뉴스토픽 발생량을 LDA 토픽모델링으로 거래량 예측 모델 논문구현

High Quality Topic Extraction from Business News Explains Abnormal Financial Market Volatility

YoonSeok_Choi

목차

- 주제 선정의 이유
- 프로젝트 라인
- 프로젝트 결과
- 데이터 수집
- 데이터 LDA 토큰화
- 데이터 활용
- LASSO 회귀분석
- 제한점 및 향후 개선 방향
- 참고문헌

주제 선정의 이유

프로젝트 정의 : 논문을 기반으로 기업 '에스원'의 변동성에 대한 토픽분석을 수행하는 것이 목표이다.

논문 내용 : 비즈니스 뉴스에서 고품질 주제(토픽)를 추출하여, 비정상적인 주식시장 변동성(거래량)을 설명하는 내용이다.

주제 선정 이유

01 2023 다발적 흉기 난동 사태

2023년 7월부터 이어진 대한민국의 흉기 난동 사태가 발생 1회 성으로 끝나지 않고, 인터넷을 통한 추가적인 범행 예고 CCTV를 통해 범행 현장의 언론 보도와 실제 사건 발생에 결과적으로 국민 혼란과 불분명한 여론이 확산되었다.

이때 반사 이익을 보았을 거라 예측한 기업과 뉴스 보도 간 주식시장 변동성을 설명한다.

가설

01 기업 매출액 상승

전국적인 CCTV 설치 여론. 보안 업체의 전반적인 실적 상승을 예상함.

02 모회사 후광효과

에스원의 삼성그룹 지분 31.6% 안정적인 모기업의 후광 효과 기대

SAMSUNG 에스원

프로젝트 라인

자료수집

기간 : 2018. 10. 24 ~ 2023. 10. 24
기술 : python(VScode, google colab)
대상 : 에스원 거래량 및 뉴스
sk쉴더스+kt텔레캅+하이트론+인콘+ITX+아이디스 뉴스 데이터

LDA 토픽화

형태소 분석 ► 정규표현식 ► 한글자 제외 ► 불용어 처리 ► 시각화

Komoran Hannanum Mecab Okt	특정 형태 제거 ‘ ‘ 형식, [ ] 형식, 숫자, E-mail, 출판사 등- 제거	len(word) > 1	불필요한 단어 약 248개 제거

