ESAA PROJECT

2024

NEWS Analysis

05.27

Published by OB2조

소식 01.
Garbage in! Garbage Out!
텍스트 분석을 위한 길고 긴 전처리

소식 02.
한눈에 보는 총선 전 정치 이슈
기간별 정치이슈와 언론사 별 정치이슈

소식 03.
진보와 보수 신문사가 논평하는 우리 사회,
그 시각의 차이는?

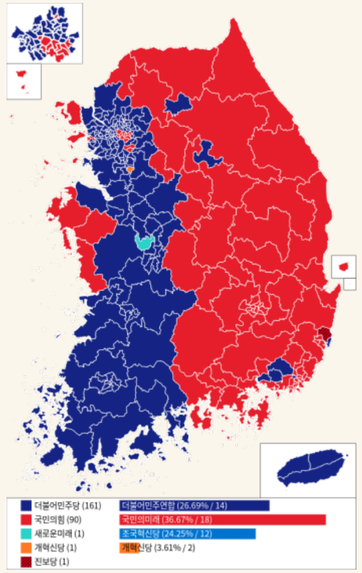
ESAA PROJECT

2024

CONTENTS

05.27

04/10 Result



POLITICAL POLARIZATION

01 크롤링을 통한 신문 기사 수집

02 텍스트분석 데이터 전처리

03 워드 클라우드로 보는 기간별 정치 이슈

04 사설 분류로 보는 진보,보수 신문사

05 사설 Topic과 성향에 따른 연관 단어 보기

06 4.10 총선 정치 기사 분석 프로젝트 의의

Through News Analysis

~24.04.10

한겨레

한국일보

경향

ChosunMedia 조선일보

ESAA PROJECT

2024

Before Crawling

05.27

프로젝트에서 사용한 크롤링을 하기 전까지의 문제점

1.네이버 크롤링 - 메이저사 제외

네이버 뉴스를 통해 검색되는 기사들
→ 모두 html 구조가 달라 함수 통일을 위해
네이버 뉴스만 남기는 구조로 진행
→ 메이저 언론사 다 제외되는 문제 발생

```
#NAVER 뉴스만 남기기
final_urls = []
for i in tqdm(range(len(news_url_1))):
    if "news.naver.com" in news_url_1[i]:
        final_urls.append(news_url_1[i])
    else:
        pass
```

```
검색할 키워드를 입력해주세요:총선

크롤링할 시작 페이지를 입력해주세요. ex)1(숫자만입력):1

크롤링할 시작 페이지:  1 페이지

크롤링할 종료 페이지를 입력해주세요. ex)1(숫자만입력):2
```

```
content = news_html.select("#articleBody")

# 기사 본문 텍스트 가져오기 및 합치기
content = ''.join([str(p) for p in content])

# HTML 태그 제거 및 텍스트 다듬기
title = re.sub(pattern1, '', str(title))
content = re.sub(pattern1, '', content).replace(pattern2, '')

news_titles.append(title)
news_contents.append(content)

try:
    html_date = news_html.select_one(
        "div#ct > div.media_end_head.go_trans
                > div.media_end_head_info.nv_notrans >div.media_end_head_info_datestamp > div > span")
    news_date = html_date.attrs['data-date-time']
except (AttributeError, TypeError):
    news_date = news_html.select_one("#content > div.end_ct > div > div.article_info > span > em")
    news_date = re.sub(pattern1, '', str(news_date))

# 날짜 가져오기
news_dates.append(news_date)
```

2.뉴스 빅데이터 DB 빅카인즈 사용

이낙연 "총선 끝나고 더 심각한 위기"...

