

STOCK CHANGE PREDICTION

Sentiment analysis of news articles and YouTube texts and
prediction of stock price fluctuations using deep learning models

2021. 10. 08.

UPbitu

주의

본 서비스에서 제공되는 투자정보는 단지
투자판단에 대한 참고임을 밝혀드립니다.
서비스에서 제공하는 주식의 가치, 가격 상승 및 가격 하락에 대해서는
업 빛 투 가 보장하지 않습니다.

Contents



PART

01

1. 프로젝트 개요

01. 프로젝트 기획 배경 및 목표 05P

02. 구성원 및 역할 09P

PART

02

1. 감성 분석

01. 감성사전 23P

02. 감성점수 27P

PART

2. 데이터

01. 데이터 수집 11P

02. 데이터 전처리 20P

03. 분석 모형 21P

PART

03

1. 웹 페이지 구현

01. 웹 페이지 프로세스 46P

02. 웹 페이지 화면 49P

PART

04

1. 발전방향 및 느낀점

01. 한계점 및 발전방향 56P

02. 느낀점 57P

03. 참고 자료 58P

2. 데이터 분석

01. 시계열 분석 및 강화 학습 31P

02. 앙상블 40P

PART 1-1

프로젝트 개요

- 프로젝트 기획 배경 및 목표
- 구성원 및 역할

Sentiment analysis of news articles and YouTube texts and
prediction of stock price fluctuations using deep learning models

1-1. 프로젝트 개요

프로젝트 기획 배경

'코로나 패닉' 1년, 코스피 1400→3000으로..."동학개미 61조원 순매수"

코스피 상승률 109.5%...코스닥도 428에서 925로 뛰어

G20 중 상승률 '최고'

동학개미, 코스닥시장서 18조원 사들여

+965.38 (47.00%) ↑ 지난 5년

10월 1일 오후 6:03 GMT+9 · 면책조항

1일 | 5일 | 1개월 | 6개월 | 연증 | 1년 | **5년** | 최대

최악 시기에 최고 뚫은 코스피..."아직 더 간다" 세가지 이유

중앙일보 | 입력 2020.11.23 17:33 업데이트 2020.11.23 21:32

장원석 기자

최악의 시기에 최고의 기록을 썼다. 미국 대선이 끝난 후부터 질주하던 코스피가 마침내 2600선에 등정했다. 23일 코스피는 전날보다 1.92% 오른 2602.59로장을 마쳤다. 종가 기준으로 역대 최고치였던 2018년 1월 29일(2598.19) 지수를 넘어섰다.