데이터 활용

가장 많은 토픽의 선정, 연간 거래량과의 비교 분석

LassoCV

미래 거래량에 영향을 미치는 토픽 예상

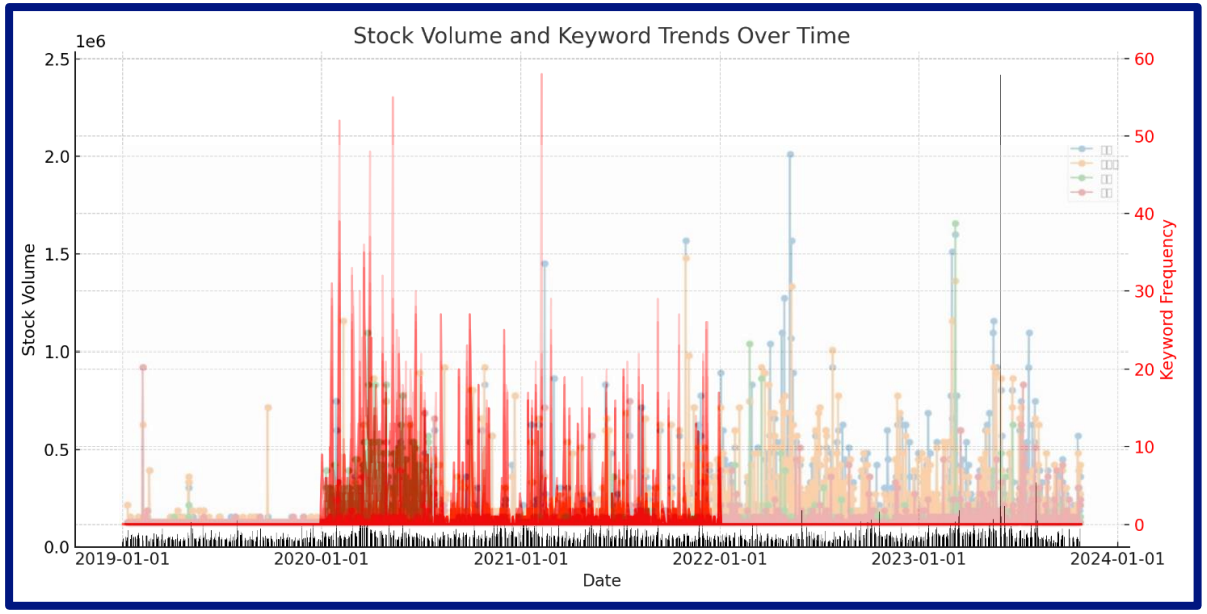
