Text Data Analytics - Assignment(1)

네이버 뉴스 본문 크롤링 & 전처리 & 워드클라우드 및 LDA
4 topic < 정치, 기술, 경제, 환경 >

AI빅데이터융합경영학과
20212548

김지은

목 차

수집 – 1. 관심 keyword 및 topic 설정 & data source

< 4 topic >

1. 정치 및 법 2. 기술 3. 경제 4. 환경

※ 선정이유:
매일 아침에 눈을 뜨면, 크게 4가지 분야의 기사를 하나씩 읽음.
정치, 기술, 경제, 환경 분야의 기사
따라서 이번 과제를 통해 평소에 보던 기사들을 크롤링하여
분석하면 흥미로울 것 같아 크게 4가지로 잡게 되었음.

< 보다 세분화한 keyword의 사용 >

1. (선거, 정부, 법, 검찰, 살인)
2. (빅데이터, AI, 자율주행, 가상현실, GPT, 로봇)
3. (투자, 기업, 부자, 주식, 실업, 보험)
4. (날씨, 재활용)

➡ 19개의 topic으로 세분화하여
크롤링 진행, 네이버 뉴스 본문 추출

수집 – 2. 함수 정의

< 크롤링을 위한 함수 정의 >

In [8]:
```
def remove_tag(my_str):
    ## 태그를 지우는 함수
    p = re.compile('(<([^>]+)>)')
    return p.sub('', my_str)

def sub_html_special_char(my_str):
    ## 특수문자를 나타내는 &apos;, &quot를 실제 특수문자로 변환
    p1 = re.compile('&lt;') #lt를 <로 바꿔줘
    p2 = re.compile('&gt;')
    p3 = re.compile('&amp;')
    p4 = re.compile('&apos;')
    p5 = re.compile('&quot;')

    result = p1.sub('₩<', my_str)
    result = p2.sub('₩>', result)
    result = p3.sub('₩&', result)
    result = p4.sub("₩'", result)
    result = p5.sub('₩"', result)
    return result
```

In [9]:
```
base_url = 'https://openapi.naver.com/v1/search/news.json'

def getresult(client_id,client_secret,query,n_display,start,sort='sim'):
    encQuery = urllib.parse.quote(query)

    url = f'{base_url}?query={encQuery}&display={n_display}&start={start}&sort={sort}'
    my_request = urllib.request.Request(url)
    my_request.add_header("X-Naver-Client-Id",client_id)
    my_request.add_header("X-Naver-Client-Secret",client_secret)
    response = urllib.request.urlopen(my_request)
    rescode = response.getcode()
    if(rescode==200):
        response_body = response.read()
        search_result_str = response_body.decode('utf-8')
        search_results = json.loads(search_result_str)
    else:
        print("Error Code:" + rescode)
    return search_results['items']
```

In [10]:
```
#데이터 수집 과정 중 일부는 배민성 학우분과 같이 협업하였습니다.
#slack에 이수인 학우분이 올려주신 코드도 참고하였습니다.
def search_news_with_link (result):
    article_ids = ['dic_area']
    titles = []
    links = []
    pubdates = []
    contents = []

    p = re.compile('https://n.news.naver.com/.+')
    for i, item in enumerate(result):
        if p.match(item['link']): ## <link>태그의 문자열이 n.news.naver.com/으로 시작하는 결과만 추출
            title = sub_html_special_char(remove_tag(item['title']))
            link = item['link']
            pubdate = item['pubDate']
            titles.append(title)
            links.append(link)
            pubdates.append(pubdate)

            html = urllib.request.urlopen(link)
            bs_obj = BeautifulSoup(html, 'html.parser')
            for article_id in article_ids:
                print(article_id)
                content = bs_obj.find_all('div', {'id':article_id})
                if len(content) > 0:
                    contents.append(content[0].text)
                    break
                else:
                    contents.append(0)
                    #연예뉴스와 같은 뉴스들을 수집하지 않기 위해 위와 같은 코드를 작성함
                    #dic_area가 아닌 본문 id들은 0으로 채워준 후, 모든 데이터 수합 후 삭제함
                    #데이터 프레임에 추가할 때, 0과 같은 값을 채워주지 않을 경우, 데이터 프레임이 만들어지지 않음(길이 문제 발생)

    result_dict = {'title': titles, 'link': links, 'pubdate': pubdates, 'content': contents}
    df = pd.DataFrame.from_dict(result_dict)
    return df
```

- 네이버 뉴스 기사들만 크롤링을 하고 싶어, dic_area에 해당하지 않는 기사 본문들은 0으로 content에 추가해주었습니다.

- 추가적으로 continue를 사용해서 해당하지 않는 id는 넘어가게 하고 싶었지만 해결이 잘 되지 않았고, 이에 따라 dictionary 형태로 변환할 때 title은 나왔지만, id가 없는 기사이기에 content에 빈 값으로 나와 dataframe이 만들어지지 않는 문제가 발생했습니다.

- 따라서 위와 같은 방식을 통해 0값으로 채우고 해당 행을 삭제하는 방향으로 전처리를 진행하였습니다.

수집 – 2. 데이터 수집

```
In [11]: client_id = '03xivz4Z_LC7mSS5tk06'
         client_secret = 'uvknoLe1Jq'
         n_display=100
```

➡ display : 100으로 설정
➡ 네이버 개발자에서 ID와 secret 설정

대선 ¶

```
In [22]: query = '대선' ①
```

```
In [23]: #일부 코드는 배민성 학우분과 협업하였습니다.
         #slack에 올려주신 이수인 학우분의 코드도 일부 참고하였습니다.
         total_results = pd.DataFrame()
         for i in range(10):
             start=1+n_display*i ②
             print(start)
             result= getresult(client_id,client_secret,query,n_display,start,sort='sim')
             up_result = search_news_with_link(result) ③

             total_results = pd.concat([total_results, up_result]) ④
             print(len(total_results)) ⑤
```

```
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
dic_area
797
```

```
In [24]: total_results.reset_index(drop=True, inplace=True)
```

➡ 일부 코드는 배민성 학우분과 협업하였습니다.
➡ slack에 올려주신 이수인 학우분의 코드도 일부 참고하였습니다.

① query에 뽑고자 하는 keyword 작성

② 한 번 크롤링 진행 시 최대 1000개까지만 가능하여,
start = 1 + n_display*i 로 설정하여 for문 작성

③ 위에서 정의한 search_news_with_link() 함수를 통해 크롤링 진행

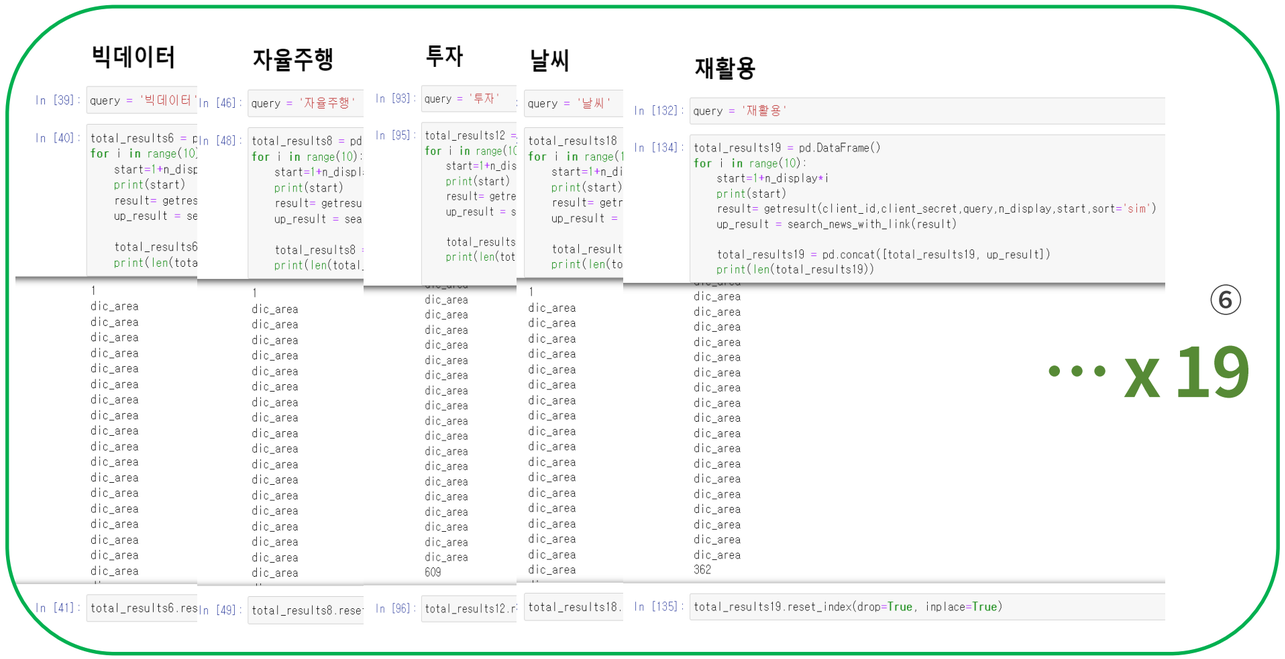
④ total_results에 이전까지 dataframe 값에 새로운 data를 concat 시킴

