네이버 웹툰 웹크롤링

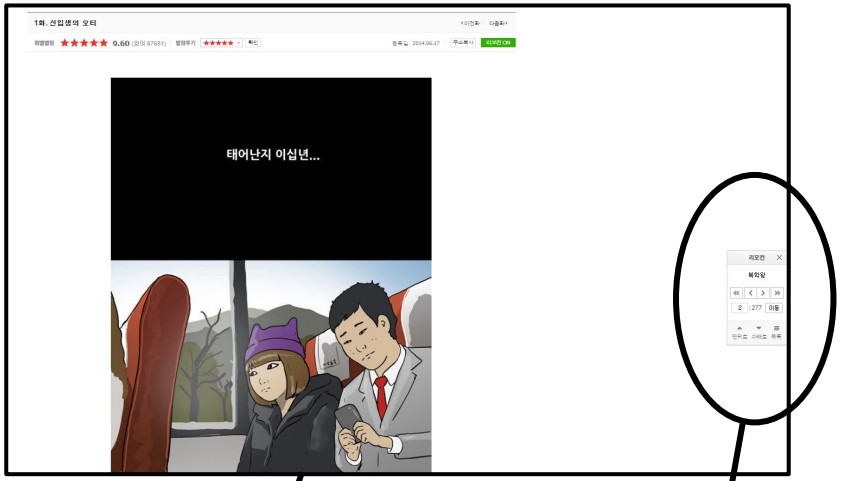
1. 웹크롤링 타깃 사이트

- 네이버 웹툰
  - 재밌어보여서 (예!지! 빼!고!)
  - 평점과 댓글 내용간 유의미한 상관관계 발견
  - 저조한 평점 발생 원인 궁금
- 선정 웹툰

- <복학왕> 선정 이유
  - 작가의 부정적 이슈, 작품 퀄리티에 따른 여론이 평점과 댓글에 가장 잘 드러나는 작품
  - 다양한 이슈와 논란이 있었던 웹툰이며 낮은 평점의 빈도수가 높았던 웹툰이기에

1.1 페이지 구조

웹툰 카테고리 바
& 해당 웹툰 소개



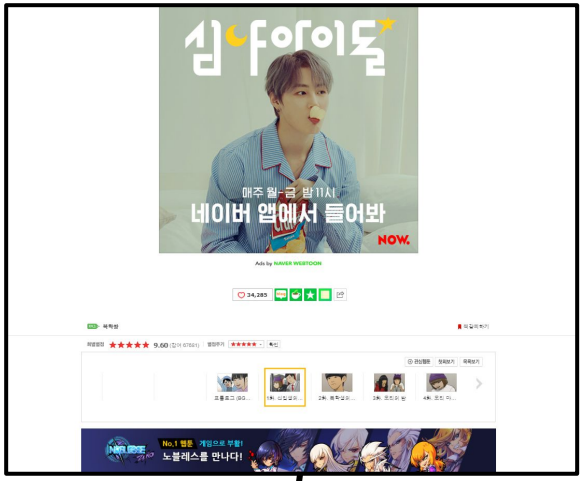
웹툰 내용

리모컨

1.1 페이지 구조


회차별 제목
1화. 신입생의 오티
‹ 이전화
다음화 ›
회별별점
9.60
(참여 67682)
별점주기
확인
등록일 2014.06.17
주소복사
리모컨 ON
평점
등록일

1.1 페이지 구조



광고
& 이전 및 다음 회차

실시간 웹툰 순위

추천 웹툰

댓글

1.2 url 구조

베이스 URL

① 작품 id

https://comic.naver.com/webtoon/detail.nhn?titleId=626907&no=2

② 회차

2. 수집할 데이터/내용

- **회차별 등록일(웹툰 발행일)**
- **제목**
- **평점**
  - 분석: 평균, 최소값, 25%, 50%, 75%, 최대값
  - 평점 8.00 미만을 받은 회차 및 날짜
- **댓글**

3. 웹페이지 데이터 크롤링 계획

- 데이터 수집
  a. 제목, 등록일, 평점, 베스트댓글 데이터 수집
  b. Requests와 BeautifulSoup 활용
  c. 댓글 데이터 수집은 용량 문제로 인하여 베스트댓글 15개만 가져오기로 결정