프로젝트 결과

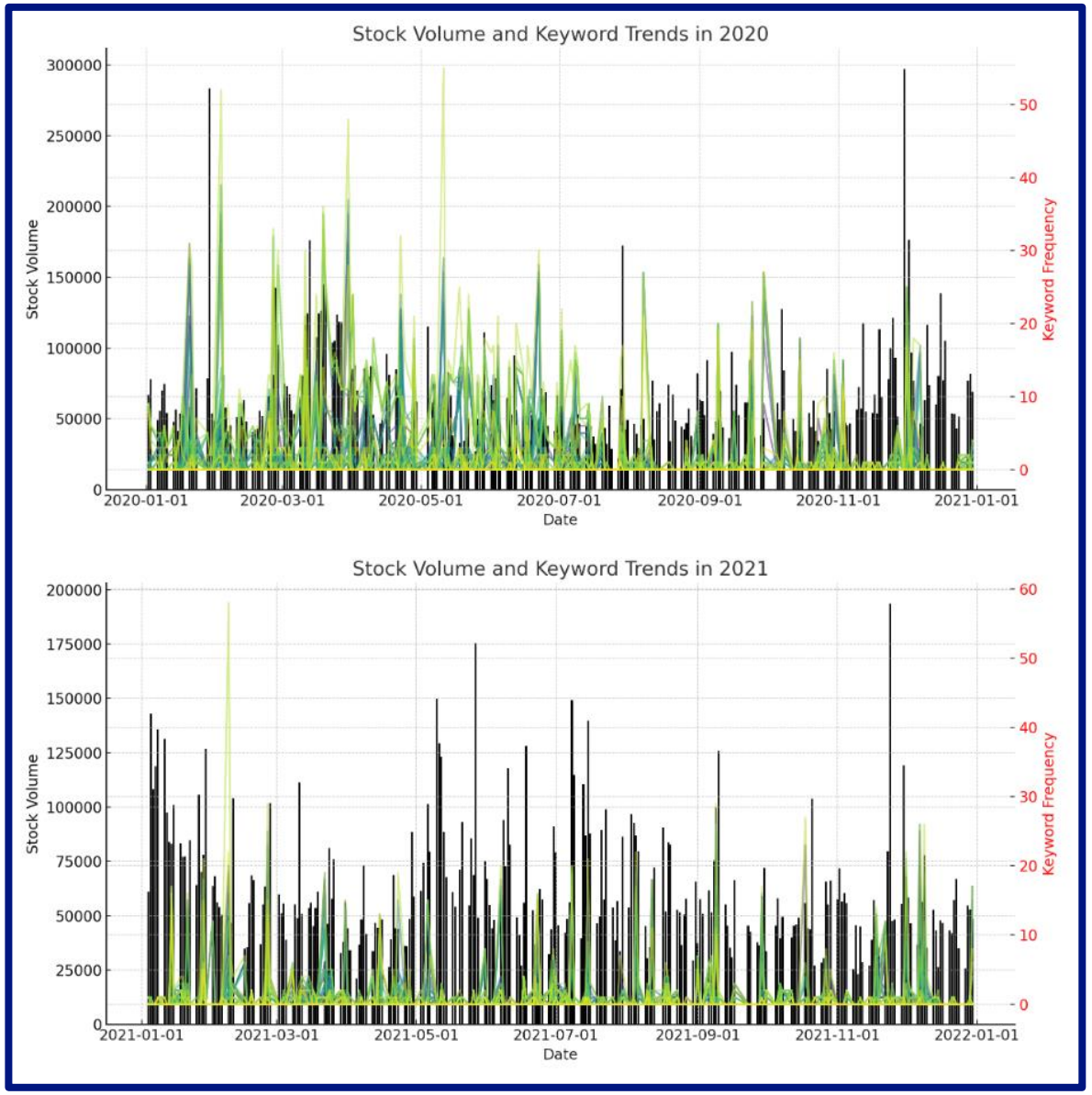
▶ 에스원의 거래량 변화에 '23년 다발적 흉기 난동 사태', '안전','보안' 이슈는 포함되지 않았다.

