경기도보도자료

2020 공공 빅데이터 청년 인턴십
송준영

Global Inspiration
세계 속의 경기도

목차

우리는 공공기관에서 어떤 업무를 맡게 될까?

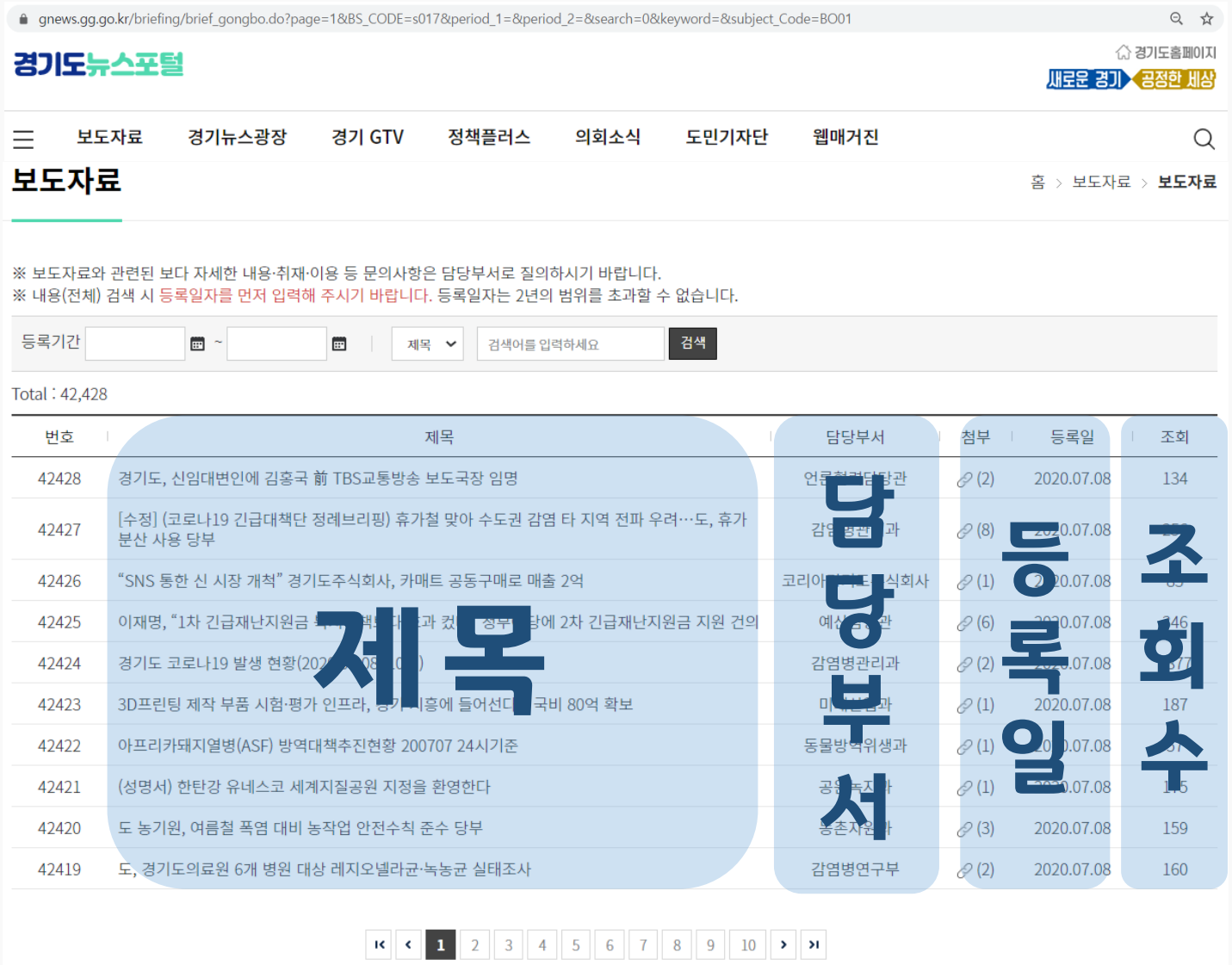
사람들이 **관심 있어하는 경기도의 공지**는 무엇일까?

