[23-2] KUBIG CONFERENCE

국내/해외주식 데이터를 활용한 숨은 투자 기회 찾기

금융 팀

목차

1.주제 설명

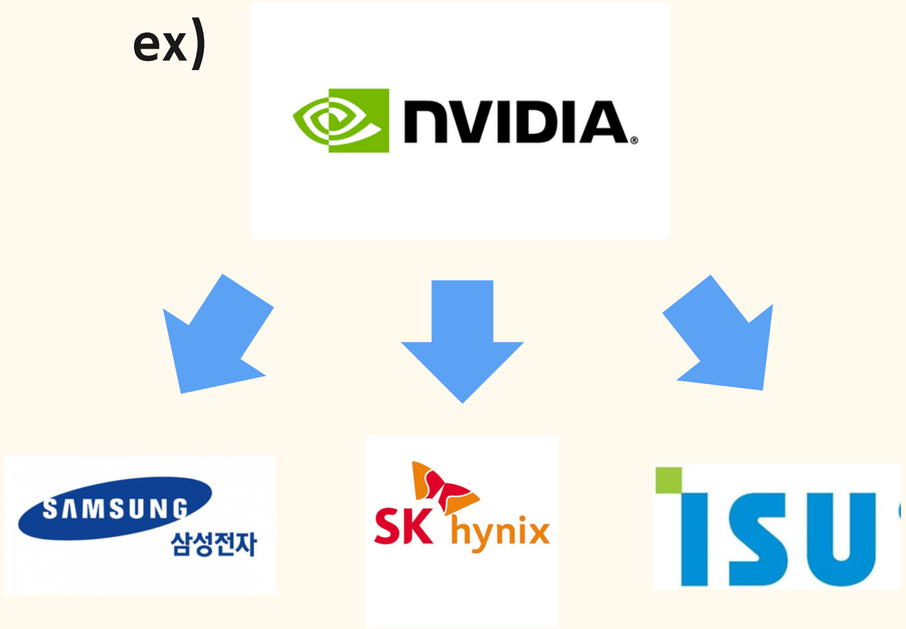
CONTENTS

소개

CONTENTS

동향

ex)



“동조화 현상‘

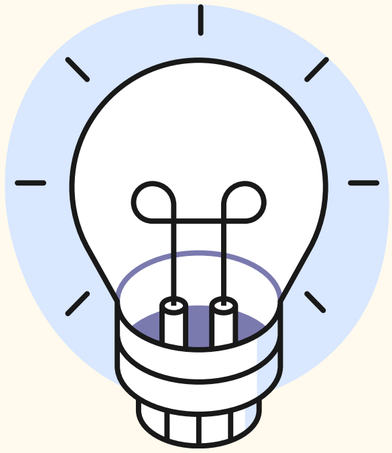
미국 시장 → 국내 시장

“금융권 DT“

비정형 데이터 분석

CONTENTS

“데이터 속에 숨은 투자 기회”



1. 주식 시세 데이터를 활용한 국내/해외 종목 관계 분석

2. 뉴스 데이터를 이용한 감성분석

→ 해외 주식 기반 국내 주식 종목 추천 시스템

2. 데이터 소개

CONTENTS

데이터 소개 – NH 제공

NASDAQ_FC_STK_IEM_IFO

2023년 미국 나스닥 거래소에서 시세를 제공하는 주문 가능한 종목 정보

ISIN_IEM_CD : ISIN 코드

TCK_IEM_CD : 종목 티커 코드

FC_SEC_KRL_NM : 해외주식 종목 한글명

FC_SEC_ENG_NM : 해외주식 종목 영문명.

