Big Data with Python 2/3

DataAnalysis / PreProcessing

2024.06.25
국립안동대학교
컴퓨터교육과 PhD. 조영복
(ybcho@anu.ac.kr)

Big-data with Python

교과목 명	빅데이터 with 파이썬	
교육 목표	파이썬과 Pandas 기반의 데이터 수집, 데이터 분석, 데이터 시각화 등의 개념에 대하여 학습하고, 실습을 통하여 데이터 분석 전문가로서의 역량을 키운다.	
교육 내용 - 교육 시간 (20H)	Day-1 Data Analysis	• 빅데이터 분석 개요 • Numpy - 넘파이 배열, Shape 변환 함수, 인덱싱과 슬라이싱 - 산술 연산과 브로드캐스팅 • Pandas - 파일에서 데이터 읽기, 데이터프레임 생성, - 열단위 데이터 추출, 행단위 데이터 추출 - 컬럼 추가/삭제, 그룹별 집계/요약
	Day-2 Data Visualization	• Matplotlib - 라인플롯, 스타일 변경(라인, 축, 레이블 등), 산점도 - 히스토그램, 박스 플롯 • Seaborn
	Day-3 Data Analysis Pre-Processing	• 타이타닉 데이터 개요 (캐글 데이터 이용) • 결측치 처리/이상치 처리, 피쳐 엔지니어링 • 성별, 객실등급, 나이별 생존율 시각화

SKT Fly AI가 시작한날 무슨일?

SKT Fly AI가 시작한날 무슨일?

```
!apt-get update -qq
!apt-get install fonts-nanum* -qq
```

```
import requests
import pandas as pd
from bs4 import BeautifulSoup
import matplotlib.pyplot as plt
import nltk
from konlpy.tag import Kkma
from konlpy.tag import Twitter
from wordcloud import WordCloud
```

```
date='20240624'
```

```
news_url = 'https://news.naver.com/main/ranking/popularDay.nhn?
date={}'.format(date)
headers = {'User-Agent' : 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/89.0.4389.90 Safari/537.36'}
req = requests.get(news_url, headers = headers)
```

SKT Fly AI가 시작한날 무슨일?

```
soup = BeautifulSoup(req.text, 'html.parser')
news_titles = soup.select('.rankingnews_box > ul > li > div > a')

crowled_title = []
for i in range(len(news_titles)):
    crowled_title.append(news_titles[i].text)
    print(i+1, news_titles[i].text)
```

```
1 다리에 실지렁이가…롱부츠 신던 20대女 '경악'한 까닭 [건강!톡]
2 "없는 거 빼고 다있소"…다이소, 이번엔 '이것'까지 내놨다
3 "노량진서 사온 것 보고 경악"…'썩은 대게' 항의했더니
4 "지금이 기회" 현금부자들 우르르…'이 동네' 방긋 웃었다
5 95년생 '천재 소녀'에 美 발칵…6개월 만에 3900억 '대박' [조아라의 IT's fun]
6 19시간 밤샘 조사 이선균 “공갈범 진술 신빙성 있나”
7 한동훈 “많이 도와달라”…권영세·이양수 등에 전화
8 친구랑 킥복싱하다 갈비뼈 부러뜨린 10대…“700만원 배상해야”
9 이원욱 “이재명에게 필요한 건 ‘권력욕’ 아닌 ‘혁신’”
10 송영길 구속 ‘1주일’ 허탕…검찰, 내일 출석 통보
11 설마 다시 팬데믹?…전세계 코로나 확진자 ‘4주간 52% 증가’
12 꽁꽁 언 새벽, 내복 바람 4살 아이는 ‘아파트 천사’를 만났다
13 불길 피해 젖먹이 끌어안고 뛰어내린 30대 아빠…끝내 숨져
14 한국이 경제성적 2위? 윤 대통령이 말하지 않은 ‘뒤집힌 세계순위들’
15 대전서 식당 폭발로 12명 중경상…“굉음과 함께 건물 흔들려”
16 "뷔·손흥민만 좋은 일"…저가커피 가맹점주들 '부글'
17 제일 위험하다는 '초피거래', 왜 할까요? [현장 써머리]
18 이준석 "김건희 명품백 몰카?...국민수준 너무 얕본다"
19 '딸기시루' 케이크가 뭐길래…크리스마스 앞두고 '3배 되팔이'
20 지방 청약 저조했지만…전혀 달랐던 '두 도시'
21 “전화로 주문해야 쌉니다”...매장보다 비싼 배달앱 방문포장
22 ‘비밀번호’ 없는 세상이 온다… 삼성·애플·구글, 도입한 ‘패스키’는
23 車 사지 않는 20대… 2009년 이후 최저치
24 [벤처하는 의사들] 칠 송 스트레스 호르몬은 알고 있었네...누가 얼마나 우울하지
```

