2023 DATA AI 분석 경진대회

모델 개발 매뉴얼

팀명 - D.P.

문제 유형 - 사회현안 문제형

선택 문제 - 뉴스와 소셜데이터 기반 이슈 분석 및 시각화

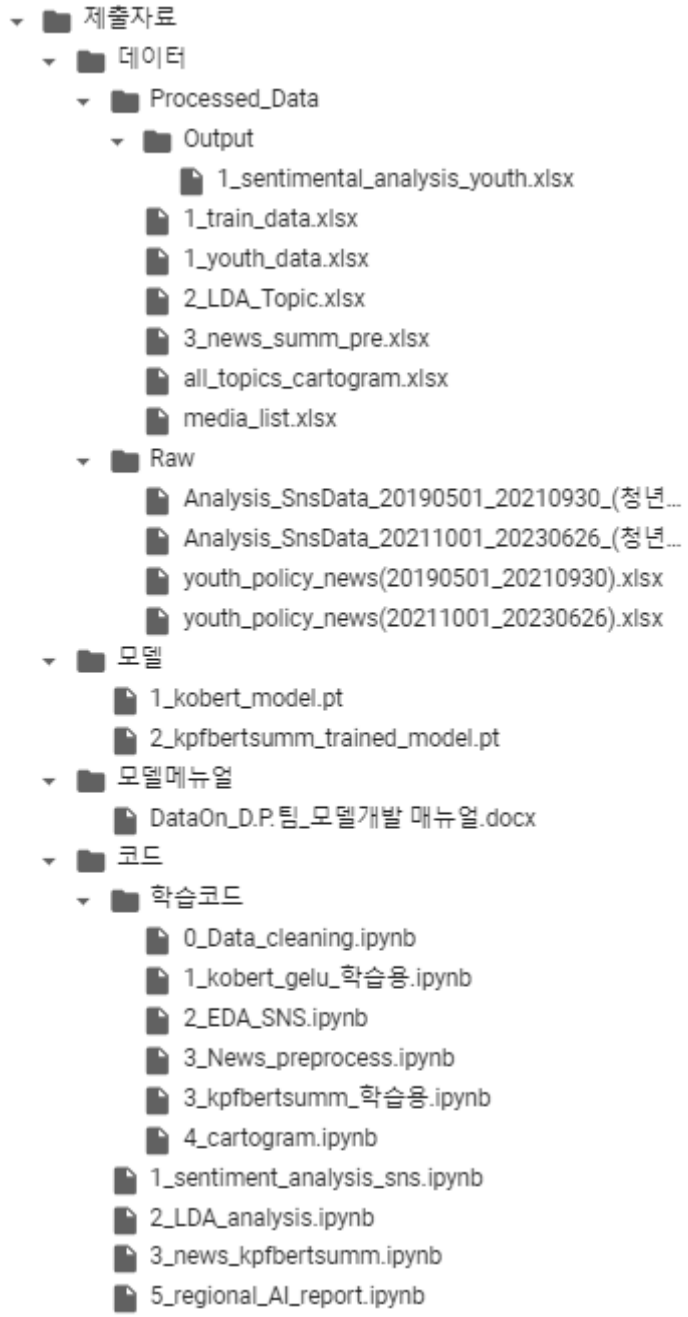
목차

Ⅰ. 학습·배포를 위한 SW 및 HW

- **프레임워크 및 라이브러리:**
  - koBERT, kpfBERTsumm: TensorFlow, PyTorch, Transformers, koBERTTokenizer, tqdm, kpfbert, kss
  - LDA: pyLDAvis, gensim, OpenAI API GPT4

    그 외 인공지능을 위한기본적인 패키지 (소스 내 import 참조)
- **GPU:** Google Colab (A100 GPU, 고용량 RAM) 환경에서 작성되었음
- **배포:** Tableau Desktop (version 2023.2.2)