NASDAQ_DT_FC_STK_QUT

2023년 종목의 시세 정보

TRD_DT : 거래일자

TCK_IEM_CD : 티커종목코드

IEM_ONG_PR : 종목시가

IEM_HI_PR : 종목고가

IEM_LOW_PI : 종목저가

IEM_END_PR : 종목종가

ACL_TRD_ATY : 누적거래수량

SLL_CNS_SUM_QTY : 매도체결합계수량

BYN_CNS_SUM_QTY : 매수체결합계수량

NASDAQ_RSS_IFO

RGS_DT : 발행일자

TCK_IEM_CD : TCK_IEM_CD

TIL_IFO : 제목정보

CTGY_CFC_IFO : 카테고리분류정보

MDI_IFO : 미디어정보

NEWS_SMY_IFO : 뉴스요약정보

RLD_OSE_IEM_TCK_CD : 관련해외종목티커코드

URL_IFO : URL정보

INTERVIEW
PRESENTATION

- 소개
- 지원직무
- 경력 및 역량
- 수행프로젝트
- 향후 계획

데이터 크롤링 – 뉴스 원문

	title	date	category	key_points	text	url
0	'I work just 5 hours a week': A 39-year-old wh...	2023-01-01	Success	N/A	Graham Cochrane, Founder of The Recording Revo...	https://www.cnbc.com/2023/01/01/39-year-old-wh...
1	Chinese state media seek to reassure public ov...	2023-01-01	Asia-Pacific News	Chinese state media sought to reassure the pub...	Revelers prepare to release balloons to celebr...	https://www.cnbc.com/2023/01/01/chinese-state-...
2	Should you get creative with your resume? Expe...	2023-01-01	Land the Job	N/A	Mature businessman congratulating young profes...	https://www.cnbc.com/2023/01/01/cv-will-a-crea...
3	Market misery deals sovereign wealth funds his...	2023-01-01	Markets	Heavy falls in stock and bond markets over the...	A trader works on the floor of the New York St...	https://www.cnbc.com/2023/01/01/market-misery-...
4	More social media regulation is coming in 2023...	2023-01-01	Tech	Days after Congress passed a bipartisan spendi...	The U.K.'s Online Safety Bill, which aims to r...	https://www.cnbc.com/2023/01/01/more-social-me...

NEWSPAPER LIBRARY 활용하여 기사본문 크롤링

+

selenium & bs4를 활용하여 미국 경제 뉴스 "CNBC" 크롤링

INTERVIEW
PRESENTATION

데이터 소개 – 외부 크롤링

국내 주식 시세 데이터 – KRX 정보데이터

http://data.krx.co.kr/contents/MDC/MAIN/main/index.cmd

date	A098120	A009520	A095570	A006840	A282330	A027410	A138930
2023-01-02	5120.0	7710	5720.0	16250	202000.0	4115.0	6300.0
2023-01-03	5160.0	7740	5790.0	16200	199000.0	4100.0	6340.0
2023-01-04	5480.0	7760	5760.0	16250	197500.0	4155.0	6550.0
2023-01-05	5570.0	7540	5760.0	16400	192000.0	4230.0	6720.0
2023-01-06	5670.0	7570	5720.0	16050	191500.0	4225.0	6790.0

재무제표 데이터

yahoo! finance

전자공시시스템 DART

NASDAQ 종합지수

date	
2023-01-03	10386.99
2023-01-04	10458.76
2023-01-05	10305.24
2023-01-06	10569.29
2023-01-09	10635.65