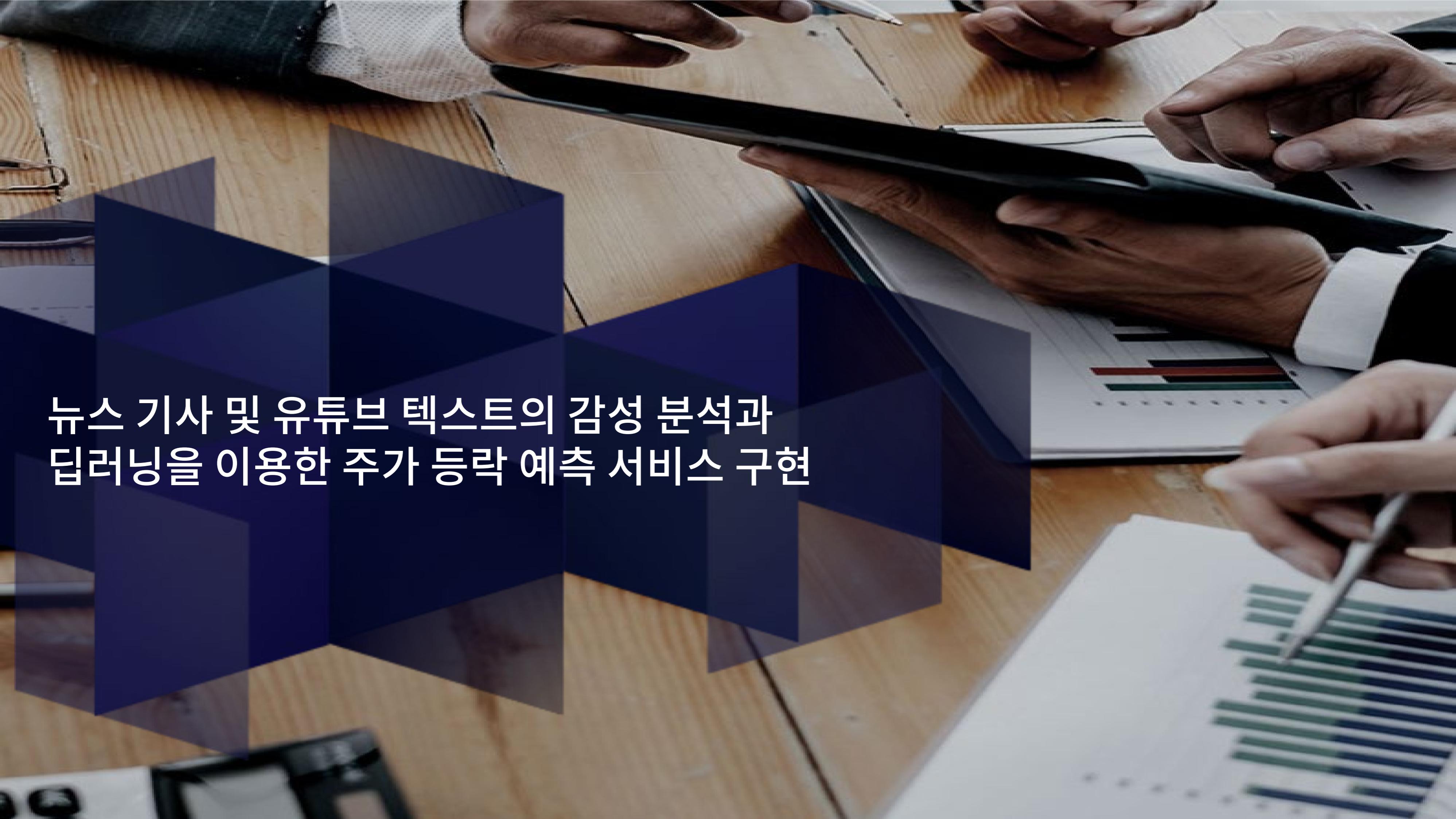


시가	3,056.21	최고	3,062.60	52-주 최고	3,316.08
전일 종가	3,068.82	최저	3,015.01	52-주 최저	2,266.93