Ⅱ. 파일 저장 구조 - 가공/처리된 데이터셋과 학습 모델 포함



III. 모델 실행 방법 - 모델 입력/출력 정보 포함

1. 감정분석 (koBERT)

- **모델 소개** : koBERT는 한국어에 특성화된 BERT 모델이다. 데이터 전처리 후, 약 5000개의 감성분석 레이블링이 완료된 Train (“1_rent_data.xlsx”), Test 데이터로 모델을 (“1_kobert_model.pt”) 학습했다.
- **실행 파일명** : 1_sentiment_analysis_sns.ipynb
- **모델 실행 방법** :

  1. 전처리된 SNS 데이터셋(“/데이터/Processed_Data/1_rent_data.xlsx”)을 준비한다.
  2. 감성분석 코드를 실행한다. (“/코드/1_sns_sentiment_analysis.ipynb”)

- **모델링 불러오는 법**

```
!pip install transformers
```

```
import pandas as pd
import torch
import transformers
from transformers import BertTokenizer, BertForSequenceClassification
from torch.utils.data import DataLoader, Dataset
```

```
# 모델 로드
model_path = '# 모델 파일의 경로/1_kobert_model.pt'

# Torch GPU 설정
device_type = 'cuda' if torch.cuda.is_available() else 'cpu'
device = torch.device(device_type)
```

```
# 모델 아키텍처를 생성
model = BertForSequenceClassification.from_pretrained("monologg/kobert", num_labels=2)  # num_labels는 분류 클래스 수에 따라 설정
model.to(device)
model.eval()
```

```
# 토크나이저 로드
tokenizer = BertTokenizer.from_pretrained("monologg/kobert")
```

- **모델링 함수**

```
def perform_sentiment_analysis(text):
    encoded_text = tokenizer(text, padding=True, truncation=True, return_tensors='pt')
    encoded_text = {k: v.to(device) for k, v in encoded_text.items()}
    with torch.no_grad():
        output = model(**encoded_text)
    logits = output.logits
    predicted_class = torch.argmax(logits, dim=1).item()
    return predicted_class
```

- **적용 예시**

- **입력 문장** : 'Summary' 컬럼에 대한 감성 분석 수행 및 결과 저장

```
df['감성분석'] = df['Summary'].apply(perform_sentiment_analysis)
```

- **결과값**("output/1_sentimental_analysis.xlsx"):

	Date	User	Contents	Type	Summary	감성분석	Topic
0	2019/05/01	gyeongsangbukdo	#공예업체 _취업을 원하는 ₩n경북의 취준생들에게 꿀 정보?? ₩n#경상북도 #청년정...	트위터	공예 업체 취업 경북 취 준 정보 경상북도 청년 정책 일자리 정책 취업 지원 사업	0	청년정책
1	2019/05/01	dudtn1688	안녕하세요! ₩n처음 여러분들에게 인사를 드리네요. ₩n이번 '2019 NCS 청년...	블로그	청년 서포터 즈 청년 서포터 즈 발대식 을 다녀오 다 안...	0	청년정책
2	2019/05/01	gabriela_ran	₩n ₩n(서로의 동거인이 된 하요와 함께 찍은 사진) ₩n ₩n_ ₩n ₩n2월...	블로그	감사 한 나날 서로 동거인 하 요 사진 일자 퇴사 완주 내 완주 생각 완주 달 몽골...	0	청년정책
3	2019/05/01	giguc31	안녕하세요! 청춘 너나들이 서포터즈 권성빈입니다. ₩n청춘 너나들이에서 청년들의 취...	블로그	청년 정책 멘토링 내 가 도움 받 을 수 있 는 정책 은 무엇 일까 안녕...	0	청년정책
4	2019/05/01	sisa2007	₩n유승희 국회의원, 청년문제 돌파구 찾는다! ₩n고려대서 정책 토크콘서트 "기본...	블로그	유승희 국회의원 청년 문제 돌파구 찾 는다 유승희 국회의원 청년 ...	1	청년정책