⑤ 세부 topic 하나 당 만들어지는 dataframe의 len을 통해 데이터 개수 파악

⑥ 위와 같은 방식으로 19개의 dataframe 생성

수집 – 2. 데이터 수집

①



②

```
In [136]: crawl= pd.concat([total_results, total_results2,total_results3,total_results4,
                            total_results5,total_results6,total_results7,total_results8,
                           total_results9,total_results10,total_results11,total_results12,
                            total_results13,total_results14,total_results15,total_results16,total_results17,
                              total_results18, total_results19])

In [137]: crawl.shape
Out[137]: (11196, 4)

In [138]: crawl.reset_index(drop=True, inplace=True)
```

➡ 19개의 keyword에 대한 기사 크롤링 후, concat 시킴

➡ 중복값 제거 전 data 개수: 11196개

③

	title	link	pubdate	content
0	이재명 대선 경선 기탁금에 대장동 자금 흘러들어갔나	https://n.news.naver.com/mnews/article/053/000...	Wed, 17 May 2023 09:43:00 +0900	\n민주당 즉시 반박자료 내고 "저질 창작소설에 불과"\n\n\n\n'불법 대선 자...
1	尹 "국민 건강 최우선" 간호법 제동…민주 "대선공약이면서"	https://n.news.naver.com/mnews/article/015/000...	Tue, 16 May 2023 14:58:00 +0900	\n尹, 취임 후 두 번째 거부권 행사"간호법, 국민 건강에 불안감 초래"다시 국회...
2	검찰 "이재명 대선 때 기탁금, 대장동 일당이 준 8.5억 중 일부"	https://n.news.naver.com/mnews/article/025/000...	Wed, 17 May 2023 00:01:00 +0900	\n\t\t\t 이재명 더불어민주당 대표가 2021년 대선 후보 예비경선 기탁금으...
3	野 김성주 "尹 대선공약 간호법 거부권 행사는 정치 코미디"	https://n.news.naver.com/mnews/article/003/001...	Tue, 16 May 2023 10:49:00 +0900	\n기사내용 요약"거부권 행사, 갈등 해소가 아니라 갈등 증폭시킬 것"\n\n\n\...
4	튀르키예 대선 결과, 극우 '3위 후보' 손에 달렸다	https://n.news.naver.com/mnews/article/028/000...	Tue, 16 May 2023 19:11:00 +0900	\n"며칠 내 지지 후보 밝힐 것"\n\n\n\n시난 오안 후보가 15일 앙카라에서...

<dataframe 결과>

전처리 – 3. 데이터 전처리

①

```
In [141]: cl.drop(cl.loc[cl['content']==0].index, inplace=True)
```

➡ content가 0인 행들은 네이버 연예 뉴스로써,
연예 뉴스에 해당하는 기사들을 제거하기 위해 0으로 추가하고, 해당 행 삭제

②

```
In [142]: cl=cl.drop_duplicates()

In [143]: cl.reset_index(drop=True,inplace=True)

In [144]: cl[cl.duplicated(keep=False)]

Out[144]:  title  link  pubdate  content

In [145]: cl.shape

Out[145]: (10400, 4)

In [149]: cl.to_csv('crawling_df.csv',index=False)
```

➡ 중복 기사들 삭제
➡ 1차 크롤링 수집 crawling _df.csv 저장
➡ 1차 df.shape: (10400,4)

③

	title	link	pubdate	content
0	이재명 대선 경선 기탁금에 대장동 자금 흘러들어갔나	https://n.news.naver.com/mnews/article/053/000...	Wed, 17 May 2023 09:43:00 +0900	\n민주당 즉시 반박자료 내고 "저질 창작소설에 불과"\n\n\n\n'불법 대선 자...
1	尹 "국민 건강 최우선" 간호법 제동… 민주 "대선공약이면서"	https://n.news.naver.com/mnews/article/015/000...	Tue, 16 May 2023 14:58:00 +0900	\n尹, 취임 후 두 번째 거부권 행사"간호법, 국민 건강에 불안감 초래"다시 국회...
2	검찰 "이재명 대선 때 기탁금, 대장동 일당이 준 8.5억 중 일부"	https://n.news.naver.com/mnews/article/025/000...	Wed, 17 May 2023 00:01:00 +0900	\n\t\t\t이 재명 더불어민주당 대표가 2021년 대선 후보 예비경선 기탁금으...
3	野 김성주 "尹 대선공약 간호법 거부권 행사는 정치 코미디"	https://n.news.naver.com/mnews/article/003/001...	Tue, 16 May 2023 10:49:00 +0900	\n기사내용 요약"거부권 행사, 갈등 해소가 아니라 갈등 증폭시킬 것"\n\n\n\...
4	튀르키예 대선 결과, 극우 '3위 후보' 손에 달렸다	https://n.news.naver.com/mnews/article/028/000...	Tue, 16 May 2023 19:11:00 +0900	\n"며칠 내 지지 후보 밝힐 것"\n\n\n\n시난 오안 후보가 15일 앙카라에서...
...	...	...	...	...
10395	상상인그룹, 지구의 날 맞아 재활용 맞춤 휠체어 기부	https://n.news.naver.com/mnews/article/215/000...	Sat, 22 Apr 2023 14:32:00 +0900	\n\n\n\n\n상상인그룹은 22일 '지구의 날'을 맞아 '휠체어 사용 아동 이동...
10396	"좌측 진출로의 시설물 90% 재활용할 것"	https://n.news.naver.com/mnews/article/421/000...	Wed, 19 Apr 2023 14:28:00 +0900	\n\n\n\n\n(광주=뉴스1) 이승현 기자 = 강기정 광주시장이 19일 오후 광...
10397	중고 갤럭시, '의료기기' 변신…희토류·금·은·동 추출 재활용	https://n.news.naver.com/mnews/article/032/000...	Mon, 17 Apr 2023 21:58:00 +0900	\n삼성·애플, 폐휴대폰 업사이클·리사이클 수준 고도화\n\n\n\n삼성전자 모델들...
10398	선거 폐현수막을 어쩌나…재활용 짜내도 고작 30%	https://n.news.naver.com/mnews/article/015/000...	Sun, 23 Apr 2023 18:11:00 +0900	\n5년간 5번 선거서 1.4만t 발생탄소배출 등 지자체 처리 골머리이른바 '무제한...
10399	광주 서석동서 재활용품 수거차량이 전선 건드려 정전	https://n.news.naver.com/mnews/article/003/001...	Tue, 18 Apr 2023 10:03:00 +0900	\n\n\n\n\n[광주=뉴시스] 이영주 기자 = 18일 오전 8시 31분께 광주 ...

10400 rows × 4 columns

<1차 dataframe 결과>

전처리 – 3. 1차 불용어 처리, 어간 추출

< 1차 불용어 제거 및 어간 추출 >

```
import copy
from konlpy.tag import Okt
import pykospacing
import kss
with open ('stopwords.txt','r',encoding='utf-8')as f:
    stopwords= f.readlines()
stopwords= [x.replace('₩n','') for x in stopwords]
okt=Okt()
```

```
#교수님의 기존 stopwords.txt를 통해 1차적으로 불용어 제거 및 어간 추출 진행
def preprocess_korean(text):
    my_text=copy.copy(text)
    #₩n 제거
    my_text = my_text.replace('₩n','')
    spacer= pykospacing.Spacing() #띄어쓰기 교정
    my_text=spacer(my_text)
    sents=kss.split_sentences(my_text)

    p=re.compile('[^ㄱ-ㅎㅏ-ㅣ가-힣]') #한글과 띄어쓰기를 제외한 모든 글자
    results=[]
    for sent in tqdm(sents):
        result=[]
        tokens= okt.morphs(sent, stem=True) #어간추출
        for token in tokens:
            token=p.sub('',token)
            if token not in stopwords: #stopwords에 없는 애들만 추가해라
                result.append(token)
        results.extend(result)
    result= ' '.join(results)

    return result
```

➡ 교수님의 stopwords.txt를 활용해 1차적으로 기본 불용어 제거 후, 어간 추출
➡ 총 소요 시간: 6일

```
In [8]: #대략 6일 정도 소요되었습니다
        df1['preprocessing_content']= df1['content'].apply(lambda x: preprocess_korean(x))
```