- 데이터 분석
  d. pandas를 이용한 평점 데이터 분석
  e. 관련 기사 수집
  f. 시각화

- 데이터 저장
  g. 크롤링 결과를 csv파일로 저장
  h. 저장 format: 등록일, 제목, 평점, 베스트댓글 순으로 컬럼을 나눠서

4. 구현 - 프로그래밍 설계 및 주요 사항

```
import requests
import pandas as pd
from bs4 import BeautifulSoup
import time
from selenium import webdriver
```

```python
def NaverWebtoon(end_page=1):
    # 해당웹툰 페이지 base_url
    base_url = 'https://comic.naver.com/webtoon/detail.nhn?titleId=626907&no={}&w
    # 해당웹툰의 댓글 페이지 comment_url
    comment_url = 'https://comic.naver.com/comment/comment.nhn?titleId=626907&no=
    # Selenium을 활용, 댓글을 불러오기 위해 웹브라우저 실행
    driver = webdriver.Chrome()
    # '등록일', '제목', '평점', '베댓' 추가할 result 변수 생성
    result = []
    for page in range(1, end_page+1):
        # url 불러오기
        url = base_url.format(page)
        c_url = comment_url.format(page)
        # base_url '등록일', '제목', '평점'내용 requests로 추출
        res = requests.get(url)
        # comment_url '댓글' 내용 selenium으로 추출
        driver.get(c_url)
        # 클래스명 div.u_cbox_comment_box(댓글전체를 포함)을 담는 c_box변수 생성
        c_box = driver.find_elements_by_css_selector('div.u_cbox_comment_box')
        # 각 화의 베스트댓글을 담을 c_list 리스트 생성
        c_list = []
```

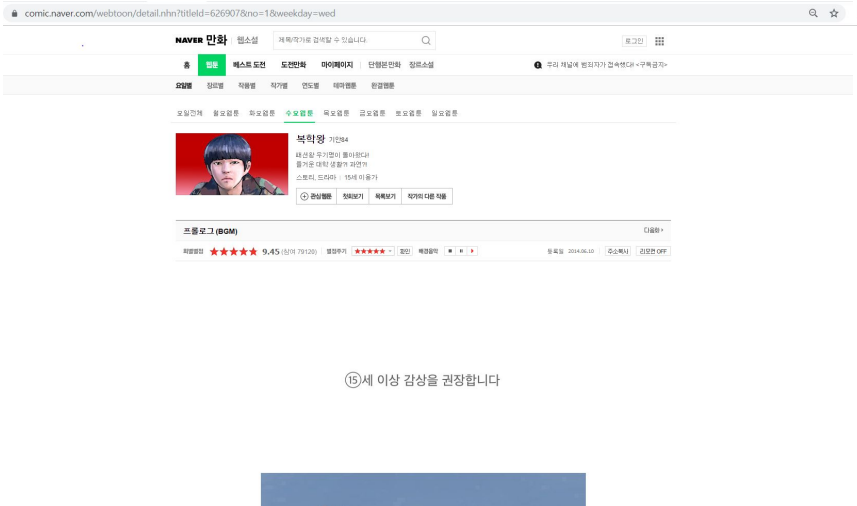
```
    for i in range(c_box.__len__()):
        # 댓글이 속한 클래스명을 try(if, else)로 c_list에 추가
        try:
            if c_box[i].find_element_by_css_selector('span.u_cbox_contents'):
                c_list.append(c_box[i].find_element_by_css_selector('span.u_c
            else:
                c_list.append(c_box[i].find_element_by_css_selector('span.u_c
        # span.u_cbox_contents, span.u_cbox_cleanbot_contents 외 댓글이 포함된 
        # 예외 이름 출력 후 continue로 계속 진행
        except Exception as ex:
            time.sleep(0.5)
            print("Unexpected error", ex)
            continue
    # BeautifulSoup을 사용해 '등록일', '제목', '평점' 추출
    if res.status_code == 200:
        soup = BeautifulSoup(res.text, 'lxml')
        title = soup.select_one('div.view h3').get_text().strip()
        point_list = soup.select_one('span#topPointTotalNumber').get_text().s
        re_date = soup.select_one('dd.date').get_text().strip()
        # 각 화의 '등록일', '제목', '평점', '베댓' result에 추가
        result.append([re_date, title, point_list, c_list])
```

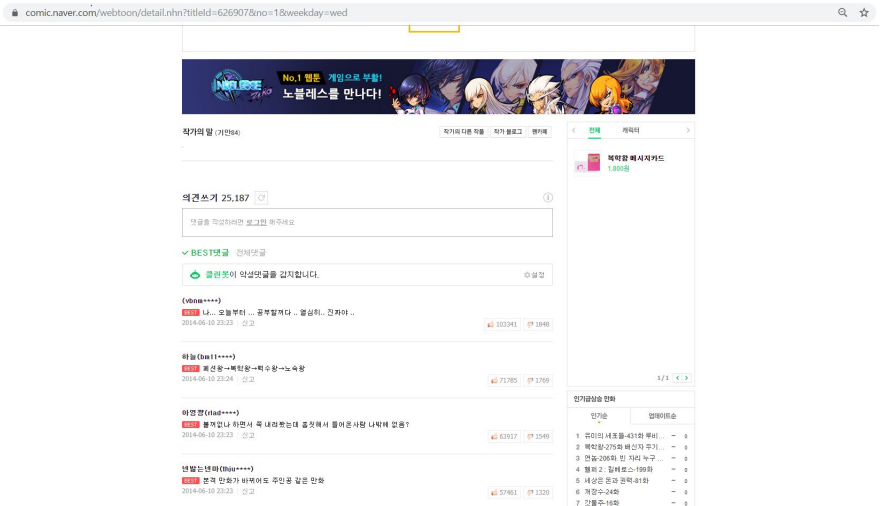
```
    # csv파일로 저장
    filename = "webtoon_{}.csv".format(end_page)
    df = pd.DataFrame(result, columns =['등록일', '제목', '평점', '베스트댓글'])
    df.to_csv("webtoon_{}.csv".format(end_page), index=False, encoding = 'UTF-8')
    return filename
if __name__=='__main__':
   NaverWebtoon(end_page=277)
```