요즘 경기도의 **정책 트렌드**는 무엇일까?

경기도 각 부서에서는 자세히 **어떤 업무**를 담당할까?

Ⅰ 개요

웹페이지



피쳐	설명
자료별 URL	보도자료별 URL. 접속 시 해당 보도자료를 상세하게 살펴볼 수 있다.
제목	해당 보도자료의 제목.
담당부서	해당 보도자료를 담당 또는 게시한 담당부서이다.
등록일	보도자료를 등록한 날짜이다. 연/월/일 까지 기재되어 있다.
조회수	게시물을 확인한 수.

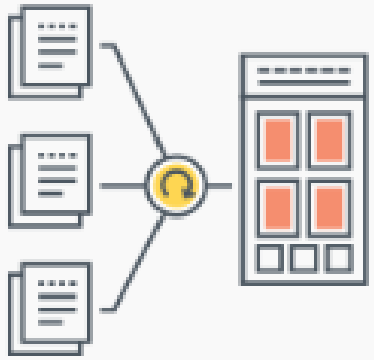
https://gnews.gg.go.kr/briefing/brief_gongbo.do?page=1&BS_CODE=s017&period_1=&period_2=&search=0&keyword=&subject_Code=BO01

Ⅰ 개요

프로세스 요약

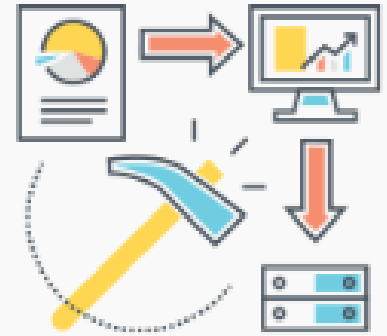
웹 크롤링

- requests, lxml, re, pandas, sqlite3, os 등의 패키지 활용.
- Xpath소스와 정규식을 활용한 크롤링.
- 경기도 보도자료 2020년 1~6월 자료 크롤링 .



형태소 분석

- Konlpy패키지의 tag클래스의 Okt함수를 활용하여 명사 추출.



워드클라우드 및 빈도분석

- Wordcloud 패키지의 WordCloud함수를 활용하여 워드클라우드.
- Collections 패키지의 Counter함수를 활용하여 단어 빈도 집계.
- Plotly 모듈을 활용하여 시각화.

경기도 보도자료 크롤링

[1]:
```
import requests
import lxml.html
import pandas as pd
import sqlite3
from pandas.io import sql
import os
```

[2]:
```
#DB에 저장하기
def db_save(KYONGGI_LIST):
    with sqlite3.connect(os.path.join('.','sqliteDB')) as con: # sqlite DB 파일이 존재하지 않는 경우 파일생성
        try:
            KYONGGI_LIST.to_sql(name = 'KYONGGI_LIST', con = con, index = False, if_exists='append')
            #if_exists : {'fail', 'replace', 'append'} default : fail
        except Exception as e:
            print(str(e))
        print(len(KYONGGI_LIST), '건 저장완료..')
```

Ⅱ 주요기술 크롤링 방법1

[8]:
```
page = 10  # 시작 페이지
end_page = 45 # 끝날 페이지

while(True):
    df_list = []
    response = requests.get('https://gnews.gg.go.kr/briefing/brief_gongbo.do?page={}&BS_CODE=s017&period_1=&period_2=&search=0&keyword=&subject_Code=BO01'\
                            .format(page))
    root = lxml.html.fromstring(response.content)
    for tr in root.xpath('//*[@id="chk-table"]/tbody/tr'):
            a = tr.xpath('td[2]/a')[0]
            url = ('https://gnews.gg.go.kr/' + a.get('href'))
            dep = tr.xpath('td[3]')[0]
            date = tr.xpath('td[5]')[0]
            view =  tr.xpath('td[6]')[0]
            df_list.append(
                pd.DataFrame({

                 'url' : [url],
                 'title' : [a.text],
                 'dept'  :  [dep.text],
                 'reg_date' : [date.text.strip()],
                 'view' : [view.text],
            })
            )

    df_10 = pd.concat(df_list)

    db_save(df_10)


    if  page >= end_page:
        break;
    else:
        page = page + 1
```

```
10 건 저장완료..
10 건 저장완료..
```