2. LDA Topic Modeling

- **데이터 전처리:** "2_EDA.ipynb"을 활용하여 기초 통계 분석 실행 및 변곡점*을 추출한다.
  *변곡점의 정의는 1) 최고 언급량 시점, 2) 여론의 긍/부정 변화가 있었던 시기로 주요 시점(일)이 포함된 월(month)의 SNS와 뉴스 데이터를 추출한다.
- **실행 파일명**: 2_EDA_SNS.ipynb, 2_LDA_analysis.ipynb
- **모델 실행 방법**:

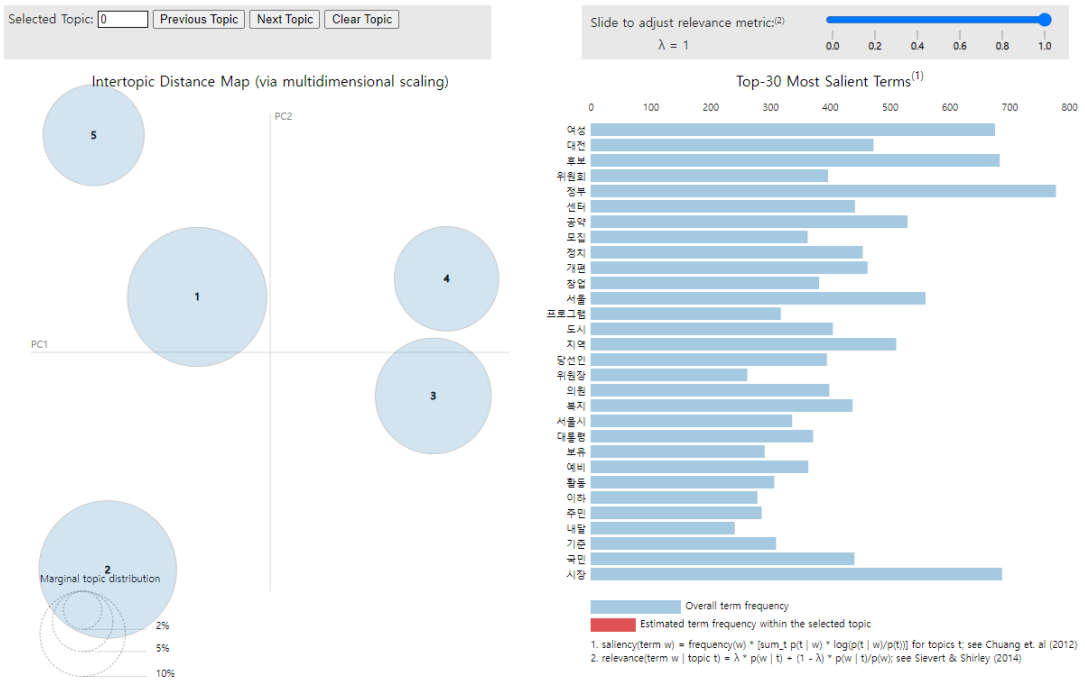
  1. 감성분석이 완료된 데이터를 준비한다.
  ("/데이터/Processed_Data/2_LDA_Topic.xlsx")
  2. "/코드/학습코드/2_LDA_analysis.ipynb"를 활용하여 다음과 같이 토픽 모델링을 실행한다.

```
import gensim
from gensim import corpora
import pandas as pd  # 필요한 패키지를 import

lda_model = gensim.models.ldamodel.LdaModel(corpus=corpus,
                                            id2word=id2word,
                                            num_topics=num_topics,
                                            random_state=50,
                                            update_every=1,
                                            chunksize=10,
                                            passes=passes,
                                            alpha='symmetric',
                                            iterations=50,
                                            per_word_topics=True)
import pyLDAvis
import pyLDAvis.gensim_models as gensimvis
```

```
pyLDAvis.enable_notebook()
vis = gensimvis.prepare(lda_model_5, corpus, dictionary=lda_model_5.id2word)
pyLDAvis.display(vis)
```