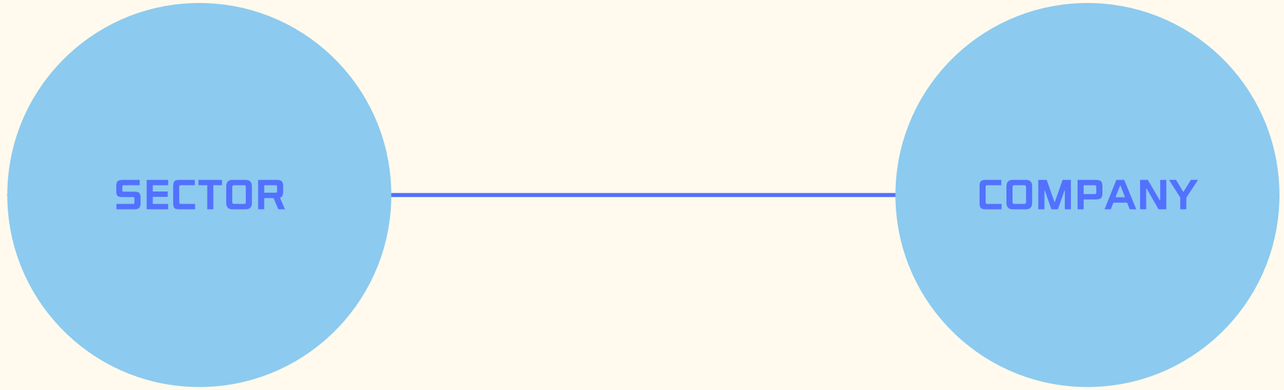
나스닥 외화주식 종목 정보 : NASDAQ.COM

https://www.nasdaq.com/market-activity/stocks/screener

3. 전처리

INTERVIEW
PRESENTATION

전처리 목표



HOT TOPIC 분야 찾기

관련 유망 기업 찾기

INTERVIEW
PRESENTATION

- 소개
- 지원직무
- 전처리
- 수행프로젝트
- 향후 계획

뉴스 데이터 자연어처리

CNBC 경제 뉴스 사이트에서 인기 있는 기사 토픽을 먼저 파악하고자 함

*월에 5번도 언급되지 않은 카테고리는 주가 분석에 있어
중요하지 않은 기사라고 판단해 삭제함.

```
# 월 별로 5번 이상 등장한 카테고리 찾기
valid_categories = monthly_counts[monthly_counts['count'] >= 5]['category'].unique()
```

토큰화 및 불용어 제거

- 크롤링 데이터의 특수문자와 문장부호 제거
- 불용어 제거(n차 토픽모델링 후 후처리 진행)
- 불용어 제거 및 토큰화 : 성능 면에서 nltk보다 **spacy**가 우수
- 긴 텍스트(기사 본문)를 처리해야 하므로 en_core_web_lg 모델을 사용했다.
- 또한 tokenizer 함수를 생성할 때는 명사만 추출하도록 했으며, 개체명은 'TIME','CARDINAL','DATE'을 제외한 모든 entity를 사용했다.

```
token.ent_type_ not in ['TIME','CARDINAL','DATE']:
```

