

경기도보도자료

5조 : 권수민 김영민 송준영 양동재 여지민 허이령

목차



I 개요

II 주요기술

III 주요기능

IV 기대효과

우리는 공공기관에서 어떤 업무를 맡게 될까?



사람들이 관심 있어하는 경기도의 공지는 무엇일까?



요즘 경기도의 정책 트렌드는 무엇일까?



경기도 각 부서에서는 자세히 어떤 업무를 담당할까?

I 개요

웹페이지

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

경기도뉴스포털

경기도홈페이지
새로운 경기 > 공정한 세상

☰ 보도자료 경기뉴스광장 경기 GTV 정책플러스 의회소식 도민기자단 웹매거진 🔍

보도자료 홈 > 보도자료 > 보도자료

※ 보도자료와 관련된 보다 자세한 내용·취재·이용 등 문의사항은 담당부서로 질의하시기 바랍니다.
※ 내용(전체) 검색 시 **등록일자**를 먼저 입력해 주시기 바랍니다. 등록일자는 2년의 범위를 초과할 수 없습니다.

등록기간 ~ | 제목 | 검색어를 입력하세요

Total : 42,428

번호	제목	담당부서	첨부	등록일	조회
42428	경기도, 신임대변인에 김홍국 前 TBS교토방송 보도국장 임명	언론협력담당관	📎 (2)	2020.07.08	134
42427	[수정] (코로나19 긴급대책단 정례브리핑) 휴가철 맞아 수도권 감염 타 지역 전파 우려...도, 휴가 분산 사용 당부	감염병관리과	📎 (8)	2020.07.08	177
42426	“SNS 통한 신 시장 개척” 경기도주식회사, 카메트 공동구매로 매출 2억	코리아주식회사	📎 (1)	2020.07.08	83
42425	이재명, “1차 긴급재난지원금 확대 대책 다 고려”...정부 당에 2차 긴급재난지원금 지원 건의	예산기획관	📎 (6)	2020.07.08	146
42424	경기도 코로나19 발생 현황(2020.07.08. 10:00)	감염병관리과	📎 (2)	2020.07.08	177
42423	3D프린팅 제작 부품 시험·평가 인프라, 경기·지흥에 들어선다...국비 80억 확보	미래기획과	📎 (1)	2020.07.08	187
42422	아프리카돼지열병(ASF) 방역대책추진현황 200707 24시기준	동물방역위생과	📎 (1)	2020.07.08	137
42421	(성명서) 한탄강 유네스코 세계지질공원 지정을 환영한다	공유복지과	📎 (1)	2020.07.08	145
42420	도 농기원, 여름철 폭염 대비 농작업 안전수칙 준수 당부	농촌자원과	📎 (3)	2020.07.08	159
42419	도, 경기도의료원 6개 병원 대상 레지오넬라균·녹농균 실태조사	감염병연구부	📎 (2)	2020.07.08	160

⏪ < 1 2 3 4 5 6 7 8 9 10 > ⏩

피쳐

설명

자료 별 URL

보도자료별 URL. 접속 시 해당 보도 자료를 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.

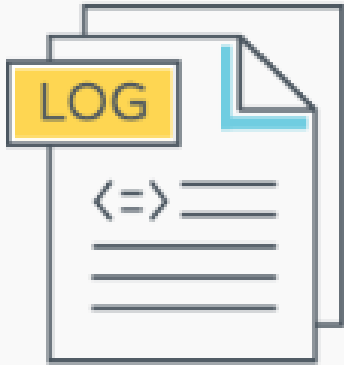
조회 수

게시물을 확인한 수.