- **적용 예시**




```
print("Topics for lda_model_5:")
print(topics_table1)
```

```
Topics for lda_model_5:
    Topic                                        Top Words
0  Topic 1   시장, 도시, 미래, 예비, 시민, 기업, 일자리, 추진, 지역, 주택
1  Topic 2  후보, 여성, 공약, 정부, 개편, 정치, 국민, 의원, 당선인, 대통령
2  Topic 3  모집, 서울, 서울시, 보유, 기준, 이하, 지난해, 올해, 대상, 화폐
3  Topic 4  대전, 센터, 창업, 프로그램, 활동, 주민, 지역, 운영, 내달, 문화
4  Topic 5  위원회, 정부, 위원장, 복지, 업무, 통상, 부처, 방안, 사항, 회의
```

2-2. OpenAI GPT API

- **실행 파일명** : 2_LDA_analysis.ipynb
- **모델 실행 방법** :

  1. “/코드/학습코드/2_LDA_analysis.ipynb”를 활용하여 생성형 문장(정책 제안)을 추출한다.

  2. OpenAI 라이브러리를 먼저 설치한다.(!pip install openai)

  3. OpenAI GPT4 API를 불러온다.

- **적용 예시**

- **입력** :

```
import openai


def generate_policy_recommendation(policy, keywords):
  question = f"한국의 {policy}을 강화하고 발전시키기 위한 방법을 찾고 싶습니다. LDA 모델링 결과로 언급된 키워드 중 {', '.join(keywords)}을 활용하여 어떤 정책을 추천하시겠습니까?"

  openai.api_key ="openAI_key"
  response = openai.ChatCompletion.create(
    model="gpt-4",
    messages=[
      {"role": "system", "content": "You are a helpful assistant."},
      {"role": "user", "content": question}
    ]
  )


  return response['choices'][0]['message']['content']
```

- **결과값** :

'청년 창업 지원 정책'을 추천드립니다. 이런 방식으로 구체화할 수 있습니다:

*1. **청년 창업 인큐베이션 센터 설립 (키워드: 시장, 도시, 기업, 추진)**: 도시 중심부에 청년들이 새로운 사업 아이디어를 모색하고, 기업을 창업하고 성장시킬 수 있도록 지원하는 인큐베이션 센터를 설립합니다. 이를 통해 시장에 새로운 일자리를 창출하고, 도시의 활력을 더합니다.*

*2. **교육과 직업 훈련 프로그램 개발 (키워드: 시민, 예비, 일자리)**: 청년들에게 현장진출을 위한 실질적인 교육과 훈련을 제공합니다. 이 개발된 프로그램은 창업에 대한 전반적인 지식과 기술을 제공할 수 있도록 구성합니다.*

*3. **청년 주택 지원 정책 (키워드: 주택, 지역)**: 청년들이 창업 초기에 안정적으로 생활할 수 있도록 청년 주택 지원 프로그램을 개발하고 실행합니다. 이는 지역 고유의 주거 문제를 해결하는 데에도 도움이 될 수 있습니다.*

*4. **미래 지향적 창업 아이디어 경진대회 운영 (키워드: 미래, 시민, 추진)**: 청년들에게 미래 지향적인 아이디어를 제시하고 이를 현실화시키는데 도움을 주는 경진대회를 운영합니다. 이를 통해 시민들의 참여를 유도하고, 미래지향적인 사업 아이디어를 발굴할 수 있습니다.*

3. Kpfbertsumm 뉴스 요약 모델

- **모델 설명**: KpfBertsumm은 Bert 사전학습 모델을 이용한 텍스트 요약 논문 및 모델인 PRESUMM모델을 참조하여 한국어 문장의 요약추출을 구현한 한국어 요약 모델이다. 본 모델은 한국언론진흥재단에서 구축하였으며, 방대한 뉴스기사 코퍼스로 학습한 kpfBERT를 이용하여 특히 뉴스기사 요약에 특화된 것이 특징이다. AI Hub의 뉴스 추출요약 데이터셋을 활용하여 학습된 모델 (“kpfbertsumm_trained_model.pt”) 사용했다.