5. 문제발생 및 해결 ①

문제 1: 베댓 왜 너만...

- 웹툰을 볼 수 있는 메인 페이지에서 다른 필요로하는 데이터는 수집 OK
- 베스트댓글 데이터는 크롤링이 안 됨

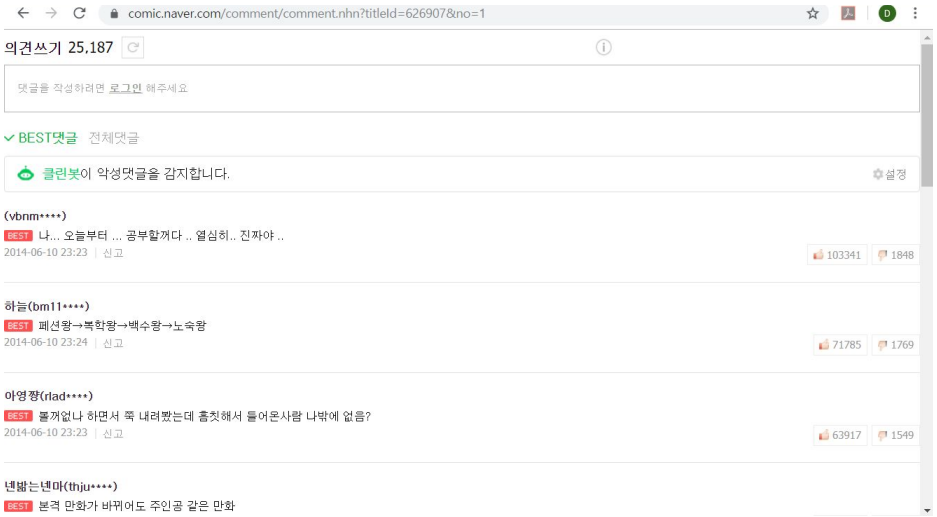




5. 문제발생 및 해결 ①

문제인식:

- 별도의 URL을 가지고 있는 베스트댓글 페이지 확인
  https://comic.naver.com/comment/comment.nhn?titleId=626907&no=1



5. 문제발생 및 해결 ①

문제인식:

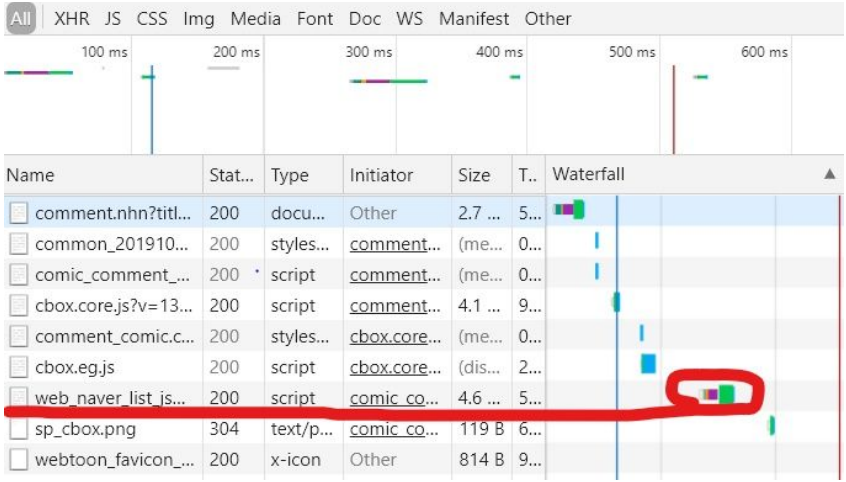
- requests 이용시 베스트댓글 데이터 수집이 안 됨
- 다른 회차의 페이지, 같은 태그를 입력해도 결과값은 계속 None 또는 없음

```
https://comic.naver.com/comment/comment.nhn?titleId=626907&no=1# <Response [200]>
None
[None, None, None, None, None, None, None, None, None, None, None, None, None, None]
https://comic.naver.com/comment/comment.nhn?titleId=626907&no=2# <Response [200]>
None
[None, None, None, None, None, None, None, None, None, None, None, None, None, None]
https://comic.naver.com/comment/comment.nhn?titleId=626907&no=3# <Response [200]>
None
[None, None, None, None, None, None, None, None, None, None, None, None, None, None]
https://comic.naver.com/comment/comment.nhn?titleId=626907&no=4# <Response [200]>
None
[None, None, None, None, None, None, None, None, None, None, None, None, None, None]
https://comic.naver.com/comment/comment.nhn?titleId=626907&no=5# <Response [200]>
None
[None, None, None, None, None, None, None, None, None, None, None, None, None, None]
```

