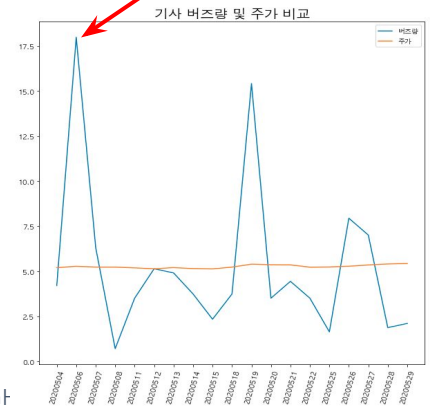
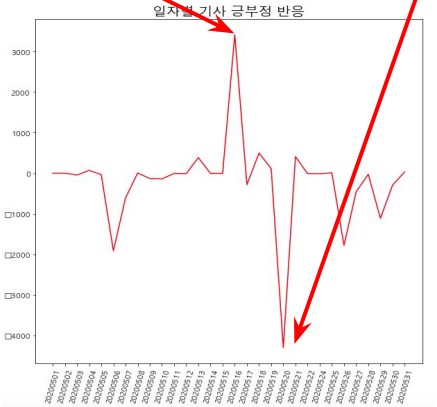
데이터 수집, 시각화 프로젝트

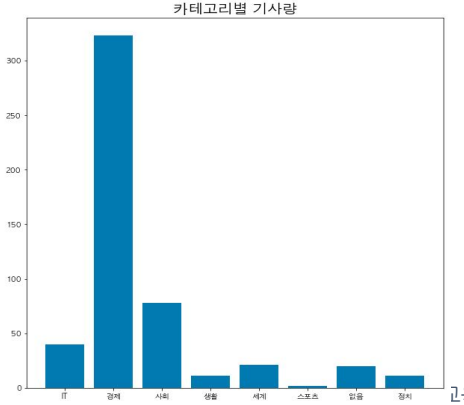
- 뉴스, 주가분석 -

최종결과물

긍부정 -> 사람들의 기사 관심도







추가학습 (CSV 파일저장)

```
import csv
file = open("outputs.csv", mode="w", encoding="utf-8", newline="")
writer = csv.writer(file)
writer.writerow(["데이터1", "데이터2", "데이터3"])
writer.writerow(["데이터1", "데이터2", "데이터3"])
writer.writerow(["데이터1", "데이터2", "데이터3"])
file.close()
```

추가학습 (워드클라우드)

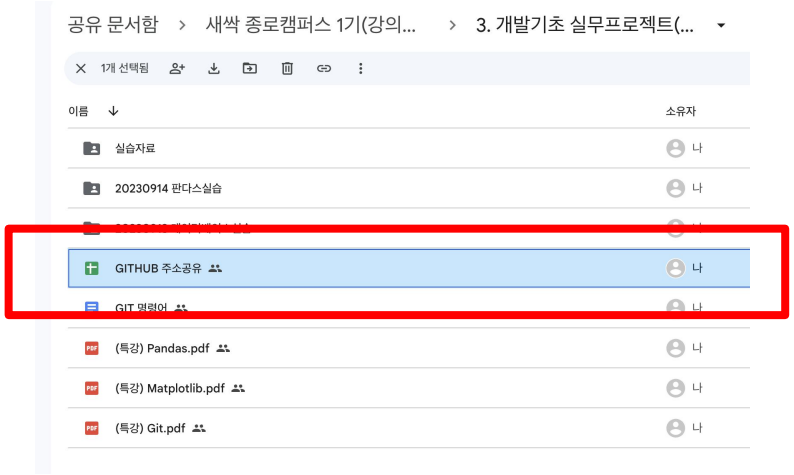
```
from wordcloud import WordCloud, STOPWORDS

text = ['안녕하세요 이진범 입니다', '안녕하세요 홍길동 입니다']
wordcloud = WordCloud(max_font_size=200,
                      font_path='./NanumGothic.ttf',
                      stopwords=STOPWORDS,
                      background_color='#FFFFFF',
                      width=1200,
                      height=800).generate(' '.join(text))

plt.figure(figsize=(5,5))
plt.imshow(wordcloud)
plt.tight_layout(pad=0)
plt.axis('off')
plt.show()
```

(1) github 레파지토리 생성

1. github 레파지토리 생성
2. github 주소공유에 레파지토리 주소 공유
3. 크롤링 코드는 파이썬코드로 분석코드는 주피터 노트북으로 진행해주세요



(2) 주가데이터 크롤링

1. 8월 한달간 삼성전자 주가 크롤링하여 stock.csv 파일로저장
(https://finance.naver.com/item/sise.nhn?code=005930)