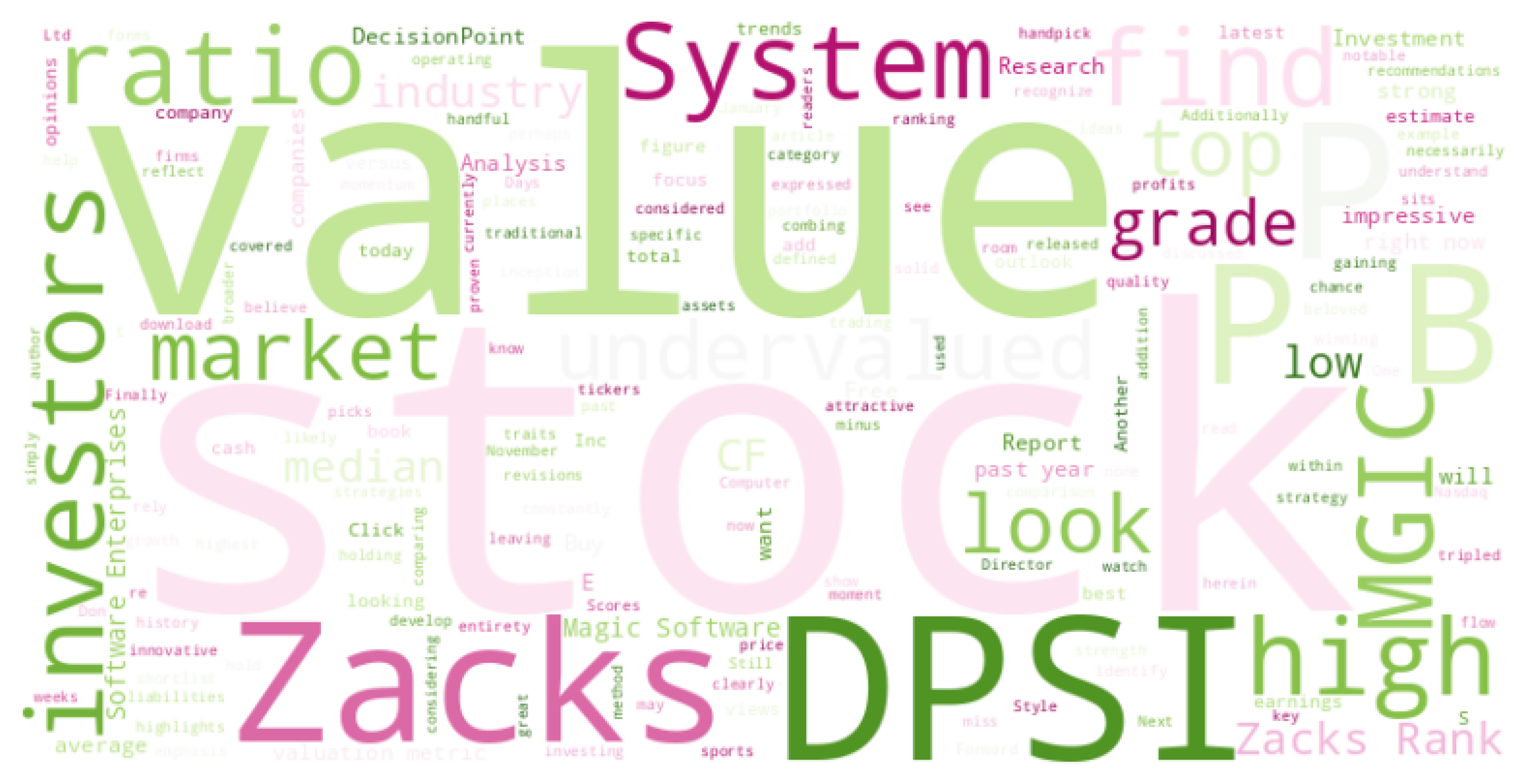
spaCy

INTERVIEW
PRESENTATION

- 소개
- 지원직무
- 경력 및 역량
- 수행프로젝트
- 향후 계획

워드클라우드 – 키워드 추출

워드 클라우드를 활용한 뉴스별 주요 단어 확인하기 : 감성분석을 적용하기 전 크롤링



INTERVIEW
PRESENTATION

주가 데이터 전처리

```python
# 기간 중 거래정지 종목 제외
trd_stop = (all_krx_adj_open == 0).sum()
trd_stop_stocks = trd_stop[trd_stop>0].index.values
trd_stop_stocks_cnt = trd_stop_stocks.shape[0]

cond_stocks = np.setdiff1d(period_listed_stocks, trd_stop_stocks)
cond_stocks_cnt = cond_stocks.shape[0]

print(f'현재 상장 종목: {all_stocks_cnt}')
print(f'- 기간 중 상장 및 폐지 종목: {all_stocks_cnt - period_listed_stocks_cnt}')
print(f'- 기간 중 거래정지 종목: {trd_stop_stocks_cnt}')
print(f'제거율: {1 - cond_stocks_cnt/all_stocks_cnt:.2%}')
print()
print(f'분석 대상 종목: {cond_stocks_cnt}')
```

```
현재 상장 종목: 2754
- 기간 중 상장 및 폐지 종목: 103
- 기간 중 거래정지 종목: 289
제거율: 13.91%

분석 대상 종목: 2371
```

INTERVIEW
PRESENTATION

성과지표(Techinical Indicators)

주식의 수익률을 측정하기 위한 지표가 필요하여 직접 구현

$$Sharpe = \frac{\mu_p - r_f}{\sigma_p}$$

$$Sortino = \frac{\mu_p - r_f}{D\sigma_p}$$

$$Calmar = -\frac{\mu_p - r_f}{MDD_p}$$

$$VaRRatio = -\frac{\mu_p - r_f}{N * VaR_{\delta,p}}$$

4. 유망산업&기업 분석

토픽 모델링

```
# 모델 생성
topic_model = create_topic_model(embedding_model, umap_model, hdbscan_model, vectorizer_model, representation_model)
```

- 임베딩 모델 선정 기준 : sbert.net의 sentencetransformer 중 가장 Performance가 높은 모델(all-mpnet-base-v2) 선정
- 5배나 빠른 속도에 정확도가 높은 all-MiniLM-L6-v2로도 시도해봤지만 성능이 좋지 않았음.

```
embedding_model = SentenceTransformer("all-mpnet-base-v2") #임베딩 모델
embeddings = embedding_model.encode(all_texts, show_progress_bar=True) #임베딩 미리 계산(파라미터 수정 용이 위함)
```