5. 문제발생 및 해결 ①

문제인식:

- **Network 확인 : 트래픽 시간차 발견 → 댓글들이 동적으로 반응하여 렌더링 된다고 판단**



5. 문제발생 및 해결 ①

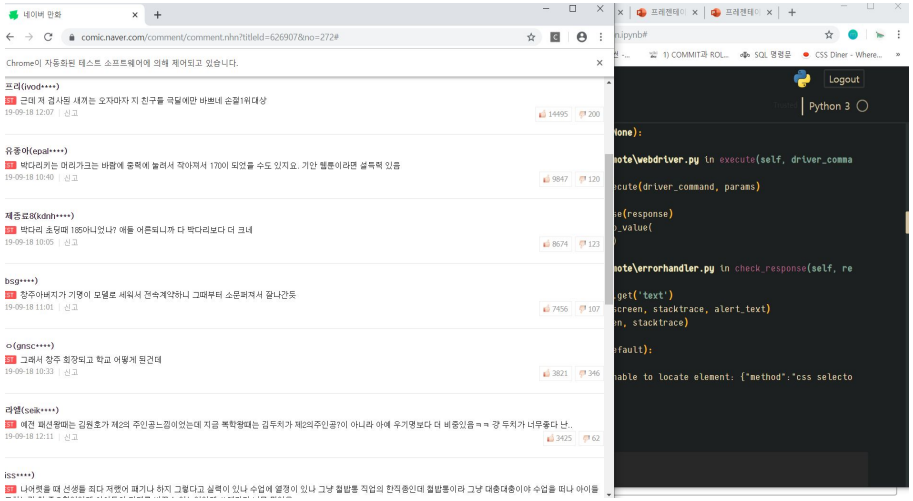
문제해결:

Selenium을 이용한
동적 크롤링 수행

5. 문제발생 및 해결 ②

문제 2: 클린봇의 방해

- page# 272 구간에서 지속적인 오류 발생
- skip 후 273부터 다시 실행해도 같은 오류 발생



5. 문제발생 및 해결 ②

문제인식:

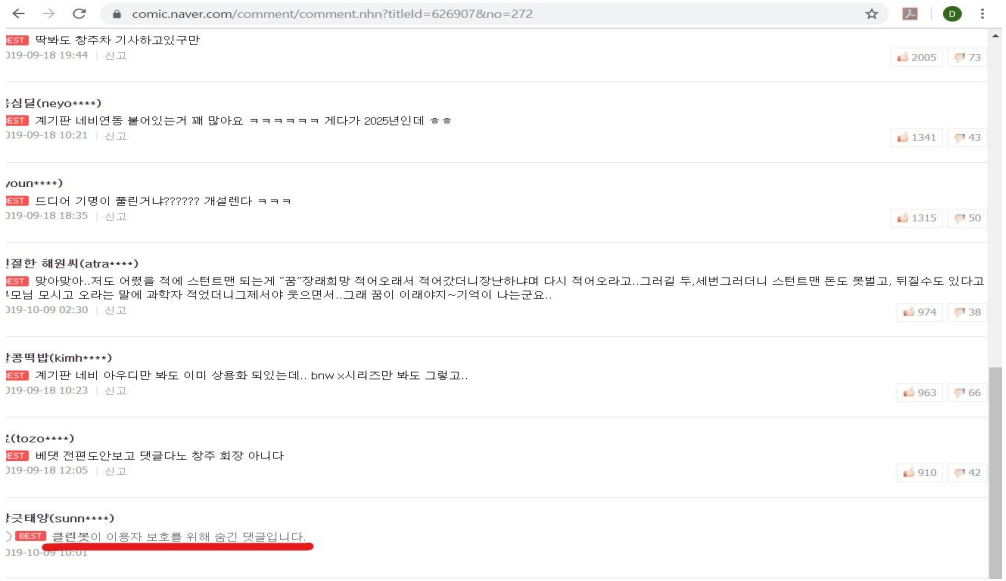
- 오류: “Unable to locate element…”
- 로딩 시간 or 지연 문제? time.sleep 사용에도 여전히 같은 문제 발생

```
    635     def find_element(self, by=By.ID, value=None):

~\Anaconda3\lib\site-packages\selenium\webdriver\remote\webdriver.py in execute(self, driver_comma
nd, params)
    319         response = self.command_executor.execute(driver_command, params)
    320         if response:
--> 321             self.error_handler.check_response(response)
    322             response['value'] = self._unwrap_value(
    323                 response.get('value', None))

~\Anaconda3\lib\site-packages\selenium\webdriver\remote\errorhandler.py in check_response(self, re
sponse)
    240                 alert_text = value['alert'].get('text')
    241             raise exception_class(message, screen, stacktrace, alert_text)
--> 242         raise exception_class(message, screen, stacktrace)
    243
    244     def _value_or_default(self, obj, key, default):

NoSuchElementException: Message: no such element: Unable to locate element: {"method":"css selecto
r","selector":"span[class=u_cbox_contents]"}
  (Session info: chrome=78.0.3904.97)
```