< 불용어 제거 및 어간 추출 결과 >

```
df1.content[0]
```

'₩n민주당 즉시 반박자료 내고 "저질 창작소설에 불과"₩n₩n₩n₩n"불법 대선 자금 수수 혐의₩"를 받고 있는 김용 민주연구원 전 부원장이 11일 서울 서초구 서울중앙지법에서 열린 불법 정치자금·뇌물 수수 관련 공판에 출석하고 있다. photo 뉴시스검찰이 김용 전 민주연구원 부원장이 대장동 일당으로부터 불법적으로 수수한 돈을 더불어민주당 이재명 대표의 2021년 대선 후보 예비경선 기탁금으로 사용했다고 의심되는 정황을 포착하고 수사 중이라고 중앙일보가 보도했다.16일 중앙일보와 법조계 등에 따르면 서울중앙지검 반부패수사3부(강백신 부장검사)는 김씨가 대장동 일당에게서 수수한 혐의를 받는 8억4700만원의 사용처를 추적하던 중 이런 정황이 담긴 통화녹음, 은행 전표 등을 확보했다고 한다.검찰은 이를 토대로 김씨가 자신이 받은 현금을 이 대표 자택에 옮긴 뒤 경기도청 비서실 직원들이 이 돈을 이 대표 계좌에 입금한 것으로 의심한다.검찰이 파악한 바에 따르면 2021년 6월 28~29일 이틀에 걸쳐 이 대표의 농협 계좌에는 총 3억2500만원이 입금됐다. 이는 이 대표가 2021년 공직자 재산공개 당시 2020년 말 기준으로 신고한 현금 액수와 같다. 이 대표는 이 중 1억원을 2021년 7월께 민주당 대선 후보 예비경선 기탁금으로 사용했다.이 대표 측은 이 돈이 선거 기탁금 등을 처리하기 위해 보유하던 현금과 모친상 조의금이라고 주장한다.하지만 검찰은 2021년 5~6월 김씨가 남욱씨가 만든 4억원을 유동규씨를 통해 현금으로 받은 만큼 이 돈의 출처가 김씨일 것으로 의심한다하고 있다.검찰은 이 대표 측근으로 알려진 전 경기도청 5급 공무원 배모씨가 이 대표 재산공개 내역에 맞춰 돈을 급히 입금한 정황으로 볼 수 있는 통화 녹음 파일도 확보했다.녹음파일에는 배씨가 "(재산공개) 현금 신고 내역을 알려주면 거기에 맞춰 입금하겠다"고 말하거나, 부하 직원을 이 대표 자택이 있는 경기 성남 분당구 수내동으로 급히 보내는 대목 등이 포함된 것으로 알려졌다.검찰은 이런 정황을 이달 4일 서울중앙지법에서 열린 김씨의 정치자금법 위반 혐의 재판에서 일부 공개했다.검찰은 김씨에게 "이재명 후보의 기탁금 출처가 무엇인지 아느냐"며 "본인이 받은 4억원 중 1억원을 이 후보에게 전달한 것 아니냐"고 추궁했다.이에 대해 김씨는 "모르는 일"이라며 "그런 점이 의심되면 수사를 하라"고 반발했다.검찰 관계자는 "현재 김씨가 대장동 일당에게서 받은 돈의 사용처를 계속 수사 중"이라고 밝혔다.이런 보도에 대해 더불어민주당 검찰독재정치탄압위원회 측은 반박자료를 내고"이 대표는 2021년 6월 28일 대선 경선을 위한 선거기탁금, 경선사무실 임차 등 2억 7000여만원의 처리를 위해, 본인 명의 농협통장에서 인출한 예금(19.3.20일 1억5000, 19.10.25일 5천 등)과 모친상(20.3.13) 조의금 등의 현금을 평소 거래하던 경기도청 농협 계좌에 입금했다"며 "이 같은 예금 변동 사실을 포함한 해당 현금 보유사실은 2020년 2021년 재산신고 하여 공직자재산신고서에도 명시되어 있고, 그 과정에서 아무런 문제가 없었다"고 주장했다. 민주당은 또한 "대장동 대선자금 설은 근거 없는 검찰의 저질 창작소설에 불과하다"고 덧붙였다. ※주간조선 온라인 기사입니다.₩n'

```
df1.preprocessing_content[0]
```

'민주당 반박 자료 내다 질 창작 소설 불과 불법 대선 자금 수수 혐의 받다 김용 민주 연구원 전 부원장 서울 서초구 서울중앙지법 열리다 불법 정치자금 뇌물 수수 관련 공판 출석 하고 뉴시스 검찰 김용 전 민주 연구원 부원장 대장동 일당 으로부터 불법 적 수수 하다 돈 더불다 민주당 이재명 대표 대선 후보 예비 경선 기탁금 사용 하다 의심 되다 정황 포착 하고 수사 중이 라고 중앙 일보 보도 하다 중앙 일보 법조 계 따르다 서 울 중앙 지 검 반 부패 수사 부 강 백신 부장 검사 늘다 김씨 대장동 일당 에게서 수수 하다 혐의 받다 억 만원 사용 처 추적 하다 중 이렇다 정황 담기다 통화 녹음 은행 전표 확보 하다 하다 검찰 은 이르다 토대 김씨 받다 현금 대표 자택 옮기다 뒤 경기 도청 비서실 직원 돈 대표 계좌 입금 하다 심하다 검찰 파악 한 바 따르다 이를 걸치다 대표 농협 계좌 에는 총 억 만원 입금 돼다 이다 대표 공직자 재산 공개 당시 말 기준 신고 한 현금 액수 대표 는 중 억원 께 민주당 대선 후보 예비 경선 기탁금 사용 하다 대표 측은 돈 선거 기탁금 처리 하다 위해 보유 하다 현금 모친상 조의 금 이라고 주장 한 다 검찰 은 김씨 남욱 씨 만들다 억원 유동 규 씨 통해 현금 받다 만큼 돈 출처 김씨 심하다 하다 검찰 은 대표 측근 알려지다 전 경기 도청 급 공무원 배 모씨 대표 재산 공개 내 역 맞추다 돈 급하다 입금 하다 정황 볼 수 통화 녹음 파일 도 확보 하다 녹음 파일 에는 배씨 재산 공개 현금 신고 내 역 알다 거기 맞추다 입금 하다 고 말다 부하 직원 대표 자택 경기 성남 분당구 수 내동 급하다 보내다 대목 포함 되다 알려지다 검찰 은 이렇다 정 황 달 서울중앙지법 열리다 김씨 정치자금 법 위반 혐의 재판 일부 공개 하다 검찰 은 김씨 이재명 후보 기 탁금 출처 인지 알다 며 본인 받다 억원 중 억원 후보 전달 한 아니다 고 추궁 하다 대해 김씨 는 모르다 라며 그렇다 점 의심 되다 수사 하라 고 반발 하다 검찰 관계자 는 현재 김씨 대장동 당 에게서 받다 돈 사용 처 계속 수사 중 이라고 밝히다 이렇다 보도 대해 더불다 민주당 검찰 독재정치 탄압 위원회 측은 반박 자료 내다 대표 는 대선 경선 위 한 선거 기탁금 경선 사무실 임차 억 만원 처리 위해 본인 명의 농협 통장 인 추다 예금 억 천 모친상 조의 금 현금 평소 거래 하다 경기 도청 농협 계좌 입금 하다 며 예금 변동 사실 포함 한 해당 현금 보유 사실 은 재산 신고 하다 공직자 재산 신 고서 에도 명시 되어다 과정 아무렇다 문제 없다 고 주장 하다 민주당 은 대장동 대선 자금 설다 근거 없다 검찰 질 창작 소설 불과 하다 고 덧붙이다 주간조선 온라인 기사 이다 '

➡ 기본적으로 \n, 숫자 및 영어 특수문자가 제거되었고,
어간이 추출된 것을 확인 할 수 있음

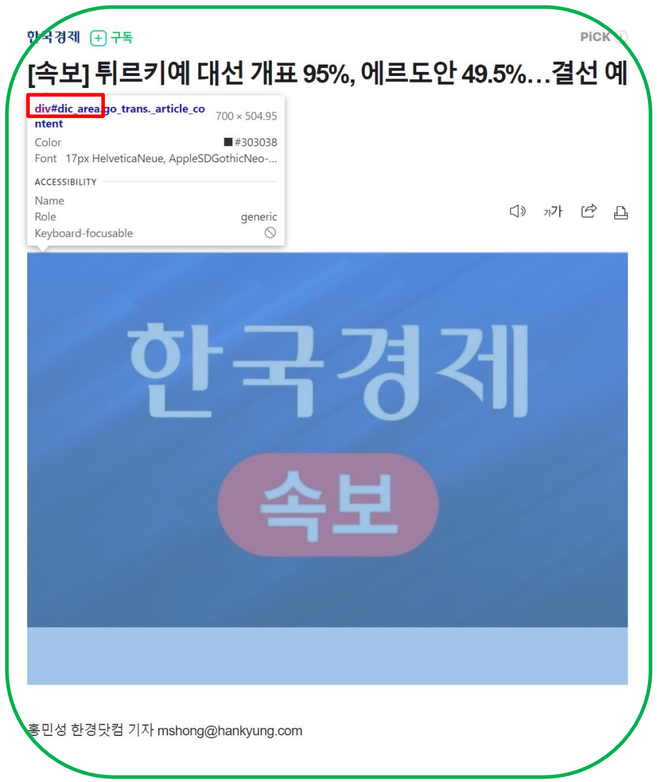
전처리 – 3. 1차 불용어 처리

```
In [231]: df.isnull().sum()

Out[231]: title                   0
          link                    0
          pubdate                 0
          content                 0
          preprocessing_content   22
          dtype: int64

In [232]: #1차 전처리 후, 본문이 비어있는 행을 확인할 수 있음
          df.iloc[[161]]
```

