Alphco 딥러닝 9기

Text mining Project

경제신문기사 댓글분석을 통한 22대 총선 여론조사

2024 . 04 . 11

Team 행보캡

권석모, 김현규, 정호석, 최용우

for i, c in enumerate(목차) : print(i, c)

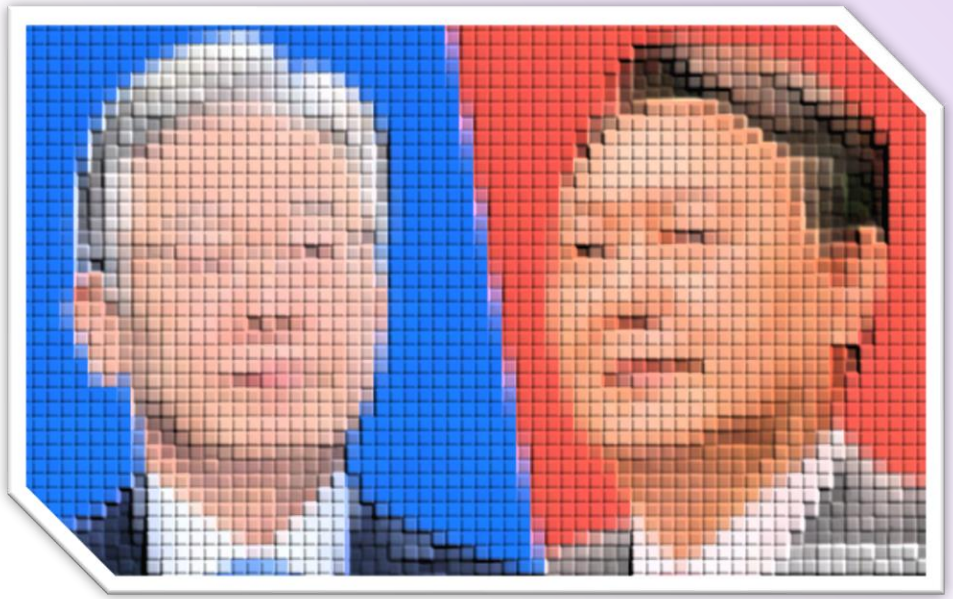
1. 개요
   - 프로젝트 목표

2. 진행 과정
   - 데이터 수집
   - 데이터 전처리
   - 시각화 및 데이터 분석

3. 프로젝트 리뷰
   - 한계점 분석
   - 자체 평가

if 프로젝트 목표:

분석 목표 “22대 총선 시즌 국민 여론의 중점 이슈 파악”



result = “분석대상”, “분석내용”

분석 대상

중도 신문사 정치 기사 댓글
- 매일경제, 머니투데이, 파이낸셜뉴스, 아시아경제

분석 내용

기간(날짜)에 따른 중점 이슈 변화
특정 이슈에 대한 언급횟수 변화 추이

print(result)

사용 예시

중점 이슈에 따른 선거 슬로건 작성 및 수정
중점 이슈 변화에 따른 적절한 선거 전략 수립

for i, c in enumerate("텍스트 전처리"): print(i, "\n", c)

1

데이터 수집

네이버 뉴스기사 크롤링 ('24.03.05 ~ '24.04.05)
- BeautifulSoup / json

호석 현규
- 매일경제
- 머니투데이

검색어 : 국민의힘/민주당

용우 석모
- 파이낸셜뉴스
- 아시아경제

정치/사회 섹션

* selenium 대비
- 링크, 스크롤/더보기 단계가 많을 수록
- 코드 복잡성↑
- but "크롤링 속도↑"

	검색어	기사날짜	기사제목	뉴스링크	댓글	댓글날짜	공감	비공감
0	국민의힘	2024.03.31	이재명 '떨어진 구두굽' 사진에...국민의미래 "안쓰럽다 못해 민망"	https://n.news.naver.com/mnews/article/009/000...	걱정마세요~~~이재명~~~교도소 가믄 새운동화 줄거~~~	2024-03-31	2867	209
1	국민의힘	2024.03.31	이재명 '떨어진 구두굽' 사진에...국민의미래 "안쓰럽다 못해 민망"	https://n.news.naver.com/mnews/article/009/000...	ㅋㅋㅋㅋ 하다하다 감성팔이 까지 하고 자빠졌네	2024-03-31	1508	54
2	국민의힘	2024.03.31	이재명 '떨어진 구두굽' 사진에...국민의미래 "안쓰럽다 못해 민망"	https://n.news.naver.com/mnews/article/009/000...	구멍난 신발 신고 후원금 구걸하고 정치 후원금 1위 찍은 코인 타짜 남국이 시즌2 ...	2024-03-31	1347	56
3	국민의힘	2024.03.31	이재명 '떨어진 구두굽' 사진에...국민의미래 "안쓰럽다 못해 민망"	https://n.news.naver.com/mnews/article/009/000...	지나가던 개도 안믿것다 ㅋㅋㅋㅋㅋㄱ	2024-03-31	682	30
4	국민의힘	2024.03.31	이재명 '떨어진 구두굽' 사진에...국민의미래 "안쓰럽다 못해 민망"	https://n.news.naver.com/mnews/article/009/000...	고인이 된 박원순 따라하기냐? 이젠 하다하다 별짓을 다하네.	2024-03-31	586	31
...	...	...	...	...	...	...	...	...
40566	국민의힘	2024.03.14	김진애 '이천수, 원희룡 몸종' 발언에...전여옥 "몸종 노릇 해봤나"	https://n.news.naver.com/mnews/article/009/000...	몸종이 아니고 내시 입니다	2024-03-14	3	1
40567	국민의힘	2024.03.14	김진애 '이천수, 원희룡 몸종' 발언에...전여옥 "몸종 노릇 해봤나"	https://n.news.naver.com/mnews/article/009/000...	[illegible]	2024-03-15	2	0

for i, c in enumerate("텍스트 전처리"): print(i, "\n", c)

2

텍스트 전처리

중점 이슈 분석을 위해 명사 추출 및 전처리 진행
- konlpy (Okt, Mecab)

for j, v in enumerate(c)

1단계

특정 정당을 지칭하는 고유명사 단어집 작성

```
1 right = ["국민의힘", "국힘", "국민의짐", "국짐", "국짐당", "빨간", "빨간색", "보수", "우파", "한동훈", "윤석렬", "윤씨", "윤가", "닭근혜", "윤재앙"]
2 left = ["민주당", "민주", "더불어공산당", "소나무", "소나무당", "파란", "파란색", "진보", "좌파", "이재명", "문재인", "문씨", "문가", "문재앙", "찢재명"]
```

2단계

특정 정당을 언급한 댓글 분류
- 정당이름 또는 지칭어가 복합명사 또는 신조어(비속어)가 많다는 문제 발생

해결방법1) 고유명사 단어집에 포함된 댓글에 별도 태깅 후 원본에서 삭제

댓글	댓글분석	좌	우
걱정마세요 교도소 가믄 새운동화 줄겨	[걱정, 교도소, 운동화, 줄겨]	Y	NaN

삭제 ← 태깅

해결방법2) 형태소추출기에 고유명사로 등록 후 명사 추출 후 원본에서 삭제

```
1  국힘,,,,NNP,*,F,국힘,*,*,*,*,*\n
2  국민의힘,,,,NNP,*,F,국민의힘,*,*,*,*,*\n
3  국짐,,,,NNP,*,F,국짐,*,*,*,*,*\n
```

```
PS C:₩mecab> .₩tools₩add-userdic-win.ps1
>>
```

for i, c in enumerate("텍스트 전처리"): print(i, "\n", c)