5. 문제발생 및 해결 ②

문제인식:

- **page# 272, 273에서 클린봇에 의해 숨겨진 댓글 발견**



5. 문제발생 및 해결 ②

문제인식:

- **클래스(class)명 차이 발견**

- 일반 댓글

```
<span class="u_cbox_contents" data-lang="ko">베댓 전편도안보고 댓글다노 창주 회장 아니다 </span>
```

- 클린봇에 의해 숨겨진 댓글

```
▼<span class="u_cbox_cleanbot_contents">
  ::before
  <em>클린봇</em>
  "이 이용자 보호를 위해 숨긴 댓글입니다."
</span>
```

5. 문제발생 및 해결 ②

문제해결: 코드 수정 후 정상적 실행

```
for i in range(c_box.__len__()):
    c_list.append(c_box[i].find_element_by_css_selector('span[class=u_cbox_contents]').text)
```

```
for i in range(c_box.__len__()):
    # 댓글이 속한 클래스명을 try(if, else)로 c_list에 추가
    try:
        if c_box[i].find_element_by_css_selector('span.u_cbox_contents'): # 코드수정
            c_list.append(c_box[i].find_element_by_css_selector('span.u_cbox_contents').text)# 코드수정
        else:
            c_list.append(c_box[i].find_element_by_css_selector('span.u_cbox_cleanbot_contents').text)# 코드수정
    # span.u_cbox_contents, span.u_cbox_cleanbot_contents 외 댓글이 포함된 클래스명 있을시 예외처리
    # 예외 이름 출력 후 continue로 계속 진행
    except Exception as ex:
        time.sleep(0.5)
        print("Unexpected error", ex)
        continue
```

6. 데이터 수집 결과 및 분석

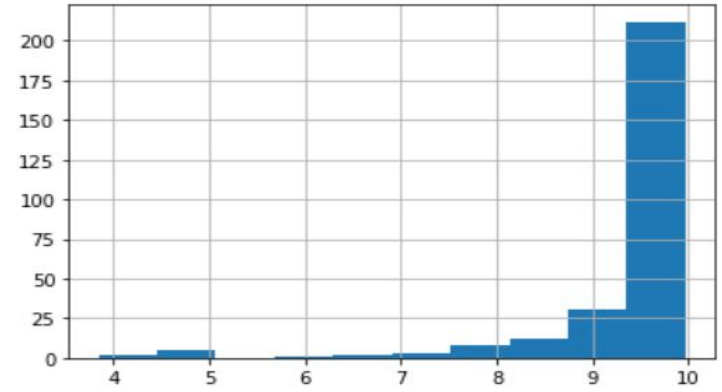
- 데이터 수집결과: **CSV** → [Google Sheet]
- 수집된 데이터 기본정보 및 분석 결과

```
import pandas

data = pandas.read_csv('webtoon_276.csv')
data.info()
```