SKT Fly AI가 시작한날 무슨일?

```
title = "".join(crowled_title)
filtered_title = title.replace('.', ' ').replace('"',' ').replace(',',' ').replace("'","
").replace('·', ' ').replace('=',' ').replace('\n',' ')
filtered_title
```

'다리에 실지렁이가…롱부츠 신던 20대女 경악 한 까닭 [건강!톡] 없는 거 빼고 다있소 . 번엔 이것 까지 내놨다 노량진서 사온 것 보고 경악 … 썩은 대게 항의했더니 지금이 기회 들 우르르… 이 동네 방긋 웃었다95년생 천재 소녀 에 美 발칵…6개월 만에 3900억 다 의 IT s fun]19시간 밤샘 조사 이선균 “공갈범 진술 신빙성 있나”한동훈 “많이 도와달라 양수 등에 전화친구랑 킥복싱하다 갈비뼈 부러뜨린 10대…“700만원 배상해야”이원욱 “이재명 건 ‘권력욕’ 아닌 ‘혁신’”송영길 구속 ‘1주일’ 허탕…검찰 내일 출석 통보설마 다시 팬데 코로나 확진자 ‘4주간 52% 증가’꽁꽁 언 새벽 내복 바람 4살 아이는 ‘아파트 천사’를 해 젖먹이 끌어안고 뛰어내린 30대 아빠…끝내 숨져한국이 경제성적 2위? 윤 대통령이 말하 집힌 세계순위들’대전서 식당 폭발로 12명 중경상…“굉음과 함께 건물 흔들려” 뷔 손흥민만 가커피 가맹점주들 부글 제일 위험하다는 초피거래 왜 할까요? [현장 써머리]이준석 백 몰카? 국민수준 너무 얕본다 딸기시루 케이크가 뭐길래…크리스마스 앞두고 3배 도 약 저조했지만…전혀 달랐던 두 도시 “전화로 주문해야 쌉니다” 매장보다 비싼 배달앱 번호’ 없는 세상이 온다… 삼성 애플 구글 도입한 ‘패스키’는車 사지 않는 20대… 2009년 치[벤처하는 의사들] 침 속 스트레스 호르몬은 알고 있었네 누가 얼마나 우울한지“새벽부 다”…성심당 ‘딸기시루’ 당근에선 10만원까지 MZ 신차 안 산다…역대 최저치중고 시장서 새벽부터 긴 줄 공항 Airport로 쓰면 위법 달라지는 배당 투자 잘 활용하면 꿀 파 쇄도… 압사당할 것 같다 제니가 입은 수영복 살래요 …최강 한파에 여름옷 매출 훌쩍 침착맨 뜬 네이버 …'

SKT Fly AI가 시작한날 무슨일?

```
tw = Twitter() # Twitter가 Okt로 변경
tokens_ko = tw.nouns(filtered_title)
tokens_ko

ko = nltk.Text(tokens_ko, name='기사 내 명사')
ko.tokens
ko.vocab()

new_ko=[]
for word in ko:
  if len(word) > 1 and word != '단독' and  word != ' ':
        new_ko.append(word)
new_ko
ko = nltk.Text(new_ko, name = '기사 내 명사 두 번째')
ko.tokens
ko.vocab()

data = ko.vocab().most_common(150)
data = dict(data)
data
```

```
/usr/local/lib/python3.10/d
  warn('"Twitter" has chang
{'한동훈': 21,
 '성탄절': 21,
 '서울': 21,
 '아파트': 20,
 '아빠': 18,
 '화재': 17,
 '크리스마스': 16,
 '안고': 15,
 '이준석': 14,
 '아이': 13,
 '대통령': 10,
 '올해': 10,
 '대게': 9,
 '김건희': 9,
 '뉴스': 9,
 '사고': 9,
 '논란': 8,
 '자녀': 8,
 '눈물': 8,
 '도봉구': 8,
 '사망': 8,
 '내년': 8,
 '노량진': 7,
 '한국': 7,
 '가격': 7,
 '산타': 7,
 '아들': 7,
 '아내': 7,
 '목욕탕': 7,
 '영상': 7,
 '결혼': 7,
 '이브': 7,
 '새벽': 6,
 '피해': 6,
```