2

텍스트 전처리

중점 이슈 분석을 위해 명사 추출 및 전처리 진행
- konlpy (Okt, Mecab)

for j, v in enumerate(c)

3단계

불용어, 영어, 특수문자 등 제거
- 정규표현식(regex) / StopWord 단어집 사용

```
1 # 공백 제거 및 한글만 남기기
2 df["댓글"] = df["댓글"].str.replace("[^가-힣.1-9]+", " ", regex = True)
```

4단계

토큰화
- 전체/성향별/공감수 등 다양한 기준별 명사 추출 (정당 지칭 단어 제외)
- Okt, Mecab 명사 추출에서는 유의미한 차이를 발견하지 못함

댓글	댓글분석
걱정마세요 교도소 가믄 새운동화 줄겨	[걱정, 교도소, 운동화, 줄겨]
하다하다 감성팔이 까지 하고 자빠졌네	[감성, 팔이]
구멍난 신발 신고 후원금 구걸하고 정치 후원금 1위 찍은 코인 타짜 남국이 시즌2 ...	[구멍, 신발, 신고, 후, 원금, 구걸, 정치, 후, 원금, 위, 코인, 타짜, ...
지나가던 개도 안믿겄다	[개도]
고인이 된 박원순 따라하기냐 이젠 하다하다 별짓을 다하네.	[고인, 박원순, 젠, 짓]

for n, v in enumerate("데이터시각화".items()): print(n, ":", v)

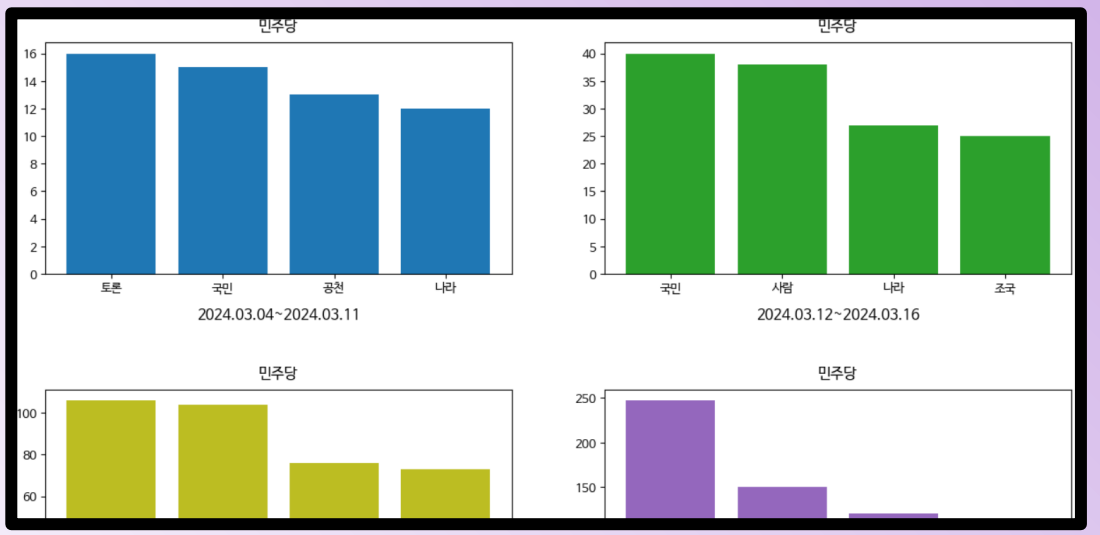
3

데이터 시각화

워드클라우드 및 그래프를 통한 데이터 시각화
- WordCloud / matplotlib.pyplot

호석 용우	국민의힘/민주당/중도 구분 별 날짜별 단어 빈도 변화 시각화 클러스터링 모델 적용	현규 석모	공감/비공감 구분 별 단어 빈도 시각화 정치 댓글 추출 방법론 구상

print(시각화.head(1))



for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

4

데이터 분석

정량적, 정성적 분석을 통한 데이터 분석
- 다양한 기준별 중점 이슈 변화 파악
- IDF 수치를 통한 이슈 분포 파악

warnings.filterwarnings("ignore")

한계 사항

1. 특정 정당을 지칭하였으나 문맥의 긍정/부정을 구분하기 어려움
2. 둘 모두 언급한 경우 주위 단어들이 어느 정당과 관련되었는지 파악하기 어려움
3. 주어가 빠진 댓글은 파악이 어려움

→ 데이터에 대한 적절한 태깅작업이 필요하지만 리소스 부족으로 수행 불가

분석 기준 수립

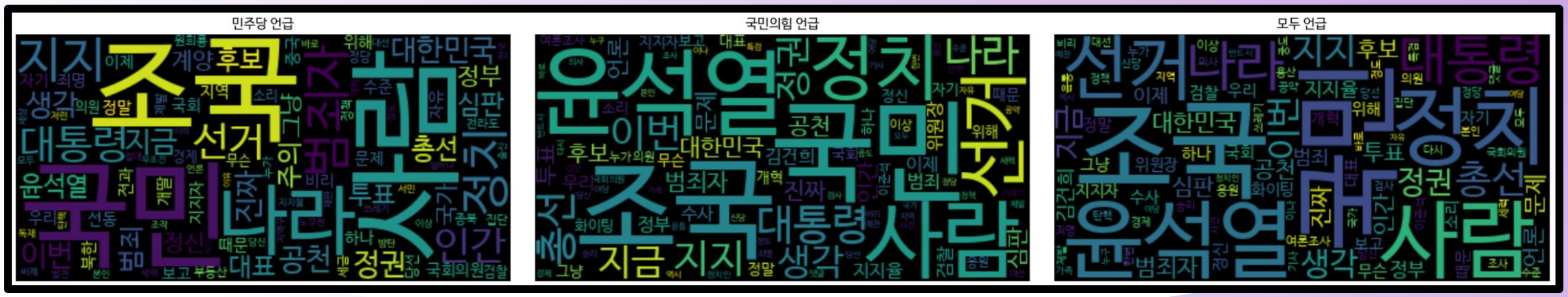
1. 긍/부정에 관계 없이 특정 정당과 함께 많이 언급된 중점 이슈를 파악
2. 중점 이슈에 대한 다각도의 분석을 통해 선거 전략에 도움이 될 수 있는 인사이트 제공
3. 주어가 빠진 댓글 제외 (주체가 명확한 댓글만 분석)

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

4 - 1

1차 분석 데이터 정제 전

plt.show()



- 민주당 : 국민, 조국, 나라, 대통령, 공천, 대한민국, 지지, 사람, 선거, 총선 등
- 국민의힘 : 윤석열, 조국, 국민, 정치, 사람, 선거, 나라, 대통령, 정권 등
- 모두 언급 : 국민, 조국, 윤석열, 정치, 사람, 정권, 총선, 대통령, 나라, 지지, 선거 등