뉴시스 이낙연 새로운미래 공동대표는 9일 윤석열 대통

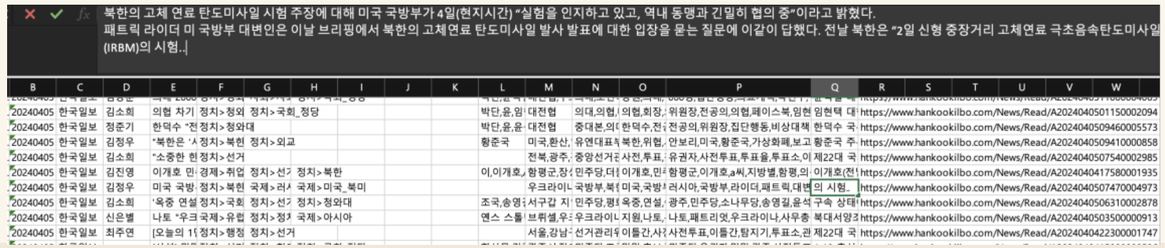
세계일보 정치>국회_정당 | 정치>선거 | 정치>청와대

北, 총선 사흘전 인천·서북도서에 고강도

북한이 이달초 서해상으로 수차례 GPS 교란 공격을 한

조선일보 정치>북한 2024/04/09 양지호 기자

3.빅카인즈 엑셀 - 뉴스본문이 생략된 채 다운



4.빅카인즈 기반 크롤링 결정!

총선 전까지 정치섹션에 한정해, 언론사 별,
다운받은 빅카인즈 엑셀 내 기사 URL
→ URL마다 들어가서 기사 내용 크롤링

	일자	언론사	제목	URL
5000	20231215	한국일보	남태평양, 얽히고설키는 안보 협정 [오늘, 세계]	https://www.hankookilbo.com/News/Read/A2023121...
5001	20231214	한국일보	푸틴 "러시아 목표 달성할 때까지 전쟁 계속할 것"	https://www.hankookilbo.com/News/Read/A2023121...
5002	20231214	한국일보	'윤심' 오독한 김기현, 체면도 실리도 다 까먹었다	https://www.hankookilbo.com/News/Read/A2023121...
5003	20231214	한국일보	'이한열 박종철' 유공자 인정하는 민주유공자법, 국회 정무위 통과	https://www.hankookilbo.com/News/Read/A2023121...
5004	20231214	한국일보	"경제성 있다" 세종시 KTX역 신설 추진...충북도는 '반발'	https://www.hankookilbo.com/News/Read/A2023121...