Out[232]:

	title	link	pubdate	content	preprocessing_content
161	[속보] 튀르키예 대선 개표 95%, 에르도안 49.5%…결선 예상	https://n.news.naver.com/mnews/article/015/000...	Mon, 15 May 2023 07:11:00 +0900	\n\n\n\n\n\n	NaN

➡ 1차 전처리 후, 본문이 없는 행 삭제



➡ 사진과 같이 본문에 아무 내용이 없고, 사진만 있는 기사들이 총 22개 존재.
➡ 해당 기사 삭제 후, “텍데분전처리1단계끝.csv” 로 저장

전처리 – 3. 2차 불용어 처리

①
```
In [238]: stopwords2= []
          for i in tqdm(range(len(df))): #0~10377
              for value in list(df.preprocessing_content[i].split(' ')):
                  stopwords2.append(value) #값 추가 -> 10400개의 본문에 대한 토큰들을 stopwords2에 저장함
```

②
```
In [240]: #가장 많이 나온 불용어들을 뽑아서 새로운 불용어 리스트에 저장 (상위 30개 중, 조사 위주로)
          word= pd.DataFrame(imsi['words'].value_counts()).head(30).index.tolist()

In [241]: # value_counts 값이 하나인 불용어들을 뽑아서 새로운 불용어 리스트에 저장 (하위 30개)
          word2=pd.DataFrame(imsi['words'].value_counts()).tail(30).index.tolist()

In [242]: stopwords2_total=word+word2
```

③
```
In [243]: # 조사 아닌 단어들 중,의미가 있을법한 단어는 불용어에서 제거함
          rm_set = {'대통령', '기업','기술','법','투자','서울'}
          # 리스트 컴프리헨션 활용: 삭제할 원소 집합 데이터와 일일이 비교
          stopwords2_total = [i for i in stopwords2_total if i not in rm_set]
          print(stopwords2_total)
```

```
['', '하다', '은', '는', '한', '되다', '이다', '도', '고', '적', '수', '하고', '인', '전', '다', '돼다', '만', '받다', '않다', '위',
'말', '늘다', '며', '밝히다', '극구', '아티브', '괴짜', '정략', '력했다', '리셔', '애슐리', '조리기구', '비비드', '아니스', '부생',
'컨벤셜홀', '쉑', '쉐이크', '명승', '단결권', '적성검사', '김말이', '호물', '바싹', '친독', '쿠라', '단체교섭권', '삼겹', '볍', '포케
포케', '팡팡', '단체행동권', '정보공학', '옥신']
```

④
```
In [244]: # 추가적으로 기사에 할당된 토큰들을 직접 찾아 보면서 (약 150개의 기사들을 예시로 찾아봄) 공통적으로 필요없는 불용어들을 리스트에 추가하
          stopwords3= ['','  ', '기자','무단','앙카라','로이터','뉴스','금지','무단','뉴스','제보','저자','방송','화면','캡쳐','사진','방송화면',
                       '연합뉴스','왼쪽부터','데일리안','현지','시각','시간','기사내용','뉴시스','뉴스데스크','카카오톡','기','다리다','이메일',
                       '앵커','자료조사','영상편집','리포트','채널','네이버','유튜브','구독','카카오','톡','전화','추가','영상','디자인',
                       '페이스북','트위터','노컷뉴스','사이트','기사','내용','요약','출처','은','는','이','가','이다','하다','돼다','에','에서',
                        '에선','라며','고','하','다','하고','하며','되다','뉴욕타임즈','오다','보다','따르다','가다','통해','에는','없다','대한',
                        '때문','관련','경우','이르다','그렇다','에서는','뿐 아니다','지다','들다','대다','보이다','에도','이나','아니다',
                        '씨','김','데','시','날','면서']
```

⑤
```
In [245]: # 조사들을 추가한 불용어 이외에 추가로 직접 찾은 불용어와의 비교를 통해 없는 불용어 추가
          for i in stopwords3:
              if i not in stopwords2_total:
                  stopwords2_total.append(i)

In [246]: stopwords2_total=' '.join(stopwords2_total)
```

① 10400개의 본문에 대한 토큰들을 stopwords2에 저장

② value_counts를 통해 상위 / 하위 30개의 값들을 새로운 불용어 리스트(stopwords2_total)에 저장

③ 조사가 아닌 단어들 중, 의미가 있을 법한 단어는 불용어 리스트에서 제거
ex) 대통령, 기업, 기술, 법, 투자, 서울 (value_counts().head(30)에 해당)

④ 추가적으로 기사에 할당된 토큰들을 직접 찾아 보면서 공통적으로 필요 없는 불용어들을 stopwords3에 추가해줌 (약 150개의 기사를 비교해봄) ex) 제보, 뉴스, 저자, 자료조사 등

⑤ stopwords2_total에 없는 stopwords3 불용어를 stopwords2_total에 추가해줌

전처리 – 3. 2차 불용어 처리

①

```
In [247]:
#https://junjun-94.tistory.com/18
#이 링크를 활용해 코드를 이해한 후, 함수로 변환하여 추가 불용어 처리 코드를 작성했습니다.
def preprocess_korean2(example):
    stop_words = stopwords2_total
    stop_words = stop_words.split(' ')

    word_tokens =word_tokenize(example)

    result=[]
    for w in tqdm(word_tokens):
        if w not in stop_words:
            result.append(w)

    result = ' '.join(result)


    return result
```

```
In [248]:
df['preprocessing_content2']= df['preprocessing_content'].apply(lambda x: preprocess_korean2(x))
```

②

```
In [250]:
# content에 내용이 없는 기사를 추가로 삭제
df1.query('preprocessing_content2==""')
```

Out[250]:

	title	link	pubdate	content	preprocessing_content	preprocessing_content2
2844	[속보] 검찰, '일감 몰아주기 의혹' KT 본사·계열사 등 압수수색	https://n.news.naver.com/mnews/article/421/000...	Tue, 16 May 2023 09:41:00 +0900	\n\n\n\n\nⓒ 뉴스1\n	뉴스	
4654	S. Korean startup Upstage unveils AI-powered s...	https://n.news.naver.com/mnews/article/001/001...	Tue, 16 May 2023 16:25:00 +0900	\n\t\t\tstartup-new AI productsS. Korean start...	...	
5608	GS25 to debut ChatGPT-developed whiskey highball	https://n.news.naver.com/mnews/article/044/000...	Tue, 16 May 2023 14:31:00 +0900	\n\n\n\n\nA model holds up cans of AskUp Lemon...	...	
5876	[Meet the President] Korea University presiden...	https://n.news.naver.com/mnews/article/640/000...	Tue, 16 May 2023 17:52:00 +0900	\n \n\n\n\nNewly-elect Korea University Presi...	...	
7782	[속보] '곽상도 부자 50억' 호반건설·산업은행 관계자 압수수색	https://n.news.naver.com/mnews/article/421/000...	Mon, 24 Apr 2023 13:56:00 +0900	\n\n\n\n\nⓒ 뉴스1\n	뉴스	

```
In [251]:
df1.drop([2844,4654,5608,5876,7782], axis=0,inplace=True)
```

```
In [252]:
df1=df1.reset_index(drop=True)
```

③

```
pd.read_csv('data/텍데분전처리2단계끝.csv').iloc[:,1:].shape
```

```
(10373, 6)
```

```
In [253]:
df1.to_csv('텍데분전처리2단계끝.csv')
```

① 위에 정의해둔 stopwords2_total 불용어 리스트를 통해 추가 불용어 처리 진행

② content에 내용이 없는 기사 추가적으로 삭제

(1차 불용어 처리한 후 생성된 '뉴스' 키워드만 있던 content에서, 2차 불용어 전처리를 통해 '뉴스'가 사라짐-> 따라서 빈 값으로 채워짐)

③ 최종 df.shape : (10373,6)
"텍데분전처리2단계끝.csv" 로 저장

모델링 – 4. 데이터 모델링 및 분석 (WordCloud)