송 준 영	:	1 월분	보도자료	크롤링,	해당 PPT 제작
김 영 민	:	2 월분	보도자료	크롤링,	EDA, Visualization
양 동 재	:	3 월분	보도자료	크롤링,	각 코드 및 크롤링 결과 병합
허 이 령	:	4 월분	보도자료	크롤링,	워드클라우드
여 지 민	:	5 월분	보도자료	크롤링,	워드클라우드
권 수 민	:	6 월분	보도자료	크롤링,	형태소 분석

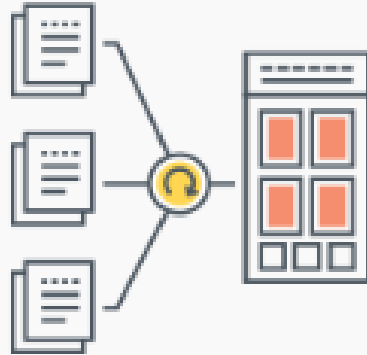
팀원 모두 크롤링 프로세스 습득
및 파트 분배를 통한 시간 단축 !

역할 분담을 통한
프로젝트 효율화!



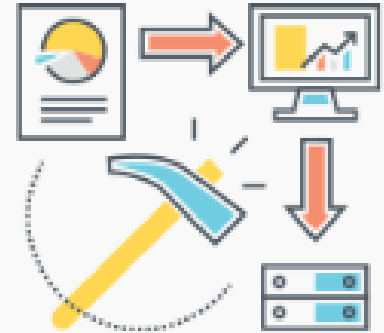
웹 크롤링

- 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]: #DB 에 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=B001'\
                            .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..

II 주요기술

크롤링 방법1

```
[9]: # 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]: # DB 데이터 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()
```

```
[ ]: #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 쉬다. 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

```
#EXCEL 에 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('{i}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

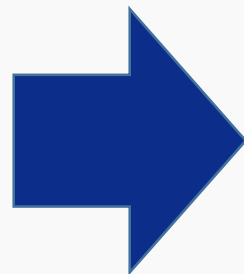
```
# !pip install Jupyter
# !pip install konlpy

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

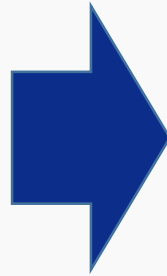
```
# 필요없는 단어 제거
def remove_values_from_list(the_list, val):
    return [value for value in the_list if value != val]
#noun.remove('경기도')
nn = remove_values_from_list(noun, '경기도')
#print(noun)
count = Counter(nn)
```