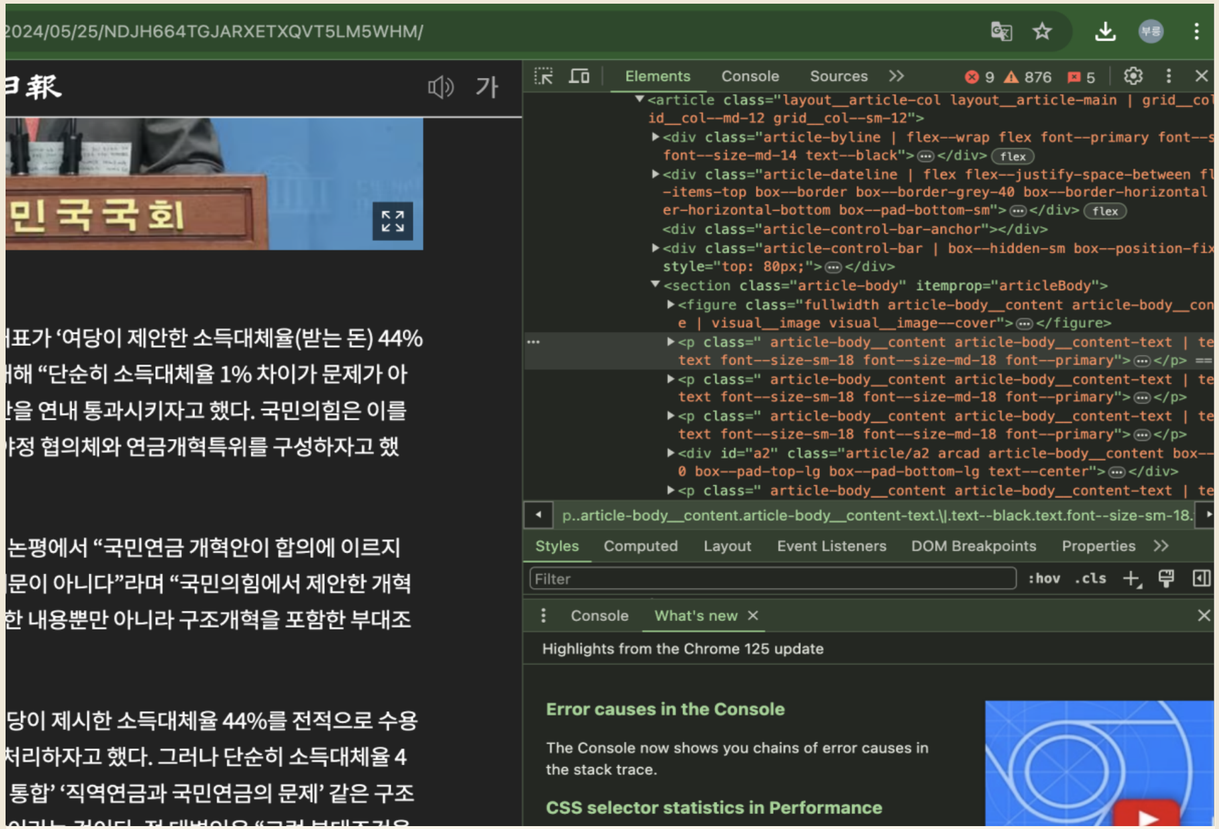
ESAA PROJECT

20
24

Crawling

05.
27

크롤링 시작! (동적 크롤링과 정적 크롤링)



01기사 위치 확인하기

```python
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from webdriver_manager.chrome import ChromeDriverManager
from bs4 import BeautifulSoup
import pandas as pd

# WebDriver 설정 (Chrome 사용)
s = Service(ChromeDriverManager().install())
driver = webdriver.Chrome(service=s)

# 기사 내용을 저장할 리스트
news_contents = []

# WebDriver가 페이지의 요소를 로드할 때까지 최대 10초간 기다리도록 설정
wait = WebDriverWait(driver, 10)


for url in raw_df['URL']:
    try:
        driver.get(url)
        # 지정된 요소가 DOM에 존재하고 보이게 될 때까지 기다립니다.
        wait.until(EC.visibility_of_element_located((By.CSS_SELECTOR,
                                                     'p.article-body__content.article-body__content-text'))

        # BeautifulSoup으로 페이지의 HTML 분석
        soup = BeautifulSoup(driver.page_source, 'html.parser')
        news_paragraphs = soup.select('p.article-body__content.article-body__content-text')
        clean_text = ' '.join([p.get_text(strip=True) for p in news_paragraphs])

        # 리스트에 기사 내용 추가
        news_contents.append(clean_text if clean_text else '내용 없음')
    except Exception as e:
        print(f"Error fetching {url}: {str(e)}")
        news_contents.append('크롤링 실패')

driver.quit()  # WebDriver 종료
```

02위치에 따른 동적 크롤링

2024

PROCESSING

05.27

데이터 전처리

A. 데이터 결합&삭제

1) 데이터 결합

경향신문

+

조선일보

+

한국일보

+

한겨레

−

2) 데이터 삭제

크롤링 실패

+

NaN

=

3) 데이터프레임 완성

	일자	언론사	제목	URL	기사내용
1					
2					
3					
.					
.					
.					
50071					

- 5 x 50071 의 데이터프레임이 생성되었다.

2024

PROCESSING

05.27

데이터 전처리

B. 데이터 클리닝

1) 단어 통일

국힘, 국민의힘, 국민의 힘
이재명, 이대표
윤 대통령, 대통령
김여사, 김건희
...

→ 하나의 단어로 통일!