1) 정치 및 법 (선거, 정부, 법, 검찰, 살인) 약 4000개

```
#슬라이싱의 경우, 토픽별 기사 개수를 활용해 다음 토픽과의 경계 인덱스를 직접 찾아주었습니다
politics= df.iloc[:3848,:].reset_index(drop=True)
```

```
all_data = ''
for _, row in politics.iterrows(): #행의 정보를 담은 객체
    all_data += row['preprocessing_content2']
```

2) 기술 (빅데이터, AI, 자율주행, 가상현실, GPT) 약 2800개

```
tech= df.iloc[3848:6458,:].reset_index(drop=True)
```

```
all_data = ''
for _, row in tech.iterrows(): #행의 정보를 담은 객체
    all_data += row['preprocessing_content2']
```

3) 경제 (투자, 기업, 부자, 주식, 실업, 보험) 약 3000개

```
money=df.iloc[6458:9375,:].reset_index(drop=True)
```

```
all_data = ''
for _, row in money.iterrows(): #행의 정보를 담은 객체
    all_data += row['preprocessing_content2']
```

4) 환경 (날씨, 재활용) 약 1000개

```
env=df.iloc[9375:,:].reset_index(drop=True)
```

```
all_data = ''
for _, row in env.iterrows(): #행의 정보를 담은 객체
    all_data += row['preprocessing_content2']
```

```
mask= np.array(Image.open('투명png/politic.png'))

cloud = WordCloud(font_path = font_path,
                  colormap='gist_rainbow',
                  background_color = 'black',
                  collocations=True,
                  width=2000, height=1000,
                 mask=mask)
my_cloud1 = cloud.generate_from_text(all_data)

arr1 = my_cloud1.to_array()

fig = plt.figure(figsize=(10, 10))
plt.imshow(arr1)
plt.axis('off')
plt.show()
fig.savefig('politics.png') #생성한 그림 저장하기
```

➡ 4개의 topic에 대한 WordCloud 진행

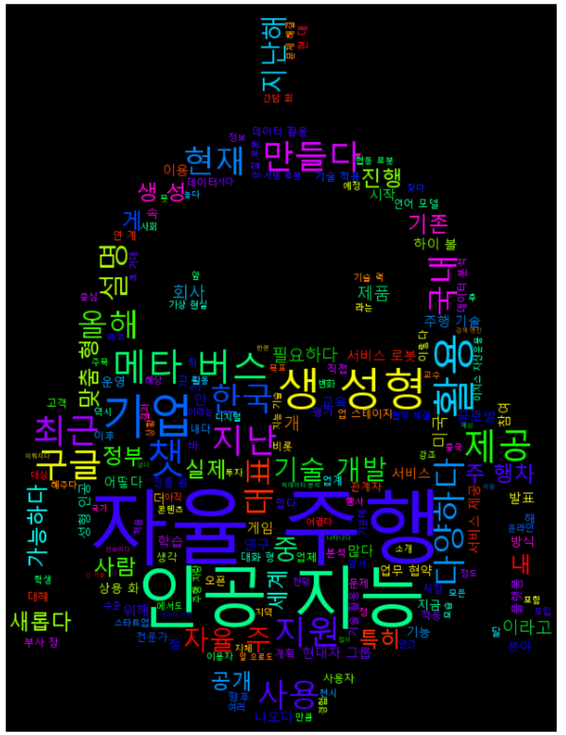
➡ 슬라이싱의 경우, 토픽 별 기사 개수(대략적 정보-len)를 통해 다음 토픽과의 경계 인덱스를 직접 찾아줌

모델링 – 4. 데이터 모델링 및 분석 (WordCloud)

< 정치 및 법 > - 선거, 정부, 법, 검찰, 살인

➡ 윤 대통령, 국민 힘, 거부권 행사, 간호, 법이 대표 keyword로 추출됨

< 기술 > - 빅데이터, AI, 자율주행, 가상현실, GPT, 로봇



➡ 자율 주행, 인공지능, 기업, 메타버스가 대표 keyword로 추출됨

1) 아무래도 현 대통령인 윤 대통령에 대한 keyword가 압도적임을 알 수 있음.
2) 생각했던 keyword와 달리 '거부권 행사' 와 '간호' 에 대한 keyword가 상당히 언급됨.
   조사해본 결과, 현재 국회에서 간호법 거부권 행사에 대한 제정안 논의가 많이
   언급되고 있어 상위 keyword로 나타난 것으로 추정됨.

1) 추가적으로 "생성, 개발, 제공, 활용, 지원, 만들다" 와 같이 소프트웨어 업계에서
   쓰이는 용어가 많이 나옴.
2) "현재, 올해, 최근, 지난, 기존, 새롭다" 와 같은 시간 관련 용어도 많이 나오는
   것을 통해 기술 분야는 시간에 밀접한 관계가 있음을 도출할 수 있음.

모델링 – 4. 데이터 모델링 및 분석 (WordCloud)

< 경제 > - 투자, 기업, 부자, 주식, 실업, 보험

➡ 투자 증권, 억원, 기업, 미국, 가상 화폐가 주 keyword로 추출됨

➡ 생각했던 것과 같이 경제와 관련된 카테고리이다 보니, 경제 용어가 많이 나옴.
➡ 또한 시간의 흐름에 민감한 분야이다 보니, “최근, 지난, 올해, 현재” 와 같은 키워드가 많이 도출된 것을 확인할 수 있음.

< 환경> - 날씨, 재활용

➡ 날씨와 관련된 “기온, 낮, 최저, 최고 기온”과
재활용과 관련된 “폐 플라스틱, 폐 배터리, 자원 순환”이 주 keyword로 추출됨

➡ 날씨에 대한 기사가 약 650개, 재활용에 대한 기사가 약 360개 정도로, 날씨 keyword에 편향되어 나타남.

모델링 – 4. 데이터 모델링 및 분석 (LDA)

```
# LDA를 위해서는 리스트 형태로 바꿔주어야 한다
df['preprocessing_content2'] = df['preprocessing_content2'].apply(lambda x: x.split())
```

```
word_dict = corpora.Dictionary(df['preprocessing_content2'])
```

```
corpus = [word_dict.doc2bow(text) for text in df['preprocessing_content2']]
```

➡ LDA를 진행하기 위해 데이터 형태를 변경함
➡ List로 변경
➡ 이후, 형태에 맞게 BOW로 변환하는 과정 수행

```
df['preprocessing_content2']
```

```
0        민주당 반박 자료 내다 질 창작 소설 불과 불법 대선 자금 수수 혐의 김용 민주 연...
1        취임 후 두 번째 거부권 행사 간호 법 국민 건강 불안감 초래 다시 국회 재의 결 ...
2        이재명 더불다 민주당 대표 대선 후보 예비 경선 기탁금 납부 억원 김용 민주 연구원.
3        거부권 행사 갈등 해소 갈등 증폭 서울 이영환 박광온 더불다 민주당 원내대표 오전 .
4        며칠 내 지지 후보 난 오안 후보 통신 인터뷰 치러지다 튀르 키 대통령 선거 결선 ...
                              ...
10368    상상 그룹 지구 맞다 휠체어 사용 아동 이동성 향상 프로젝트 확보 유 휠체어 여대 ..
10369    광주 이승현 강기 정 광주시 장이 오후 광주 순환도로 지산 진출 대안 발표 현장 설..
10370    삼성 애플 폐 휴대폰 업 사이클 리 사이클 수준 고도화 삼성 모델 브라질 상파울루 ..
10371    간 번 선거 서다 발생 탄소 배출 지자체 처리 골 머리 이른바 무제한 현수막 법 불...
10372    광주 이영주 오전 분 께 광주 동구 서 석 동 이면 도로 재활용품 수거 차량 도로 ...
Name: preprocessing_content2, Length: 10373, dtype: object
```

```
# LDA를 위해서는 리스트 형태로 바꿔주어야 한다
df['preprocessing_content2'] = df['preprocessing_content2'].apply(lambda x: x.split())
```

```
df['preprocessing_content2']
```

```
0        [민주당, 반박, 자료, 내다, 질, 창작, 소설, 불과, 불법, 대선, 자금, 수...
1        [취임, 후, 두, 번째, 거부권, 행사, 간호, 법, 국민, 건강, 불안감, 초래...
2        [이재명, 더불다, 민주당, 대표, 대선, 후보, 예비, 경선, 기탁금, 납부, 억...
3        [거부권, 행사, 갈등, 해소, 갈등, 증폭, 서울, 이영환, 박광온, 더불다, 민...
4        [며칠, 내, 지지, 후보, 난, 오안, 후보, 통신, 인터뷰, 치러지다, 튀르, ...
                              ...
10368    [상상, 그룹, 지구, 맞다, 휠체어, 사용, 아동, 이동성, 향상, 프로젝트, 확...
10369    [광주, 이승현, 강기, 정, 광주시, 장이, 오후, 광주, 순환도로, 지산, 진출...
10370    [삼성, 애플, 폐, 휴대폰, 업, 사이클, 리, 사이클, 수준, 고도화, 삼성, ...
10371    [간, 번, 선거, 서다, 발생, 탄소, 배출, 지자체, 처리, 골, 머리, 이른바...
10372    [광주, 이영주, 오전, 분, 께, 광주, 동구, 서, 석, 동, 이면, 도로, 재...
Name: preprocessing_content2, Length: 10373, dtype: object
```