에스원의 실적은 19년부터 유의미하게 증가하였다, 23년 실제 CCTV 설치율의 증가, 뉴스 보도와 거래량의 증가 그러나 거래량 변화에 유의미한 토픽이 아니었다.

▶ 에스원의 모회사 삼성의 거래량 변동과 비슷한 동향을 보여 주었다.

거래량에 영향을 미치는 토픽은 '보안', '서비스', '무인', '거래', '채용' 과 같은 경제와 관련된 이슈였다.

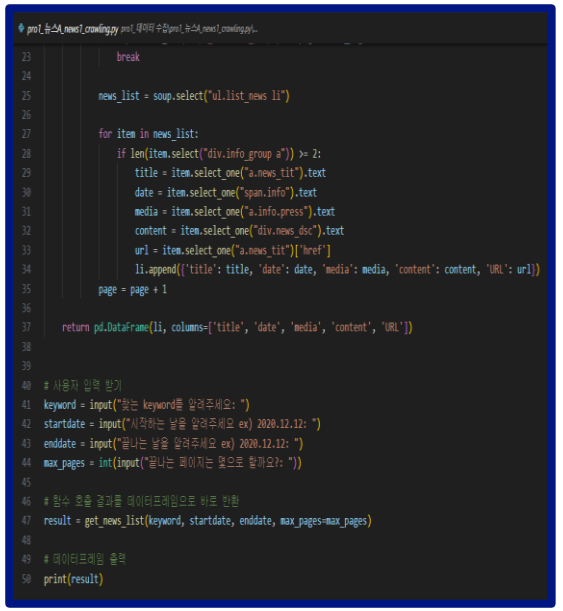




데이터 수집

함수	기능
get_news_list	**수집을 위한 항목을 지정** Keyword : 검색할 뉴스의 키워드 Startdate : 검색 시작 날짜 (형식: yyyy.mm.dd) Enddate : 검색 끝 날짜 (형식: yyyy.mm.dd) max_pages : 최대로 검색할 페이지 수
pandas.date_range	날짜별로 뉴스를 검색하기 위해 사용 지정된 시작 날짜부터 끝 날짜까지의 날짜 범위를 생성
requests.get	지정된 URL로부터 HTML 페이지를 가져오기 위해 사용 이 프로젝트에는 네이버 뉴스 검색 결과 페이지만을 요청
BeautifulSoup	HTML 문서를 파싱하고, 문서 내에서 데이터를 추출하거나 조작하기 위해 사용
pd.DataFrame	pandas DataFrame을 생성 지정한 키워드에서의 뉴스 제목, 날짜, 본문, URL을 정리

```python
import requests
import pandas as pd
from bs4 import BeautifulSoup

def get_news_list(keyword, startdate, enddate, max_pages=50):
    li = []
    h = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrom

    for d in pd.date_range(startdate, enddate):
        str_d = d.strftime("%Y.%m.%d")
        page = 1
        print(str_d)

        while True:
            start = (page - 1) * 10 + 1
            print(page)
            URL = "https://search.naver.com/search.naver?where=news&sm=tab_pge&query={0}&sort=2&photo=0&field=4

            res = requests.get(URL, headers=h)
            soup = BeautifulSoup(res.text, "html.parser")

            if soup.select_one(".api_noresult_wrap") or page > max_pages:
                break

            news_list = soup.select("ul.list_news li")

            for item in news_list:
                if len(item.select("div.info_group a")) >= 2:
                    title = item.select_one("a.news_tit").text
                    date = item.select_one("span.info").text
                    media = item.select_one("a.info.press").text
                    content = item.select_one("div.news_dsc").text
```



```python
                break

            news_list = soup.select("ul.list_news li")

            for item in news_list:
                if len(item.select("div.info_group a")) >= 2:
                    title = item.select_one("a.news_tit").text
                    date = item.select_one("span.info").text
                    media = item.select_one("a.info.press").text
                    content = item.select_one("div.news_dsc").text
                    url = item.select_one("a.news_tit")['href']
                    li.append({'title': title, 'date': date, 'media': media, 'content': content, 'URL': url})
            page = page + 1

    return pd.DataFrame(li, columns=['title', 'date', 'media', 'content', 'URL'])


# 사용자 입력 받기
keyword = input("찾는 keyword를 알려주세요: ")
startdate = input("시작하는 날을 알려주세요 ex) 2020.12.12: ")
enddate = input("끝나는 날을 알려주세요 ex) 2020.12.12: ")
max_pages = int(input("끝나는 페이지는 몇으로 할까요?: "))

# 함수 호출 결과를 데이터프레임으로 바로 반환
result = get_news_list(keyword, startdate, enddate, max_pages=max_pages)

# 데이터프레임 출력
print(result)
```

에스원 “무인매장 절도 절반은 10대, 추석연휴 앞두고 주말·새벽 대비해야”