2) 필요 없는 내용 제거

이메일, url, 바이라인, 기업공시 등등

3) 한글, 알파벳, 숫자만 남기기

(예) 첫 번째 기사

```
article['기사내용'][0]
```

'한동훈 국민의힘 비상대책위원장은 4·10 22대 총선을 하루 앞둔 9일 서울 중구 청계광장에서 마지막 유세를 했다. 한 위원장은 "대한민국이 경악스러운 혼돈으로 무너질지, 아니면 위기를 극복할지를 결정할 운명의 시간이 다가왔다" 며 "무도하고 뻔뻔한 야당을 견제할 수 있는 최소한의 의석을 달라" 고 했다. 한 위원장은 이날 서울 격전지 등 15곳을 도는 강행군을 펼쳤다. 한 위원장은 이날 저녁 청계광장 유세에서 "더불어민주당이 샴페인을 터뜨리며 조롱하듯 말하는 200석이 만들 혼돈과 퇴행을 생각해달라" 며 "충무공(이순신)이 남은 12척 배로 나라를 구했다. 나라를 구하기 위해 12시간이 남아 있다. 딱 한 표가 부족하다" 고 강조했다. 여당이 열세라는 전망에 투표를 포기하려는 지지층에 투표장으로 나와달라고 호소한 것이다. 한 위원장은 "(비대위원장으로 일한) 지난 100일 동안 정부와 여당에 대해 여러분이 잘못이라고 지적하면 바로 바꾸고 바로잡았다" 며 "앞으로 더 그렇게

↓

```
article['기사내용_clean'][0]
```

'한동훈 국민의힘 비상대책위원장은 4 10 22대 총선을 하루 앞둔 9일 서울 중구 청계광장에서 마지막 유세를 했다 한 위원장은 대한민국이 경악스러운 혼돈으로 무너질지 아니면 위기를 극복할지를 결정할 운명의 시간이 다가왔다 며 무도하고 뻔뻔한 야당을 견제할 수 있는 최소한의 의석을 달라 고 했다 한 위원장은 이날 서울 격전지 등 15곳을 도는 강행군을 펼쳤다 한 위원장은 이날 저녁 청계광장 유세에서 더불어민주당이 샴페인을 터뜨리며 조롱하듯 말하는 200석이 만들 혼돈과 퇴행을 생각해달라 며 충무공 이순신 이 남은 12척 배로 나라를 구했다 나라를 구하기 위해 12시간이 남아 있다 딱 한 표가 부족하다 고 강조했다 여당이 열세라는 전망에 투표를 포기하려는 지지층에 투표장으로 나와달라고 호소한 것이다 한 위원장은 비대위원장으로 일한 지난 100일 동안 정부와 여당에 대해 여러분이 잘못이라고 지적하면 바로 바꾸고 바로잡았다 며 앞으로 더 그렇게 하겠다 민심만 따르겠다 제가 책

- 특수문자가 모두 제거된 것을 확인할 수 있다.