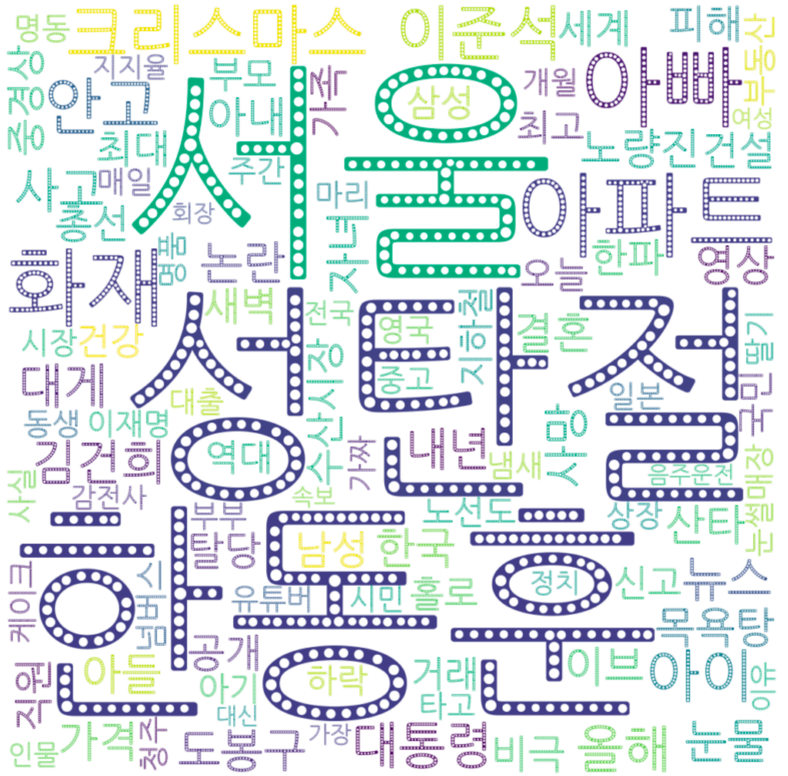
SKT Fly AI가 시작한날 무슨일?