```
[ ]: 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')
```

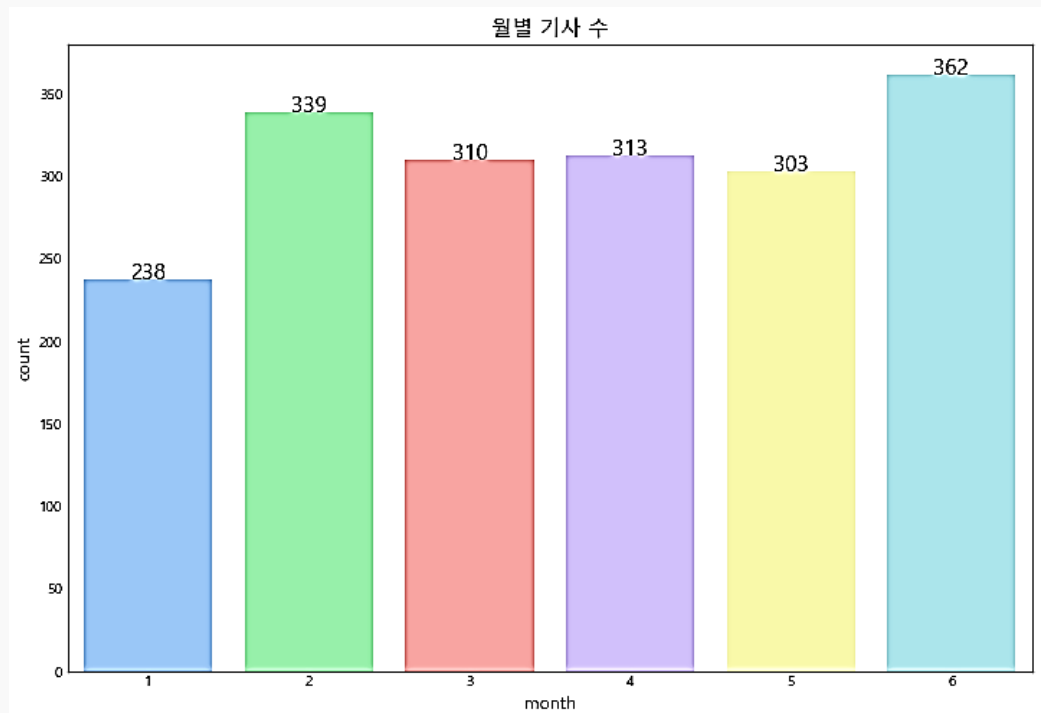


조회수 상위 10개 보도물



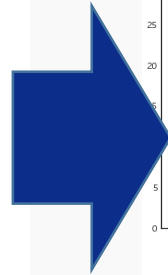
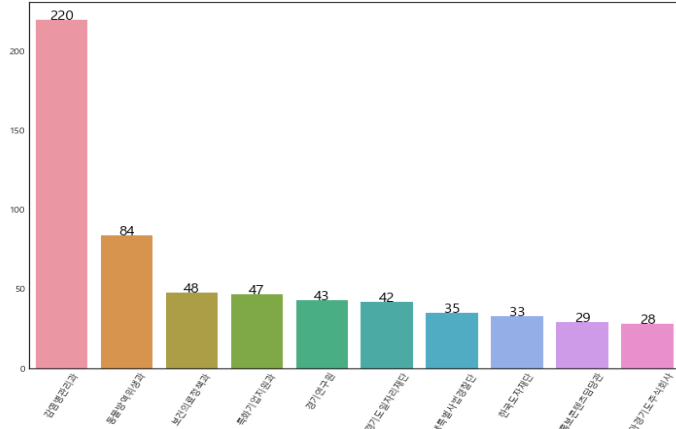
TITLE	VIEWS
(브리핑) 이재명, 4월부터 도민 1인당 10만원씩 '경기도 재난기본소득' 지급	306643
(브리핑) 이재명, “경기도 재난기본소득, 기존 경기지역 화폐 · 신용카드로 사용 가능”	89532
경기도 코로나19 발생 현황(2020.03.05) 10시	27413