일별시세

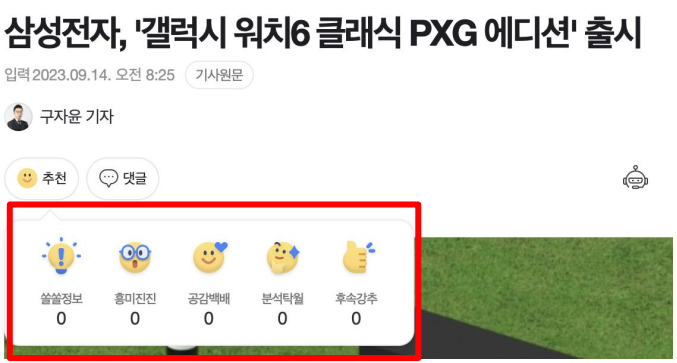
날짜	종가	전일비	시가	고가	저가	거래량
2020.05.22	48,750	▼ 1,200	49,600	49,800	48,600	19,706,284
2020.05.21	49,950	▼ 50	50,300	50,400	49,850	14,949,266
2020.05.20	50,000	▼ 300	50,000	50,200	49,800	14,896,899
2020.05.19	50,300	▲ 1,500	50,100	50,500	49,700	25,168,295
2020.05.18	48,800	▲ 950	47,950	49,100	47,600	20,481,981
2020.05.15	47,850	▼ 150	48,400	48,450	47,700	18,463,118
2020.05.14	48,000	▼ 550	47,750	48,100	47,650	19,305,974
2020.05.13	48,550	▲ 650	47,250	48,550	47,200	20,223,277
2020.05.12	47,900	▼ 500	48,400	48,500	47,550	23,433,590
2020.05.11	48,400	▼ 400	48,900	49,250	48,300	16,357,743

‹‹ 맨앞 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 다음 › | 맨뒤 ››

(3) 네이버뉴스 크롤링

1. 2023년 8월 한달동안의 연합뉴스 크롤링
2. 크롤링 항목은 날짜, 타이틀, 기사본문, 기사반응, 카테고리
3. 기사내용 전처리 진행
4. 기사는 news.csv 파일로 저장

추가적인 웹서비스로 데이터 가져옴
만약에 카테고리가 여러개로 분류될경우
가장 앞에있는 카테고리로



스포츠뉴스와 연예뉴스는 형식이 다름
스포츠뉴스와 연예뉴스는 뉴스의 형식이 다르고 아래 카테고리가 없어
카테고리를 각각 연예, 스포츠로

추가적인 웹서비스로 데이터 가져옴
각각 개수 크롤링

https://news.naver.com/main/list.nhn?mode=LPOD&mid=sec&oid=001

(4) 기사 데이터 정리 및 통계

1. 전체기사 일자별 카운트
2. 본문이 비어있는 기사등을 제외하고 크롤링한 기사중 본문에 삼성전자 글자가 들어간 뉴스분류
3. 삼성 주가데이터 합체 및 주말같이 주가데이터가 없는날짜 정리
4. 수치를 백분위로 변경 (날짜별합 / 전체합 * 100)

전체기사 :

날짜	
20200501	1692
20200502	696
20200503	942
20200504	1808
20200505	1268
20200506	2760
20200507	2901
20200508	2537
20200509	491
20200510	1442
20200511	2401
20200512	2733
20200513	2736
20200514	2925
20200515	2244
20200516	465
20200517	1334
20200518	2653
20200519	2682
20200520	3640
20200521	2875
20200522	2332
20200523	588
20200524	1221
20200525	2414
20200526	2581
20200527	3063
20200528	3005
20200529	2284
20200530	536
20200531	1232

삼성전자기사 :

날짜	
20200501	2
20200502	2
20200503	12
20200504	18
20200505	11
20200506	77
20200507	27
20200508	3
20200509	1
20200510	7
20200511	15
20200512	22
20200513	21
20200514	16
20200515	10
20200516	7
20200517	4
20200518	16
20200519	66
20200520	15
20200521	19
20200522	15
20200524	10
20200525	7
20200526	34
20200527	30
20200528	8
20200529	9
20200530	6
20200531	16

최종 :

날짜	본문	삼성전자주가	본문%	삼성전자주가%
20200504	18	48500.0	4.205607	5.198842
20200506	77	49200.0	17.990654	5.273877
20200507	27	48800.0	6.308411	5.231000
20200508	3	48800.0	0.700935	5.231000
20200511	15	48400.0	3.504673	5.188123
20200512	22	47900.0	5.140187	5.134527
20200513	21	48550.0	4.906542	5.204202
20200514	16	48000.0	3.738318	5.145246
20200515	10	47850.0	2.336449	5.129167
20200518	16	48800.0	3.738318	5.231000
20200519	66	50300.0	15.420561	5.391789
20200520	15	50000.0	3.504673	5.359631
20200521	19	49950.0	4.439252	5.354272
20200522	15	48750.0	3.504673	5.225640
20200525	7	48850.0	1.635514	5.236360
20200526	34	49250.0	7.943925	5.279237
20200527	30	49900.0	7.009346	5.348912
20200528	8	50400.0	1.869159	5.402508
20200529	9	50700.0	2.102804	5.434666