[9]:
```python
# 데이터 출력
def db_select():
    with sqlite3.connect(os.path.join('.','sqliteDB')) as con: # sqlite DB 파일이 존재하지 않는 경우 파일생성
        try:
            query = 'SELECT * FROM KYONGGI_LIST'
            KYONGGI_LIST = pd.read_sql(query, con = con)
        except Exception as e:
            print(str(e))
        return KYONGGI_LIST
db_select()
```

[7]:
```python
# DB데이터 삭제
def db_delete():
    with sqlite3.connect(os.path.join('.','sqliteDB')) as con: # sqlite DB 파일이 존재하지 않는 경우 파일생성
        try:
            cur = con.cursor()
            sql = 'DELETE FROM KYONGGI_LIST'
            cur.execute(sql)
        except Exception as e:
            print(str(e))
db_delete()
```

[ ]:
```python
#DB 삭제
def db_DROP():
    with sqlite3.connect(os.path.join('.','sqliteDB')) as con: # sqlite DB 파일이 존재하지 않는 경우 파일생성
        try:
            cur = con.cursor()
            sql = 'DROP TABLE KYONGGI_LIST'
            cur.execute(sql)
        except Exception as e:
            print(str(e))
db_DROP()
```

Ⅱ 주요기술 크롤링 방법2

```
In [ ]: # 2월 크롤링 코드
def crawling(start_page,end_page):
    url = [] # url 받는 리스트
    title = [] # title 받는 리스트
    dept = []  # department 받는 리스트
    reg_date = [] # register date 받는 리스트
    view = [] # 조회수 받는 리스트
    df_list=[] # dataframe 받는 리스트
    sleep_time = 0
    for i in range(start_page,end_page):
        ad = 'https://gnews.gg.go.kr/briefing/brief_gongbo.do?page={}&BS_CODE=s017&period_1=&period_2=&search=0&keyword=&subject_Code=B001'₩
        .format(str(i))
        response = requests.get(ad)
        root = lxml.html.fromstring(response.content)
        if (sleep_time % 10 == 0) & (sleep_time !=0): # ip 막히지 않게 잠깐 쉰다. 10페이지당 한 번
            time.sleep(5)
            print('sleep..')

        for t in root.xpath('//*[@id="chk-table"]/tbody/tr'):
            a = t.xpath('td[2]/a')[0]
            url = 'https://gnews.gg.go.kr'+a.get('href') # url
            dep = t.xpath('td[3]')[0] # 부서
            date = t.xpath('td[5]')[0] # 날짜
            view =  t.xpath('td[6]')[0] # 조회수

            df_list.append(
                pd.DataFrame({

                    'url' : [url],
                  'title' : [a.text], # 제목
                  'dept'  :  [dep.text],
                  'reg_date' : [date.text.strip()],
                  'view' : [view.text],
            })
            )
#         if df_list:
#             df_ev = pd.concat(df_list)
#             db_save(df_ev) # db 저장


        sleep_time+=1

    df_10 = pd.concat(df_list)
    return df_10.reset_index(drop=True)
feb = crawling(140,174)# 140~173 -> 2월 데이터
data.to_excel('2월 데이터.xlsx',index=False)
```

Ⅱ 주요기술 크롤링 방법3

```
title=[]
url=[]
dep=[]
date=[]
views=[]
df_april=[]
df_10=[]
page=2
endpage=33
max_page = 0
while(True):
    df_april = []
    response=requests.get('https://gnews.gg.go.kr/briefing/brief_gongbo.do?page={}&BS_CODE=s017&period_1=2020-04-01&period_2=2020-04-30&search=6&
keyword=&subject_Code=B001'#
                          .format(page))
    root = lxml.html.fromstring(response.content)

    for article in root.xpath('//*[@id="chk-table"]/tbody'):
        for a in article.xpath('tr/td[2]/a'):
            url.append('https://gnews.gg.go.kr/'+a.get('href'))
            title.append(a.text)
        for a in article.xpath('tr/td[3]'):
            dep.append(a.text)
        for a in article.xpath('tr/td[5]'):
            date.append(a.text)
        for a in article.xpath('tr/td[6]'):
            views.append(a.text)
    df_article=pd.DataFrame({
        'url':url,
        'title':title,
        'dept':dep,
        'reg_date':date,
        'number':views
    })

    df_april.append(df_article)

    if df_april:
        df_10 = pd.concat(df_april)
        db_save(df_10)

    page=page+1
    if page==endpage:
        break

df_10.to_excel('4월 데이터.xlsx',index=False)
```