- **실행 파일명**: 3_news_kpfbertsumm.ipynb
- **모델 실행 방법**:

  1. 전처리된 뉴스 데이터 “/데이터/Processed_Data/3_news_summ_pre.xlsx”을 준비한다.

  2. 뉴스 요약을 위해 “/코드/3_news_kpfbertsumm.ipynb” 모델을 실행한다.

  3. 상위 토픽 모델링 단계에서 추출된 주요 변곡점 기간(월기준)의 뉴스 파일(“3_news_summ_pre.xlsx”)을 불러온다.
- **모델링 불러오는 방법**

```
PATH = '#모델파일의경로/2_kpfbertsumm_trained_model.pt'

device = torch.device("cuda")
trained_model = Summarizer()
trained_model.load_state_dict(torch.load(PATH, map_location="cuda:0"))  # 사용할 GPU 장치 번호를 선택합니다.
trained_model.to(device)
```

- **모델링 함수**

```
def summarize_test(text):
    data = data_process(text.replace('₩n',''))

    #trained_model에 넣어 결과값 반환
    _, rtn = trained_model(data['src'].unsqueeze(0), data['segs'].unsqueeze(0), data['clss'].unsqueeze(0))
    rtn = rtn.squeeze()

    # 예측 결과값을 받기 위한 프로세스
    rtn_sort, idx = rtn.sort(descending = True)

    rtn_sort = rtn_sort.tolist()
    idx = idx.tolist()

    end_idx = rtn_sort.index(0)

    rtn_sort = rtn_sort[:end_idx]
    idx = idx[:end_idx]

    if len(idx) > 3:
        rslt = idx[:3]
    else:
        rslt = idx

    summ = []
    # print(' *** 입력한 문단의 요약문은 ...')
    for i, r in enumerate(rslt):
        summ.append(data['sents'][r])
        print('[', i+1, ']', summ[i])

    return summ
```

- **적용 예시**

- **입력문장**: 'Summary' 컬럼에 대한 뉴스요약 및 결과 저장

```
df['Summary'] = df['Text'].apply(lambda x: summarize_test(x))
```

"가르치려 한 오만함, 청년과 단절 원인" 반성 "박원순 피해자 제대로 된 사과와 반성 이뤄지지 않았어" 이소영 의원을 비롯한 더불어민주당 2030의원들이 9일 오전 서울 여의도 국회 소통관에서 '더불어민주당 2030의원 입장문' 발표를 기자회견을 하고 있다. 2021.4.9/뉴스1 © News1 박세연 기자 (서울=뉴스1) 서혜림 기자 = 더불어민주당 2030대 의원들은 9일 입장문을 내고 "돌아선 국민의 마음의 원인은 저희를 포함한 민주당의 착각과 오판에 있었음을 자인한다"고 말했다. 민주당 오영환·이소영·장경태·장철민·전용기 의원은 이날 오전 국회 소통관에서 기자회견을 열고 "선거 유세 현장과 삶의 현장에서 만난 20대 30대 청년들은 민주당에 싸늘하고 무관심했고 지난 1년 동안 많은 분의 마음이 돌아섰음을 현장에서 느꼈다"고 말했다……다만 장 의원은 "선거 책임은 모두가 져야 한다고 본다. 특정 인물을 지목하거나, 그분들에게만 책임이 있다고 생각하지는 않는다. 그동안의 관행이나 오만과 독선으로 비춰질 수 있는 부분에 대해서 스스로 문제를 제기하는 반성이 담긴 것"이라고 설명했다. suhhyerim777@news1.kr”