모델링 – 4. 데이터 모델링 및 분석 (LDA) : N_TOPICS = 4

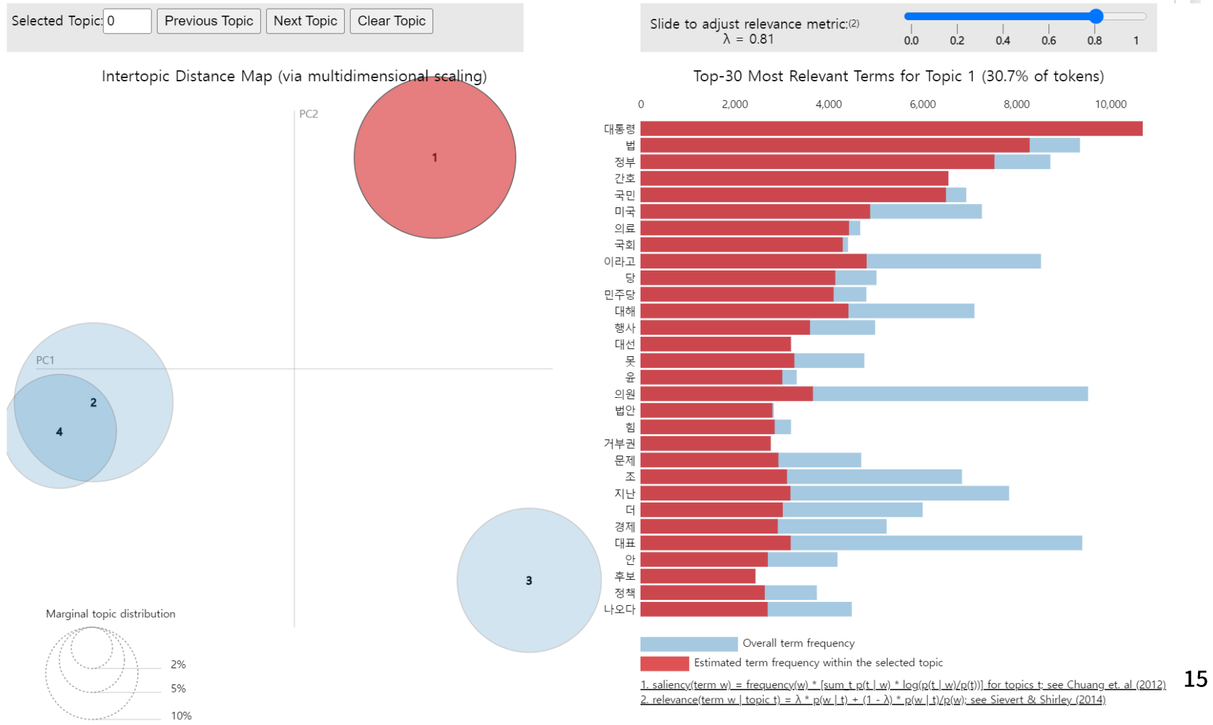
```
N_TOPICS = 4
ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics = N_TOPICS, id2word=word_dict, passes = 15)

topics = ldamodel.print_topics(num_words=4)
for topic in topics:
    print(topic)
```

```
(0, '0.012*"재활용" + 0.009*"기술" + 0.008*"폐" + 0.007*"기업"')
(1, '0.015*"대통령" + 0.012*"법" + 0.012*"정부" + 0.010*"국민"')
(2, '0.012*"보험" + 0.011*"의원" + 0.008*"검찰" + 0.008*"혐의"')
(3, '0.006*"투자" + 0.006*"미국" + 0.005*"기온" + 0.005*"억원"')
```

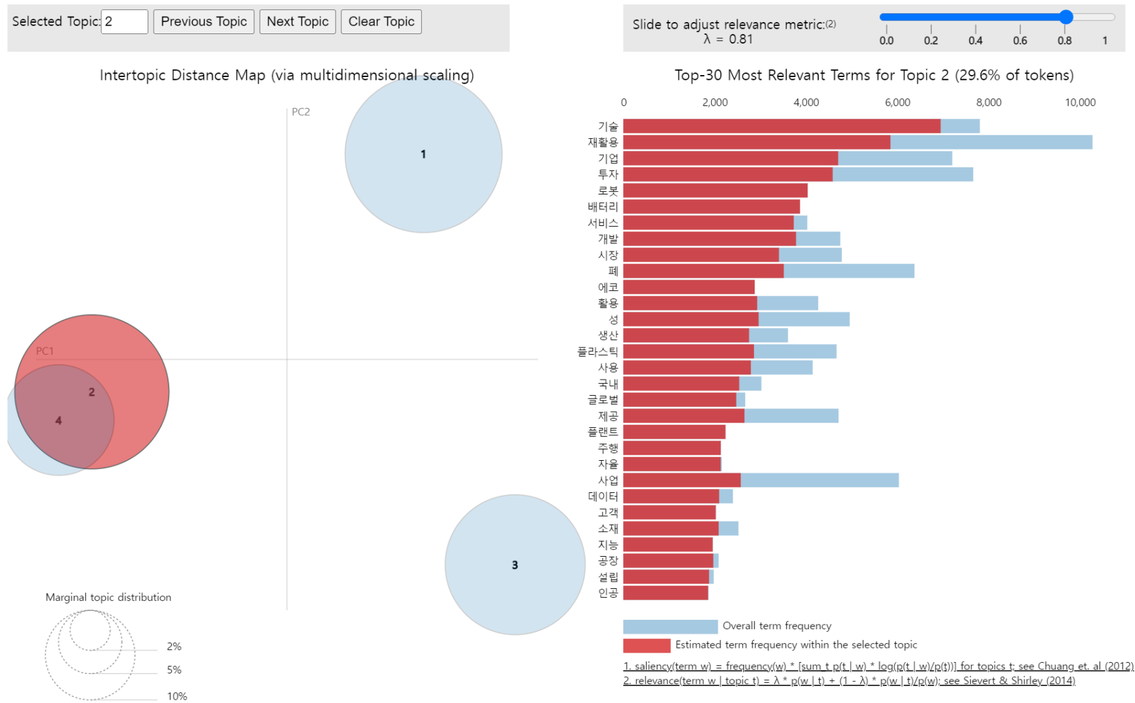
```
pyLDAvis.enable_notebook()
vis = pyLDAvis.gensim_models.prepare(ldamodel, corpus, word_dict)
pyLDAvis.display(vis)
```