경기도 코로나19 발생 현황(2020.03.07.10시)	27293
경기도 코로나19 발생 현황(2020.03.04) 10시	25419
(브리핑) 이재명 “경기도 재난 기본 소득, 18개 시군과 함께 지급”	25242
경기도 코로나19 발생 현황(2020.03.06.10시10분)	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

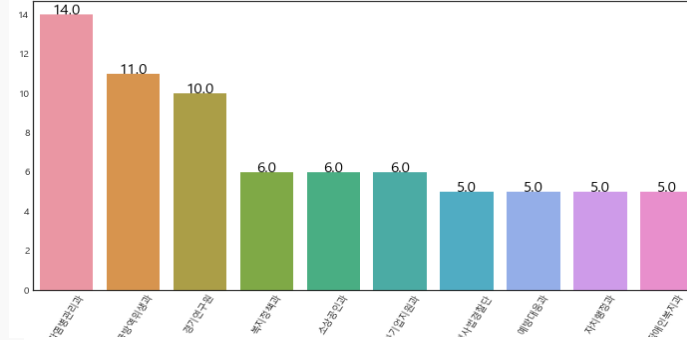


담당부서별 보도자료 수

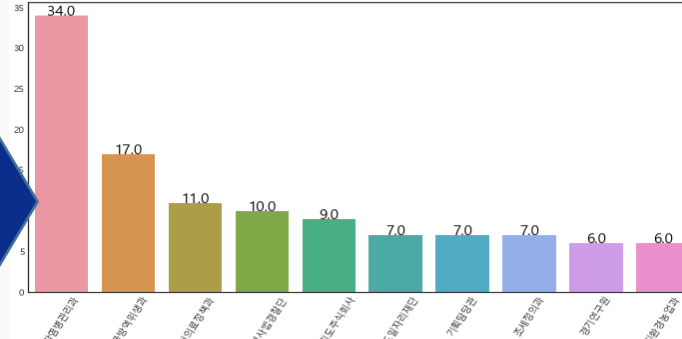
담당 부서 빈도표



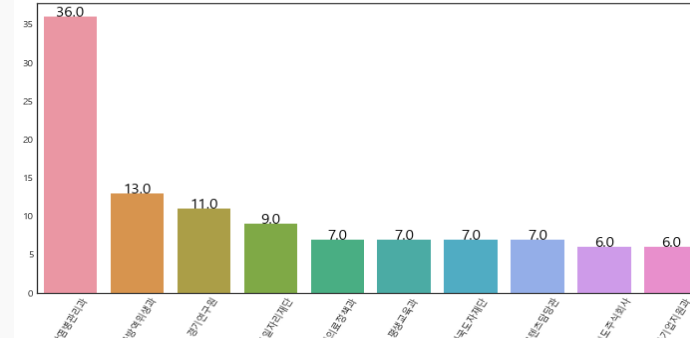
부서별 활동량 (1월)



