()
            store(title, content)
            return bs.find('div', {'id': 'bodyContent'}).findAll('a', href=re.compile('^(/wiki/)((?!:).)*$'))

        links = getLinks('/wiki/Kevin_Bacon')
        try:
            while len(links) > 0:
                newArticle = links[random.randint(0, len(links)-1)].attrs['href']
                print(newArticle)
                links = getLinks(newArticle)
        finally:
            cur.close()
            conn.close()

/wiki/List_of_gothic_festivals
/wiki/Moldova
/wiki/Taracia_District
/wiki/Romani_people
/wiki/Turkey
/wiki/Karachays
/wiki/Soyot
/wiki/Chechens
/wiki/Berbers_in_Belgium
/wiki/Aghul_people
/wiki/Archil_people
/wiki/Rutul_people
/wiki/Poles_in_Azerbaijan
/wiki/Azerbaijan
/wiki/United_Nations_Development_Program
/wiki/UNICEF
/wiki/Chapter_XIV_of_the_United_Nations_Charter

```

그림 83. DML 사용

그림 81과 동일하게, conn 과 cur 변수를 활용하여 pyMySQL 의 데이터에 접근합니다. 기본적으로 위키백과의 데이터를 하나 참조하되, 데이터 내부의 href 를 모두 찾아서 ref 의 제목과 내용을 모두 가져와서 DB에 저장하는 코드를 작성해 보았습니다. 이는, 추후에 네이버 블로그나 티스토리 등에서 작품을 검색하고, 그 내용을 수집하는 과정과 매우 유사할 것이라 판단합니다. 특히, 네이버 블로그 위키백과와 동일하게 request 를 활용합니다.

이때, DML 명령어를 사용하기 위해서 execute 문을 사용하여 insert 문을 실행하고 있음을 확인할 수 있습니다.

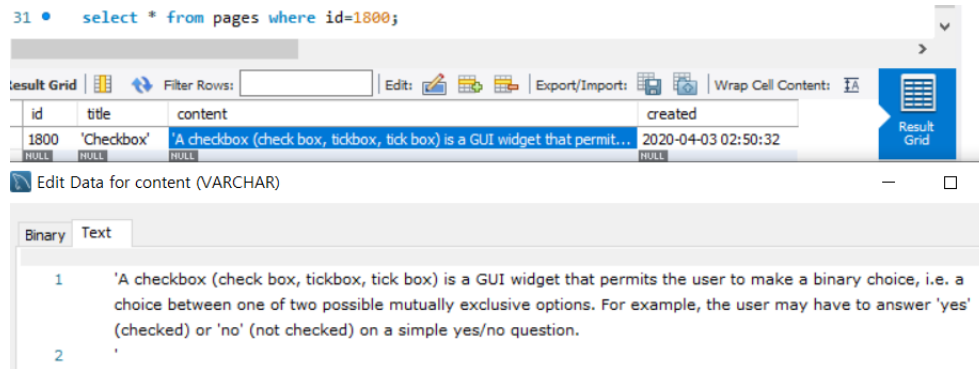


그림 84. 구축 예시 결과

컴퓨터의 속도에 비해 데이터의 길이가 크다고는 생각을 안 했는데, 생각보다 긴 시간이 소요되는 것을 확인하였습니다. 웹페이지당 1초 정도 소요되긴 하지만 추후에는 더 많은 데이터를 저장할 것이라 판단하므로 빠른 작업이 필요할 것 같습니다.

5. UI Development

사용자의 접근성 편리를 위해 웹사이트 구축으로 결정하였다.

웹사이트 구축은 조예슬, 김성종을 주축으로 진행한다.

(진행도가 있을 때 추가 작성예정)

III. 과제 평가

1. 개선방안

- 이번 주차는 각 플랫폼에서 정보를 추출할 크롤러와 분석에 앞서 분석 내용 크롤링을 위한 크롤러 제작을 진행하고 있으며 다음 분석과 DB, 마지막 UI 제작이 서로 연계되는 부분이 많아 역할을 분담을 통하여 각자 맡을 부분을 정하였다.
- 다음 주차는 본격적으로 분석 작업에 들어간다. 분석 작업에 앞서 각 웹사이트들의 지속적인 태그 변경이 확인되는 바 크롤러 수정과 분석 작업에 들어가며 DB 구축에 대한 초기 작업을 진행한다.

2. 기대효과

ㄱ. 기업적 측면

즉각적인 피드백이 필요한 문화 산업에서 소셜미디어와 커뮤니티 같은 독자층의 실시간 반응이 보이는 곳의 리뷰를 통합적으로 확인 가능함으로써 앞으로의 홍보, 제작, 투자 방향 선택에 도움이 되는 지표가 될 것이다.

ㄴ. 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.

- 현재 작품에 대한 주요 평가가 어떻게 되는지 시각적으로 확인 가능합니다

<참고문헌>

- [1] 파이썬을 활용한 클로러 개발과 스크레이핑 입문 (카토 카츠야, 요코야마 유우키, 위키북스, 2019)
- [2] 파이썬 데이터 수집 자동화 한방에 끝내기 한입에 웹크롤링 (김경록, 서영덕, 비제이퍼블릭, 2018)
- [3] 파이썬을 이용한 웹크롤링과 스크레이핑 (카토 코타, 위키북스, 2018)
- [4] 파이썬을 이용한 머신러닝, 딥러닝 실전 개발 입문 (쿠지라 히코우즈쿠에, 위키북스, 2019)
- [5] Web Scraping with Python (라이언미첼, 한빛미디어, 2019)
- [6] 잡아라! 텍스트 마이닝 with 파이썬 (서대호, 비제이퍼블릭, 2019)
- [7] <https://www.crummy.com/software/BeautifulSoup/bs4/doc.co/>
- [8] 오피니언 마이닝 기술을 이용한 효율적 상품평 검색 기법 (윤홍준, 김한준, 장재영, 2010)
- [9] 한글 텍스트의 오피니언 분류 자동화 기법 (김진옥, 이선숙, 용환승, 2011)
- [10] 상품평가 텍스트에 암시된 사용자 관점추출 (장경록, 이강욱, 맹성현, 2013)
- [11] 텍스트 마이닝을 이용한 2012년 한국대선 관련 트위치 분석 (배정환, 손지은, 송민, 2013)
- [12] 한글 감성어 사전 api구축 및 자연어 처리의 활용 (안정국, 김희웅, 2014)
- [13] 한글 음소단위 trigram-signature 기반의 오피니언 마이닝 (장두수, 김도연, 최용석, 2015)
- [14] 소셜네트워크서비스에 활용할 비표준어 한글처리 방법연구 (이종화, 레환수, 이현규, 2016)
- [15] 인공지능을 활용한 오피니언 마이닝 - 소셜 오피니언 마이닝은 무엇인가?6 (윤병운, 2017)
- [16] 한국어 비정형 데이터 처리를 위한 효율적인 오피니언 마이닝 기법 (남기훈, 2017)
- [17] A study on Sentiment Analysis with Multivariate ratings in Online Reviews (임소현, 2020)

⁶ https://www.samsungsds.com/global/ko/support/insights/1195888_2284.html