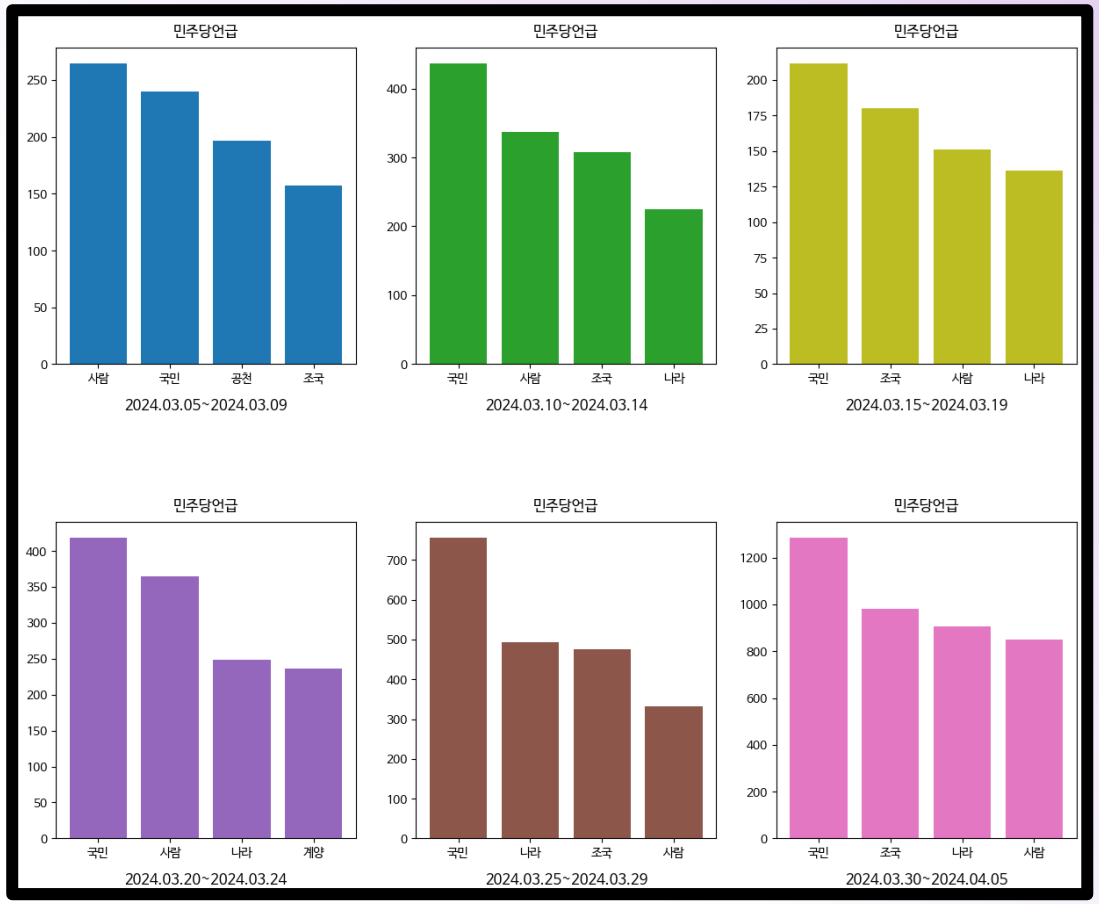
기간 내 상위 언급 단어 빈도 수

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

4 - 1

1차 분석 데이터 정제 전

plt.show() if name == "민주당언급"



기간별 중점이슈 사람, 국민, 공천, 나라, 조국, 계양

문제점

(국민에게는 중요한 단어들이지만..)
분석 주체 입장에서 크게 의미 없는 단어 다수 존재

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

4 - 1

1차 분석 IDF 수치 분석을 통한 해당 단어들의 분포도 파악

print(sorted(idf, key = lambda x: x[1])

```
[('국민', 2.066282672443657 4),
 ('조국', 2.313812801033323 7),
 ('사람', 2.414642027392787 8),
 ('나라', 2.461844374923316 5),
 ('지지', 2.833173327249151 7),
 ('범죄자', 2.910216004370467 7),
 ('정치', 2.913896538521593),
 ('선거', 2.939289866103022),
 ('이번', 2.953766429897088 5),
 ('대한민국', 2.956840990351334 4),
 ('대통령', 2.983363605860313 2),
 ('인간', 2.99369345536189),
 ('생각', 3.007364757959856),
 ('진짜', 3.079543729173178),
 ('지금', 3.107789207952733),
 ('정권', 3.119495780184893 3),
 ('총선', 3.171552142140946 3),
 ('대표', 3.202562378883506 7),
 ('범죄', 3.269729022427347 4),
 ('윤석열', 3.288858192336342 7),
 ('공천', 3.307266941746412 7),
 ('심판', 3.311648109761975 4),
 ('주의', 3.318256063322007 5),
 ('후보', 3.320468453604948 2),
 ('투표', 3.393980300554727 8),
 ('그냥', 3.464336266476623),
 ('정신', 3.544165868129107 3),
 ('정부', 3.551115166422313 5),
 ('하나', 3.605569879463238),
 ('국가', 3.646146445095281),
 ('때문', 3.650758290317843 7),
 ('국회의원', 3.650758290317843 7),
 ('수준', 3.663161550101263 8),
```

IDF

특정 단어가 전체 문서 중 몇 개의 문서에서 나타나는지에 대한 수치
- 낮을 수록 흔하게 사용되는 단어로 해석 가능
- 다만 중점 이슈 또한 흔하게 언급되는 경향이 있기 때문에 참고자료로만 활용

1. 크게 의미 없어보이는 단어들이 IDF도 낮을 경우 중점 이슈와는 거리가 멀 가능성이 높음
2. 단어(이슈)의 IDF가 높다면 특정 기간에 집중된 이슈일 가능성이 높음

※ 정성적 분석을 중심으로 하되, 정량적 데이터를 참고자료로 활용

→ 의미 없는 단어들을 삭제하고 2차 분석 진행

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

4 - 2

2차 분석 의미 없는 단어 삭제 후 분석

plt.show("한달 간 중점이슈")

민주당 이슈

- 대표 리스크 (범죄자, 범죄, 법카, 전과, 비리 등)
- 친중, 친북 성향에 대한 리스크 (중국, 북한 등)
- 계양 선거구 관심사 (계양 등)

국민의힘 이슈

- 대통령 리스크 (윤석열, 김건희, 범죄)
- 검찰 리스크 (검찰, 언론, 수사 등)
- 정권 심판론 리스크(심판)

특이사항

- 조국 신당 관련 이슈 상위
- 각 정당 모두 대표 및 배우자 리스크 다수

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

4 - 2

2차 분석 의미 없는 단어 삭제 후 분석

plt.show("공감/비공감이 많은 중점 이슈")