Ⅱ 주요기술 크롤링 결과

```
#EXCEL에 저장
KYONGGI_LIST = db_select()

def save_excel(KYONGGI_LIST):
    excel = pd.ExcelWriter('경기도보도자료.xlsx')
    KYONGGI_LIST.to_excel(excel,'.',index=False)
    excel.save()

save_excel(KYONGGI_LIST)
```

```
# 엑셀파일 합치기
data = pd.DataFrame()
for i in range(1,7):
    df = pd.read_excel('{}월 데이터.xlsx'.format(i))
    data = pd.concat([data,df])
print('data shape: ' ,data.shape)
data.head()
```

```
data shape:  (1870, 5)
```

	url	title	dept	reg_date	view
0	https://gnews.gg.go.kr/briefing/brief_gongbo_v...	경기도 신종 코로나바이러스감염증 현황(2020.2.1. 10시30분)	감염병관리과	2020.02.01	4681
1	https://gnews.gg.go.kr/briefing/brief_gongbo_v...	이재명, 마스크 매점매석 형사고발 검토 '초강경 대응'... 마스크 최고가격 제한도 정...	공정소비자과	2020.02.01	694
2	https://gnews.gg.go.kr/briefing/brief_gongbo_v...	경기도 신종 코로나바이러스감염증 현황(2020.1.31. 18시)	감염병관리과	2020.01.31	2024
3	https://gnews.gg.go.kr/briefing/brief_gongbo_v...	이재명 "동물이 행복해야 사람도 행복, 생명존중사회 걸맞은 정책 전환 필요해"	동물보호과	2020.01.31	256
4	https://gnews.gg.go.kr/briefing/brief_gongbo_v...	경기도, 감염병 우선관리계층 대상 '신종 코로나' 집중 관리 나서	보건의료정책과	2020.01.31	435

III 주요기능 형태소분석

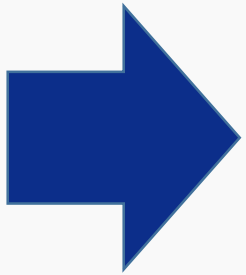
```
# !pip install JPype1
# !pip install konlpy

from konlpy.tag import Okt, Kkma, Komoran, Hannanum
tagger = Okt()

#파일 읽어오기
import pandas as pd
gg = pd.read_excel('./6월.xlsx')
# DataFrame -> List
title = list(gg['title'])
# List -> String
title_st = ' '.join(title)
# 특수문자 제거
import re
body = []
punc = '[!" “” ‘’ #$%&₩'()*+,-/:;<=>?[₩₩]^_`{|}…~· ··]'
body.append(re.sub(punc, '' , title_st))
full_body = ' '.join(body)
full_body

#형태소 분석
nouns = tagger.nouns(full_body)
from collections import Counter
count = Counter(nouns)
nouns_list = count.most_common(100)

#저장
import pandas as pd
data = pd.DataFrame(nouns_list)
data.head()
data.to_excel('빈도.xlsx')
```



	A	B	C
1		0	1
2	0	경기도	145
3	1	도	100
4	2	코로나	74
5	3	현황	44
6	4	지원	38
7	5	등	38
8	6	발생	34
9	7	경기	33
10	8	추진	27
11	9	대책	24
12	10	사업	22
13	11	도민	21
14	12	소방	20
15	13	현장	18
16	14	위	18
17	15	방역	18
18	16	일자리	18
19	17	긴급	18
20	18	브리핑	17

III 주요기능 워드클라우드

```
# 필요없는 단어 제거
def remove_values_from_list(the_list, val):
    return [value for value in the_list if value != val]
#noun.remove('경기도')
nn = remove_values_from_list(noun, '경기도')
#print(noun)
count = Counter(nn)
```

