

디지털인문학입문 II 17분반

코로나 19 발생 이후 달라진 페스트 리뷰

크롤링 코드와 gephi 시각화를 이용하여

PAGE_001

연구문제

PAGE_002

데이터 수집

PAGE_003

데이터 시각화

PAGE_004

해석

Finish

Presentation

노어노문학과 이수현

불어불문학과 오혜민

불어불문학과 최조은

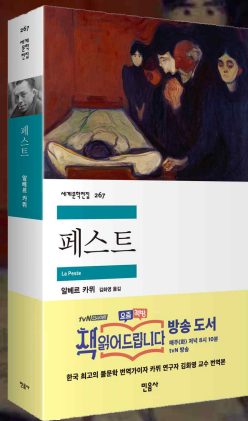
2022

PAGE_001

연구문제

불문학 고전으로 손꼽히는 알베르 카뮈의 <페스트>는 COVID19의 대유행으로 다시 한번 주목받았다. 실제로 코로나 전후로 <페스트>에 대한 사람들의 생각에는 유의미한 차이가 있을까?

온라인 서점 리뷰 크롤링



PAGE_002

데이터 수집

대표적인 온라인 서점 사이트 3곳을 대상으로,
민음사에서 출판한 <페스트>의 리뷰를 크롤링
함

총 587개 리뷰를 수집한 후, 각 리뷰를 형태소로
분리하였음

1. 단어 동시출현 분석

2. 단어 빈도 분석

3. 리뷰 추이 분석



414리뷰



115리뷰



115리뷰

```
[ ] act = ActionChains(driver)
element1 = driver.find_element(By.XPATH,
act.click(element1).perform()

[ ] commentReviewList > div.review_list_w
commentReviewList > div.review_list_w

'*[@id="divReviewPageMore"]

[ ]

evview_data=[]

or i in range(1, 30):
time.sleep(3)

html = driver.page_source
soup = BeautifulSoup(html, 'html.pa

cr = driver.find_elements(By.XPATH,

for r in cr:
r = r.find_elements(By.CSS_SELECT

for t in r:
review = t.text
review_data.append(t.text)
```

크롤링 코드-python

	A	B	C	D	E	F
1	words					
2	불어					
3	페스트					
4	섬세하다					
5	가독성					
6	오랜만					
7	코로나					
8	카워					
9	김화영					
10	페스트					
11	번역					
12	번역					
13	이방인					
14	이번					
15	최근					
16	인생					
17	불문학					
18	이제					
19	글					
20	오지다					
21	좋다					
22	전형					
23	이방인					
24	투쟁					
25	코로나					
26	번역					
27	나이					
28	다양					
29	카워					
30	번역					
31	페스트					
32	요즘					

형태소 분리-excel, 스프레드시트

```
[ ] text = '유명 작품 사다 생각 날 읽다'
words = text.split()
list(zip(words, words[1:]))

review =[]
review.append(list(zip(words[

[ ] review

[[('유명', '작품'), ('작품', '사다'), ('사
다', '생각'), ('생각', '날'), ('날', '읽
다')]]

[ ] f = open('/content/drive/MyDrive/kyo
f

[ ] word_network = []

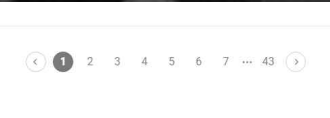
text = f

for i in f:
words = i.split()
list(zip(words, words[1:]))

word_network.append(list(zip(words
```

단어 동시출현 분석-python

동적 크롤링



```
[ ] #Colab에선 웹브라우저 창이 뜨지 않으므로 별도 설정한다.

options = webdriver.ChromeOptions()
options.add_argument('--headless') # Head-less 설정
options.add_argument('--no-sandbox')
options.add_argument('--disable-dev-shm-usage')
driver = webdriver.Chrome('chromedriver', options=options)
driver.get('https://product.kyobobook.co.kr/detail/S000000620404')

act = ActionChains(driver)
element1 = driver.find_element(By.XPATH, '//*[@id="contents"]/div[2]/div')
act.click(element1).perform()
```

```
act = ActionChains(driver)
element1 = driver.find_element(By.XPATH, '//*[@id="contents"]/div[2]/div')
act.click(element1).perform()
```

