

트위터 크롤링을 통한 백신별 감성분석 & FAST COVID-19 LIVE



# 데이터 수집

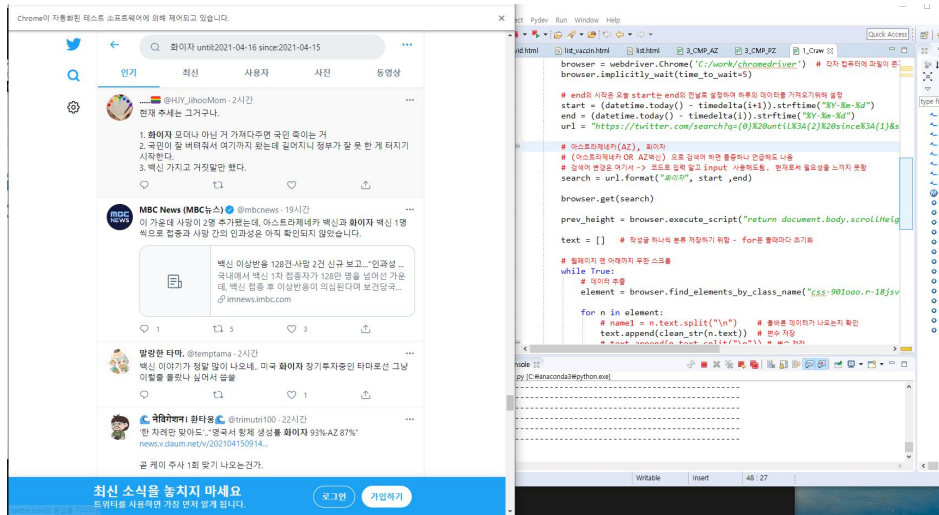


Seleniun 라이브러리를 사용해 트위터 페이지를 크롤링하여 백신별(영어/한글) 따로 수집

수집한 데이터를 가공해 감성분석



# 데이터 수집



Selenium을 이용한 백신별 Twit수집



# 데이터 가공

▶관련 링크 국경면 자유무역 관역간 인프라를 효율적으로 운영할 수 있도록 터미널 하는척 대항마르에서 물건 사고있는데 사람들 갑자기 정보도 알고 정보 좀 거기에 포기심때문에 다가가면 검역 되는거임 무사히 검역 피해서 편선 도착해야만 찾을 수있음 바이스1 ; 오세훈 한국 집중 속도 아프리카 평균보다 느리 5세후시 시장날 보자있는데 경찰중의 주 리버풀 교민들로 구성됐다 1천만 개대 도시 수장이 대분과 거지할 해도 되나구 /2021-04-14

▶이탈리아 집중 60 명은 50 명인 정원은 2 억 달러도 불행한 대한민국 국민들 조종 받은 것 같아 하루 이틀도 무사히 통과된 것 같아 수급 불안인 듯했는데 코로나19 발생은 세계 100여국을 뚫고든 문재인 임은 재영인임 하중인가 /2021-04-14

▶주요행사 코로나19 돌고 전역학적으로 나타났다 원상실정 중일 때는 미리 계약을 취소했다고 일에 가뭄을 풍고 나리더니 예상 구약한다나까 문재인 대통령이 온 국민을 실행대상으로 삼는다고 지난 물론 이 자 놓자 와 추종자들은 집중을 언할 것이 분명하니 현재 80세 이모와 통화 1차 맞았는데 똑같은 때를 맞아 똑같더라고 같은 부인만 조금 더 아프고 하루 지나니나까 괜찮더라고 동생이도 안 맞았어 지금 더 아프고 시합 몇 안맞아도 정부가 노인을 생각해서 준비해준다고 고집해 생각하고 있는것이 코로나가 무성지 주자가 뭐지 무성지고 그 코앞을 집중률 1도 안되면서 슬럼프 강행 다음 뉴스 20210414154555337 한국 기제기들은 꼭꼭 숨기는 뉴스/2021-04-14

▶27만와 영국의 메인 은 이다/2021-04-14

▶프랑스인 보류금 집중률 세계 111위 기회는 분석기 시대 과정은 분석기 시대 결과는 청동기 시대/2021-04-14

▶협동 관에서 크리스티안 타형하는 알반도 이스타엘이 팔레스타인 집중 안해놓는거는 조항하네 이런 심지에 걸 가듯 같은 이스타엘 갖버힐해도 나타나후 가는지었는데 하도 악역에서 2월에 이스타엘에서 알하는 팔레스타인 노동자 10만명은 해준다고 했는데 팔레스 독점 의 진 화자 부사장 걸 최고 과학자인 박사야 따르면 이 19 은 수백만명의 사람들에게 큰 행을 끼치고 심지어 사망까지 이르게 했다고 경고 하고 있습니다 /2021-04-14