- **결과값:**

[1] "가르치려 한 오만함, 청년과 단절 원인" 반성 "박원순 피해자 제대로 된 사과와 반성 이뤄지지 않았어" 이소영 의원을 비롯한 더불어민주당 2030의원들이 9일 오전 서울 여의s도 국회 소통관에서 '더불어민주당 2030의원 입장문' 발표를 기자회견을 하고 있다.
[2] 민주당 오영환·이소영·장경태·장철민·전용기 의원은 이날 오전 국회 소통관에서 기자회견을 열고 "선거 유세 현장과 삶의 현장에서 만난 20대 30대 청년들은 민주당에 싸늘하고 무관심했고 지난 1년 동안 많은 분의 마음이 돌아섰음을 현장에서 느꼈다"고 말했다.
[3] 이들은 "이번 재보궐선거를 치르게 된 원인이 우리 당 공직자의 성 비위 문제였음에도 불구하고 우리 당은 당헌·당규를 개정해 후보를 내고 피해자에 대한 제대로 된 사죄도 없었다"며 "당내 2차 가해를 적극적으로 막는 조처를 하지 않았다. 이 문제를 회피하고 외면할 수 있지 않을까 하는 오만함이었다"고 비판했다.

4. 지도 시각화 및 AI 리포트

- **실행 파일명**: 5_regional_AI_report.ipynb
- **모델 실행 방법** :

  1. 데이터를 다음과 같이 준비한다.

     1) 지역별 뉴스데이터 (“/데이터/Processed_Data/4_youth_news.xlsx”),

     2) 뉴스 요약을 위한 뉴스 데이터 (youth_policy_news(20190501_20210930).xlsx, youth_policy_news(20211001_20230626).xlsx)

2. “코드/5_regional_AI_report.ipynb”파일을 실행하여, 지역별 뉴스 토픽 모델링, 생성형 정책 제안, 뉴스요약 모델을 실행한다.

- **적용 예시**

- **입력 데이터**

Date	Publisher	Title	Text	Contents	Region	Type	topic_region
2019/05/01	노컷뉴스	경남 떠난 청년들 '부산으로'...일자리 찾으러 '수도권으로'	경남발전연구원 심인선 선임연구위원 연구보고서 유출 청년 30.7% 부산 이동, 일자...	경남 청년 부산 일자리 수도 경남 발전 연구원 심인 선 선임 연구 위원 연구 보고서...	전국	전국	부산
2019/05/02	노컷뉴스	당정청 '2030 청心 잡기에 올인'...'청년 기구' 구성	종리실 '청년 컨트롤타워' 담당 청년정책조정위원회 의장 이 종리 겁기로 [CBS노컷...	당정 청 출인 청년 기구 구성 종리 실 청년 컨트롤 타워 담당 청년 정책 조정 위원...	전국	전국	전국
2019/05/03	노컷뉴스	당정청이 힘싣는 '올지로路'...민생 선점 전략	文대통령 '올지로위원회' 공약 수정·실행..."민생문제 해결 방점" '삭발식' 한국당,...	당정 청 힘 을 지로 민생 선점 전략 대통령 지로 위원회 공약 수정 실행 민생 문제...	전국	전국	서울
2019/05/08	노컷뉴스	마스터키로 문 '벌컥'...“주인님, 그러시면 아니되옵니다”	집주인의 마스터키 메너, 세들어사는 대학생 생활불편 1위 쓰레기무단 투기·소음, ...	마스터 문 주인 집 주인 마스터키 매너 세 어사 대학생 생활 불편 위 쓰레기 무단 ...	전국	전국	전국
2019/05/09	노컷뉴스	[영상] 이재준 고양시장 "미래 세대까지 행복한 고양시를 만들 것"	'피플앤리더' 이재준 고양시장 인터뷰 [CBS노컷뉴스 고태현 기자] 경기도 고양시가...	영상 이재준 고양 시장 미래 세대 행복 양시 피플 앤 리더 이재준 고양 시장 인터뷰...	전국	전국	전국