국민의힘 관련

공감 : 수사, 가족, 개혁

비공감 : 정치, 선거, 의사

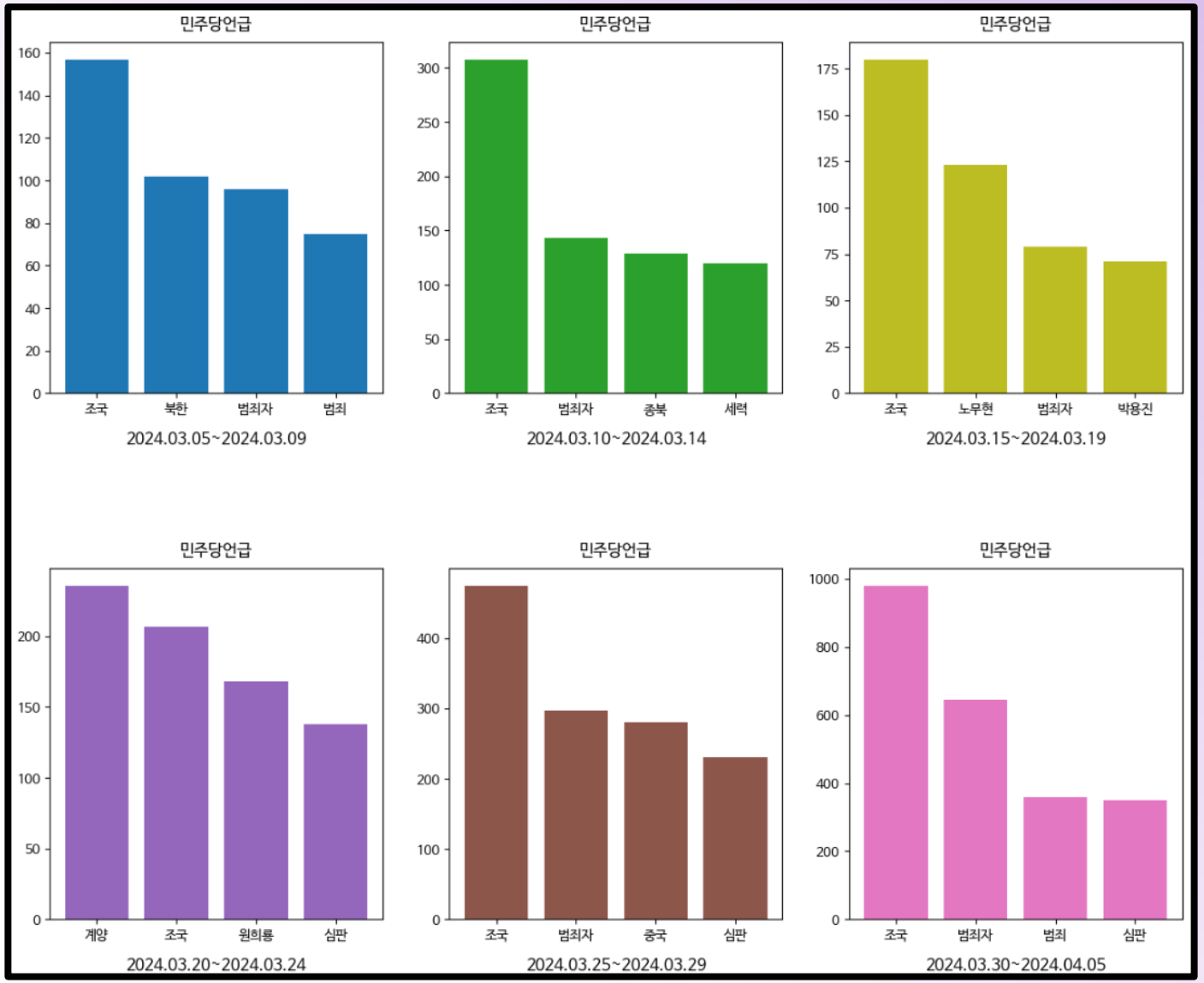
민주당 관련

공감 : 수사, 검찰, 개혁, 승리

비공감 : 선거, 대표, 후보

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

plt.show("기간 별 중점 이슈 분석") if name == "민주당"



전반적 이슈

- 조국, 범죄, 범죄자

→ 조국 신당 이슈 지속
→ 대표자 리스크가 지속되고 있음

기간 이슈

- 박용진, 원희룡

→ 특정기간, 특정 인물들에 대한 이슈 발생

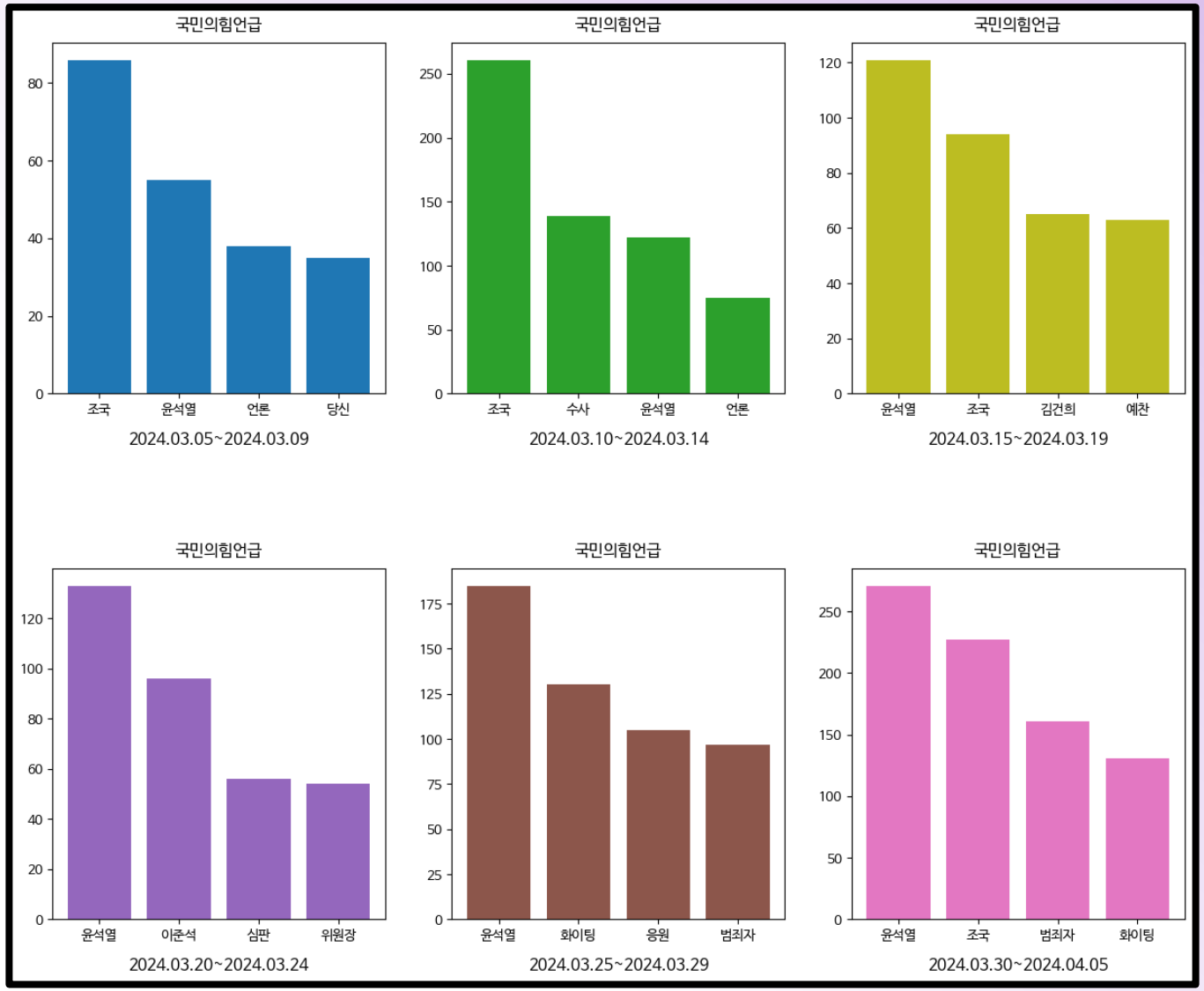
특이사항

- 심판

→ 총선이 다가올 수록 정권 심판에 대한 언급이 증가

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

plt.show("기간 별 중점 이슈 분석") if name == "국민의힘"