2024

PROCESSING

05.27

데이터 전처리

C. 불용어 제거, 품사 태깅

1) 불용어 사전 불러오기

불용어 사전 다운로드

어, 나, 우리, 저희,
따라, 의해, 을 등

+

사전에 없는 불용어 추가

수, 말, 전

2) 사용자 지정 사전 생성

국민의힘 → 국민 + 의 + 힘
조국혁신당 → 조국 + 혁신 + 당
더불어민주당 → 더불어 + 민주당

→ **방지**

3) 품사태깅 : OKT 이용

<형태소 분석기 선택>

Stanza , Kkma : 느리다.
→ 속도가 빠른 okt와 바른 중 선택

↓

okt vs 바른
(예) 한국은행
- okt : 한국은행
- 바른 : 한국+은행

↓

본래의 뜻을 살리기 위해
단어를 더욱 큼지막하게 나누는 **okt 선택**

<품사 태깅 및 불용어 제거 함수 적용 전/후 비교>

기사내용_clean	기사내용_tagged
한동훈 국민의힘 비상대책위원장은 4 10 22대 총선을 하루 앞둔 9일 서울 중구 ...	[한동훈, 국민의힘, 비상, 대책, 위원장, 총선, 하루, 서울, 중구, 청계, 광...
이재명 더불어민주당 대표가 서울 용산역 광장에서 4 10 22대 총선 전날인 9일 ...	[이재명, 더불어민주당, 대표, 서울, 용산역, 광장, 총선, 전날, 마지막, 유세...
어엿한 대한민국 국민의 한 사람으로서 투표에 참여할 수 있다는 것 자체가 자랑스럽고...	[대한민국, 국민, 사람, 투표, 참여, 자체, 연해주, 항일, 무장, 투쟁, 독립...
박정희 기념사업 조례안 에 대한 대구시민의 반발이 큰 것으로 나타났다 대구시는 지난...	[박정희, 기념, 사업, 조례, 안, 대한, 시민, 반발, 대구시, 지난달, 대구광...
방송통신심의위원회 방심위 방송심의소위원회 방송소위 가 윤석열 대통령의 바이든 날리면...	[방송통신, 심의, 위원회, 방심위, 방송, 심의, 위원회, 방송, 소위, 윤석열,...

2024

PROCESSING

05.27

데이터 전처리

D. 워드 클라우드 확인 및 추가 처리

1) 워드 클라우드 확인

2) 불용어 추가 제거후 확인

- 불용어 추가 제거
  많이 등장하는
  며, 라며, 더, 때문, 대해
  등의 단어를 제거하였다.

- 기사에서 중요한 의미를
  가진 명사들이
  워드클라우드를
  이루고 있다.

2024

WORD CLOUD

05.27

기간별 워드클라우드

A. 100일 워드 클라우드

<전체>

언급량TOP5

'대통령': 36475,
'의원': 31415,
후보': 23683,
'민주당': 21740,
위원장': 19109

<보수>

언급량TOP5

'의원': 15302,
'대통령': 14693,
'후보': 12047,
'민주당': 10622,
'위원장': 8078

<진보>

언급량TOP5

'대통령': 21782,
'의원': 16113,
'후보': 11636,
'민주당': 11118,
'위원장': 11031

ESAA PROJECT

2024

WORD CLOUD

05.27

기간별 워드클라우드

A. 100일 워드 클라우드 불용어 처리 후

<전체>

대표
총선
정부
선거
공천
정치
대한
국회
지역
문제
위원회
정책
윤석열
북한
서울
이재명
지난
장관
정당
비판
국민
민주당
국가
생각
사진
주장
미래
경기
검찰
조사
사회
인사
상황
노란
트럼프
지지
대선
지난해
결과
요구
김건희
김건희
김건희