(5) 관심도, 카테고리 데이터 통계

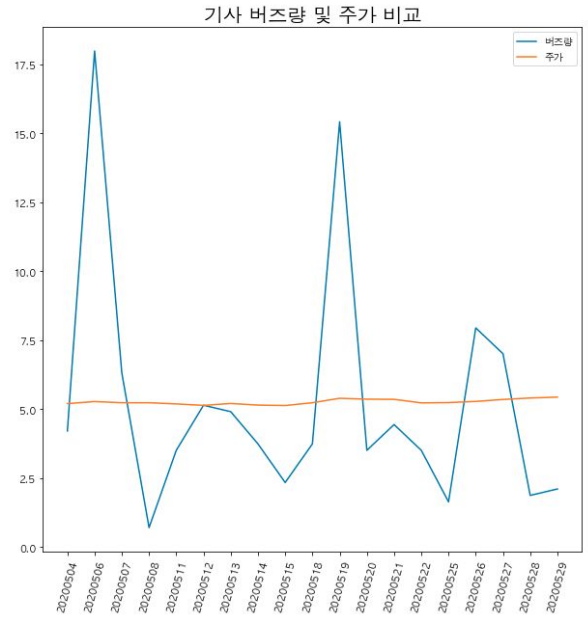
1. 삼성전자기사 날짜별 관심도 정리
2. 삼성전자기사 카테고리별합계

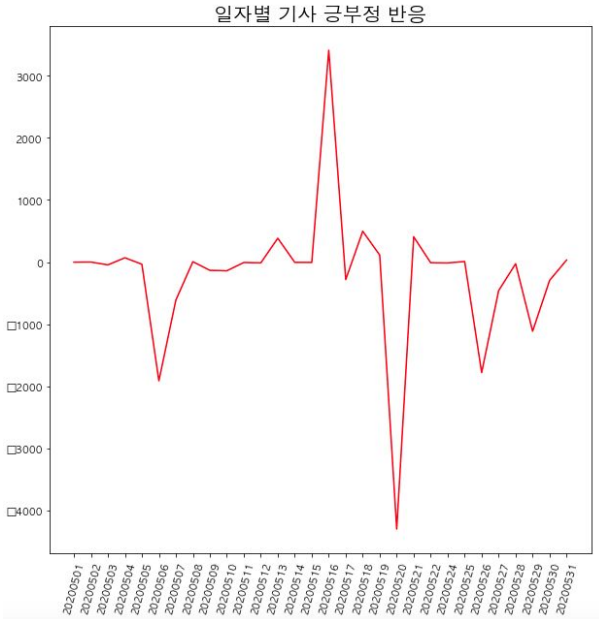
```
날짜
20200501     -2
20200502      0
20200503    -45
20200504     69
20200505    -35
20200506  -1912
20200507   -612
20200508      6
20200509   -133
20200510   -140
20200511     -7
20200512    -13
20200513    384
20200514     -5
20200515     -6
20200516   3411
20200517   -282
20200518    496
20200519    112
20200520  -4300
20200521    407
20200522    -11
20200524    -14
20200525      9
20200526  -1779
20200527   -460
20200528    -27
20200529  -1114
20200530   -295
20200531     33
Name: 감정, dtype: int64
```

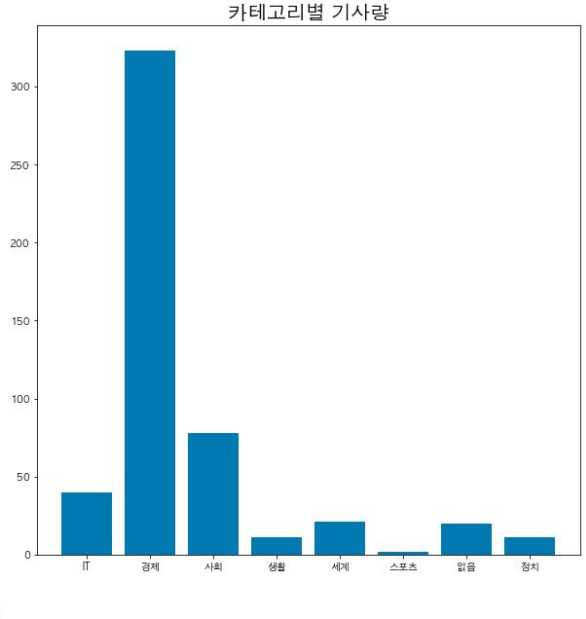
```
카테고리
IT       40
경제      323
사회       78
생활       11
세계       21
스포츠       2
없음       20
정치       11
Name: 본문, dtype: int64
```

(6) 시각화

1. 위 정리한 데이터를 기반으로 아래 그래프 작성







(7) 워드클라우드

1. 기사 버즈량이 가장 많은날 워드 클라우드
2. 관심도 반응이 가장 많은날 워드클라우드

(8-심화) 댓글 워드클라우드

1. 기사 버즈량이 가장 많은날 댓글 수집 후 워드 클라우드 (기사별 최대 100개씩)
2. 관심도 반응이 가장 많은날 댓글 수집 후 워드클라우드 (기사별 최대 100개씩)