전반적 이슈

- 조국, 윤석열, 김건희

→ 조국 신당 이슈 지속
→ 대표자 리스크가 지속되고 있음

기간 이슈

- 이준석

→ 특정기간, 특정 인물에 대한 이슈 발생

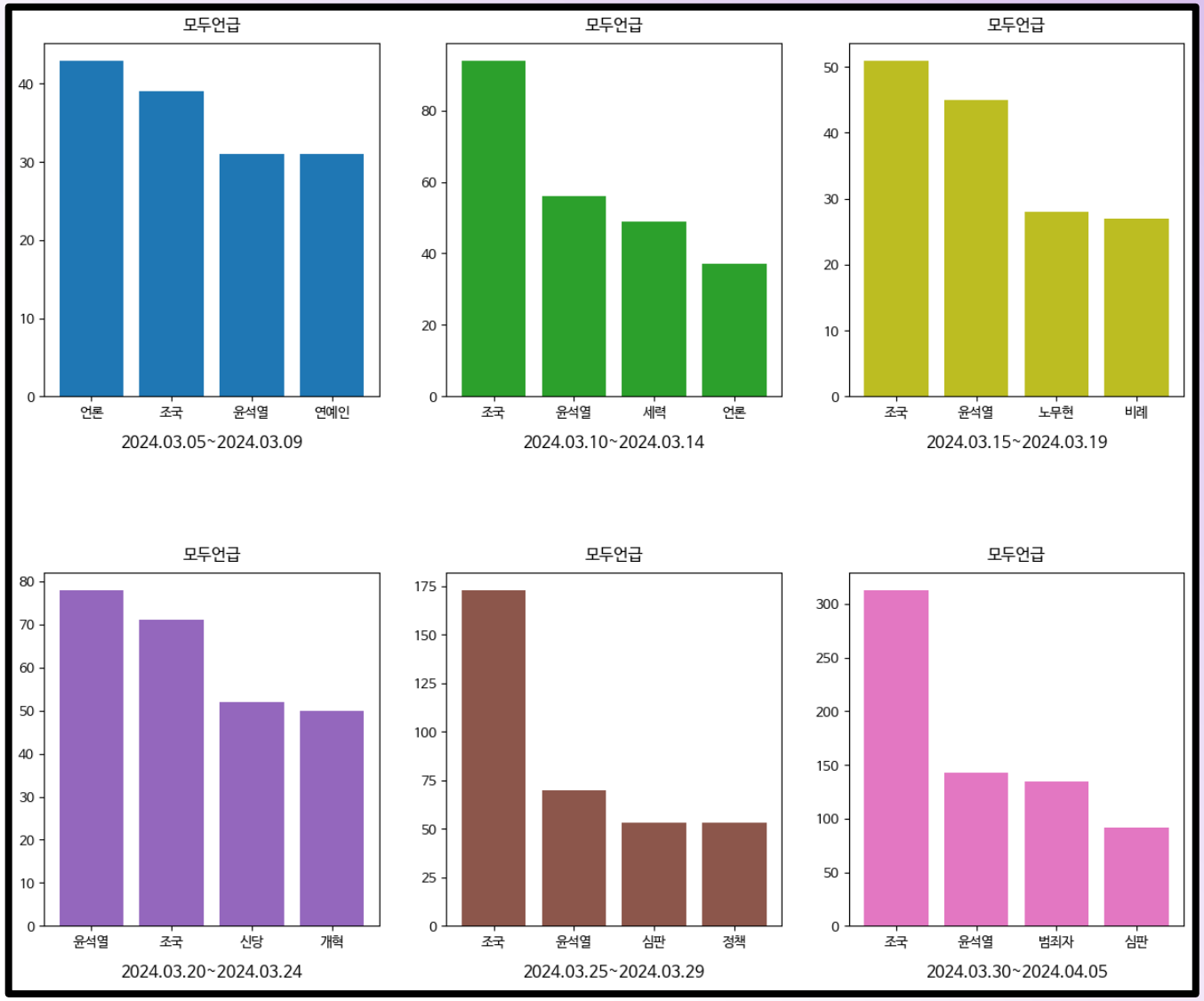
특이사항

- 심판, 화이팅

→ 총선이 다가올 수록 정권 심판과 응원이 동시에 증가추세

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

plt.show("기간 별 중점 이슈 분석") if name == "모두 언급"