이정구 기자

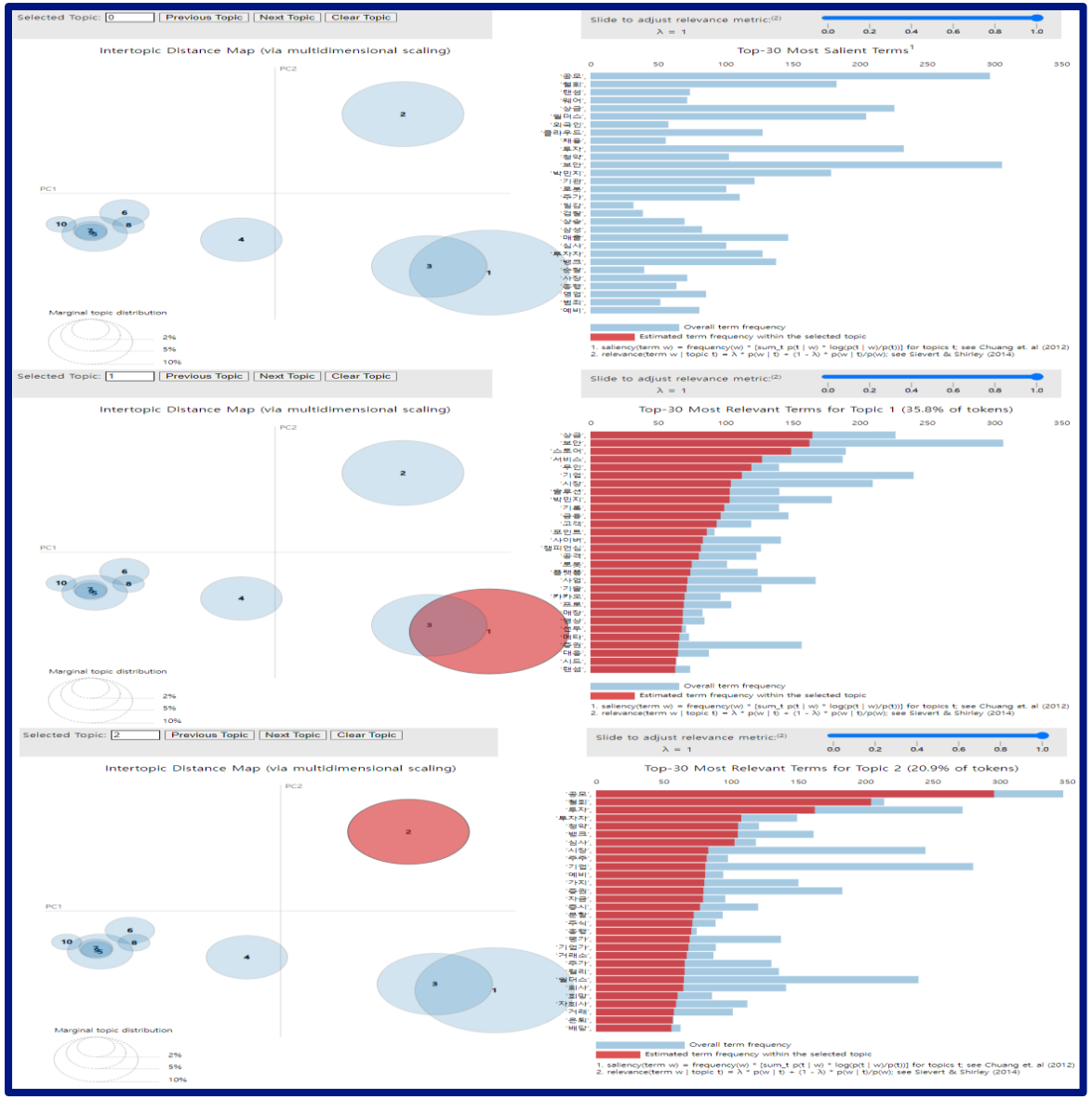
업데이트 2023.09.25. 10:34

코로나 팬데믹 기간 급증한 무인매장을 노리는 범죄는 주말과 심야 시간 집중된 것으로 조사됐다. 무인 매장을 노린 절도 범죄자는 10대가 가장 많았고, 주로 매장 안 현금을 노린 경우가 많았다.

보안 기업 에스원은 무인매장이 본격적으로 확산한 2019년부터 올해 6월까지 무인매장 절도 범죄 동향을 분석한 결과 보고서를 25일 발표했다. 보고서에 따르면, 범죄자 연령대는 10대가 가장 많았고, 무인 매장이 범죄에 가장 취약한 시간대는 주말과 심야 시간대였다. 범죄 피해 물품은 ‘매장 내 현금’이 가장 높은 것으로 나타났다.