▶북보 덴마크 아스트라제네카 코로나바이러스 사용을 영구적으로 중단/2021-04-14

▶일고 지고 일어난 느낌 아닐까 정도도 알함 그냥 필만기 한번 지나고 일어난 정도도 알함 사신 출근하려고 할 수 있을 것 같은데 쉬라고 하셔서이 이만 남에 중 서서안이나 귀해 어지러 주어 /2021-04-14

▶이렇게 듣고 부작용으로 집중어 중단하는데 이는 7백만 도스 중 6개 케이스로 이런 부작용을 경험할 가능성은 반개에 10만 맞는 것과 같다고 0 0001 하지만 조사를 위해 중단했는데 이미 우리 가 같은 미국 사회 분위기에 악영향을 미칠까 우려 /2021 연도분 이따리가 한바의 햇살을 다 머금고 있는 듯 빛난다 2차 맞고 나오는 걸 난 날 아를까 /2021-04-14

▶중공산 을 잊고 싶으면 정상적인 일상생활 살아갈수없는 공산사회주의식 여권사회가 바로 눈앞에 다가온다 개인의 사생활과 일거수 일투족이 공산주의 중앙정부에 의해서 통제되는 줄같은 세상이 현실이 되어간다 과연 생존을 위해 순응해야 할것인가 아니면 개인도 용버야 하로운 준비자 아스트라 제네카 마저도 더 구하기 힘들어질 것보 예상함 /2021-04-14

▶아제 오세훈이 국무회의에서 방역지원 완화를 거론하자 오늘 700명대 확진자로 확대한다 설 추석이 다가오면 증가폭도 돌아서고 시정당대는 진희정선언 평화문공사를 집행하고 큰 소리 뻔뻔 친 은 김강무사식이 모든게 까마귀 날자 뿔여지는 사견 특성이다 문재인 코로나 공급 및안 가면 지원 문제안 방역 성공했다 자와자만 했지만 국민은 집중 못해 하루하루 불안한 고층으로 살고 있다 문재인 정부가 수입을 늦춘 이유는 중봉시위 같은 반정부 시위가 일어날까 두려워 그렇지 않았을까 문재인 정부에서는 국민 생활은 안고 안고의 3 4배 어치의 술 선구매 계약해 놓고 기간진 사자하라 그렇다 생각 알만 몰랐어 말리던 어디서 이상한 소리 나출지 모르니 화병겨놓고 나중에 물건들이올때쯤 되면 될 이거까지 이유로 돈 안출수 있는 방법은 뭐 무궁무진 하지말 거다 같은진 3월은 임박하며 가꾸어가면서 국민 기쁨 의뢰인성 도 집중못하고 임어중 울미함 조국 추미애 같은 정치스태기들이 세금 타먹고 대응성이나 직원들 월급도 세금인데 내부정보로 방투기나 하고 비슷한 집단이 있는데 그게 조국외도 가뭄개울은 죽도록 상남이나 해 정부를 맡고 울 맞아 주세로 정통위에서 역언어에 관한 두들음 밀고있고 /2021-04-14

▶여러분 큰일났습디다 최파장부가 드려 알을벌일 모양입니다 중공산 을 들여와 집중시키고 중공산의 여권을 도입하여 중공산 을 집중받지 않은 국민의 사회생활을 원형분해할 계획들을 세우고 있습니다 모두 이같은 반행을 인지하고 널리알려 경제합시다 /2021-04-14

▶이러분 큰일났습디다 최파장부가 드려 알을벌일 모양입니다 중공산 을 들여와 집중시키고 중공산의 여권을 도입하여 중공산 을 집중받지 않은 국민의 사회생활을 원형분해할 계획들을 세우고 있습니다 모두 이같은 반행을 인지하고 널리알려 경제합시다 /2021-04-14

▶이러분 큰일났습디다 최파장부가 드려 알을벌일 모양입니다 중공산 을 들여와 집중시키고 중공산의 여권을 도입하여 중공산 을 집중받지 않은 국민의 사회생활을 원형분해할 계획들을 세우고 있습니다 모두 이같은 반행을 인지하고 널리알려 경제합시다 /2021-04-14