- **결과값 1.**

	Topic	Top Words
0	Topic 1	분야, 확대, 주거, 정신, 경남, 조성, 조사, 예산, 계획, 올해
1	Topic 2	장애, 경제, 도시, 친화, 조직, 행정, 업무, 여성, 아동, 산업
2	Topic 3	센터, 대상, 운영, 대학, 취업, 선정, 정보, 활동, 고용, 기업
3	Topic 4	후보, 국민, 문제, 의원, 대표, 정치, 선거, 위원장, 정부, 지방
4	Topic 5	공간, 문화, 인구, 위원회, 조례, 행사, 소통, 개최, 계획, 기본

- **결과값 2.**

[1] 국민의힘이 윤석열 대선 후보가 참석하는 '전국 청년 간담회' 화상회의를 5일 개최했으나, 예정과 달리 윤 후보는 통화로만 참석해 참가자들의 분노를 샀다.
[2] 국민의힘은 청년들의 의견을 수렴하겠다며 5일 오후 4시 중앙선대위 국민소통본부 전국 청년간담회를 열었다.
[3] 윤 후보가 참석한다는 소식에 몰려든 300명에 가까운 참가자들 사이에서는 즉시 욕설이 터져 나왔다.
[1] 이 대표는 이날 송파구 올림픽공원에서 열린 청년의 날 기념식에 참석해 "국민의힘은 지방선거와 국회의원 선거에 출마할 수 있는 피선거권 연령 제한을 선거권과 동일하게 조정, 연령 제한을 철폐하겠다"고 밝혔다.
[2] 국민의힘 이준석 대표가 6일 오후 서울 송파구 올림픽공원 평화의광장에서 열린 제5회 대한민국 청년의 날 기념식에서 축사하고 있다.
[3] 연합뉴스 국민의힘 이준석 대표는 6일 지방선거와 국회의원 선거에 출마할 수 있는 피선거권 연령을 만 18세 이상으로 낮추겠다고 밝혔다.
[1] 24일 인터넷방송 플랫폼 '아프리카TV'는 "'민심쇼' 아프리카TV가 묻고 이준석이 답하다!"라는 제목의 공지글을 통해 25일 오후 8시에 이 대표가 참여하는 생방송이 진행된다고 밝혔다.
[2] 연합뉴스 국민의힘 이준석 대표가 25일 윤석열 대선 후보가 나서는 TV토론 시간대에 인터넷방송에 출연한다.
[3] 이를 두고 일각에서는 윤 후보가 참여하는 TV토론이 열리는 시간대에 당 대표가 인터넷 방송에 출연하는 것이 부적절하다는 지적이 나오고 있다.
[1] 변성완 더불어민주당 부산시장 예비후보가 부산형 청년 기초자산 도입을 첫 공약으로 내놨다.
[2] 부산에서 태어나는 모든 아이에게 1000만 원씩을 입금해준 뒤 만 20세가 되면 이자를 더해 2000만 원을 찾을 수 있도록 해 대학 학비나 창업을 위한 종잣돈으로 활용할 수 있도록 하겠다는 취지다.
[3] 변 후보는 4일 부산시의회에서 '1호 공약'으로 청년 정책을 발표하고 "청년도시 부산을 만들어 사람들이 찾아오는 도시로 바꾸겠다"고 밝혔다.
[1] 19일 부산에서 "전두환 (전) 대통령이 잘못한 부분이 있지만, 군사 쿠데타와 5·18만 빼면 정치는 잘했다고 말하는 분들이 많다" 고 말해 당 안팎에서 거센 비판과 함께 시민사회의 사과 요구를 받은 지 이틀 만이다.
[2] 이날 윤 전 총장은 "전두환 전 대통령 옹호 발언과 관련한 비판을 겸허히 수용하고 유감을 표한다" 고 말했다.
[3] 다만 윤 후보는 송구하다는 글을 남기면서도 "정치인이라면 '자기 발언이 늘 편집될 수 있다' 는 생각까지 해야 한다는 지적을 받아들인다" 며 자신의 진심을 몰라주는 기성 정치 환경을 탓하는 핏맛을 남겼다.