ESAA PROJECT

2024

WORD CLOUD

05.27

기간별 워드클라우드

B. 2024년 1월 이전 워드 클라우드

<전체>

업급량 TOP5

'대통령': 19183,
'의원': 16971,
'정부': 11042,
'민주당': 10561,
'장관': 10090

<보수>

언급량TOP5

'의원': 8074,
'대통령': 7870,
'민주당': 5113,
'장관': 4787,
'북한': 4477

<진보>

언급량TOP5

'대통령': 11313,
'의원': 8897,
'정부': 6631,
'날': 5603,
'위원장': 5508

ESAA PROJECT

2024

WORD CLOUD

05.27

기간별 워드클라우드

C. 2024년 1월 워드 클라우드

<전체>

언급량TOP5

'대통령': 12145,
'의원': 8820,
'위원장': 6684,
'날': 5280,
'민주당': 5223

<보수>

언급량TOP5

'대통령': 4980,
'의원': 4804,
'북한': 3248,
'위원장': 2943,
'민주당': 2800

<진보>

언급량TOP5

'대통령': 7165,
'의원': 4016,
'위원장': 3741,
'정부': 3074,
'날': 2812

ESAA PROJECT

2024

WORD CLOUD

05.27

기간별 워드클라우드

D. 2024년 2월 워드 클라우드(-대통령, -의원)

<전체>

언급량TOP5

'민주당': 5866,
'공천': 5763,
'대표': 5041,
'날': 4736,
'총선': 4109

<보수>

언급량TOP5

'민주당': 2998,
'공천': 2444,
'대표': 2428,
'북한': 2082,
'이재명': 2049

<진보>

언급량TOP5

'공천': 3319,
'민주당': 2868,
'날': 2728,
대표': 2613,
위원장': 2364

ESAA PROJECT

2024

WORD CLOUD

05.27

기간별 워드클라우드

E. 2024년 3월 워드 클라우드(-대통령, -의원)

<전체>

언급량TOP5

'후보': 11319,
'민주당': 8410,
'공천': 6757,
'국민의힘':6427,
'위원장': 5935

<보수>

선거
이재명
정치
후보
공천
민주당
북한
정부
윤석열
대한
총선
한국
대표
문제
논란
국민의힘
지역
위원장
경선
미국
변호사
결과
조국
진행
치단해
경우
비례대표
위원회
러시아
의혹
이후
사건
당시
지역구
민주
수사
사실
국국
대통령실
정권
요구
중국
정권
진행
기후
호주
지역구
관계자
사실
공약
한동훈
임명
장관
위원회
전략
발언
임장
무표
대사
평가
연합
의료
주장
재판
비판
정책
경제
사회
현역
관련
생각
대책
정당
비례
인사
석
중국
국가
사건
생방
정권
수사
사실
국가
당시
지역구
민주
의혹
이후
트럼프
행의
대통령실

2024

WORD CLOUD

05.27

기간별 워드클라우드

F. 2024년 4월 워드 클라우드(-대통령, -의원)

<전체>

언급량TOP5

'후보': 5837,
'선거': 2883,
'투표': 2771,
'총선': 2703,
'국민의힘': 2468

<보수>

언급량TOP5

'후보': 3322,
'선거': 1231,
'투표': 1149,
'위원장': 1096,
'민주당': 1093

<진보>

언급량TOP5

'후보': 2515,
'총선': 1706,
'선거': 1652,
'투표': 1622,
'국민의힘': 1499

ESAA PROJECT

2024

WORD CLOUD

05.27

월별 워드 클라우드

G. 차이점

보수 진영에서만 사용된 단어

1. 100일 전
['트럼프', '출마', '중국', '이후', '재판', '당시', '사실', '대선', '입장', '인사']
2. 2024년 1월 이전
['검찰', '당시', '이후', '발사', '사건', '사실', '탄핵', '입장']
3. 2024년 1월
['대만', '당시', '지역', '출마', '재판', '전대표', '사실', '공천', '국가', '더불어민주당', '서울', '검찰', '인사', '러시아', '대선']
4. 2024년 2월
['트럼프', '재판', '이후', '인사', '당시', '현역', '대선', '사실', '경기', '상황']
5. 2024년 3월
['미국', '변호사', '출마', '경기', '재판', '트럼프', '당시', '번', '중국', '이후']
6. 2024년 4월
['주장', '미국', '논란', '사실', '중국', '의사', '이후', '선', '의대', '여론조사']

진보 진영에서만 사용된 단어

1. 100일 전
['회의', '대통령실', '지난해', '사회', '결과', '뒤', '법', '수사', '조사', '지역구']
2. 2024년 1월 이전
['군사', '비판', '합의', '후보자', '추진', '뒤', '예산', '관계']
3. 2024년 1월
['김건희', '행사', '요구', '위원회', '비판', '생각', '대통령실', '거부권', '관계', '회의', '정당', '사회', '대책', '뒤', '의혹']
4. 2024년 2월
['지난해', '회의', '비판', '전략', '발표', '뒤', '대통령실', '결과', '주장', '법']
5. 2024년 3월
['수사', '비판', '회의', '처', '대통령실', '국회', '조사', '사회', '생각', '지역구']
6. 2024년 4월
['정당', '위원회', '국회', '조국혁신당', '뒤', '사회', '조사', '방송', '정권', '상황']