▶유사지가 이런 기사를 보내보내니 적어도 언데 집중 주 요양원서 시설 확진자 85 갑스 사망자도 127명 3일 중앙 /2021-04-14

▶이 문제가 얼마나 심각한지 하만 알아야 아재집 집계 15일에 최자자 집중하는결과 확정 받고 기다리고 있었는데 어제 갑자기 최자와 최자 무기한 연기 되었고 하더래 그게 언제나나까 보건국에서 알 수 없다고 하더래 /2021-04-14

▶정부가 노인을 상대로 시가 친것이지 시교 음연등을 통해 집중 신형하는 것부터 사기였지 /2021-04-14

▶대통령 신형한 이전은 지난날 중공산 을 집법적으로 대항수입하여 우리국민에게 집중시작기 위해 의 원상지와 성분표기를 없애는 게같은데 발생이법한 발의를 추진하다 발각되어 더 다수국민들의 국책발판시사의 법안발의에 대한 반대서명운동을 일으켰던 0 대명은 뜻서민들은 문재인 대통령 조국 장관 추미애 장관 박원순 시장을 지지하고 100여명을 끌어 들인 것 이게 정상적인 인간의 사고방식인가 한번 토라임을 이날 20210414103918834 /2021-04-14

▶손은이 보스 의 침범의 위기 3일의 정형에 현안이 생기고 그로 안에서 뇌출혈과 뇌경색이 생긴다 /2021-04-14

▶현재의 공급 부족은 결국 적년 여름의 오만이 맞지만 합성인대 뽀뽀는 정부 정책을 비난 지지 오프라인에 앞서 이 정책을 수행할 정부가 반대 정부인 경우를 가정해 보는 게 낫지 않나 싶다 물론 그런 반성적 사고가 가능하더라 문해나 골부수가 되지 말 이상한부 부족을 면 회자보다 3배 더 많이 집중 다음 뉴스 20210414175153379 아스트라 산상만 하고 싶었다는 미국은 올해엔 1755명 사망 보고되었는데 미 부작을 보고 싶습디다 /2021-04-14

▶중공산 을 잊고 싶으면 정상적인 일상생활 살아갈수없는 공산사회주의식 여권사회가 바로 눈앞에 다가온다 개인의 사생활과 일거수 일투족이 공산주의 중앙정부에 의해서 통제되는 줄같은 세상이 현실이 되어간다 과연 생존을 위해 순응해야 할것인가 아니면 개인도 용버야 하로운 준비자 아스트라 제네카 마저도 더 구하기 힘들어질 것보 예상함 /2021-04-14

▶어떻게 죽을수있어 이리도 많아서 슬러노 /2021-04-14

▶스가 풀리는 미국어 구하려 가는 것으로 대가는 미국적 매일 확대 /2021-04-14

▶스카파투 코로나19 감염자 수 2일 연속 1천명 넘어 1130명으로 과거 최대 210414 2104140040 1 /2021-04-14

▶수급 자신있고 뒤돌고 회망고분 하나 5 부등산만큼은 자신있고 단 문재인 부등산 어떻게 됐나 / 수급 자신있고 2021-04-14

▶코로나 극복 소중함 일상 회복의 중요성만 하면 집중 두려움과 간헐은 개고 알다못 해야만 되지만 생각하면 승기 강소정분부 방위산업에서 고생하시는 모든분들께 감사드리며 더 많은 분들의 집중으로 코로나로부러 안전한 강원도를 만들어나갔습니다 한국에 4000만분쯤 혼자던 모나리 문 대통령과의 약속 이 드로날 수있음 뿐이다 은 소로 얻어서는게 아니다 /2021-04-14

## 수집한 데이터를 txt파일 형식으로 1차 가공



# 데이터 가공

negative\_words\_twit - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

혈전증  
혈전  
사망  
부작용  
불가피  
더럽고  
사망자  
사퇴  
차질  
개소리  
불신  
개짓거리  
개탄  
고통  
이상반응자  
헛소리  
대유행  
근육통  
오보  
미열  
가짜  
위험  
근육통  
논란  
여지려고

positive\_words\_twit - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

호응  
돌파  
좋은  
효과  
확보  
성과  
예방  
면역  
안전한  
자궁경부암  
혈장공여  
건강한  
백신확보  
신속  
개발  
안정적  
대박  
안아픔  
대박이네  
접종  
면역  
최고  
자유  
좋다  
고피