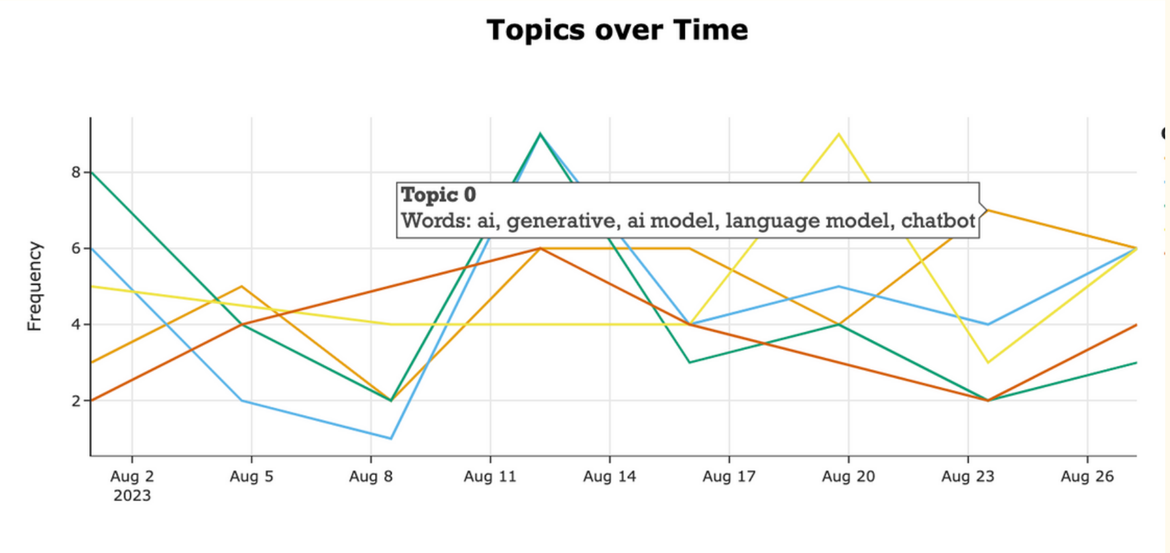
- 파라미터 튜닝 결과 아래의 파라미터로 하는 것이 가장 토픽을 잘 찾는다고 판단함.
- UMAP(n_neighbors=8, min_dist=0.1, n_components=2)
- HDBSCAN(min_cluster_size=5)
- TfidfVectorizer 사용 이유 : 단어의 빈도 뿐만 아니라, 그 단어가 전체 문서 집합에서 얼마나 중요한지를 고려하기에 토픽모델링 시에 해당 모델을 사용하는 것이 적합함.
- MaximalMarginalRelevance 사용 이유 : 토픽의 키워드를 통해 관련주를 찾아내야 하므로, 토픽 키워드를 추출하는 것이 정교해야 한다고 판단함. 또한 diversity를 0.2로 설정해 토픽과 관련된 키워드를 다소 다양하게 뽑고자 했음.

INTERVIEW
PRESENTATION

- 소개
- 지원직무
- 경력 및 역량
- 수행프로젝트
- 향후 계획

모델링 시각화 & 키워드 제시

ex) AI 테마 키워드



Representation
[bitcoin, resume, oracle, market, sec, etf, ap...
[analyst refinitiv, analyst, cramer, price tar...
[ai, chatgpt, google, ai model, openai, chatbo...
[retailer, walmart, foot locker, merchandise, ...
[pfizer, vaccine, pharmacy, medication, obesit...
[cnn, disney, microsoft, activision, espn, sal...
[election, president donald, indictment, case,...
[iphone, apple, ipad, smartphone, apple iphone...
[playlist, schwartz, taylor, feedback, billion...
[china, beijing, chinas, economist, yuan, peop...
[twitter, musk, app, meta, elon musk, fda, bot...
[happiness, gate, harvard, brain, lifestyle, p...
[oil, vessel, port, sailing, coast, gulf, ocea...
[inflation, ecb, european central, rate hike, ...
[climate, carbon, emission, gigawatt, climate ...
[debt, expense, finance, spending, survey, car...
[xpeng, yuan, malaysia, toyota, tesla, volkswa...
[rent, city, new york, housing, cost living, m...
[hayes, menu, saudi, restaurant, breakfast, or...
[treasury, treasury yield, fed, inflation, bas...