ESAA PROJECT

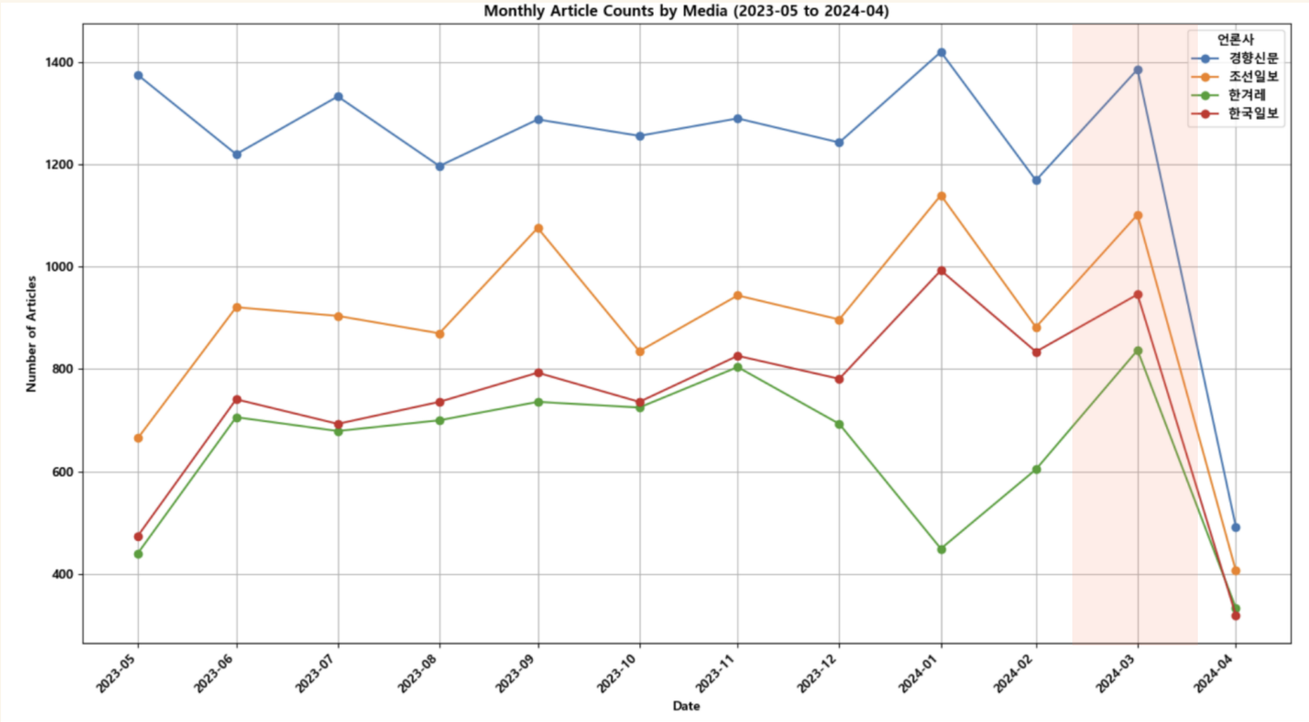
2024

ANALYSIS

05.27

시계열 기사 분석

A. 언론사별 월별 기사량 변화 추이



2024

ANALYSIS

05.27

시계열 기사 분석

B. 기사가 많았던 날 (24.03.28)

총선으로 인해 기사가 많았음을 확인 가능

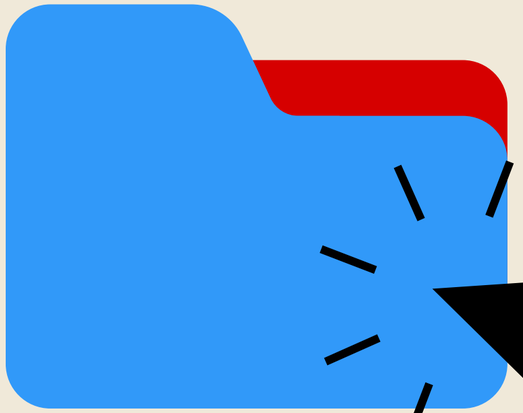
2024

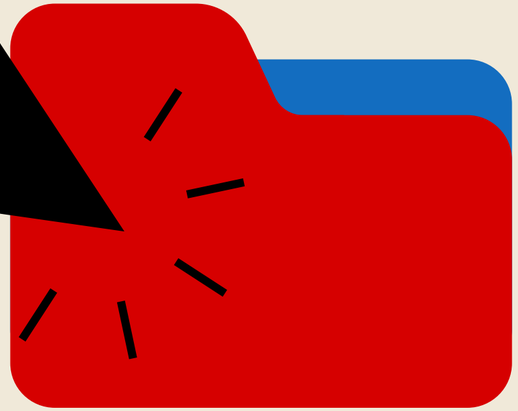
Classification

05.27

진보사설, 보수사설 그 차이는?