긍정, 부정, 중립 감성별 키워드로 데이터 재가공



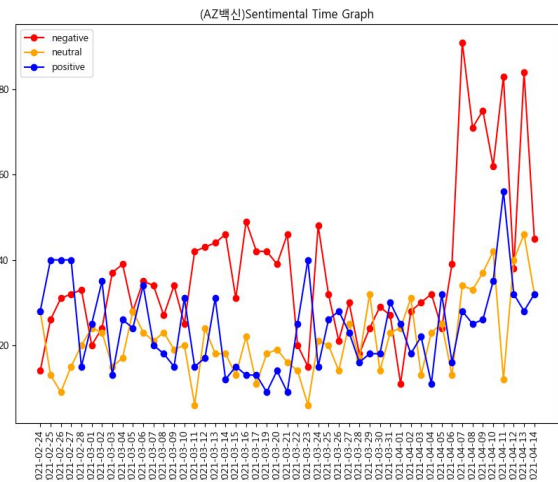
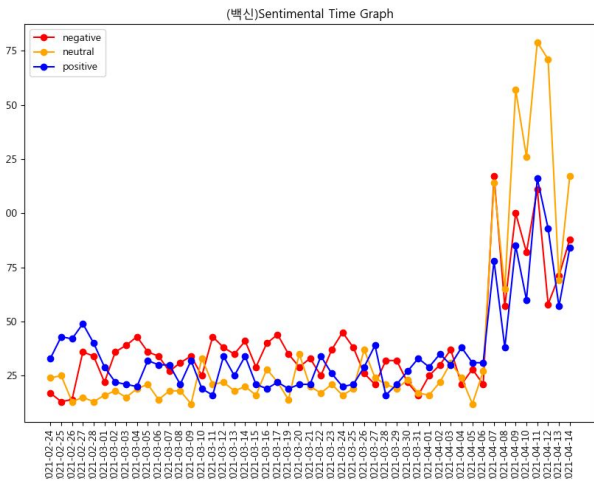
# 데이터 가공

## 키워드별 빈도수 분석

freq_AZ - Windows 메모장				freq_AZ - Windows 메모장				freq_AZ - Windows 메모장					
파일(F)	편집(E)	서식(O)	보기(V)	파일(F)	편집(E)	서식(O)	보기(V)	도움말(H)	파일(F)	편집(E)	서식(O)	보기(V)	도움말(H)
운빨:1				자다:97					백신:6467				
유추:1				입원:97					아스:4741				
번지:1				종합:97					접종:3774				
파스카:1				세계:97					맞다:2004				
소아:1				코로나바이러스:95					있다:1351				
최고경영자:1				분기:94					되다:1266				
실장:1				수출:94					화이자:1134				
상탭니:1				내일:93					코로나:1094				
주작:1				회의:93					에서:981				
후자:1				환자:92					뉴스:779				
한일:1				진자:92					이상:705				
아시:1				보고:92					없다:705				
맨뒤:1				위탁:92					이다:680				
분란:1				공장:91					효과:648				
다툼:1				사이언스:91					보다:516				
처지다:1				사실:90					생산:496				
긴밀하다:1				사망자:90					혈전:484				
래서:1				증상:89					대통령:447				
멕이기:1				일부:89					않다:446				
초도:1				이나:89					유럽:445				
버거운:1				계속:89					아니다:444				
판매승인:1				아직:87					국내:444				
기왕:1				쓰다:87					영국:423				
권영미:1				시키다:86					부터:423				
데이터베이스:1				회분:86					부작용:393				
변국:1				하나:85					한국:388				
개르:1				개발:85					국민:383				
징발:1				계약:85					출처:370				
대공황:1				중증:84					문재인:356				
소득세:1				중사:84					중단:352				
물수:1				청장:84					공급:345				
				주사기:83					하고:332				
				제외:83					까지:317				
									다음:317				
									정부:315				



# 분석 결과



시간별 부정, 중립 긍정 빈도수 변화(한글 트윗)