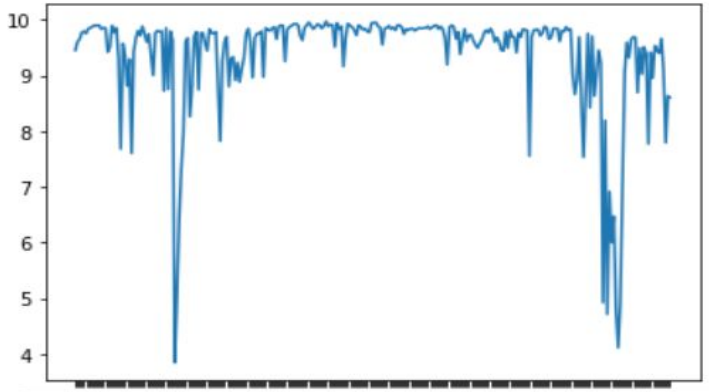
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 276 entries, 0 to 275
Data columns (total 4 columns):
등록일    276 non-null object
제목      276 non-null object
평점      276 non-null float64
베댓      276 non-null object
dtypes: float64(1), object(3)
memory usage: 8.7+ KB
```

	평점
count	276.000000
mean	9.371848
std	0.998056
min	3.840000
25%	9.410000
50%	9.750000
75%	9.840000
max	9.970000

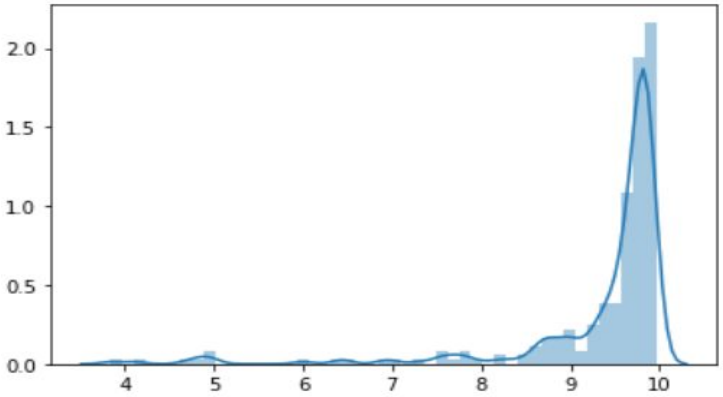
데이터 빈도, 평균, 편차, 최소값, 25%, 50%, 75%, 최대값 출력



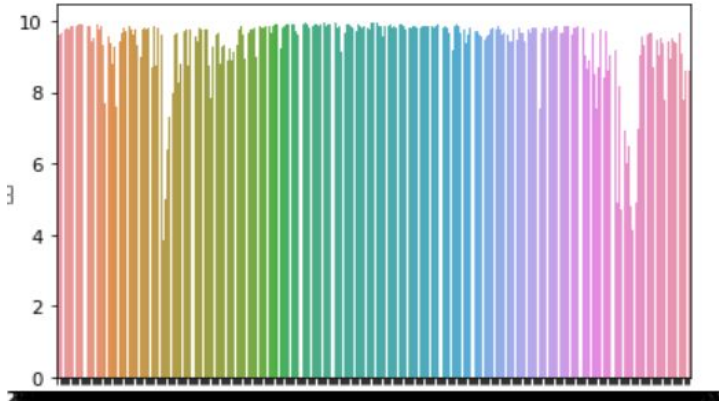
<10구간으로 나눈 히스토그램>



<시계열 꺾은선 그래프>



<혼합 분포도 그래프>



<시계열 막대 그래프>

- **평점 8.00 미만 회차 정보**

```
low_rate = data[data.평점 < 8.00]
low_rate.sort_values(by='평점', ascending=1).head(276)
```

	등록일	제목	평점	베댓
46	2015.04.28	46화. 바락 우바마 1	3.84	['오바마대통령이 이 만화보면 뭔가 불쾌할꺼같다..', '이건솔까오바다 내존경의대상...
251	2019.05.21	250화 세미나 3	4.11	['아 전개 이렇게 가는것도 짜증나고 독자 기만하는거 같은데아무생각없이 보다가 발 ...
246	2019.04.16	245화 장례식 2	4.71	['ㅋㅋ...
250	2019.05.14	249화 세미나 2	4.81	['생산직 무시도 문제지만 인종차별이 너무 노골적이다. 캅캅거리면서 더러운 숙소보고...
252	2019.05.28	251화 세미나 4	4.91	['아니ㅋㅋ245화장례식2에서 최존잘이 오디션2등했다며근데 왜 김원호 과거회상에서 ...
244	2019.04.02	243화 신년 마무리	4.92	['이젠 보면서 좀 짜증이 나려 그런다. 현실적인 면을 보여주려는건 알겠지만 굳이 ...
47	2015.05.05	47화. 바락 우바마 2	5.00	['이번화 빨리끝내라. 국제적 망신내지 말고', '도데체 오바마 대통령을 나오게 해...