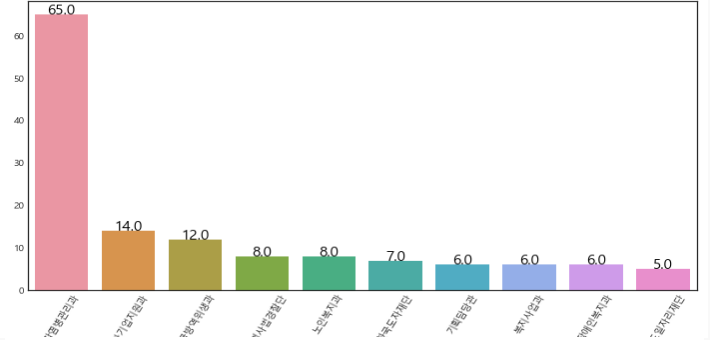
부서별 활동량 (3월)



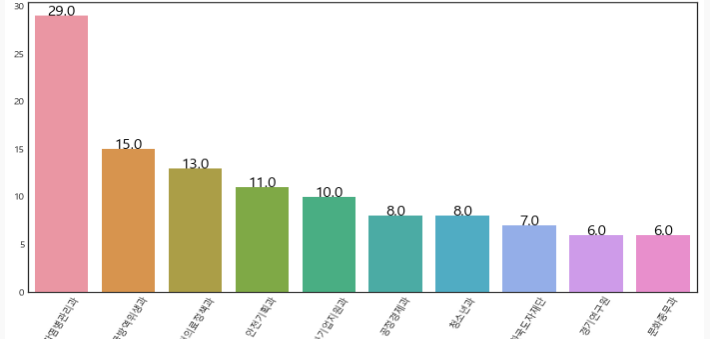
부서별 활동량 (5월)



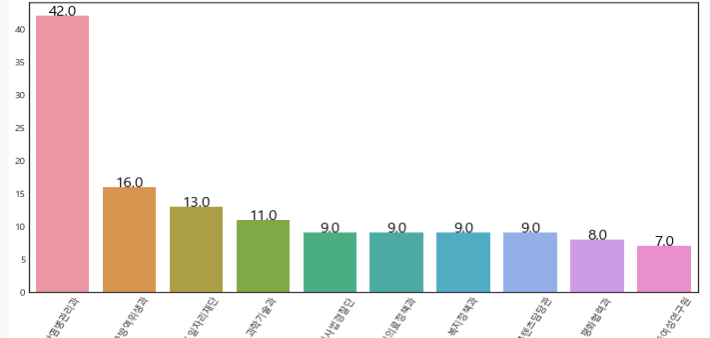
부서별 활동량 (2월)



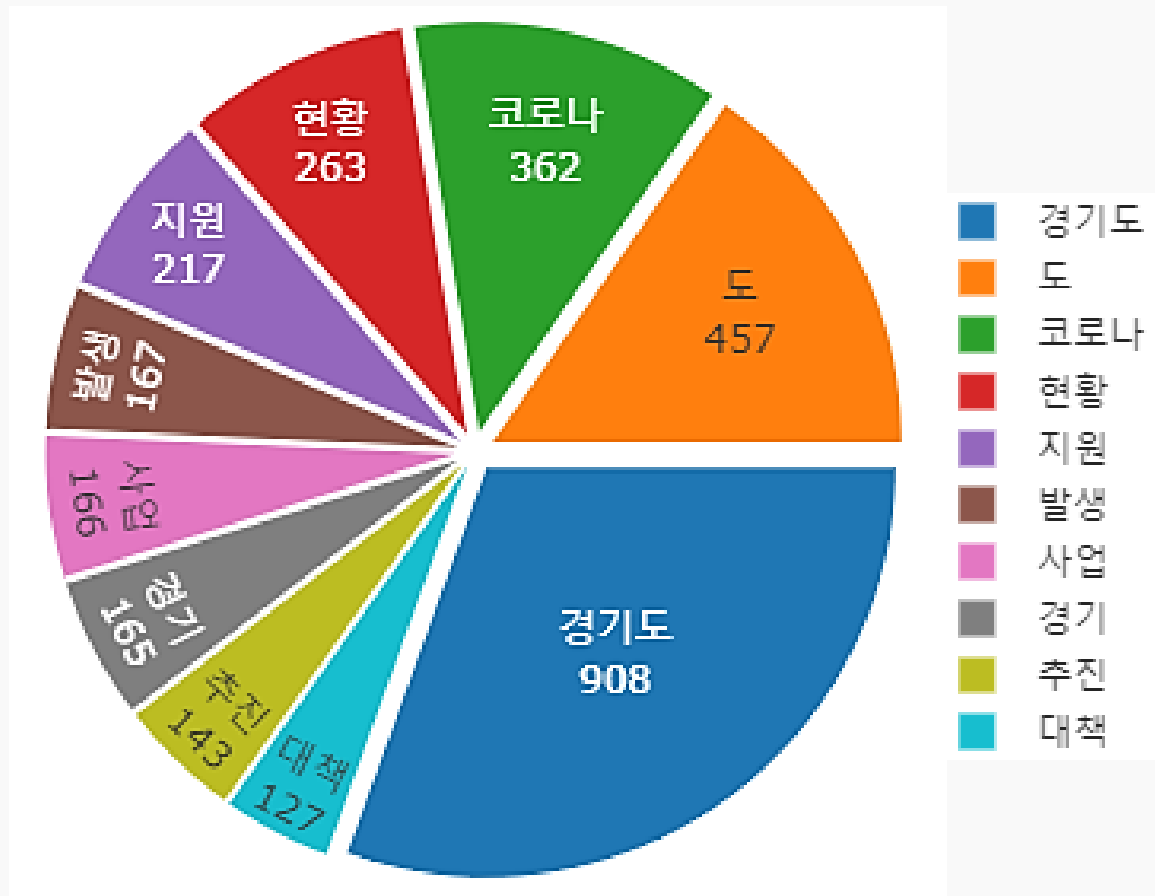
부서별 활동량 (4월)



부서별 활동량 (6월)



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.**
- 코로나 19, 아프리카 돼지 열병, 청년 대책 관련 보도자료가 다수를 이루었음.
▶ **2020년 상반기, 경기도청의 주요 이슈 파악.**

Thank you!



Q&A