# 분석 결과

```
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.font_manager as fm

fm.get_fontconfig_fonts()
font_location = 'C:/Windows/Fonts/malgun.ttf'
font_name = fm.FontProperties(fname=font_location).get_name()
plt.rc('font', family=font_name)

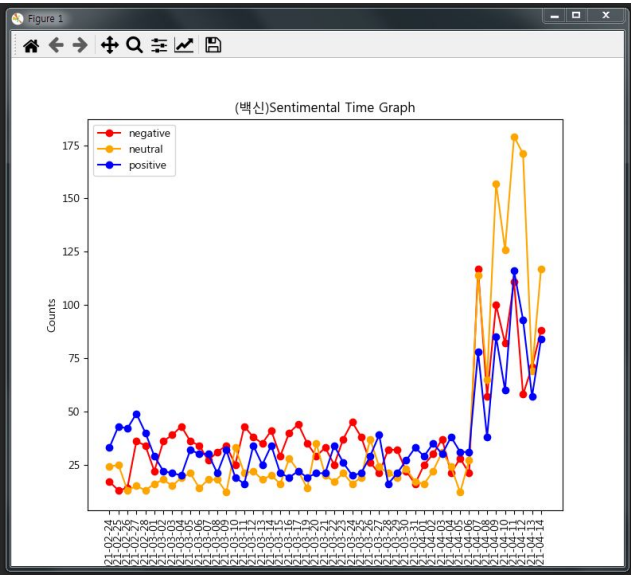
df_Vac = pd.read_excel('./data/result(백신).xlsx')
df_Vac = df_Vac.loc[:, ['date', 'Label']]

# '백신' 에 관련된 timegraph
df_Vac2 = df_Vac['Label'].groupby([df_Vac['date'], df_Vac['Label']]).count()
print(type(df_Vac2))
df_Vac2 = df_Vac2.reset_index(name='count')
print(df_Vac2)

neg_df = df_Vac2.query('Label=="-1"')
days = neg_df['date']
neu_df = df_Vac2.query('Label=="0"')
pos_df = df_Vac2.query('Label=="1"')

fig = plt.figure(figsize=(10,7)) ## 캔버스 생성
fig.set_facecolor('white') ## 캔버스 색상 설정
ax = fig.add_subplot() ## 그림 해대(프레임) 생성

ax.plot(days, neg_df['count'], markers='o', label='negative', c='r') ## 선그래프 생성
ax.plot(days, neu_df['count'], markers='o', label='neutral', c='orange')
ax.plot(days, pos_df['count'], marker='o', label='positive', c='b')
plt.title('백신)Sentimental Time Graph')
plt.xlabel('Date')
plt.ylabel('Counts')
plt.legend()
plt.xticks(rotation=90)
plt.savefig('백신)Sentimental Time Graph')
plt.show()
```

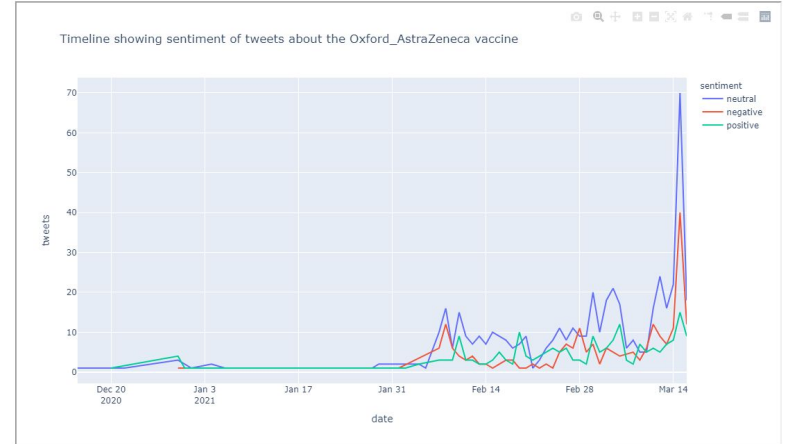
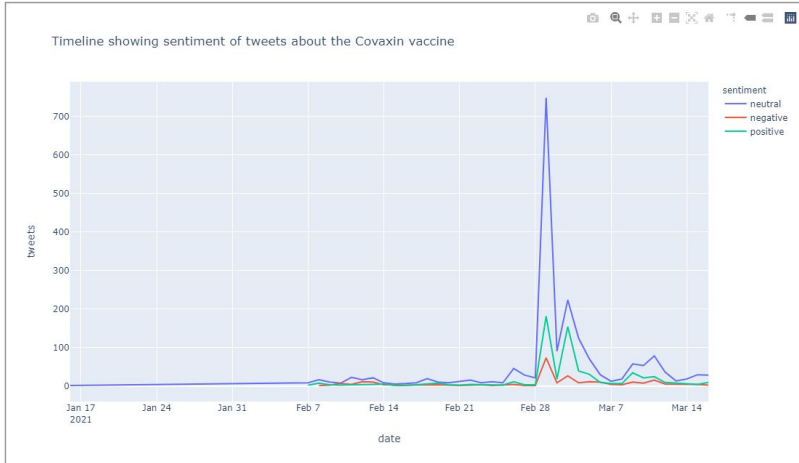


시간별 부정, 중립 긍정 빈도수 변화(한글 트윗)