248	2019.04.30	247화 성장	6.00	['기안아. 너도 니만화가 이제 감당이 안되지?', '다음주 나혼자산다예상 박나래:...
48	2015.05.12	48화. 바락 우바마 3	6.41	['넹 3주동안 다 까놓고 이제와서 착한척 하시는 베댓님들도 수고 많으십니다!', ...
249	2019.05.07	248화 세미나 1	6.47	['기안84 그냥 의식의 흐름대로 그리는구나? 전엔 그래도 뭔가 깊게 생각한번 할수...
247	2019.04.23	246화 장례식 3	6.91	['만우절 아니고 부활절인가', '유품이었던 2억 공중분해시켰는데 아는 척을 하네'...
253	2019.06.04	252화 세미나 5	6.99	['애들아 베댓좀보내줘라 기안옆집산다 베댓되면 쿠키값받으러간다', 'ㅋㅋㅋㅋㅋ 난 ...
49	2015.05.19	49화. 바락 우바마 4	7.31	['미국을 부러워하는 한국인, 그러나 한국인의 교육방식을 따라가고 싶어하는 미국. ...
235	2019.01.29	234화 김대리 마무리	7.53	['그냥 우기명 자살로 끝내자.', '또 봉지은 스멀스멀 기어나오네', 'ㅋㅋㅋㅋㅋ...
210	2018.08.07	209화 대학생 봉지은 4화	7.55	['계속 참다 한 번 못참아서 우기 인생 다시 내려가네ㅋㅋㅋㅋㅋㅋㅋ 이제 오디션도 ...
26	2014.12.10	26화. 박뽈테 3	7.60	['거기 긁고 봉지은 얼굴 만지는거 진심 더러웠다 손', '뽈테 집에서 혼자 헤헤 ...
21	2014.11.05	21화. 채플 1	7.68	['하나님... 기안이 갈수록 마감을 늦게 합니다. 초심으로 돌아가는 것을 막아주소...
265	2019.08.27	264화 인생 김창주 10	7.77	['두치 한 번도 진 적 없다 했음', '세븐이라 평점이 7.77인거냐;;', 'ㅋ...
273	2019.10.22	272화 배신자 우기명1화	7.79	['기안이 뭐했는데 댓글이 다 이모양이야', '우기명이 어떻게 배신했는지 알려줘야지...
67	2015.09.30	67화. 인생 봉지은 10	7.82	['요즘초딩들 놀토모름', '어제부터 상황정리1. 사람들이 11시쯤에 웹툰을 보려고...
50	2015.05.26	50화. 바락 우바마 5	7.97	['그냥 웹툰으로 받아들이자 제발', '참 댓글들도 웃긴게 에피소드초반에는 대통령 ...

7. 이슈/논란 관련 기사 수집 및 평점 영향 분석

- 평점이 낮은 회차의 원인을 사회적 이슈 및 논란을 통한 분석
- 출력한 ‘평점 **8.00** 미만 회차 정보’를 활용
- 베스트댓글 속 특정 단어들을 키워드로 활용하여 관련 기사 데이터 수집
    - 기사제목, 기사내용