전반적 이슈

- 조국, 윤석열

→ 진보는 민주당보다 조국 관련 언급이 많음
→ 보수는 윤석열 언급이 가장 많음

기간 이슈

- 연예인

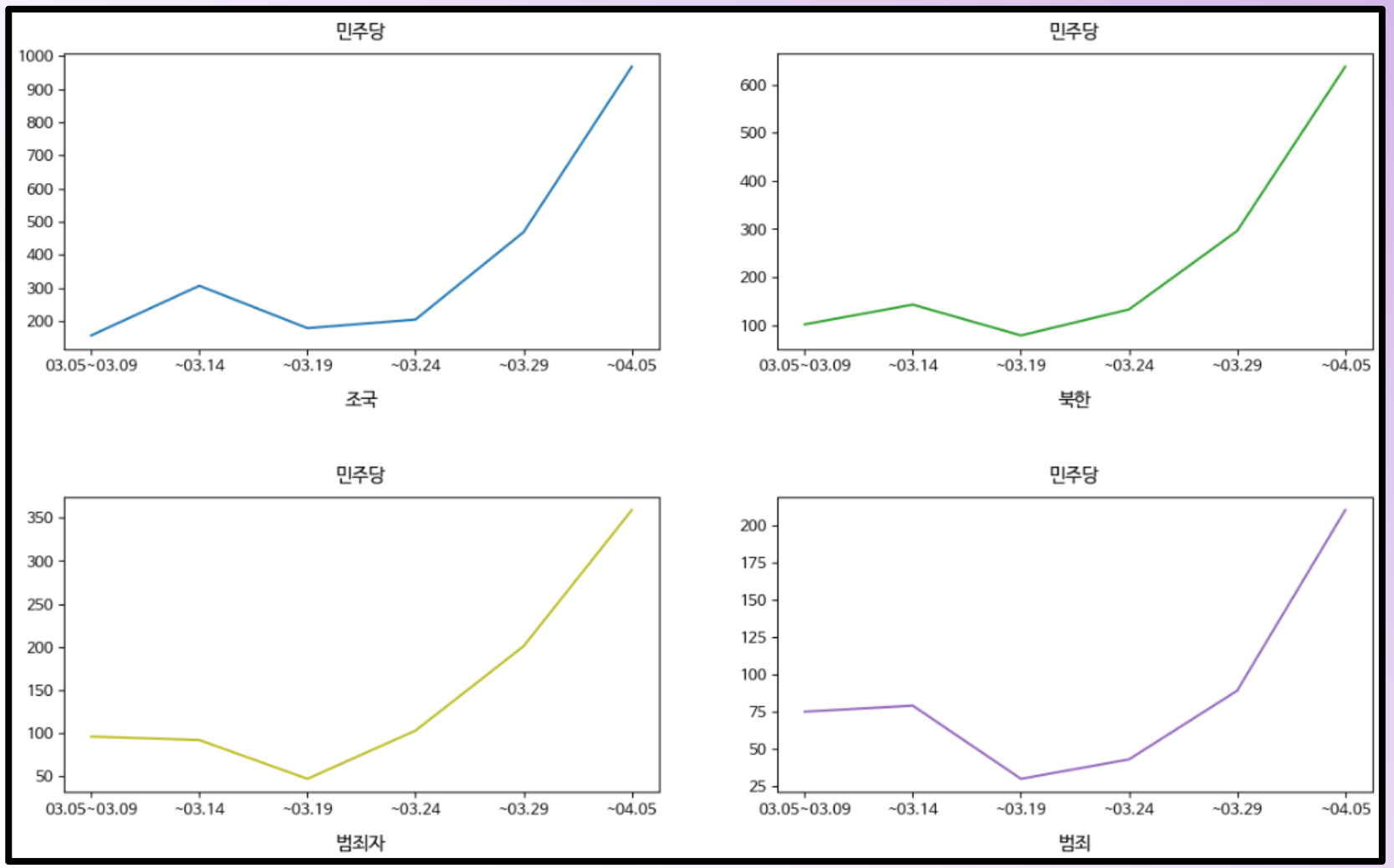
→ 기간 내 특정 연예인 파동 이슈가 정치적 이슈로 인식

가장 큰 이슈는 조국신당, 그 외에는 각 대표 리스크가 가장 크다는 것을 알 수 있음.

→ 선거기간임에도 정책보다 특정 인물들에 대한 관심도가 높음

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

plt.show("상위 단어 빈도 추이 변화") if name == "민주당"



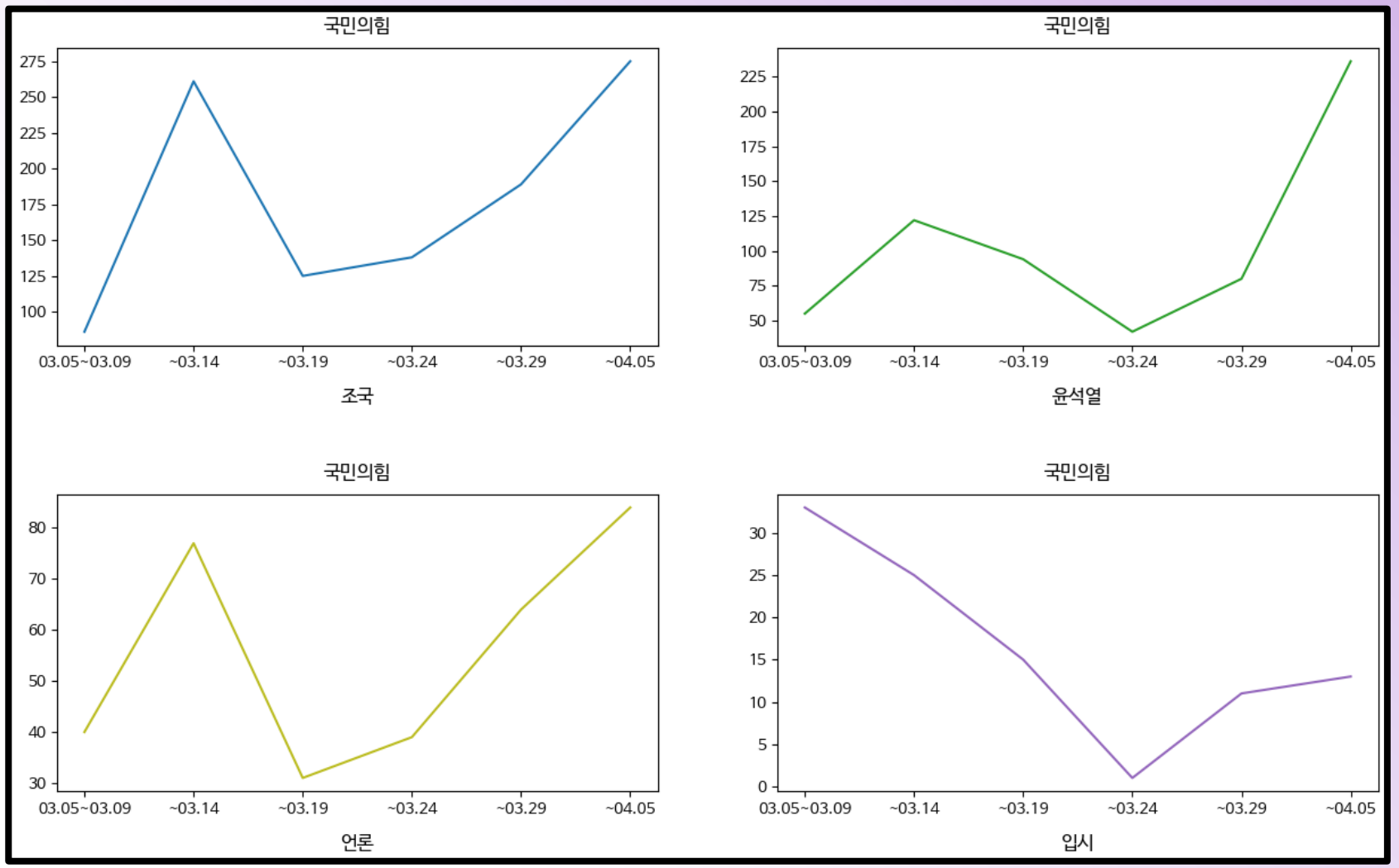
이슈 사항

총선이 다가올 수록 조국에 대한 언급 증가

부정적인 이슈 언급과 증가율이 높음

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

plt.show("상위 단어 빈도 추이 변화") if name == "국민의힘"