➡ 처음에 선정했던 4개의 토픽으로 모델링이 잘 될까? 를 역으로 확인해보기 위해 N_TOPICS를 4로 설정함

< Topic 1 >

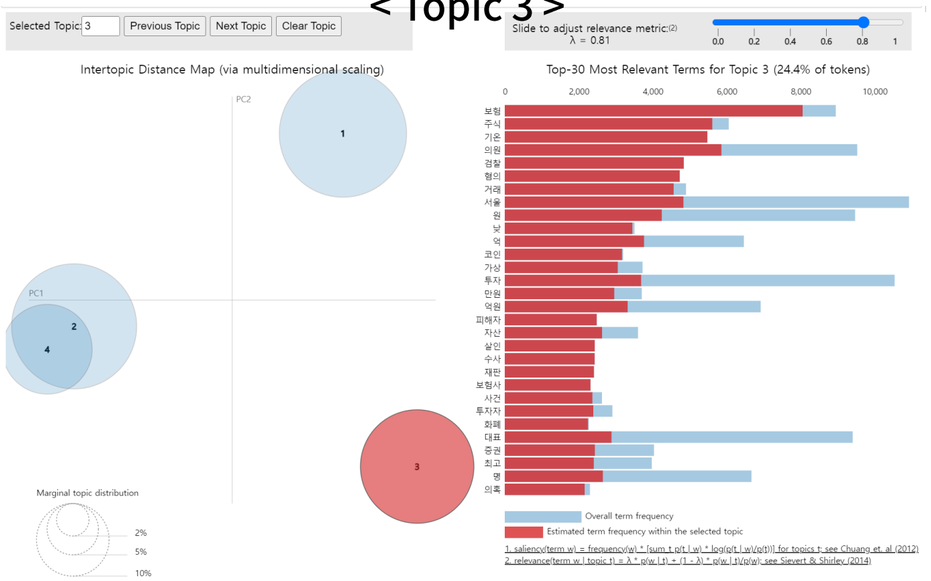


< Topic 2 >

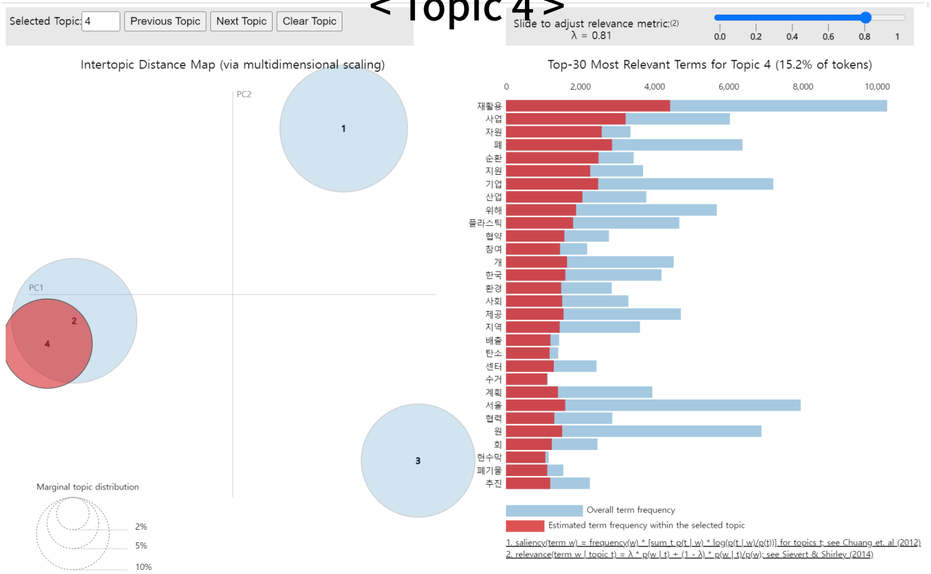


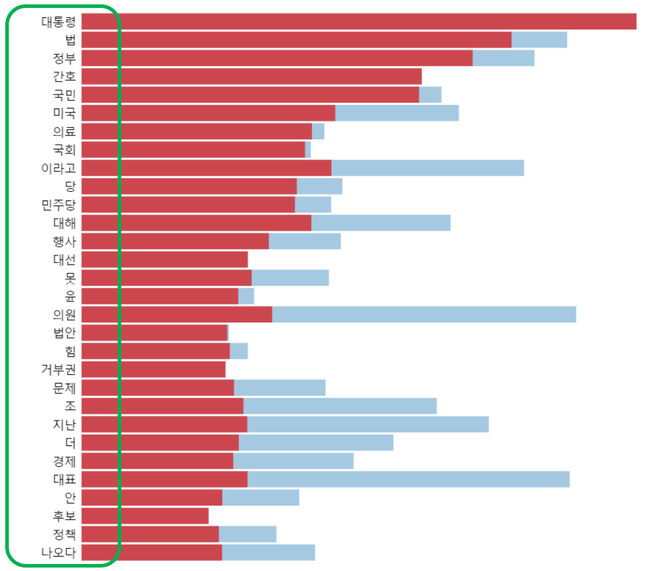
모델링 – 4. 데이터 모델링 및 분석 (LDA) : N_TOPICS = 4

< Topic 3 >



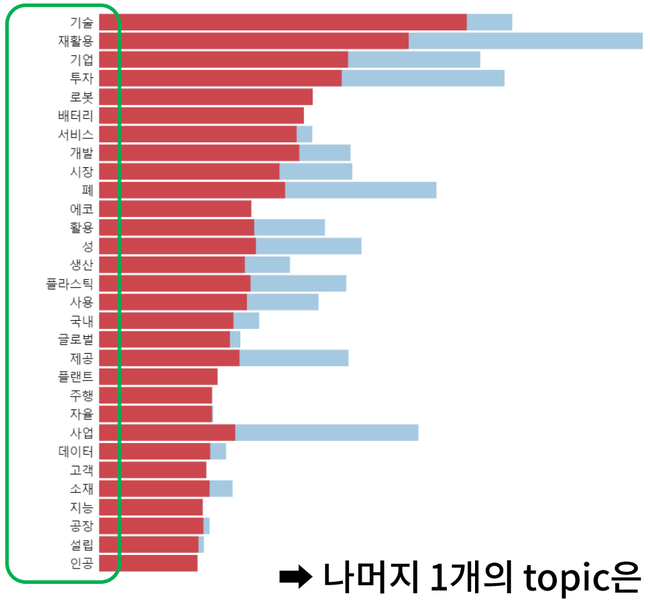
< Topic 4 >





< Topic 1에 대한 결과 >

➡ 정치 및 법에 관련된 keyword들이 많이 할당되어 있음을 확인할 수 있음

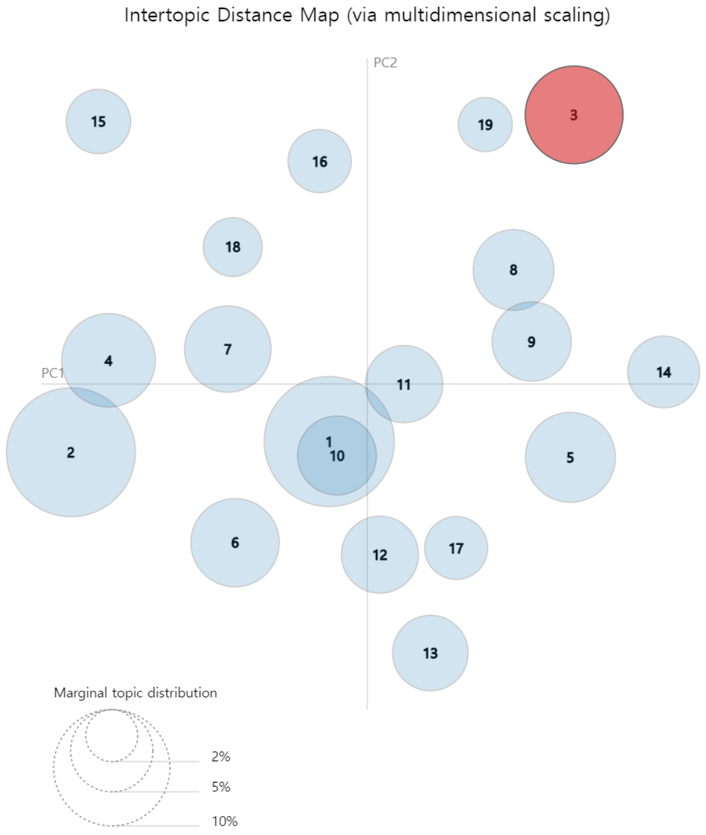


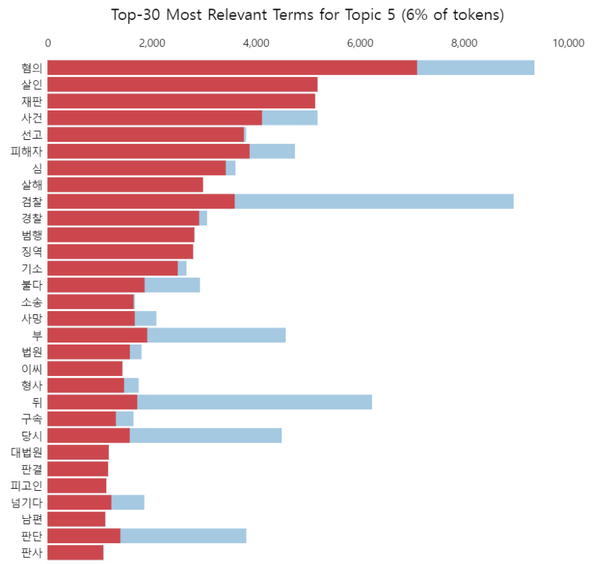
< Topic 4에 대한 결과 >

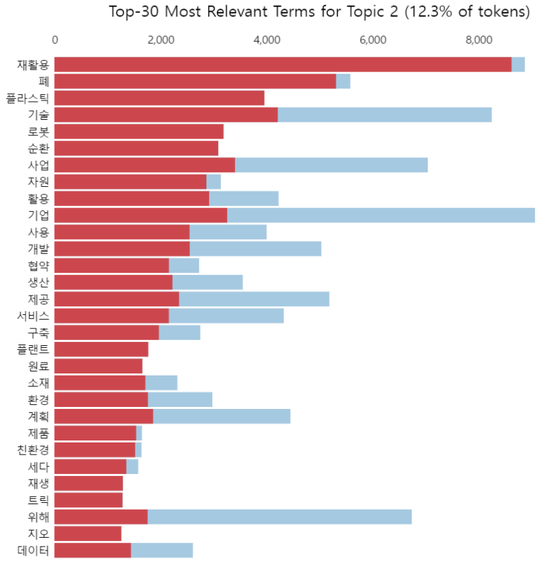
➡ 환경에 관련된 keyword들이 많이 할당되어 있음을 확인할 수 있음 (Topic 2도 유사)