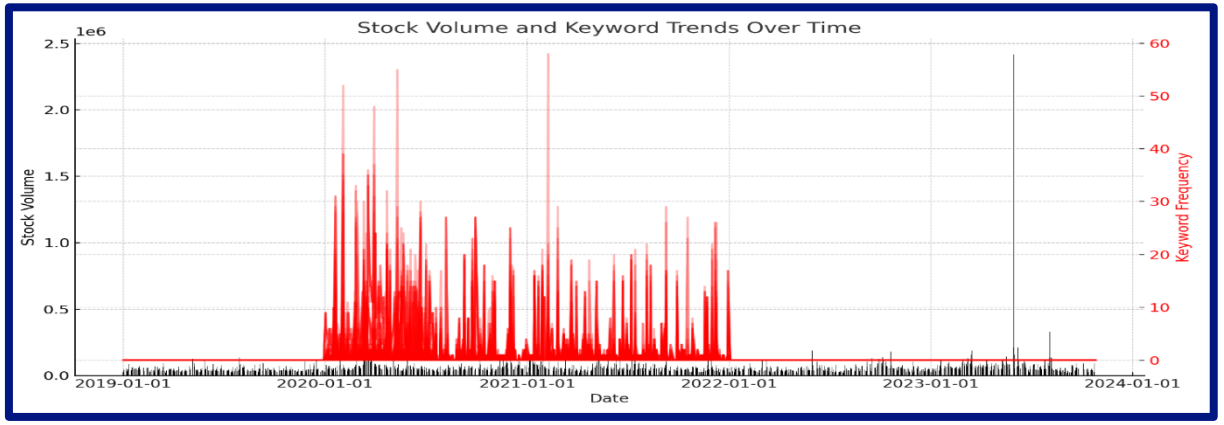
데이터 LDA 토큰화

함수	기능
from konlpy.tag import Okt	**형태소 분석** 한국어 텍스트인 뉴스 기사의 내용을 형태소 단위로 분석
import re regex_pattern	**정규표현식** 텍스트에서 HTML 태그, 이메일 주소, 특정 날짜 형식, 저작권 기호 등 불필요한 정보를 제거
if len(word) > 1	**한글자 제거** 한 글자로 된 단어는 대부분 중요한 정보를 담고 있지 않다고 가정하고 제거
stop_word = ["불용어", ...] def preprocess(text):	**불용어 처리** 분석에 필요하지 않은 특정 단어들(불용어)를 제거
def make_tokens(df):	**토큰리스트 생성** 전처리 된 텍스트 결과를 다시 문자열로 저장
def make_tokens(df): for i, row in df.iterrows():	**Gensim 패키지** 각 뉴스 기사의 내용을 분석에 더 적합한 형태로 가공
import pyLDAvis	**시각화** LDA(잠재 디리클레 할당) 모델을 사용한 토픽 모델링 결과를 시각화



데이터 활용

뉴스 토픽 시각화

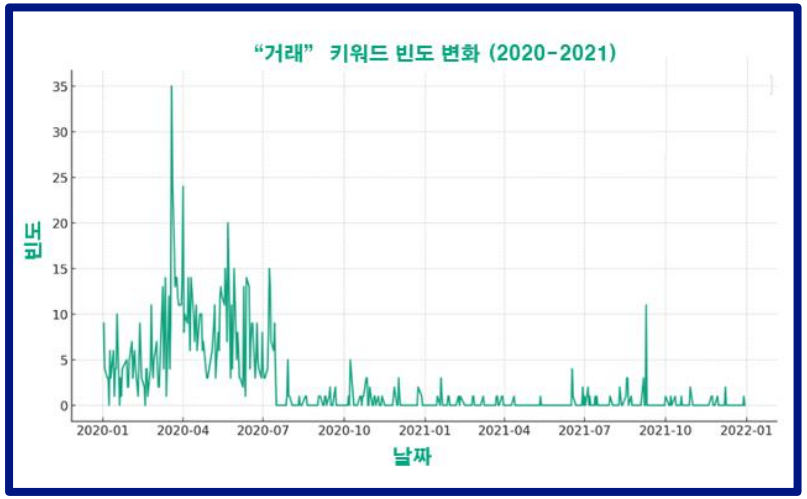


경제 관련 뉴스 발생량과 거래량

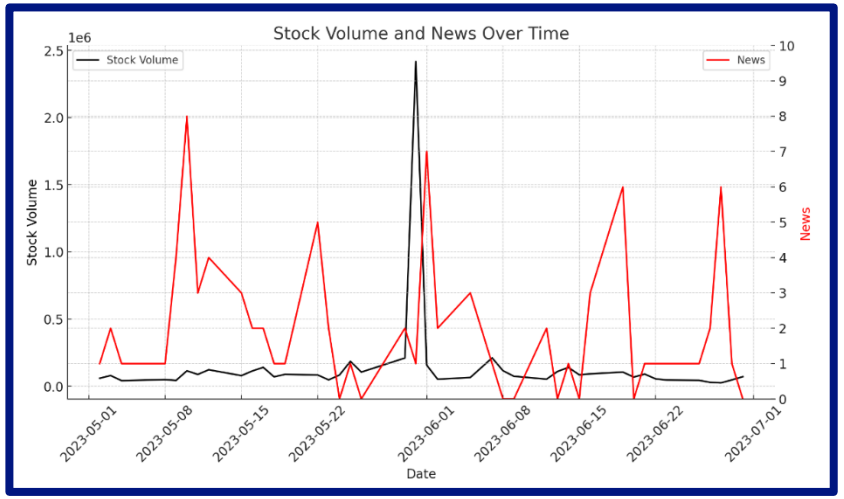
보안	서비스	무인	케어	채용	연속	대회	유해
187	100	69	58	52	33	28	27
캡스	클라우드	상장	안전	매장	텔레콤	교육	일감
27	23	20	20	19	15	14	9
스토어	마지막	시즌	공모	최종	선수	뱅크	로봇
4	4	2	2	2	1	1	1