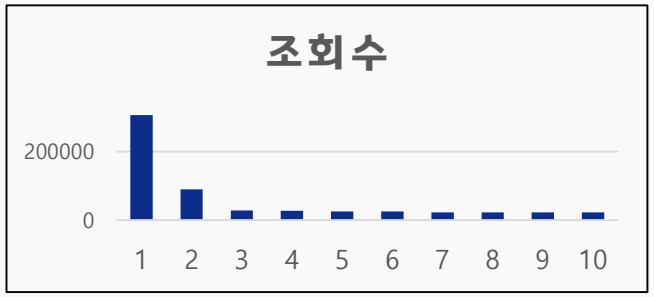
```
[ ]: import sys
     from wordcloud import WordCloud

     filename = sys.argv[1]

     #font_path 설정 : font의 위치
     wc = WordCloud(font_path='c:/Windows/Fonts/malgun.ttf',\
                    background_color="white",\
                    width=1000,\
                    height=1000,\
                    max_words=200,\
                    max_font_size=300
                    )
```

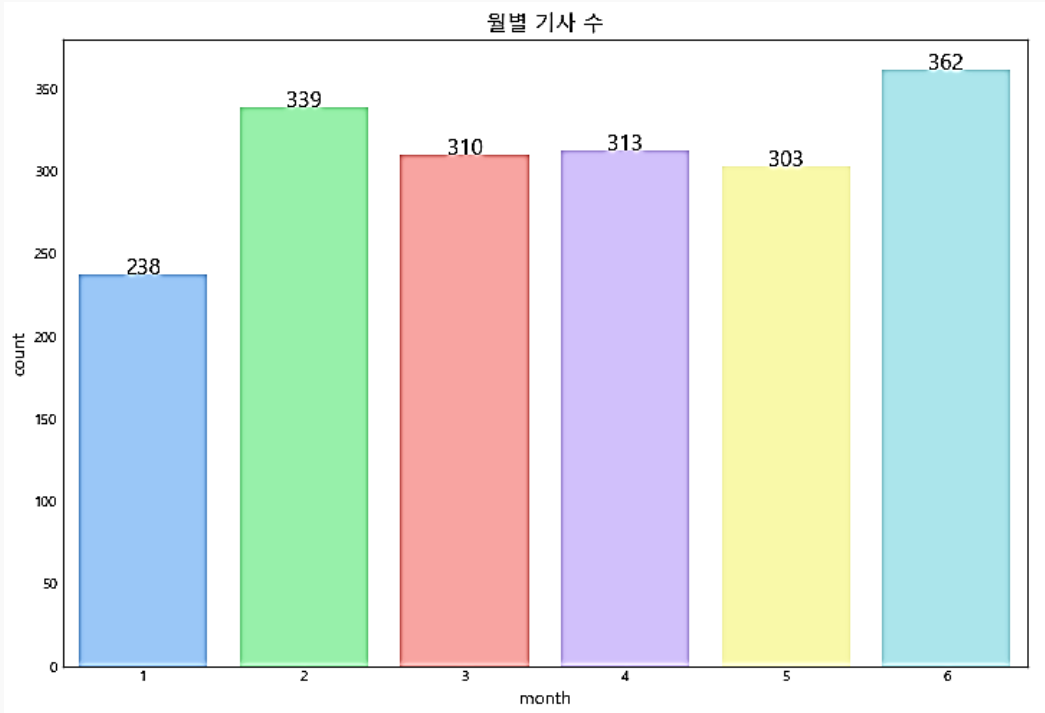
```
[ ]: # 이미지 저장
     wc.generate_from_frequencies(dict(nouns_list))
     wc.to_file('wordcloud.png')
```