INTERVIEW
PRESENTATION

- 소개
- 지원직무
- 경력 및 역량
- 수행프로젝트
- 향후 계획

연관성 높은 기업 필터링

```python
# 토픽 관련 기업 찾기
def retrieve_companies_by_keywords(keywords):
    keywords_set = set([word.lower() for word in keywords])

    # 기업 리스트
    cp = []

    for i, row in stock_info_df.iterrows():
        description = row['description']

        #단어 추출
        if isinstance(description, str):
            description_words = set(description.replace(",", '').lower().replace('.', '').split('

            # 토픽 키워드와 description이 겹치는 기업 찾기
            if description_words & keywords_set:
                cp.append(row['tck_iem_cd'])

    return list(set(cp))
```

```python
# 키워드 언급횟수 count
def count_companies_by_keywords(keywords):
    keywords_set = set([word.lower() for word in keywords])

    # 키워드 언급횟수 딕셔너리 생성
    keyword_counts = {word: 0 for word in keywords_set}

    for i, row in stock_info_df.iterrows():
        description = row['description']

        if isinstance(description, str):
            description_words = set(description.replace(",", '').lower().replace('.', '').split(" "))

            # 키워드 횟수 카운트
            for word in keywords_set:
                if word in description_words:
                    keyword_counts[word] += 1

    for keyword, count in keyword_counts.items():
        print(f"{keyword}: {count} companies")
```

```
computings: 0 companies
chatgpts: 0 companies
advanced micro: 0 companies
ai model: 0 companies
generative ai: 0 companies
micro device: 0 companies
chatbots: 0 companies
processing units: 0 companies
googles: 0 companies
openais: 0 companies
gpu: 1 companies
gpus: 1 companies
ais: 0 companies
chatgpt: 1 companies
ai models: 0 companies
aw: 0 companies
vmwares: 0 companies
chatbot: 0 companies
graphic processings: 0 compani
computing: 52 companies
czech: 4 companies
amds: 0 companies
advanced micros: 0 companies
language model: 0 companies
micro devices: 0 companies
graphic processing: 0 companie
processing unit: 0 companies
vmware: 0 companies
amd: 3 companies
openai: 0 companies
language models: 0 companies
google: 9 companies
generative ais: 0 companies
czechs: 0 companies
ai: 40 companies
aws: 2 companies
```

INTERVIEW
PRESENTATION

XGBOOST 기반 기업 Selection

dmlc XGBoost

- XGBoost를 사용한 이유 : 해당 데이터에는 결측치가 많아 결측치를 효과적으로 처리하는 모델이 필요함. 또한 분류 모델 중 가장 성능이 높다고 알려진 XGBoost로 해당 분류를 진행함.
- 재무제표 데이터, 성과지표 등 정량 데이터를 활용해 8월 대비 9월의 평균 주가가 오를 것으로 예상되는 기업을 추출함.

```
# 현금 창출, 매출 관련 (ttm)
'totalRevenue': '총매출액',
'grossProfits': '매출총이익',  # 매출이익(매출액 - 매출원가)
'revenuePerShare': '주당매출액',
'ebitda': 'EBITDA',  # 감가상각 등의 부가비용을 차감하기 전의 금액, 영업 활동을 통한 현금 창출 능력. 유형자산의 가치까지 포함하는 지표
'ebitdaMargins': 'EBITDA마진',  # 유형자산의 유지비용을 고려한 기업의 현금 창출 능력

# 재무 상태 관련 (mrq)
'debtToEquity': '부채자본비율',
'operatingCashflow': '영업현금흐름',  # 영업현금흐름 : 영업이익 - 법인세 - 이자비용 + 감가상각비
'freeCashflow': '잉여현금흐름',  # 기업의 본원적 영업활동을 위해 현금을 창출하고, 영업자산에 투자하고도 남은 현금
'totalCashPerShare': '주당현금흐름',
'currentRatio': '유동비율',  # 회사가 가지고 있는 단기 부채 상환 능력
'quickRatio': '당좌비율',  # 회사가 가지고 있는 단기 부채 상환 능력
'overallRisk': '위험 점수',

# 경영 효율 관련
'returnOnAssets': '자기자본이익률',  # mrq : 간단히 말해, 얼마를 투자해서 얼마를 벌었냐
'returnOnEquity': '총자산순이익률',  # mrq : ROE와 비교하여 기업이 가지고 있는 부채의 비중을 볼 때
'grossMargins': '매출총이익률',  # ttm : 매출이익(매출액 - 매출원가) / 매출액 : 매출이익률, Gross Profit Margin (GPM)
'operatingMargins': '영업이익률',  # ttm : 매출총이익 - 판관비 - 감가상각비
'profitMargins': '순이익률',  # ttm : Net Income(순이익) / Revenue(총수익) : 순이익률, Net Profit Margin (NPM)
```