보안	서비스	무인	케어	디지털	사업자	범죄	캡스
187	100	69	58	46	39	32	27
고객	프로	클라우드	인상	혜택	안전	피싱	영상
26	23	23	23	22	20	18	11
롯데	하반기	콘텐츠	하락	누구	인수	유치	로봇
10	10	9	8	3	2	2	1

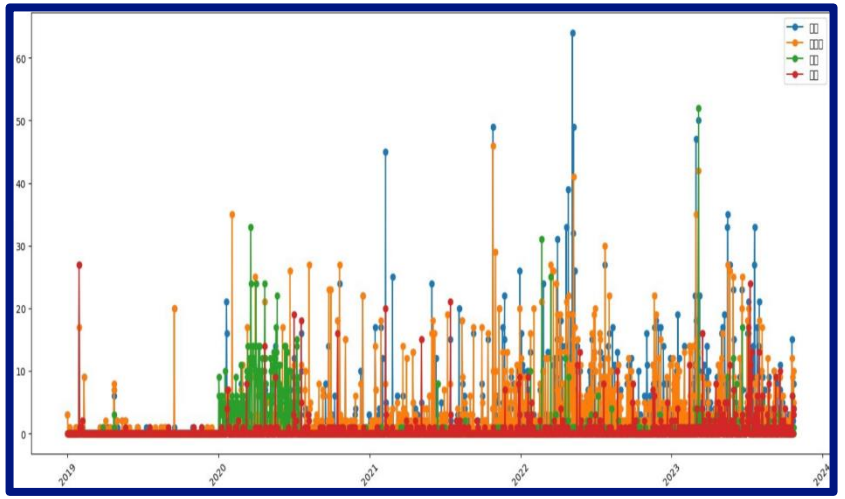
전체 토픽의 비율



20년~ 21년 특정 키워드 발생량



23년 뉴스 발생량과 거래량

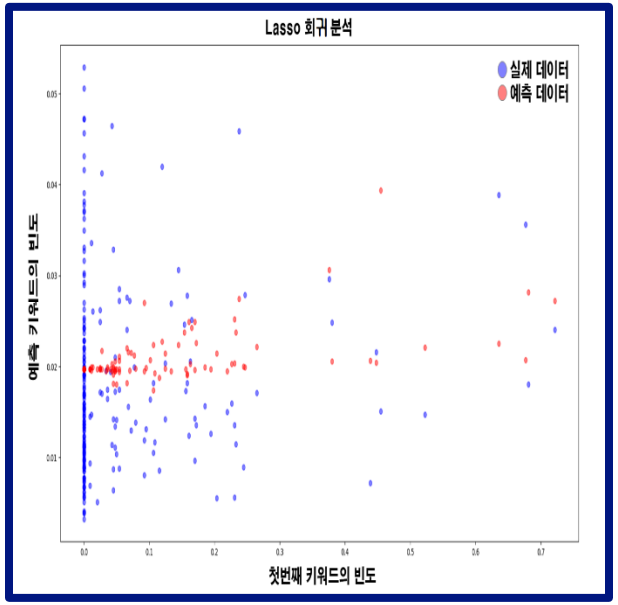


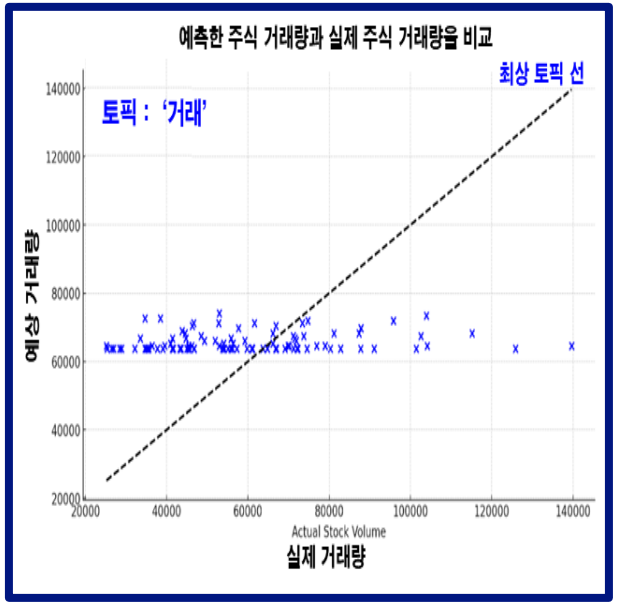
상위 토픽 키워드의 발생비율

LASSO 회귀분석

함수	기능
pd.to_datetime():	**날짜형식 변경** 문자열 형태의 날짜 데이터를 pandas의 DateTime 형식으로 변환
MinMaxScaler():	**데이터 정규화** 데이터의 특성을 0과 1 사이의 범위로 스케일링 (정규화)하여 모델 성능 향상
train_test_split():	**데이터를 세트 분리** 훈련세트와 데이터 세트 분리 test_size=0.2 : 테스트 세트 20% 설정
LassoCV():	**알파 값 교차 검증 수행** 최적의 알파(정규화 강도) 값 교차검증 cv=cv 매개변수 : 차 검증의 폴드 수를 지정
Lasso():	**Lasso 회귀 모델 설정과 훈련** L1 규제를 사용하여 일부 회귀 계수를 정확히 0으로 만들어, 변수 선택의 효과를 가지며, 과대적합을 방지
mean_squared_error() mean_absolute_error() r2_score():	**모델의 성능(예측 정확도)을 평가** MSE(mean squared error) : 평균 제곱 오차 MAE(mean absolute error) : 평균 절대 오차 R^2(R-squared) : 결정 계수

```python
import pandas as pd
import numpy as np
from sklearn.linear_model import LassoCV, Lasso
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt

file_path = r'C:\Users\HOME\Desktop\새싹_교육\GitHub_CHOI\project_1_Relevance-between-news-topics-and-tradin