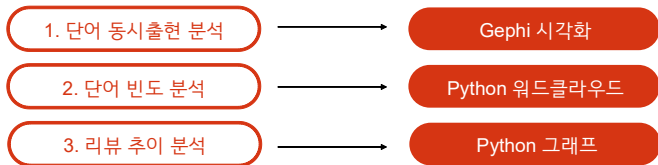
```
[ ] # 다음페이지 버튼 클릭
WebDriverWait(driver, 3).until(
```

```
[ ] date_list=[ ]
```

정적 크롤링

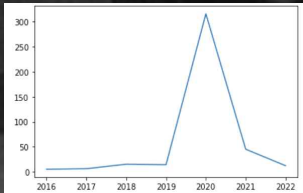
```
[ ] review_data = []  
    date_data = []  
  
from urllib.request import urlopen  
from bs4 import BeautifulSoup  
  
for i in range(1,24):  
    a = urlopen('https://www.yes24.com/Product/communityModules/GoodsReviewList/4827619?Sort=1&PageNumber='+str(i)+'&Type=ALL&_=1669898842373'  
  
    soup = BeautifulSoup(a.read(), 'html.parser')  
    b = soup.find_all("div", "review_cont")  
    c = soup.find_all("div", "review_etc")  
  
    for i in b:  
        review = i.get_text()  
        review_data.append(review)  
  
    for i in c:  
        date = i.get_text()  
        date_data.append(date)
```