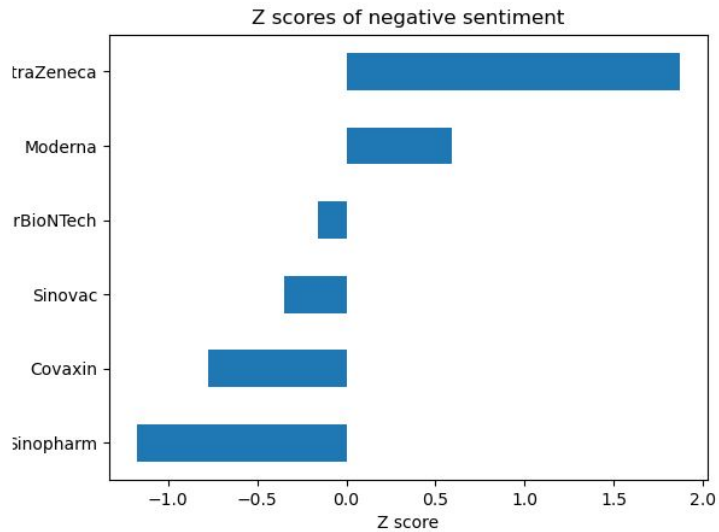
# 분석 결과



시간별 부정, 중립 긍정 빈도수 변화 (영어 트윗)



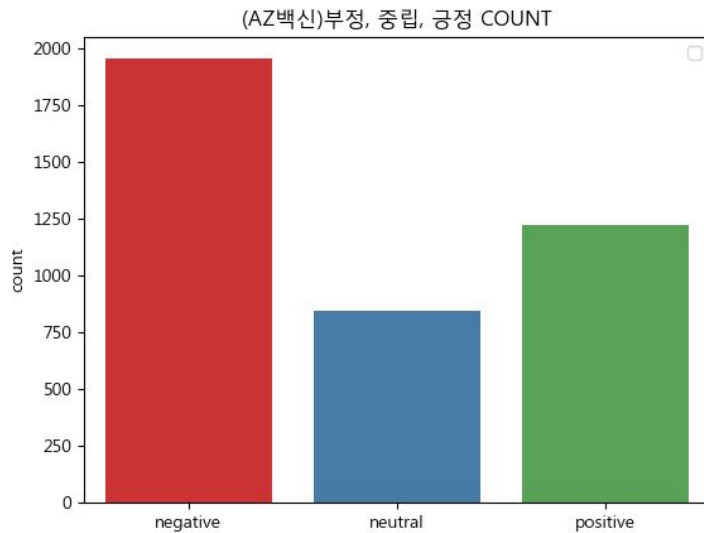
# 분석 결과



백신별 부정적인 지수 비교



# 분석 결과



부정 / 중립 / 긍정 지수비교



# 분석 결과

```

# 부정 단어
negative_data = pnn_data['Label'] == -1
negative_data = pnn_data[negative_data]
negative_data = negative_data.drop("Label", axis=1)
print(negative_data)
# 중립 단어
neutral_data = pnn_data['Label'] == 0
neutral_data = pnn_data[neutral_data]
neutral_data = neutral_data.drop("Label", axis=1)
print(neutral_data)
# 긍정 단어
positive_data = pnn_data['Label'] == 1
positive_data = pnn_data[positive_data]
positive_data = positive_data.drop("Label", axis=1)
print(positive_data)
...

#####
# 트위터 (빈도수기준) 워드클라우드
twit_coloring = np.array(Image.open('twit.png'))
if search_text == '백신':
    twit_coloring = np.array(Image.open('v_image.png'))
from wordcloud import ImageColorGenerator
image_colors = ImageColorGenerator(twit_coloring)

covid_wc = WordCloud(font_path = font_location, background_color='white',width=1000, height=500, random_state=1)

fig, ax = plt.subplots(figsize=(12,6))
plt.imshow(covid_wc, interpolation='bilinear')
plt.axis('off')
plt.title('(' + search_text + ')wordcloud')
plt.savefig('wordcloud(' + search_text + ').png')
plt.show()

...
# (은감 자체에서 단어를 뽑아 클라우딩)

```



백신별 워드 클라우드



# 개선 사항

## 모델 평가(AZ)

정확도 : 0.9529120326042175

손실값 : 0.3152697682380676

## 모델 평가(PZ)

정확도 : 0.9603710174560547

손실값 : 0.2377401441335678

Ashley 애슐리는



Has 가지고 있다..  
무얼 가지고 있을까?



A great figure  
훌륭한 몸매를



영어는 결론을 먼저 내리고

나는 ...