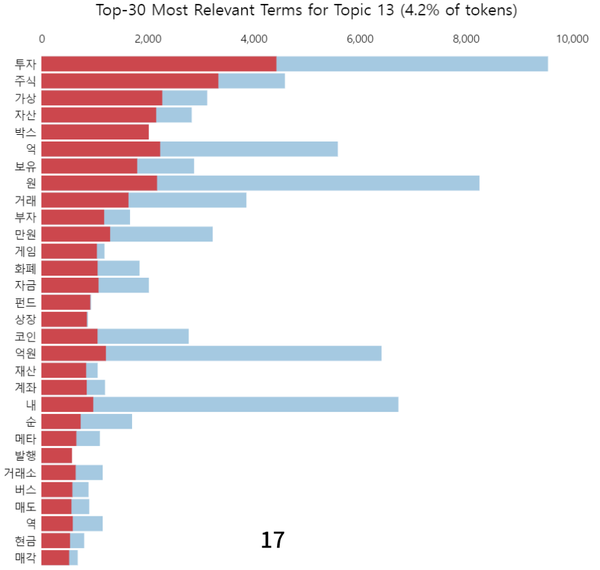
➡ 나머지 1개의 topic은 어떤 연관성을 지니는지 파악하기 어려움

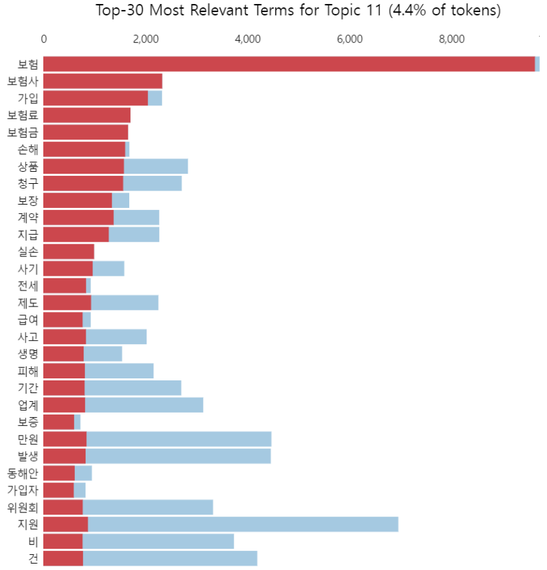
모델링 – 4. 데이터 모델링 및 분석 (LDA) : N_TOPICS = 19



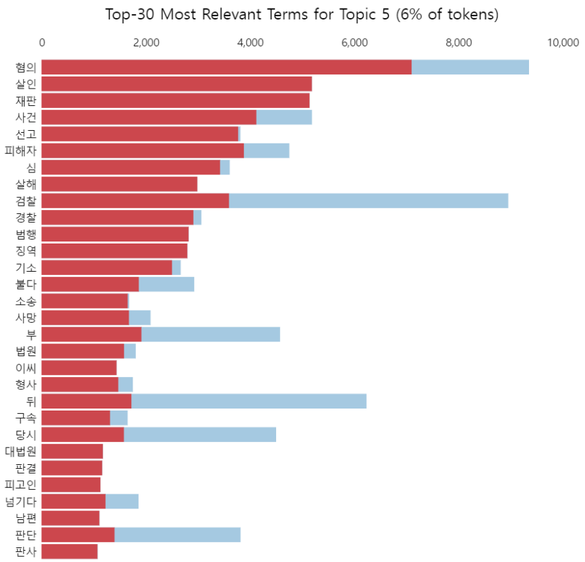


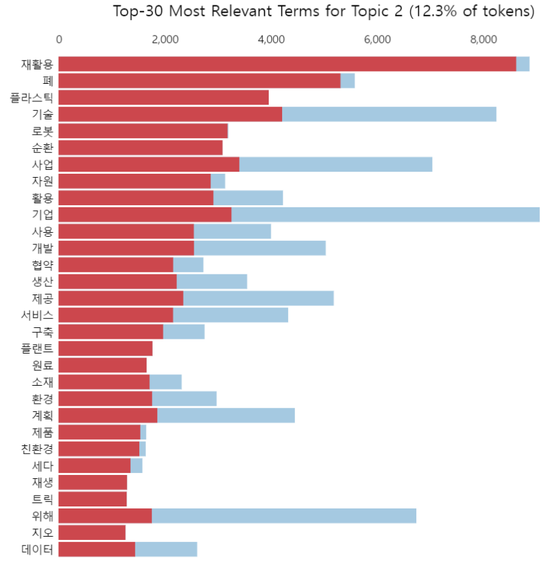


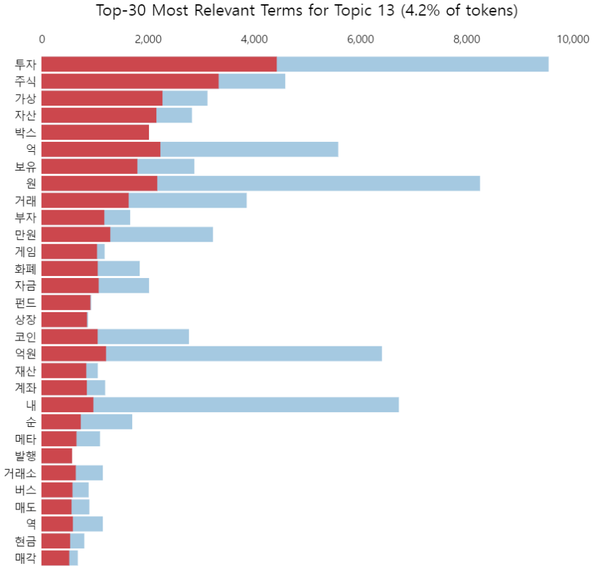


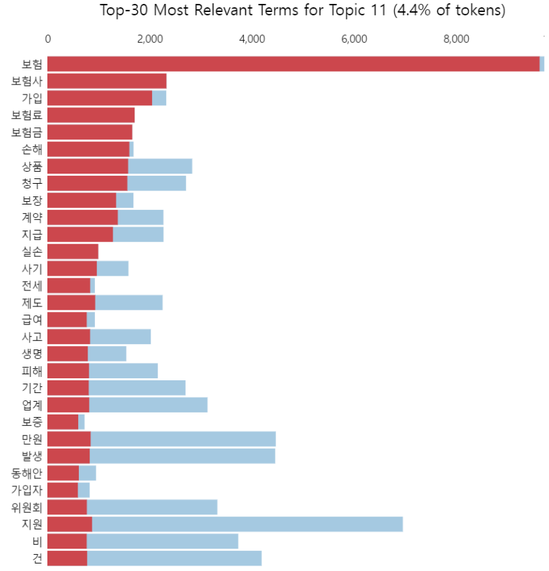


모델링 – 4. 데이터 모델링 및 분석 (LDA) : N_TOPICS = 19









1) 19개로 topic을 나눠본 결과, 4개로 나눴던 것보다 좀 더 확연하고, 구체적으로 토픽 모델링이 된 것을 확인함

2) 정치 분야와 경제 분야는 겹치는 것들이 많아, 두 분야에 해당하는 keyword들이 합해져 토픽이 형성된 것들이 다수 보임

3) 반면, 환경(날씨, 재활용)과 같은 분야는 다른 분야와 이질적인 형태로써, 구분이 보다 확실하게 된 것을 확인할 수 있음

모델링 – 5. 한계점 및 보완점, 느낀점

< 한계점 및 보완점, 느낀점 >

1. 양 대비 부족한 시간

➡ 시간 여유가 조금 더 있었다면, 불용어 리스트에 더 많은 단어들을 추가하고 모델을 돌리면, 더 깔끔하고 의미 있는 토큰들이 추출되었을 것 같은데 당장 기본적인 처리만 했을 때도 6일이라는 시간이 걸린 것에 있어서 더 많은 불용어를 추가하지 못했다는 점이 조금 아쉬웠습니다.

2. 크롤링의 다양성

➡ 네이버 뉴스 본문만 크롤링을 진행했었지만, 다음 기회에는 유튜브 댓글이나 네티즌의 댓글의 단어 분포들도 전처리하여 분석을 진행해보고 싶습니다.

3. WordCloud 및 LDA – 수집 및 전처리의 중요성

➡ 예상했던 결과와 달리, 생각보다 분석이 잘된 것 같아 앞에서 진행했던 과정들을 정리하면서 느낀 것이 있는데, 수집과 전처리가 분석 과정에서 가장 핵심적이고 중요한 단계라는 점입니다. 이번 과제를 수행하며 크롤링 기술 뿐만 아니라 데이터의 초기 단계에서의 중요성 또한 다시 한 번 깨닫는 시간이 되었습니다.