조회수 상위 10개 보도물

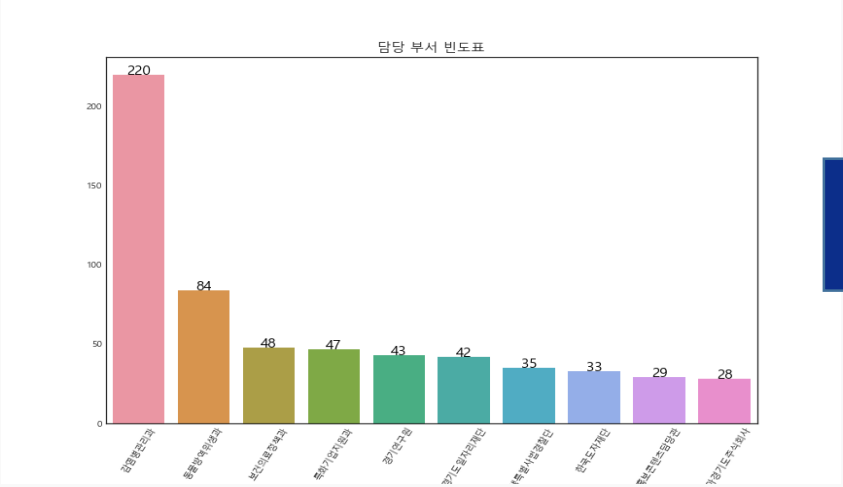


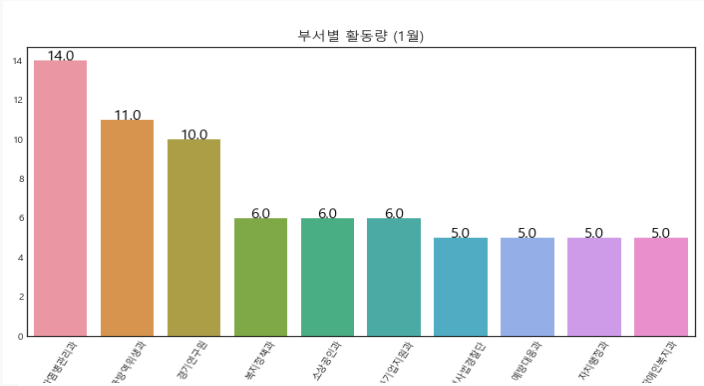
TITLE	VIEWS
(브리핑) 이재명, 4월부터 도민 1인당 10만원씩 '경기도 재난기본소득 ' 지급	306643
(브리핑) 이재명, "경기도 재난기본소득, 기존 경기지역 화폐 · 신용카드로 사용 가능"	89532
경기도 코로나19 발생 현황(2020.03.05) 10시	27413
경기도 코로나19 발생 현황(2020.03.07.10시)	27293
경기도 코로나19 발생 현황(2020.03.04) 10시	25419
(브리핑) 이재명 "경기도 재난 기본 소득, 18개 시군과 함께 지급"	25242
경기도 코로나19 발생 현황(2020.03.06.10시10분)	22415
경기도 코로나19 발생 현황(2020.03.08. 10시)	22685
"경기도 재난 기본 소득, 이렇게 사용하세요"	22469
경기도 재난기본소득 첫날 83만여명. 1,359억 원 신청	22292