나이브 베이즈 분류기 통해 진보, 보수 분류 시, 가중치를 두는 단어들





방통위, 노동자, 이태원,
카르텔, 감세, R&D, 유가족
극우, 전세, 성찰
방송, 언론, 시민, 기후

조총련, 이화영,
납침, 탈북민, 김정은, 도발
한미, 한일, 한중, 사드
쌍방울, 천안함, 북송

ESAA PROJECT

2024

Topic Modeling

05.27

24.03.01~24.04.09

외교

Topic 1

대통령	23.81%
북한	22.17%
정부	
미국	
중국	
러시아	
문제	
정책	
국가	

선거의제

Topic 2

정치	21.73%
선거	14.35%
여성	
투표	
사람	
사회	
정당	
총선	
국민	
대한	

의료

Topic 3

대통령	14.29%
정부	12.61%
의료	
의대	
증원	
명	
의사	
국민	
윤석열	
대화	

정치의혹

Topic 4

대통령	40.18%
선거	50.87%
공천	
민주당	
후보	
총선	
정치	
대사	
대표	
수사	

ESAA PROJECT

2024

Topic Modeling

05.27

24.03.01~24.04.09

조수진
남성, 증오,혐오
세월호

가장 많은 비중을 차지한 '정치의혹'에 대한 진보/보수 워드 클라우드

개딸,
이화영, 박은정
천안함
이명박

2024

ANALYSIS

05.27

24.03.01 ~ 24.04.09

A. PMI 이용한 연관 분석

'Pointwise Mutual Information'

두 단어가 같이 등장할 확률과 각 단어가 개별적으로 등장할 확률의 비율을 측정

```
# PMI 계산 함수 정의
def calculate_pmi(word, word_freq_with_context, word_freq_all, total_words, total_words_with_context):
    p_word_and_context = word_freq_with_context[word] / total_words
    p_word = word_freq_all[word] / total_words
    p_context = total_words_with_context / total_words

    if p_word == 0 or p_context == 0:
        return 0
    pmi = math.log2(p_word_and_context / (p_word * p_context))
    return pmi
```

2024

ANALYSIS

05.27

24.03.01 ~ 24.04.09

B. PMI 이용한 연관 분석 ('민주당')

부정적인 사건 위주

2024

ANALYSIS

05.27

24.03.01 ~ 24.04.09

C. PMI 이용한 연관 분석 ('이재명')

부정적인 사건 위주

2024

ANALYSIS

05.27

24.03.01 ~ 24.04.09

D. PMI 이용한 연관 분석 ('윤석열')

검찰 개혁, 보수적 이념

정책, 진보적 시각

2024

ANALYSIS

05.27

24.03.01 ~ 24.04.09

E. PMI 이용한 연관 분석 ('여당')

추진 정책에 대한 비판

2024

ANALYSIS

05.27

24.03.01 ~ 24.04.09

F. PMI 이용한 연관 분석 ('야당')

2024

SIGNIFICATION

05.27

프로젝트 의의

1) 텍스트 분석을 위해 데이터 수집 기법 중 하나인 크롤링을 사용

2) 여러 형태소 분석기 사용과 단어 처리 등을 통해 전처리 단계 중요성 확인

3) 총선 전까지 정치적 사건의 신문사별 보도 경향을 텍스트 분석 통해 체계적으로 분석

- 빈도분석을 통해 기간별 정치 이슈를 워드클라우드로 생성
- 진보/보수 신문사 분류 가중치 확인을 통해 그 차이점 확인
- 토픽모델링과 관련 토픽에 대한 진보/보수 신문사 차이를 워드 클라우드로 확인
- PMI 방식을 통해 특정 키워드에 대한 연관단어를 진보/보수 신문사 별 분석