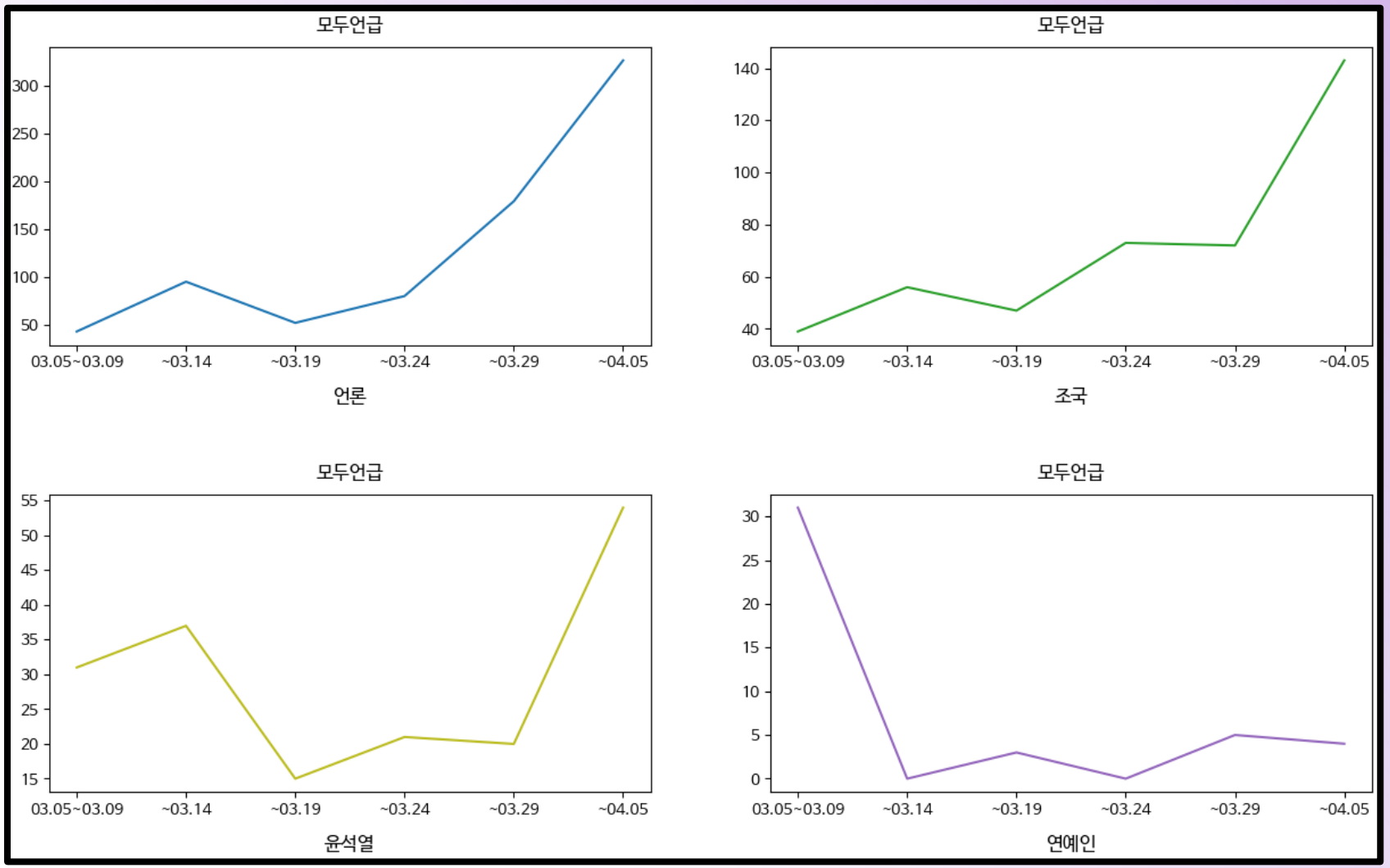
이슈 사항

총선이 다가올 수록 조국에 대한 언급 증가
언급빈도는 민주당에 비해 낮음

언론, 입시 등 정책 이슈 존재
총선이 다가올 수록 언론 중립성에 대한
언급 증가

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

plt.show("상위 단어 빈도 추이 변화") if name == "모두 언급"



이슈 사항

두 정당을 모두를 지칭한 댓글에서 언론에 대한 언급이 증가하는 것으로 보아 총선이 다가올 수록 언론의 중립성에 대해 의문을 표하는 여론이 높아짐을 유추할 수 있음

연예인 이슈는 생명주기가 그리 길지 않아 일시적임을 알 수 있음

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

for "정치댓글" in "타 섹터기사"

정치 섹터 외 다른 섹터에 있는 정치 댓글 분류	정치 전문 섹터의 뉴스기사와는 다른 이슈가 언급될 가능성이 존재

for "방법" in "비지도학습 클러스터링"

코사인 유사도	정치기사 댓글과 타 섹터 기사 댓글과의 코사인 유사도의 평균을 측정하여 유사값의 기준점을 구해 정치 댓글 추출 진행
K-meas 클러스터링	각 섹터 댓글들을 모아 클러스터링 진행

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

for word in "댓글" if "코사인유사도" ⟩ 기준수치

정치 섹터 정치 댓글
댓글1…
댓글2…
댓글3…
샘플 5,000개

정치기사댓글 5,000개와
타 섹터 댓글 1개 간의
코사인 유사도 평균을 모두 구함

타 섹터 모든 댓글	코사인유사도 평균
정치기사 댓글…	0.00q…
IT기사 댓글…	0.00z…
사회기사 댓글…	0.00x…
5개 섹터 별 샘플 100개씩	

10885	기사는 제대로 읽고 혐오성댓글들 다는거냐 연체율 평균 프로정도 올랐다잖아 고작 그거가지고 젊은애들 소비를 다 싸잡아서 욕하는 난독증 환자들 왜이렇게 많냐 남 신경끄고 너네나 잘살아 정부에서 뭘 지원하주겠다는 기사도 아니고 통계기사 보고도 혐오질이냐	경제	0.005597
44228	무능무식이라는 병이 사회곳곳에 번져서 큰일이다 가뜩이나 철밥통 깡통들 대가리 안돌리는데 급도 안되는 대통에 정부를 만났으니 년만에 국가경쟁력이 북한보다 못하단다 이것들아	IT	0.005556
10863	월급쟁이들 욕먹어가매 월급 원천떼인 세금으로 탈세 하는 것들 빚갚는다고 세금쓴다하면 그 정권 뽑지말고 직장인들 다 같이 세금 안 내야 한다	경제	0.005300
45181	가격 살벌한거 봐라 대체 얼마나 해처먹은거지 기업도 기업인데 정책만든 놈은 어째 책임지는 놈이 단 한놈도 없냐 결과적으로 사람들 통신비는 빨아먹힐대로 빨아먹히고 기술개발같은 거 존도없이 끝났으면 누구라도 나와서 뭔 말이라도 해야되는 거 아님	IT	0.005239

코사인 유사도가 큰 순으로 정렬하여 정치 댓글로 분류할 수 있는 임계값 추출 == 0.0052

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

print "타섹터 정치댓글" if "코사인유사도" ⟩ 0.0052

	댓글	코사인유사도	분류
	역쉬 권의주의 정권에서나 가능한 창의적 해법	0.021858	세계
	도라이 천지삐까리구마	0.017537	사회
	코스피 삼스피 코스닥 코스라 로 바뀌 밸류 지랄하네 야	0.016858	경제
	러시아 답다	0.016287	세계
	아직 두덕이 살아있는 갚네	0.014920	사회
	차칸 선생님이시내옇	0.014865	세계
	무슨 행정소송 아주 특권층 젤 꼭대기에 있으신분들인가 점점 더 신물이 난다	0.014694	사회
	사지마요 평생 무주택 추천	0.014311	경제
	억 만명 중에 명 걸린걸 무서워하라는거지 지금	0.013831	세계

Error: 정확도 낮음!