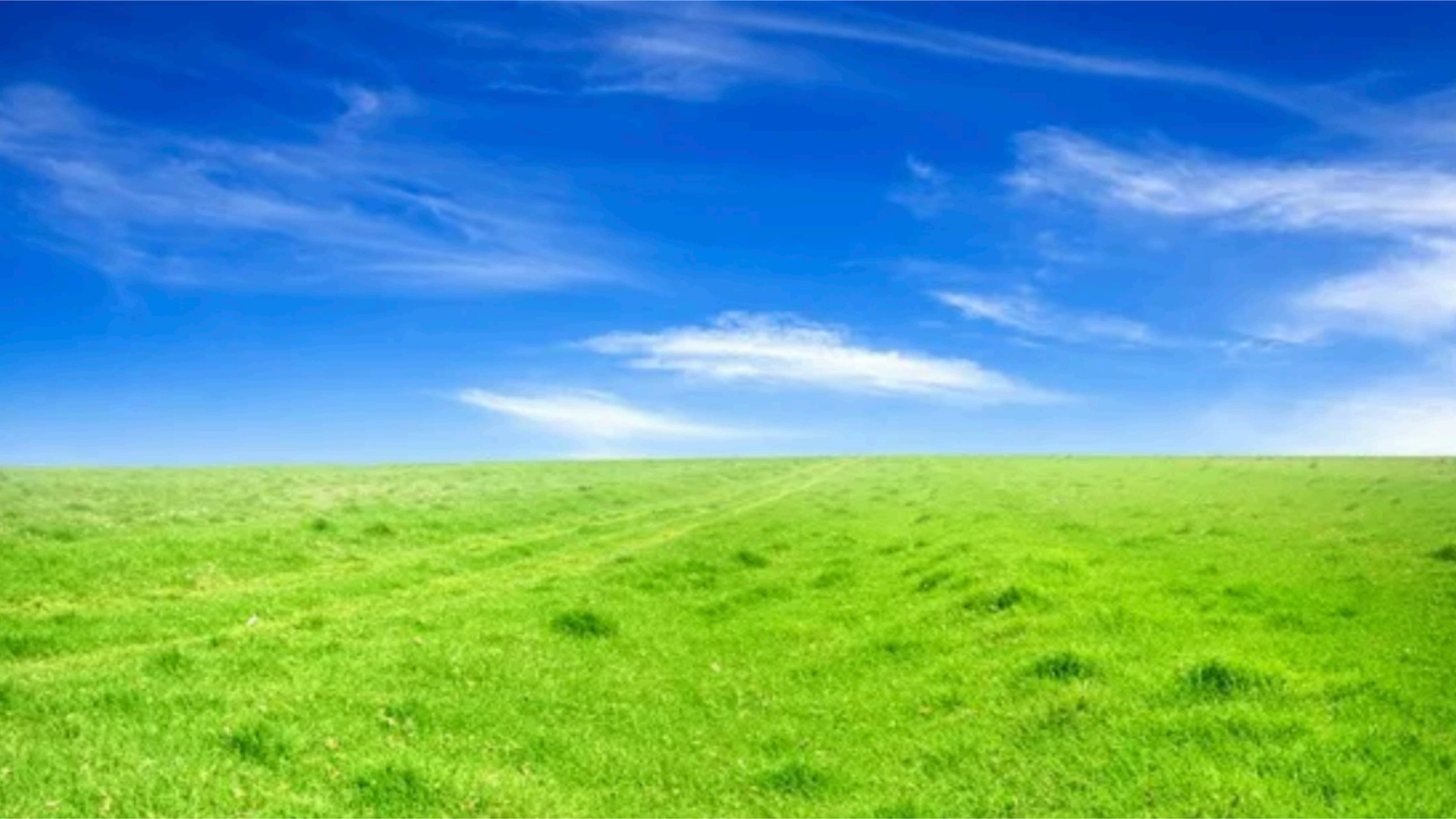
```
wordcloud = WordCloud().generate(filtered_title)

font = '/usr/share/fonts/truetype/nanum/NanumGothicEco.ttf'

wc = WordCloud(font_path=font,\
    background_color="white", \
    width=1000, \
    height=1000, \
    max_words=100, \
    max_font_size=300)
wc = wc.generate_from_frequencies(data)

plt.figure(figsize=(10,10))
plt.imshow(wc, interpolation='bilinear')
plt.axis('off')
plt.show()
```





멜론챠트 크롤링

- 멜론 챠트 1-100위

```
from bs4 import BeautifulSoup as bs
import requests as req
import pandas as pd

# header 만들기
url = 'https://www.melon.com/chart/index.htm'
header = {'user-agent':'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/
537.36(KHTML, like Gecko) Chrome/109.0.0.0 Safari/537.36'}

# 헤더 확인
res =req.get(url, headers=header)
res

#파이선 객체 데이터로 변환

html = bs(res.text,'lxml')
html
```

```
<span class="bg_album_frame"></span>
</a>
</div></td>
<td><div class="wrap">
<a class="btn button_icons type03 song_info"
href="javascript:melon.link.goSongDetail('36892797');" title="음악
보"><span class="none">곡정보</span></a>
</div></td>
<td><div class="wrap">
<div class="wrap_song_info">
<div class="ellipsis rank01"><span>
<a href="javascript:melon.play.playSong('1000002721',36892797);"
악의 신 재생">음악의 신</a>
</span></div><br/>
<div class="ellipsis rank02">
<a href="javascript:melon.link.goArtistDetail('861436');" title="
(SEVENTEEN) - 페이지 이동">세븐틴 (SEVENTEEN)</a><span class="checkEl
style="display:none"><a
href="javascript:melon.link.goArtistDetail('861436');" title="세븐
(SEVENTEEN) - 페이지 이동">세븐틴 (SEVENTEEN)</a></span>
</div>
</div>
</div></td>
<td><div class="wrap">
<div class="wrap_song_info">
<div class="ellipsis rank03">
<a href="javascript:melon.link.goAlbumDetail('11348980');"
title="SEVENTEEN 11th Mini Album 'SEVENTEENTH HEAVEN' - 페이지 이
동">SEVENTEEN 11th Mini Album 'SEVENTEENTH HEAVEN'</a>
</div>
</div>
</div></td>
```