데이터 시각화

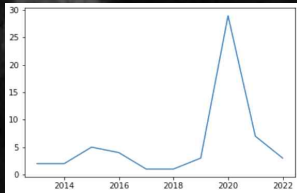


리뷰량 추이

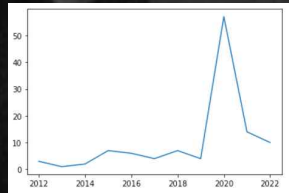
교보문고



알라딘



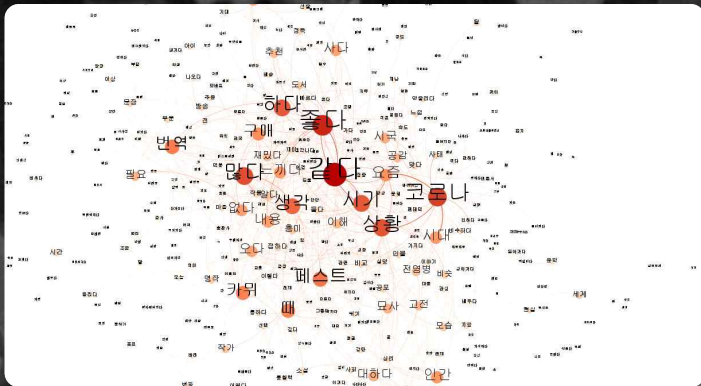
예스24



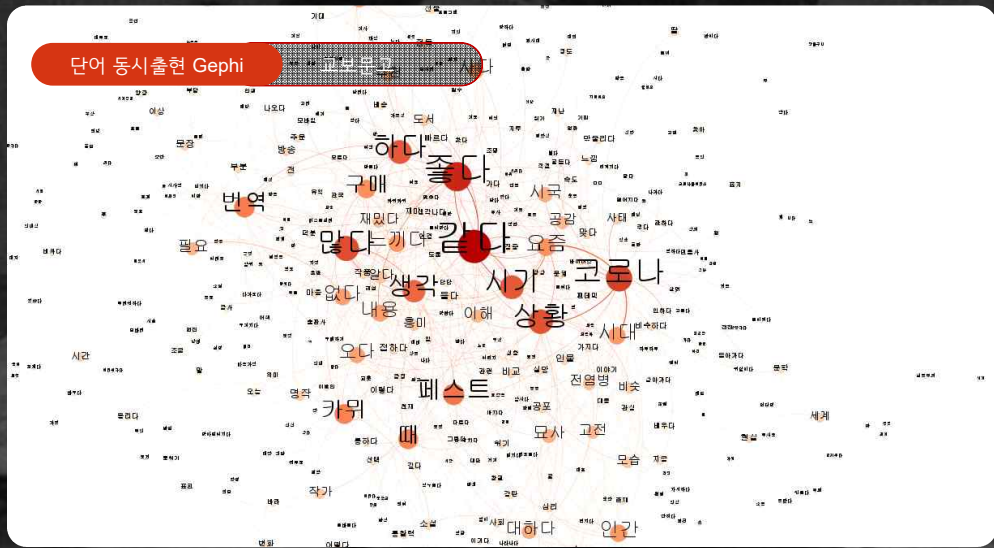
단어 동시출현 Gephi 교보문고

단어 동시출현 Gephi 교보문고

코로나 19 이후



단어 동시출현 Gephi



단어 동시출현 Gephi



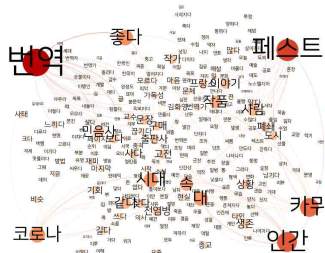
단어 동시출현 Gephi

알라딘

코로나 19
이전

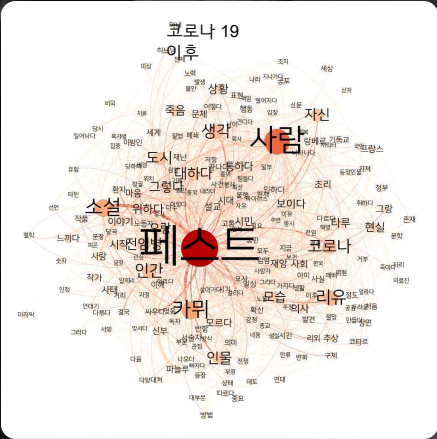
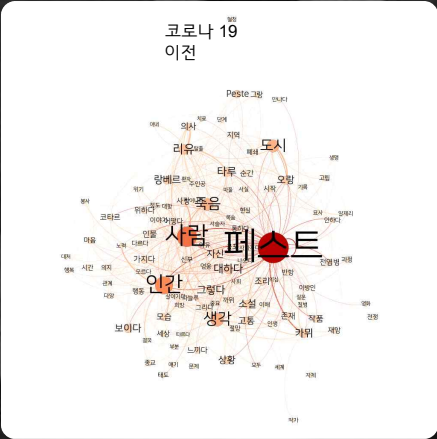


코로나 19
이후



단어 동시출현 Gephi 알라딘

알라딘



단어 동시출현 Gephi

```
[ ] 주요품사 = ["NNG", "NNP", "VV", "VA", "XR", "SL"]
    용언품사 = ["VV", "VA"]
    stop_words = "읽다 알베르"

## 형태소 분석 결과를 위해서 원하는 형태소만 찍어쓰기한 결과를 수집한다.

def read_documents(df, col):
    문서리스트 = []
    for index, row in df.iterrows():
        if row[col]:
            필터링결과 = [(token.form, token.tag) for token in row[col] if token.tag in 주요품사 and len(str(token.form)) > 1 and token.form not in stop_words]
            필터링결과 = [form+"다" if tag in 용언품사 else form for form, tag in 필터링결과]
```

정제 코드

- 1) '읽다, '알베르'를 불용어(stopword)로 지정했음
- 2) 2음절 이상의 단어만 추출했음



워드클라우드

교보문고/알라딘/예스24 리뷰 (2012년~2022년)



PAGE_004

해석

3곳의 사이트 모두 리뷰량이
기존의 10배 이상으로 증가

‘코로나’, ‘시대’, ‘페스트’, ‘같다’, ‘공감’ 의
동시출현 증가→ 코로나 시국을 소설 속
페스트의 창궐과 견주어 보는 사람들

고전이 시대의 거울이
되다



끝으로
프로젝트 소감

시계열 분석

감정어 분석

자료 정리의 중요성



Digital Humanities

Python & Gephi

디지털인문학입문 II 17분반

감사합니다

Presentation

노어노문학과 이수현

불어불문학과 오혜민

불어불문학과 최조은

2022