원인 분석

- 댓글의 특성 상 문맥이 아닌 단어만으로 유사도를 판단하기 어려움 (모든 섹터 간 겹치는 표현이 많음)

- 특정 주제를 모아둔 섹터 내에서는 TF-IDF를 통한 가중치 적용도 어려움 (주제에 관련된 주요 단어까지 패널티를 받을 확률이 높음)

→ 더 높은 수준의 텍스트 전처리가 필요

→ 문맥까지 파악할 수 있는 모델링 필요

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

print "5개 섹터 정치댓글" if "k-means 클러스터링" == 1

	댓글	분류	그룹
7000	인형같네 근데 이렇게 외모에 집착이 강한 사람은 나이들면 또하게 됨	세계	1
7001	이런 뉴스에는 찍 댓글부대들은 입 꾹 닫음	세계	1
7002	우리는 같은번호로 개씩사서 당첨되는데	세계	1
7003	오타니라면 그랬을거같다	세계	1
7004	민주당 만행을 늘어놓으면 이사건은 그닥 뭐	세계	1
7005	인체의 신비전 왜 안하냐 다시 보고싶다	세계	1
7006	용와대 지령이 와 내 편은 절대 부패하지도 않고 틀리지도 않는다 내 가족은 절대 잘못을 저지르지도 않고 원도 횡령하지 안했다	세계	1
7007	동물단체는 뭐하냐 동물원에 갇혀 사는 동물의 존엄성을 생각한다면 동물원이라는 곳을 다 없애야 한다 자연에 살아야 할 동물을 동	세	1

그룹	댓글	분류
0	192	192
1	12303	12303
2	409	409
3	357	357
4	232	232

Error: 정확도 낮음!

원인 분석

- 각 카테고리 간 유사한 단어 사용이 많음
- 데이터 특성에 맞는 텍스트 전처리 미흡

→ 각 카테고리 간 데이터 수가 맞다면 TF-IDF를 이용한 벡터화를 시도해볼 수 있음

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

print "타섹터 정치댓글" if "k-means 클러스터링" == 1 and "벡터" = "TFIDFVectorizer"

그룹	index	댓글	분류	명사
0	10379	10379	10379	10379
1	160	160	160	160
2	446	446	446	446
3	164	164	164	164
4	344	344	344	344

	댓글	분류	그룹
6500	인형같네 근데 이렇게 외모에 집착이 강한 사람은 나이들면 또하게 됨	세계	0
6501	이런 뉴스에는 찍 댓글부대들은 입 꾹 닫음	세계	0
6502	우리는 같은번호로 개씩사서 당첨되는데	세계	0
6503	오타니라면 그랬을거같다	세계	0
6505	인체의 신비전 왜 안하냐 다시 보고싶다	세계	0
...	...	...	...
11488	보험료 납부가 상속인이냐 피상속인이냐의 기준이 아니라 상속보험금이 고액이냐 아니냐	사회	0
11489	국힘지지하는것들은 하나같이 정신세계가 그냥 인간 쓰레기들이네	사회	0
11490	기자들이 쓸대없는 기사 쓰지 않으면 된다	사회	0
11491	까불지말고복귀하라그래 안하면뒈지는수가나온다	사회	0
11492	광주지역 주민들도 올바른 생강이 있는 분들이 있을 것으로 생각됩니다	사회	0

Error: 정확도 낮음!

결과 약간 분류가 더 잘 되긴 했으나 유의미한 차이를 가지지는 못함.

결론 각 카테고리 특성만 담을 수 있는 전처리 또는, 수기 태깅 작업과 딥러닝 모델 필요

for n, v in enumerate("데이터 분석".items()): print(n, ":", v)

n, v for "제목으로 섹터분류".items()

	기사 제목	분류
0	한동훈, 박민식에 서울 강서을 출마 요청..."승리 위해 헌신해 달라"[2024 총선]	정치
1	한동훈 "함께 정치 하고 싶다"...김영주 "늦지 않게 답하겠다"[2024 총선]	정치
2	민주, 안태준 경기 광주을·장종태 대전 서구갑 '경선 승리'[2024 총선]	정치
3	[속보]민주, 경기 광주을 안태준·대전 서구갑 장종태 경선 승리	정치
4	[2024 총선]민주당, 전북 '경선 대진표' 확정...김윤덕·이원택·한병도 단수공천	정치
...	...	...
1870	"스미싱 여부? 카톡서 확인해보세요.. 공공기관 사칭 급증"	IT
1871	"美, 나토 정상희의에 日기시다 초대 조율 중"	세계
1872	중일, 후쿠시마 오염수 관련 전문가 협의	세계
1873	갤럭시 S24 울트라, 컨슈머리포트 '최고의 폰 카메라' 1위	IT
1874	비트코인 8000개 실수로 버렸던 그 남자..11년만에 전해진 근황	세계

뉴스 제목으로 섹터 구분하기

- 로지스틱스 회귀 모델을 사용해 기사 제목으로 섹터 구분
- 표준어 사용과 카테고리 별 특색있는 표현 사용으로 예측 정확성 74.2%
- 정치, 사회, 세계 카테고리 간 중복 표현이 많아 정확도 감소

→ 기사 내용까지 포함한 학습 필요

TF-IDF Logistic Regression 의 예측 정확도는 0.742

for I, e in enumerate("자체평가")

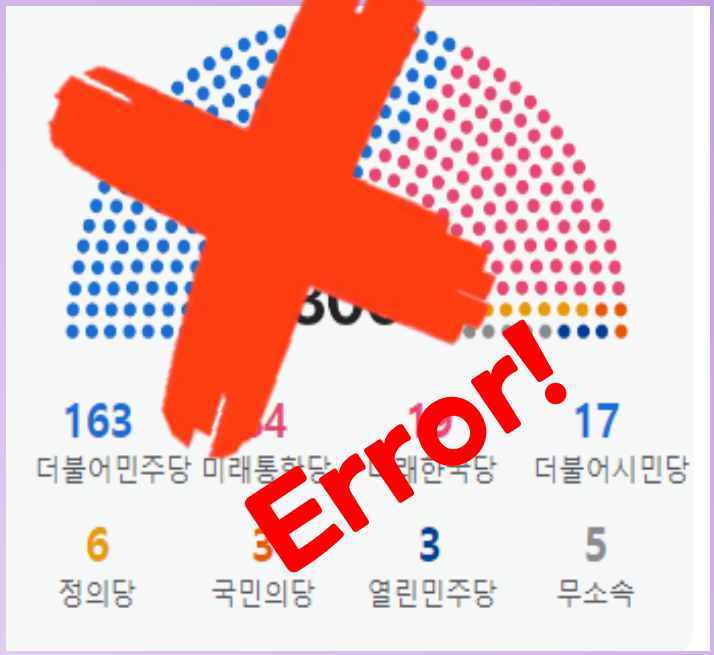
초기 최종 목표

- 댓글의 감성분석을 통해 정당에 대한 선호도 분석
- 공감, 조회수 등 가중치를 반영한 선호도 비율로 총선 결과 예측

실패 이유

- 낮은 표준어 사용율, 은어/신조어/비속어/줄임말 등의 다량 사용으로 인한 텍스트 정규화의 어려움
- 정규화되지 않은 텍스트 데이터로 감성사전 구축 불가

→ 데이터 특성을 고려하지 않고 불가능한 목표를 정한 것이 가장 큰 실패 요인



가장 먼저 "데이터 자체의 특성"을 먼저 고려해 목표의 기술적 실현 가능성을 판단하고
프로젝트 방향성과 방법론을 설계하는 것이 가장 중요!!