멜론챠트 크롤링

- 멜론 챠트 1-100위

```
song=html.select('.ellipsis.rank01>span>a')

song_list=[i.text for i in song]
song_list

singer=html.select('.ellipsis.rank02>a')
singer

singer_list=[i.text for i in singer]
singer_list

# 랭킹 리스트 생성, 각 리스트 길이 확인
rank_list =[i+1 for i in range(len(song_list))]
print(len(rank_list))
print(len(singer_list))
print(len(song_list))
```

```
100
106
100
```

```
'WAY 4 LUV',
'MAESTRO',
'Get A Guitar',
'오래된 노래',
'Ditto',
'연애편지',
'Attention',
'Spicy',
'사랑인가 봐',
'그랬나봐',
'Accendio',
'Impossible',
'너의 모든 순간',
'인사',
'청혼하지 않을 이유를 못 찾았어',
'Siren',
'파이팅 해야지 (Feat. 이영지)',
'봄눈',
'Girls Never Die',
'OMG',
'Perfect Night',
'우리 영화',
'Midas Touch',
'Right Now',
'보금자리',
'MANIAC',
'From',
'Dynamite',
'EASY',
'퀸카 (Queencard)',
'Lucky Girl Syndrome',
'Watch Me Woo!',
'사막에서 꽃을 피우듯',
'주저하는 연인들을 위해',
'LOVE DIVE',
'사건의 지평선',
'Run Run',
'Love Lee',
"I Don't Think That I Like Her",
'봄날',
'손오공',
'Smart',
'어떻게 이별까지 사랑하겠어, 널 사랑하는 거지',
'모든 날, 모든 순간 (Every day, Every Moment)',
'꿈',
'취중고백',
```