data = pd.read_csv(file_path)

# 데이터 전처리
data['Stock_Volume'] = data['Stock_Volume'].str.replace(',', '').astype(int)
data['Date'] = pd.to_datetime(data['Date'])

# MinMaxScaler를 사용한 데이터 표준화
scaler = MinMaxScaler()
X = data.drop(['Date', 'Stock_Volume'], axis=1)
y = data['Stock_Volume']
X_scaled = scaler.fit_transform(X)
y_scaled = scaler.fit_transform(y.values.reshape(-1, 1))

# 데이터 분할
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y_scaled, test_size=0.2, random_state=42)

# 라쏘 회귀 모델 설정 및 훈련
cv = 100
lasso_cv = LassoCV(cv=cv, random_state=0)
lasso_cv.fit(X_train, y_train.ravel())
best_alpha = lasso_cv.alpha_

# 라쏘 모델로 예측
lasso_model = Lasso(alpha=best_alpha)
lasso_model.fit(X_train, y_train.ravel())
```





제한점 및 향후 개선 방향

▪ 제한점

1. 불용어처리한 단어 중 이상 거래량을 나타내는 토픽이 있을 가능성
2. 구분한 토픽 단어가 거래량을 완전히 대변하지 못함.
3. 지정한 기업의 주도적인 활동이 거래량에 미치는 영향이 적음.

▪ 향후 개선 방향

1. 데이터의 수집량을 늘려 이상 거래량을 보이는 데이터를 탐색
2. 대상 기업을 3개 이상 지정해 동일 증가, 동일 감소를 탐색

참고문헌

참고	세부
기술적 지원	자연어처리 교제, 파이썬 기초 교제, ChatGPT
데이터 수집	VScode, google colab
시각화 작업	PPT, pyLDAvis, KCC-한빛. 서평원 꺾깎체 TTF
참고 웹페이지	팀원 GitHub # https://github.com/kangsik10 # https://github.com/eminoart # https://github.com/heedahan16