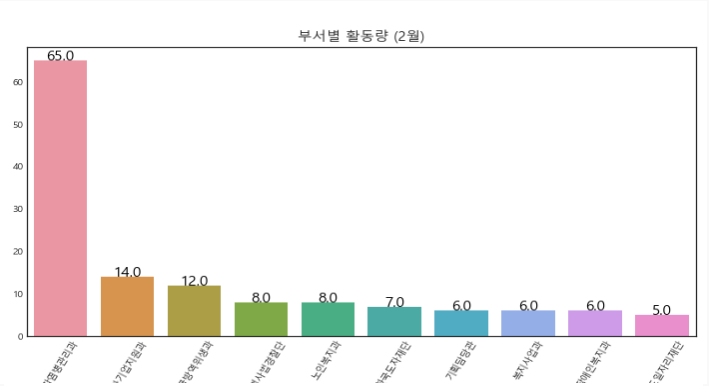
월별 보도자료 수

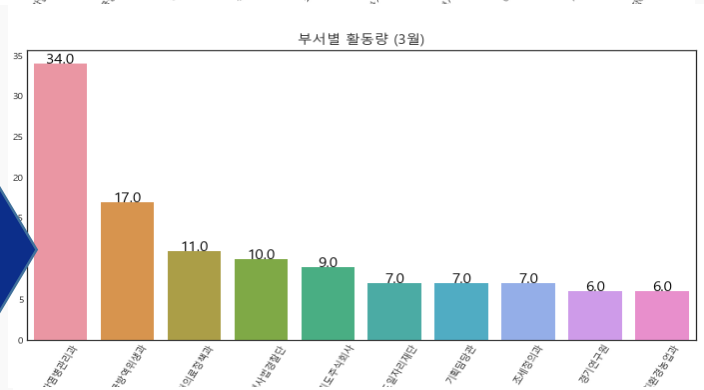


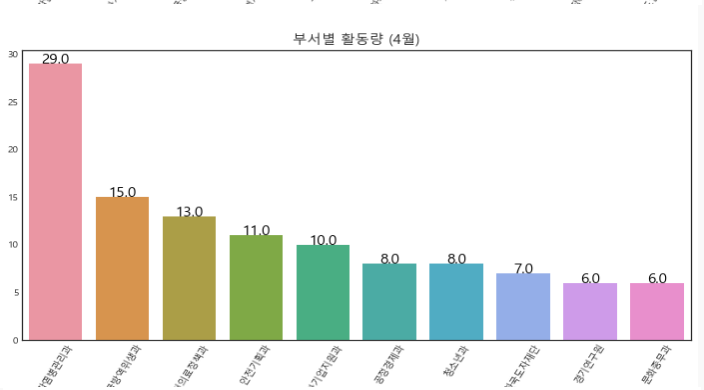
담당부서별 보도자료 수

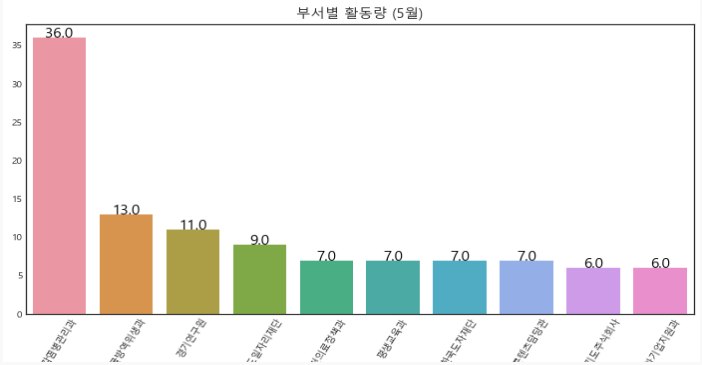

담당 부서 빈도표
220
84
48
47
43
42
35
33
29
28

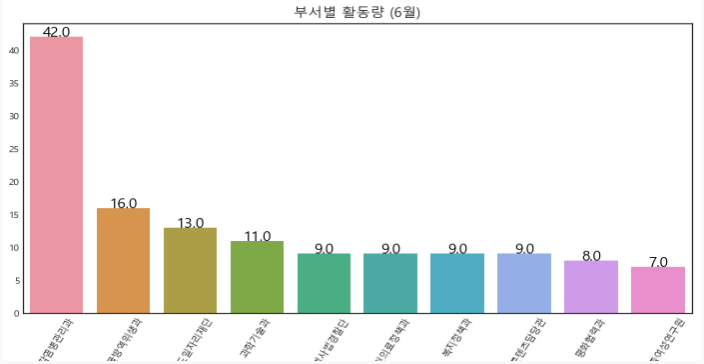

부서별 활동량 (1월)
14.0
11.0
10.0
6.0
6.0
6.0
5.0
5.0
5.0
5.0


부서별 활동량 (2월)
65.0
14.0
12.0
8.0
8.0
7.0
6.0
6.0
6.0
5.0

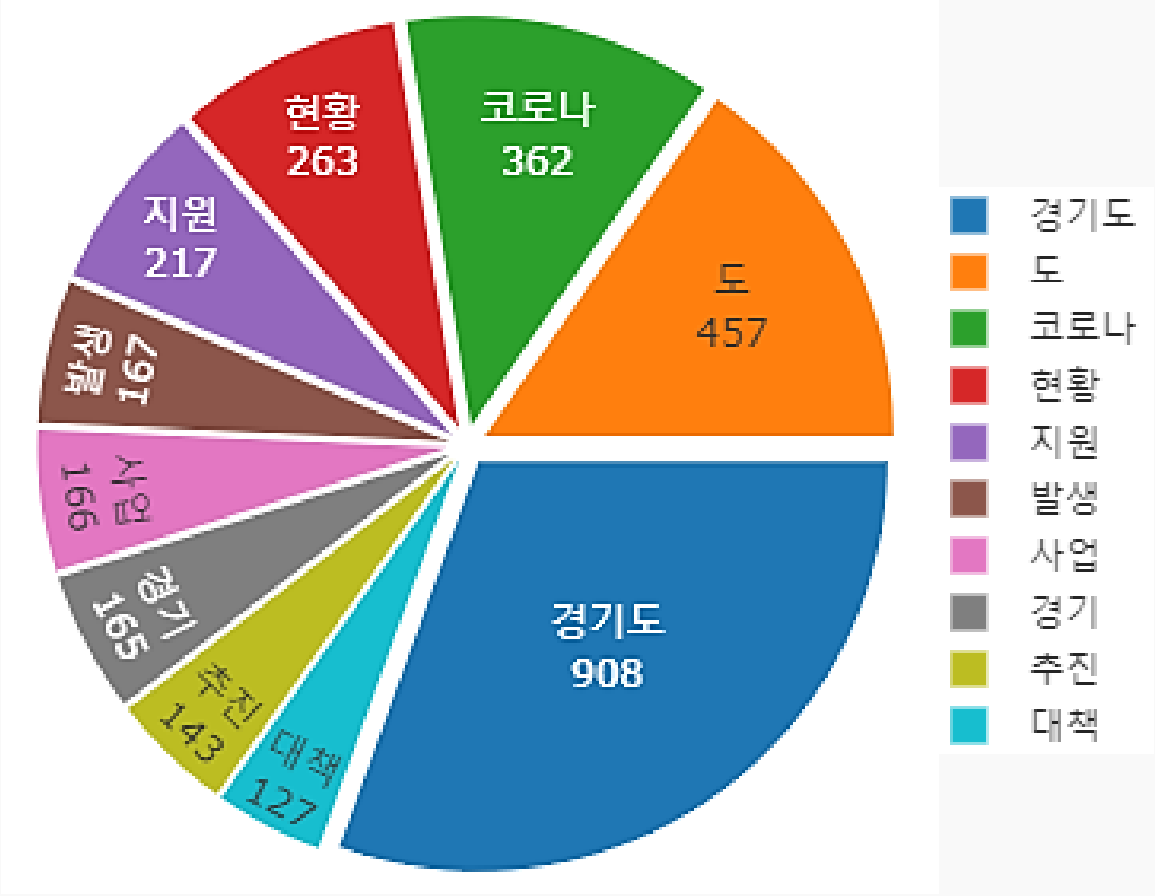

부서별 활동량 (3월)
34.0
17.0
11.0
10.0
9.0
7.0
7.0
7.0
6.0
6.0


부서별 활동량 (4월)
29.0
15.0
13.0
11.0
10.0
8.0
8.0
7.0
6.0
6.0


부서별 활동량 (5월)
36.0
13.0
11.0
9.0
7.0
7.0
7.0
7.0
6.0
6.0


부서별 활동량 (6월)
42.0
16.0
13.0
11.0
9.0
9.0
9.0
9.0
8.0
7.0

상반기 보도자료 형태소 분석 결과



Ⅳ 기대효과 활용방향 제안

- 코로나 19관련 자료와 재난기본소득 관련 자료가 경기도민의 관심이 가장 많음.

  ▶ 관련 보도자료가 활성화된다면 도민의 편익 증진이 예상.

- 감염병관리과의 게시량이 가장 많았고 동물방역위생과와 보건의료정책과가 그 뒤를 이음.

  ▶ 관련 부서에 인원 확충 또는 적절한 업무분장을 제안.

- 코로나19, 아프리카 돼지 열병, 청년 대책 관련 보도자료가 다수를 이루었음.

  ▶ 2020년 상반기, 경기도청의 주요 이슈 파악.

Thank you!

Q&A