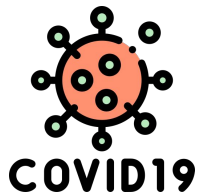
훌륭한 몸매를...



부러워한다~ ㅠ.ㅠ



우리말은 끝까지 들어보아야 안다



# FAST COVID-19 LIVE

실시간 코로나 현황을 볼 수 있는 웹을 개발해  
기존해 분석한 백신 데이터를 사이트에 접목시키기로함.



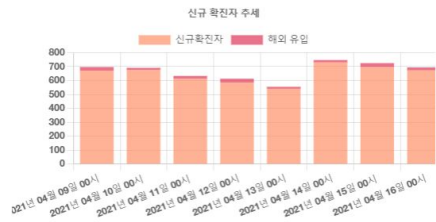
COVID-19

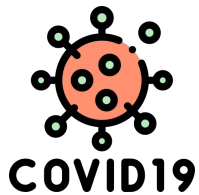
VACCIN

마지막 업데이트 : 2021-04-16 09:37:24.578

확진자	격리해제	사망자	격리중
112789	103062	1790	7937
+673	+0	+8	+1079

신규 확진자 : 해외유입 +21 / 지역발생 +652





# FAST COVID-19 LIVE

총4건

보건의료	한국지능정보사회진흥원
활용신청	[승인] 보건복지부_코로나19 연령별·성별감염_현황
신청일	2021-04-01
만료예정일	2023-04-01

보건의료	한국지능정보사회진흥원
활용신청	[승인] 보건복지부_코로나19 감염_현황
신청일	2021-04-01
만료예정일	2023-04-01

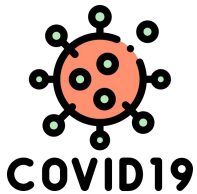
  

보건의료	한국지능정보사회진흥원
활용신청	[승인] 보건복지부_코로나19 시·도발생_현황
신청일	2021-04-01
만료예정일	2023-04-01

보건의료	한국지능정보사회진흥원
활용신청	[승인] 보건복지부_코로나19해외발생_현황
신청일	2021-04-01
만료예정일	2023-04-01

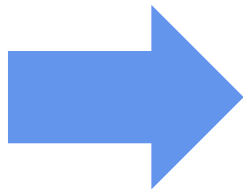
공공 데이터API를 활용해 국가별/지역별/남녀별/세대별  
감염, 예방 현황을 시각화하기로함.



# FAST COVID-19 LIVE

This XML file does not appear to have any style information associated with it. The document tree is is

```
r<response>
  <header>
    <resultCode>00</resultCode>
    <resultMsg>NORMAL SERVICE.</resultMsg>
  </header>
  <body>
    <items>
      <item>
        <createDt>2020-04-10 11:17:35.35</createDt>
        <deathCnt>0</deathCnt>
        <defCnt>352</defCnt>
        <gubun>강원</gubun>
        <gubunCh>원주</gubunCh>
        <gubunEn>Lazaretto</gubunEn>
        <incDec>4</incDec>
        <isolClearCnt>3</isolClearCnt>
        <isolIngCnt>349</isolIngCnt>
        <localOccCnt>0</localOccCnt>
        <overFlowCnt>4</overFlowCnt>
        <aurRate></aurRate>
        <seq>1014</seq>
        <stdDay>2020년 04월 10일 00시</stdDay>
        <updateDt>NULL</updateDt>
      </item>
      <item>
        <createDt>2020-04-10 11:17:35.35</createDt>
        <deathCnt>0</deathCnt>
        <defCnt>12</defCnt>
        <gubun>제주</gubun>
        <gubunCh>제주</gubunCh>
        <gubunEn>Jeju</gubunEn>
        <incDec>0</incDec>
        <isolClearCnt>4</isolClearCnt>
        <isolIngCnt>8</isolIngCnt>
        <localOccCnt>0</localOccCnt>
        <overFlowCnt>0</overFlowCnt>
        <aurRate>1.79</aurRate>
        <seq>1013</seq>
        <stdDay>2020년 04월 10일 00시</stdDay>
        <updateDt>NULL</updateDt>
      </item>
      <item>
        <createDt>2020-04-10 11:17:35.35</createDt>
        <deathCnt>0</deathCnt>
        <defCnt>115</defCnt>
        <gubun>경남</gubun>
        <gubunCh>진주</gubunCh>
```



```
def get_corona_data(startCreateDt, endCreateDt):
    params = {
        'serviceKey': serviceKey_2,
        'pageNo': '1',
        'numOfRows': 10,
        'startCreateDt': startCreateDt,
        'endCreateDt': endCreateDt,
    }

    res = requests.get(url=url, params=params)
    # print(res.url)
    # print(res.text)

    # xml -> dict
    dict_data = xmltodict.parse(res.text)
    # print(dict_data)

    # dict -> json
    json_data = json.dumps(dict_data)
    # print(json_data, type(json_data))

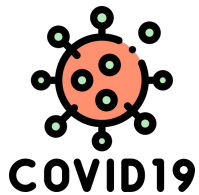
    # json -> dict
    dict_data = json.loads(json_data)
    # print(dict_data, type(dict_data))
    # pprint(dict_data['response']['body']['items']['item'])

    # total Cnt Check
    totalCount = dict_data['response']['body']['totalCount']
    if totalCount == "0":
        return False

    # 지역정보를 담은 리스트 저장
    area_data = dict_data['response']['body']['items']['item']
    area_data.reverse()
    # pprint(area_data)
    for a in area_data:
        print(a)
    return area_data
```