멜론챠트 크롤링

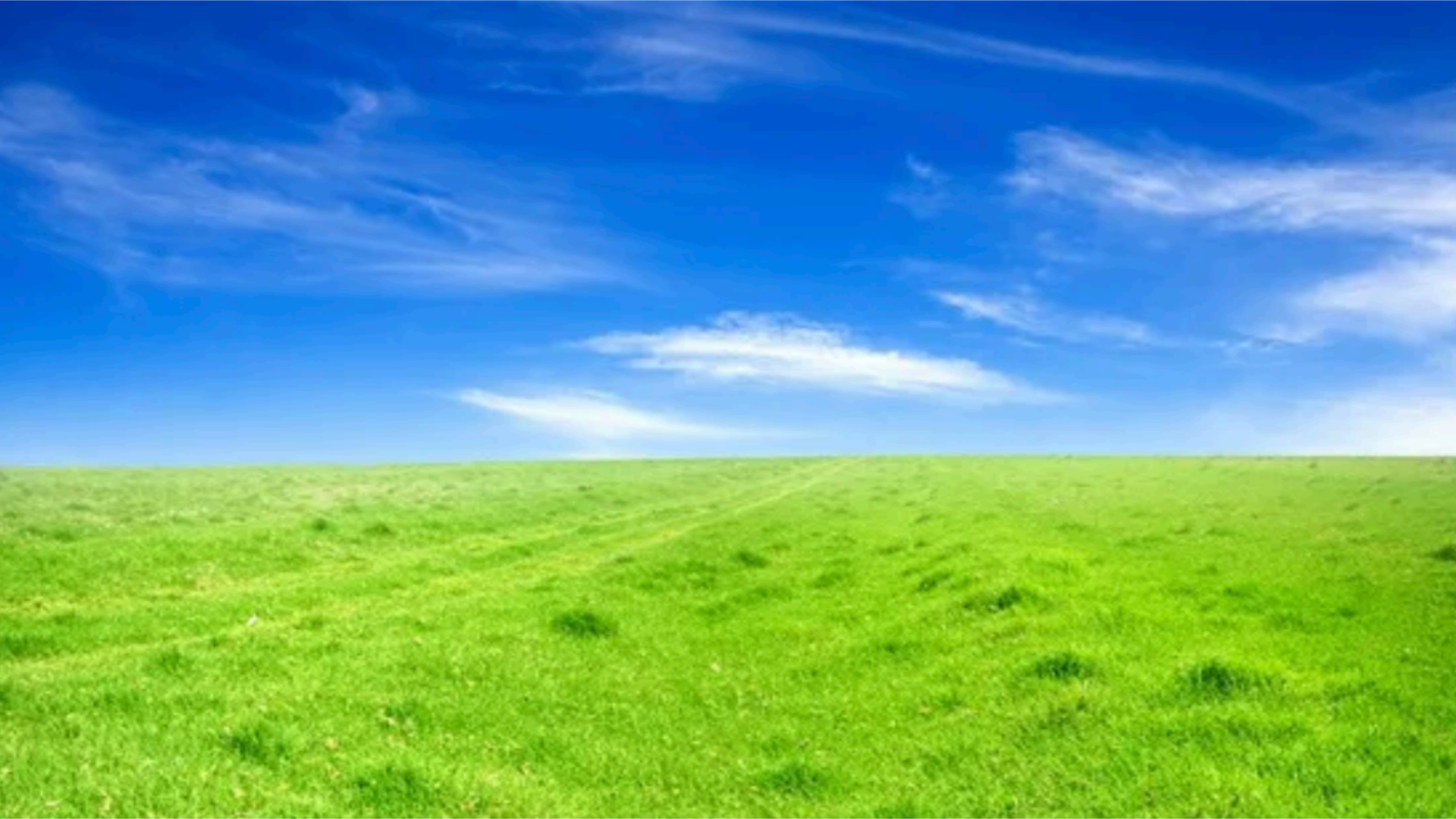
- 멜론 챠트 1-100위

```
#딕셔너리 형태로 생성하고 데이터 프레임 생성

top_100={'순위':rank_list, '가수':singer_list, '곡명':song_list}
top_100_df=pd.DataFrame(top_100)
top_100_df
```