	Company	xgboost_prob
52	VOD	0.897710
22	INTC	0.847029
33	NICE	0.794178
17	DRS	0.750374
39	PERI	0.692716
27	LNTH	0.692540
14	CSCO	0.686896
20	GOOG	0.664316
47	SOUN	0.661067
25	KTOS	0.639091
6	AOSL	0.635895
32	NEWT	0.598093
9	CCCS	0.573763
7	APLD	0.568441
31	MSFT	0.566283
15	CTSH	0.555496
43	RGTI	0.547021
40	PLTK	0.542991
55	WDC	0.518878
45	SCSC	0.513791
36	NTAP	0.511056
29	MCHP	0.494042
54	VUZI	0.493545
16	DIOD	0.492292
53	VRNT	0.487612
3	AMD	0.487412
12	CEVA	0.483946
56	XRX	0.470001
0	AEHR	0.463278

6. 감성분석

INTERVIEW
PRESENTATION

FINBERT

- Unsupervised pretraining
- generic BERT 모델보다 강력
- corporate report / conference call transcript / analysis report 사용

- 금융 뉴스데이터에서 4,840 개의 문장 포함
- 16명의 전문지식을 갖춘 연구자들에 의해 수동 라벨링 하여 만들었음.
- 감정 라벨 : positive, neutral, negative

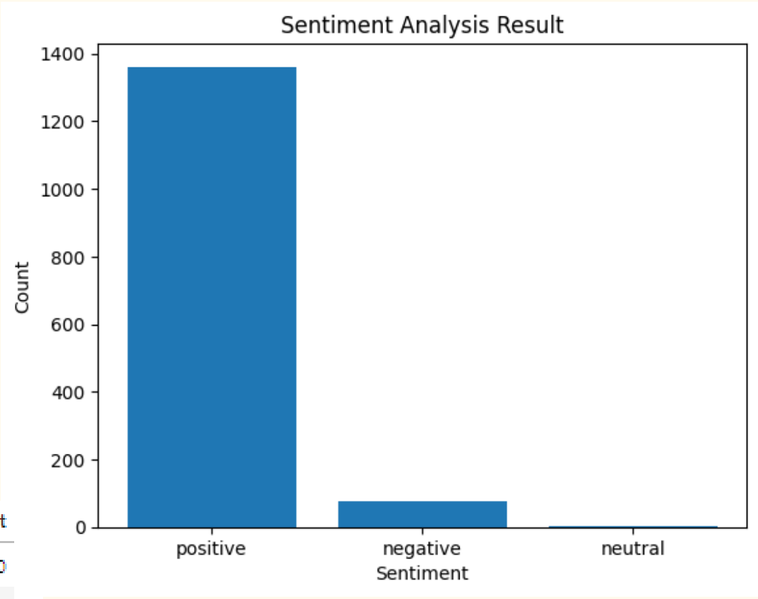
	BERT		FinBERT-BaseVocab		FinBERT-FinVocab	
	cased	uncased	cased	uncased	cased	uncased
PhraseBank	0.755	0.835	0.856	0.870	0.864	**0.872**
FiQA	0.653	0.730	0.767	0.796	0.814	**0.844**
AnalystTone	0.840	0.850	0.872	0.880	0.876	**0.887**

INTERVIEW
PRESENTATION

감성지수 도출

- XGBoost에서 선택된 기업들에, 제공된 NASDAQ 뉴스의 8월 데이터에서 FinBERT 감성분석을 실시함
- 이때, all_tck_iem_cd 열에 여러 기업이 있는 경우, 뉴스 기사에 여러 기업에 대한 언급이 포함됨
- 특정 기업에 대한 감성분석을 실시하려면 해당 기업코드 또는 기업명이 들어간 문단을 추출하는 게 적절하다고 판단함
- FinBERT를 사용한 이유는, 금융 도메인에 특화하여 pre-trained된 모델이기 때문임
- 그 중 ProsusAI의 모델을 사용한 이유는, 해당 모델이 금융 뉴스 문장들로 이루어진 Financial PhraseBank 데이터로 fine-tuning되어, 뉴스 데이터를 분석하기 적합하다고 판단했기 때문

```
기업 VOD의 기사는 6개입니다.
Mean: 0.5204 0.0188 0.4608
Count: 0.6000 0.0000 0.6000
기업 INTC의 기사는 1763개입니다.
Mean: 0.3008 0.2760 0.4217
Count: 0.3019 0.2704 0.4277
기업 NICE의 기사는 6개입니다.
Mean: 0.6676 0.0944 0.2360
Count: 0.6000 0.0000 0.2000
```