인증키를 받아 태그값을 추적해 원하는 데이터 추출 후  
활용하기 쉬운 형태로 가공.



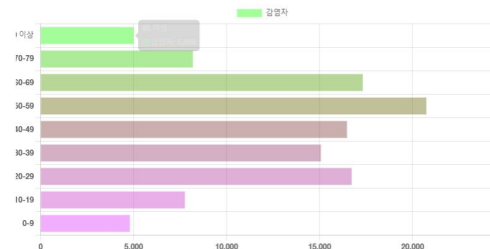


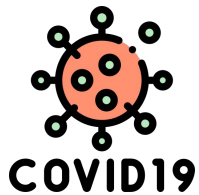
# FAST COVID-19 LIVE



## Chart.js

Chart.js를 통해 웹에서 데이터 시각화





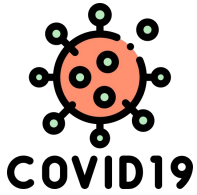
## FAST COVID-19 LIVE

This XML file does not appear to have any style information associated with it. The document tree is shown below.

---

```
▼ <response>
  ▼ <header>
    <resultCode>99</resultCode>
    <resultMsg>LIMITED NUMBER OF SERVICE REQUESTS EXCEEDS ERROR.</resultMsg>
  </header>
</response>
```

사이트의 잦은 열람 시 API 접근오류가  
걸리는 문제 발생



# FAST COVID-19 LIVE

1000

## 활용신청 상세기능정보

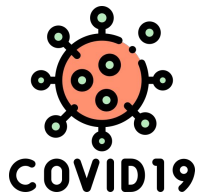
NO	상세기능	설명	일일 트래픽	미리보기
1	코로나19감염현황 조회 서비스	코로나19감염증으로 인한 일별 확진자, 원치자, 치료중인환자, 사망자 등에 대한 현황자료	1000	<input type="button" value="확인"/>

## 요청변수(Request Parameter)

[닫기](#)

항목명	샘플데이터	설명
ServiceKey	-	공공데이터포털에서 받은 인증키
pageNo	1	페이지번호
numOfRows	10	한 페이지 결과 수
startCreateDt	20200310	검색할 생성일 범위의 시작
endCreateDt	20200414	검색할 생성일 범위의 종료

일일 트래픽 1000 제한이 원인.



## FAST COVID-19 LIVE

```
url = "http://openapi.data.go.kr/openapi/service/rest/Covid19/getCovid19SidoInfStateJson"
url_gender = "http://openapi.data.go.kr/openapi/service/rest/Covid19/getCovid19GenAgeCaseInfJson"
serviceKey_1 = '2PLDIVNdhJ3pWgmyk2qXL2LekLrwfv5r8z2tq6[REDACTED]'
serviceKey_2 = 'h92s7QADkL1WM1LYJeYFz0x4Fmh+CV5s/VhL0j[REDACTED]'
serviceKey_gender1= '2PLDIVNdhJ3pWgmyk2qXL2LekLrwfv5r8[REDACTED]nqtEA%3D%3D'
serviceKey_gender2='h92s7QADkL1WM1LYJeYFz0x4Fmh+CV5s/V[REDACTED]Kw=='
url_vaccin = "https://nip.kdca.go.kr/ingd/cov19stats.do?list=all"
```

팀원들의 키값을 할당 받아 키오류시 다른 키를 받아 오도록  
개선.

**THANK YOU**

**THANK YOU**

**THANK YOU**

**THANK YOU**

**THANK YOU**