```python
# 목록에 저장된 링크의 뉴스 기사 가져오기. (기사제목,기사내용)
import requests
from bs4 import BeautifulSoup
import pickle
import pandas as pd

# 목록 읽어오기
with open('기안84 오바마_검색결과.pkl','rb') as f:
    link_list = pickle.load(f)

# 네이버 스포츠 연예 뉴스는 일반 뉴스와 페이지 구조가 다름.
result = [] #[제목,내용]
for idx, link in enumerate(link_list):
    res = requests.get(link)
    try:
        if res.status_code == 200:
            soup = BeautifulSoup(res.text, 'lxml')
            title = soup.select_one('h2.end_tit').get_text().strip()
            print(title)
            news = soup.select_one('div#articeBody').get_text().strip()
            result.append([title,news])
    except Exception as e:
        print(e)
        print('{}번 기사를 못가져옴'.format(idx),link)
        continue

#csv 파일로 저장
pd.DataFrame(result, columns=['기사제목','기사내용']).\
                to_csv('2news_content.csv',encoding='UTF-8')
```

- 관련 기사 데이터 사례
  - 기안84 46화 논란 **(평점 3.84)**

[웹툰 복학왕]잠든 여대생에 흑심 품은 우바마? 비난 폭주

*… 이번 '복학왕' 46화에서 우바마는 국방비를 술값으로 사용하고, 잠든 봉지은에게 묘한 감정을 느끼는 등 매우 희화적으로 묘사됐다. 그러나 네티즌들은 마지막 장면에서 우바마가 잠든 봉지은을 쳐다보며 이성적 또는 성적 감정을 느끼는 듯 묘사된 것에 대해 "이건 모독이다(sewu****)", "너무 실망이 크다(pink****)" 등의 부정적인 반응을 보이고 있다.아이디 'dkgu****'는 "현직과 대통령을 떠나서 그만큼 각 나라에서 존경받는 분을 이렇게 개그물로 비아냥거리면 상대편 나라에서 불쾌해할 수도 있다"고 지적했으며, 아이디 'lord****'는 "무엇보다도 이건 비판 같은 내용이 아니라 그냥 한사람에 대한 모욕일 뿐"이라고 꼬집었다. 또한 아이디 'wjdg****'는 "오바마라서가 아니라 실존인물은 원래 건드리기 조심스러운거 아닌가"라며 작가가 신중하지 못했다는 반응을 보이기도 했다.이처럼 거센 비난 여론에 따라 우바마가 잠든 봉지은을 쳐다보며 얼굴을 붉히는 장면은 현재 삭제된 상태다. … -- 데일리안 스팟뉴스팀 ⓒ (주)데일리안*

복학왕 '바락 우바마' 등장에 악플 쏟아져 '또 늑대 변신하나'

네이버 수요웹툰 '복학왕 46화 바락 우바마1'편에 대한 네티즌 반응이 심상치 않습니다.지난 27일 업데이트 된 '바락 우바마1'편은 현재 평점 3.3점을 기록하는 등 좋지 않은 평가를 받고 있습니다.'바락 우바마1'편의 내용은 미 대통령 바락 우바마가 기안대학교를 방문해 각종 문제를 해결하는 장면을 그렸습니다. 바락 우바마는 기안대 원룸촌에 월세를 얻었고 간단하게 조깅을 하려 나서는 순간 우기명과 마주쳤습니다. 이는 그들의 두 번째 만남이었고 반가운 마음에 방에서 술을 마셨습니다.우기명은 학교 후배인 봉지은을 불렀고 셋은 다른 친구들도 불러가며 함께 술을 마셨습니다. 특히 끝에는 잠에 든 봉지은을 그윽하게 바라보는 우바마의 모습이 그려졌습니다. 팬들은 "오바마 대통령이 이 만화를 보면 불쾌할 것 같다" "점점 내용이 산으로 가네" "곧 닭도 나오겠네" 등 작가의 전작 '패션왕'에서 황당한 전개로 끝을 맺었던 일이 되풀이 되는 것이 아닌가 하며 우려를 나타내고 있습니다. Copyright ⓒ MBN(www.mbn.co.kr)

- **기안84 248화 논란 (평점 6.00)**

기안84, 논란사과 "불쾌한 표현 죄송, 더욱 신중 기할 것"

기안84가 생산직과 외국인 근로자를 비하했다는 논란에 대해 사과했다.17일 YTN Star 보도에 따르면 기안84는 많은 이들이 불쾌함을 느꼈을 표현에 대해 사과한다는 입장을 전했다. 아울러 "앞으로 웹툰 내용에 더욱 신중을 기하겠다"고도 말했다.지난 15일 다수 온라인 커뮤니티를 중심으로 기안84의 인기 웹툰 '복학왕' 249화에 인종차별과 생산직 비하 시각이 담겼다는 주장이 제기됐다. 누리꾼들은 웹툰 주인공 우기명이 근무하는 '기안식품' 직원들이 세미나 숙소에 도착한 장면에 대해 지적했다. 해당 장면에서 우기명은 지저분한 숙소를 보고 '좋은 방 좀 잡아주지'라고 생각하는 반면 외국인 근로자는 "우리 회사 최고다. 죽을 때까지 다닐 것"이라며 기뻐한다. 특히 기안84는 이 외국인 근로자의 대사마다 '캅'을 붙였고, 이를 두고 동남아시아 출신 외국인 근로자를 비하는 게 아니냐는 추측도 나왔다.기안84의 논란은 이뿐만이 아니다. 기안84는 지난 10일 '복학왕' 웹툰으로 장애인 비하 논란에도 휩싸인 바 있다.당시 전국장애인차별철폐연대(전장연)는 '복학왕'이 청각장애인 희화화 내용을 담고 있다며 공개 사과를 요구하는 입장문을 발표했다. 이에 기안84는 웹툰 말미에 사과문을 게재하고 논란을 일단락 지었다. MBN스타 대중문화부 김노을 기자 sunset@mkculture.com< Copyright ⓒ MBN(www.mbn.co.kr) 무단전재 및 재배포 금지 >

"앞으로 더 신중" 기안84 '외국인노동자 비하' 논란 사과…1주 만에 2차례

인기 웹툰 작가이자 방송인인 기안84(본명 김희민)가 최근 자신의 웹툰 '복학왕'이 장애인 비하 논란에 이어 인종차별 논란에 휩싸인 것에 대해 재차 사과했다.복학왕이 연재되고 있는 네이버 웹툰은 17일 "기안84 작가가 많은 분들이 불쾌함을 느끼셨을 표현에 대해 사과드린다는 입장을 전해왔다"며 "더불어 앞으로 내용에 더 신중을 기할 것이라는 말도 덧붙였다"고 한 매체를 통해 전했다.문제가 된 장면은 지난 14일 연재된 복학왕 249화(세미나 2)다. 웹툰 속 '기안식품' 직원들이 세미나 기간 동안 묵을 숙소를 보고 반응하는 장면이다. 지저분한 숙소를 보고 한국인 직원은 '좋은 방 좀 잡아 주지'라고 반응한 반면, 외국인 노동자는 "우리 회사 최고다. 죽을 때까지 다닐 거다. 캅캅캅!!"이라며 기뻐하는 모습으로 묘사됐다.만화를 본 일부 네티즌은 "노골적으로 인종차별을 했다"고 비판했다. … 복학왕은 앞서 248화도 논란이 됐다. 만화가 청각장애인을 희화화 하고 있다는 비판이다. 이에 기안84는 지난 10일 "많은 분들이 불쾌하실 수 있는 표현이 있었던 점에 사과말씀 드린다"고 한차례 사과했다. 박태근 동아닷컴 기자 ptk@donga.com ⓒ 동아일보 & donga.com, 무단 전재 및 재배포 금지

감사합니다

박슬기 박재홍 김예지 김대영