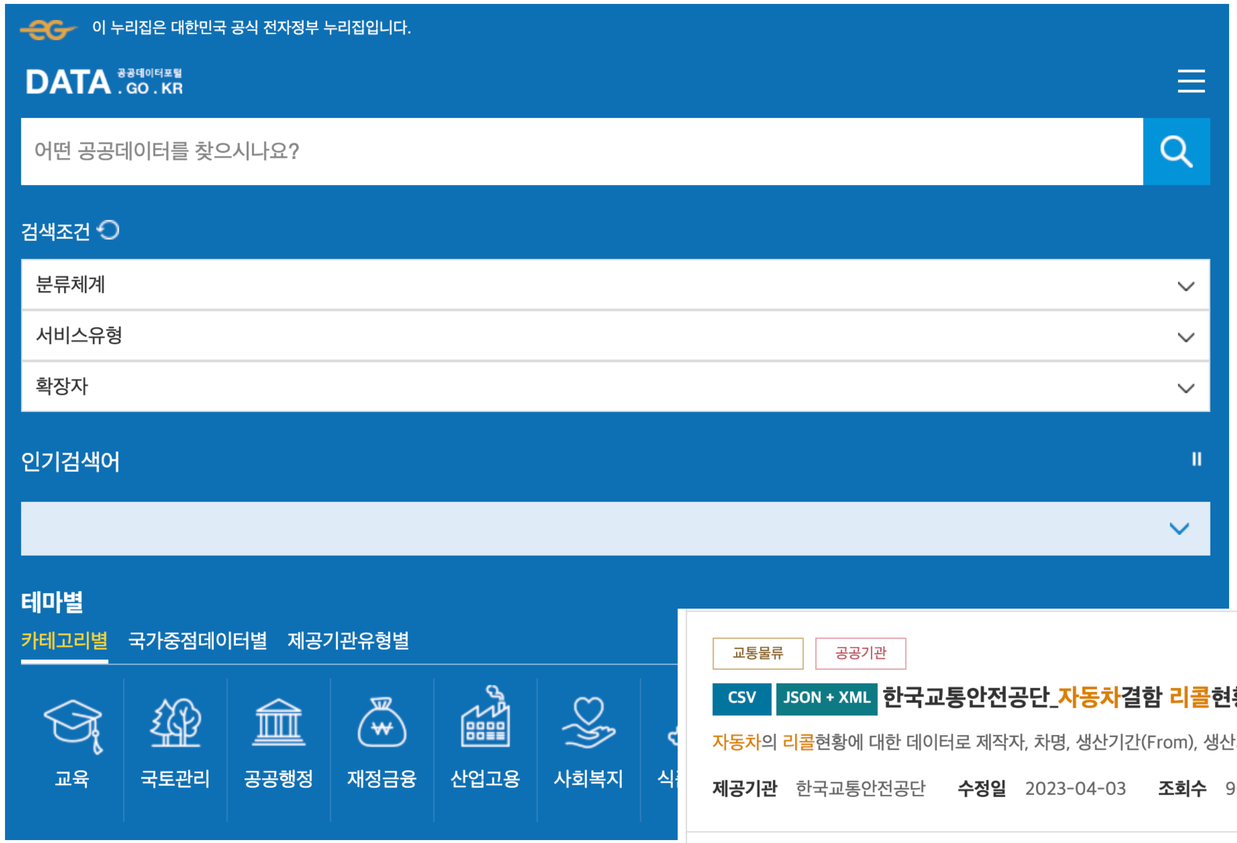
	순위	가수	곡명
0	1	aespa	Supernova
1	2	이영지	Small girl (feat. 도경수(D.O.))
2	3	NewJeans	How Sweet
3	4	이클립스 (ECLIPSE)	소나기
4	5	NewJeans	Bubble Gum
...	...	...	...
95	96	AKMU (악뮤)	청춘찬가
96	97	폴킴	사랑하지 않아서 그랬어
97	98	태연 (TAEYEON)	내가 S면 넌 나의 N이 되어줘
98	99	김민석	그대가 내 안에 박혔다(그내박)
99	100	경서예지	잘 지내자, 우리 (여름날 우리 X 로이킴)

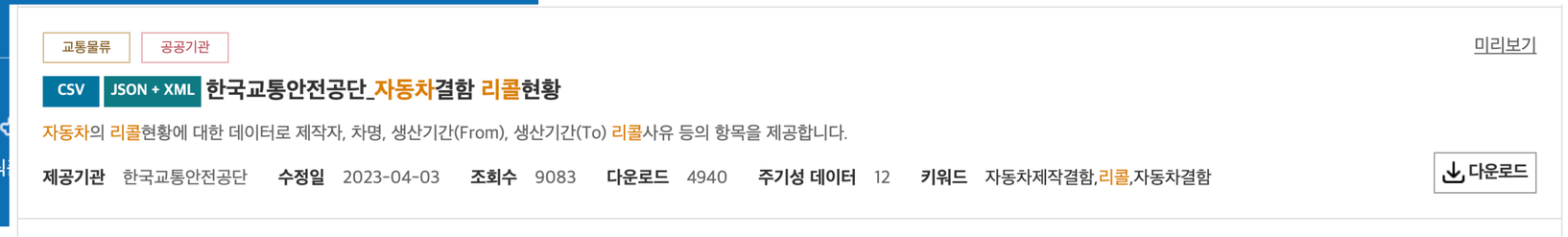
100 rows × 3 columns



공공데이터 분석

- https://www.data.go.kr/data/3048950/fileData.do





공공데이터 분석

- 1. 데이터로딩
- 2. 결측치 확인
- 3. 중복값확인
- 4. 데이터 시각화
  - 제조사별/모델별/월별/생산연도별 리콜현황
  - 워드클라우드를 이용한 리콜 사유
    - #(Q) 2022년 리콜 개시가 가장 만이 일어난 달(month)과 가장 적게 일어난 달의 차이는?