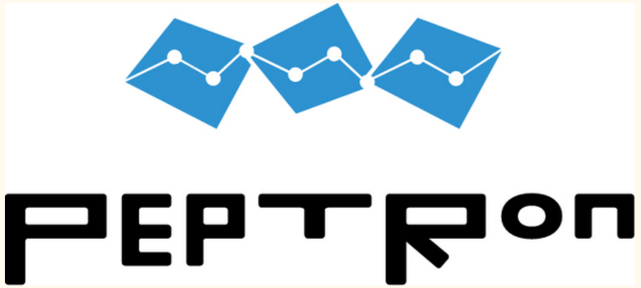
Company	Count	Score_Pos_Mean	Score_Neg_Mean	Score_Neu_Mean	Score_Pos_Count	Score_Neg_Count	Score_Neu_Count
MSFT	434	0.392048	0.147402	0.460549	0.396175	0.128415	0.475410
AMD	226	0.369931	0.331374	0.298695	0.398058	0.325243	0.276699
GOOG	184	0.315841	0.160963	0.523195	0.289855	0.152174	0.557971
INTC	176	0.300308	0.277963	0.421729	0.301887	0.270440	0.427673
CSCO	111	0.351525	0.175900	0.472576	0.323810	0.161905	0.514286
PERI	15	0.287211	0.121754	0.591035	0.214286	0.071429	0.714286
SOUN	14	0.433710	0.146551	0.419739	0.500000	0.142857	0.357143
RGTI	10	0.512812	0.308145	0.179043	0.500000	0.400000	0.100000
NEWT	8	0.593522	0.106689	0.299788	0.625000	0.125000	0.250000
AEHR	7	0.404243	0.148987	0.446769	0.285714	0.142857	0.571429

INTERVIEW
PRESENTATION

최종 결과물

```
2023-07-26: ['A010640' 'A032560' 'A042940' 'A045300' 'A060310' 'A127120' 'A317530']
2023-07-27: ['A002710' 'A007770' 'A256610' 'A266490' 'A317770' 'A355150' 'A357550']
2023-07-28: ['A007770' 'A039200' 'A064440' 'A256610' 'A317770' 'A355150' 'A357550']
2023-07-31: ['A010640' 'A032560' 'A042940' 'A045300' 'A070590' 'A162120' 'A236490']
2023-08-01: ['A010640' 'A032560' 'A045300' 'A070590' 'A093520' 'A162120' 'A236490']
2023-08-02: ['A002710' 'A007770' 'A039200' 'A169670' 'A256610' 'A317770' 'A355150']
2023-08-03: ['A002710' 'A007770' 'A039200' 'A169670' 'A256610' 'A317770' 'A355150']
```

ex) 8월 1일자 포트폴리오:



7. 그래프 기반 분석 (Similarity & LSTM)

INTERVIEW
PRESENTATION

접근 방향 설정

INTERVIEW
PRESENTATION

- 소개
- 지원직무
- 경력 및 역량
- 수행프로젝트
- 향후 계획

KRX 크롤링

INTERVIEW
PRESENTATION

소개

지원직무

경력 및 역량

수행프로젝트

향후 계획

Featuring Engineering

INTERVIEW
PRESENTATION

소개

지원직무

경력 및 역량

수행프로젝트

향후 계획

Graph Building

시각화 결과

INTERVIEW
PRESENTATION

소개

지원직무

경력 및 역량

수행프로젝트

향후 계획

LSTM #1

INTERVIEW
PRESENTATION

소개

지원직무

경력 및 역량

수행프로젝트

향후 계획

LSTM #2

8. 결론 및 한계

결론

전체적으로 간략한 정리정도.

한계

아쉬웠던 점 :

1. 분석 초기에 Sentiment 정보를 간과한 점
2. 미국의 영향력을 녹여내지 못한 점 (AAPL, NVDA의 주가 상승은 삼성전자, 하이닉스에도 영향을 줌)
3. 거시경제환경에 대한 미흡한 분석 (ex. 미중 무역전쟁, Fed의 통화정책, 기준금리 등)
4. DL에서 피쳐로 활용한 데이터 셋 내 결측치 처리에 대한 부족한 고민
5. 2023.09 ~ 현재 & 2023.01 이전 데이터들의 부재 (장기적 관점에서 분석하기엔 데이터가 부족.)
6. Similarity의 측면에서 평가 후 더 심화적으로 분석하지 못한 점
7. 종목들의 구성 비율 (portion을 정하지 않음.)
8. 포트폴리오의 분산화에 대한 작용이 미비함.

감사합니다!