1-1. 프로젝트 개요

프로젝트 기획 배경





뉴스 기사 및 유튜브 텍스트의 감성 분석과
딥러닝을 이용한 주가 등락 예측 서비스 구현

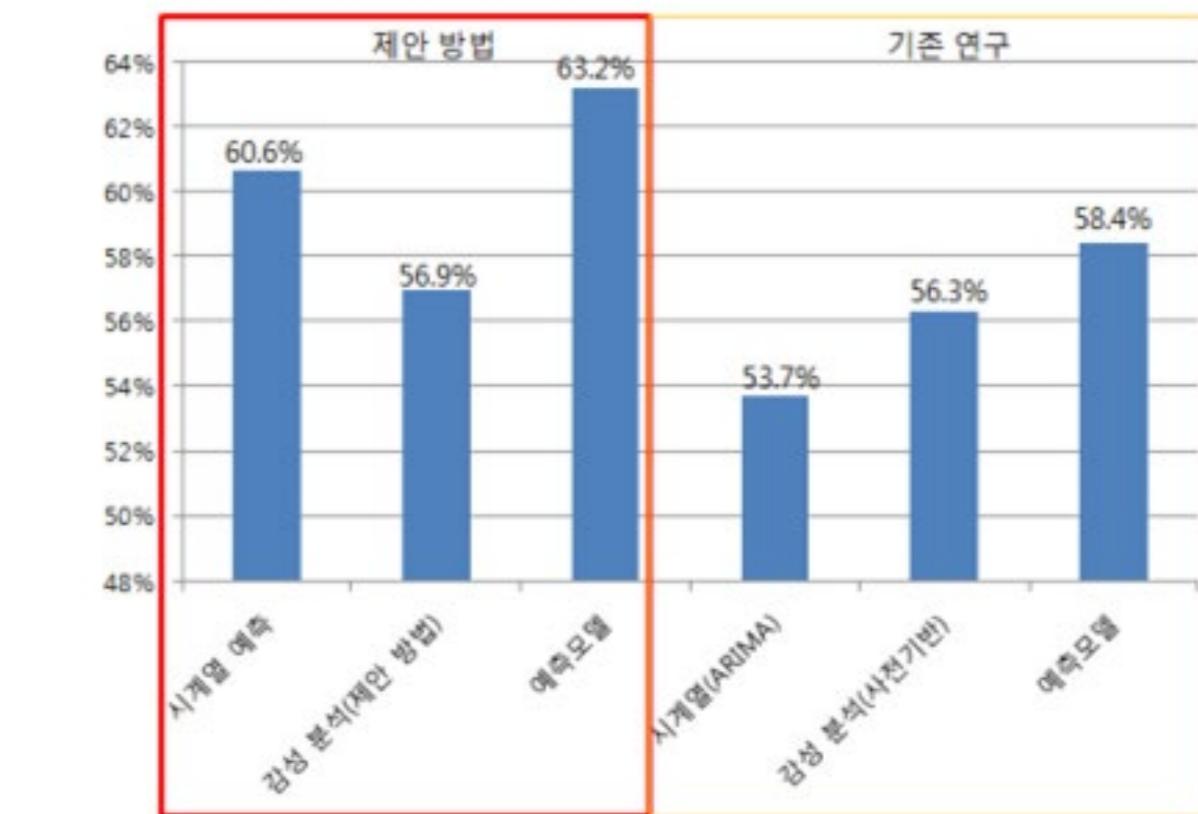
1-1. 프로젝트 개요

프로젝트 주가 등락 예측 선행연구의 예측 정확도

Accurate Multivariate Stock Movement Prediction via Data-Axis Transformer with Multi-Level Contexts

Model	ACL18 (US)		KDD17 (US)		NDX100 (US)	
	ACC	MCC	ACC	MCC	ACC	MCC
LSTM [24]	0.4987 ± 0.0127	0.0337 ± 0.0398	0.5118 ± 0.0066	0.0187 ± 0.0110	0.5263 ± 0.0003	0.0037 ± 0.0049
ALSTM [31]	0.4919 ± 0.0142	0.0142 ± 0.0275	0.5166 ± 0.0041	0.0316 ± 0.0119	0.5260 ± 0.0007	0.0028 ± 0.0084
StockNet [31]	0.5285 ± 0.0020	0.0187 ± 0.0011	0.5193 ± 0.0001	0.0335 ± 0.0050	0.5392 ± 0.0016	0.0253 ± 0.0102
Adv-ALSTM [9]	0.5380 ± 0.0177	0.0830 ± 0.0353	0.5169 ± 0.0058	0.0333 ± 0.0137	0.5404 ± 0.0003	0.0046 ± 0.0090
DTML (proposed)	0.5744 ± 0.0194	0.1910 ± 0.0315	0.5353 ± 0.0075	0.0733 ± 0.0195	0.5406 ± 0.0037	0.0310 ± 0.0193
Model	CSI300 (China)		NI225 (Japan)		FTSE100 (UK)	
	ACC	MCC	ACC	MCC	ACC	MCC
LSTM [24]	0.5367 ± 0.0038	0.0722 ± 0.0050	0.5079 ± 0.0079	0.0148 ± 0.0162	0.5096 ± 0.0065	0.0187 ± 0.0129
ALSTM [31]	0.5315 ± 0.0036	0.0625 ± 0.0076	0.5060 ± 0.0066	0.0125 ± 0.0139	0.5106 ± 0.0038	0.0231 ± 0.0077
StockNet [31]	0.5254 ± 0.0029	0.0445 ± 0.0117	0.5015 ± 0.0054	0.0050 ± 0.0118	0.5036 ± 0.0095	0.0134 ± 0.0135
Adv-ALSTM [9]	0.5337 ± 0.0050	0.0668 ± 0.0084	0.5160 ± 0.0103	0.0340 ± 0.0201	0.5066 ± 0.0067	0.0155 ± 0.0140
DTML (proposed)	0.5442 ± 0.0035	0.0826 ± 0.0074	0.5276 ± 0.0103	0.0626 ± 0.0230	0.5208 ± 0.0121	0.0502 ± 0.0214

Stock Fluctuation Prediction based on News Sentiment Analysis and Time Series Prediction

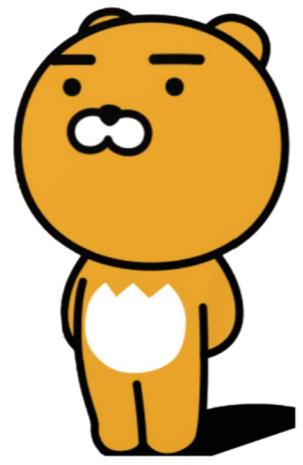


주가 등락 예측 정확도 일반적으로 45~55% 수준

1-1. 프로젝트 개요

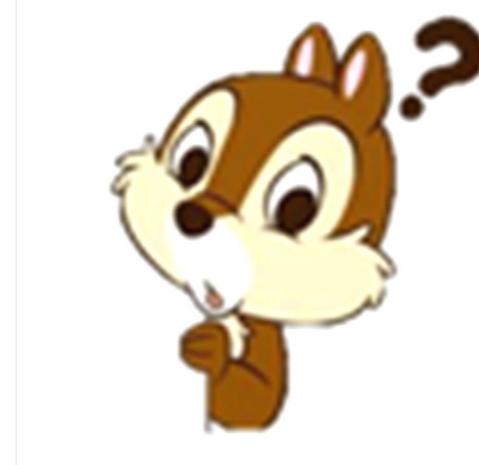
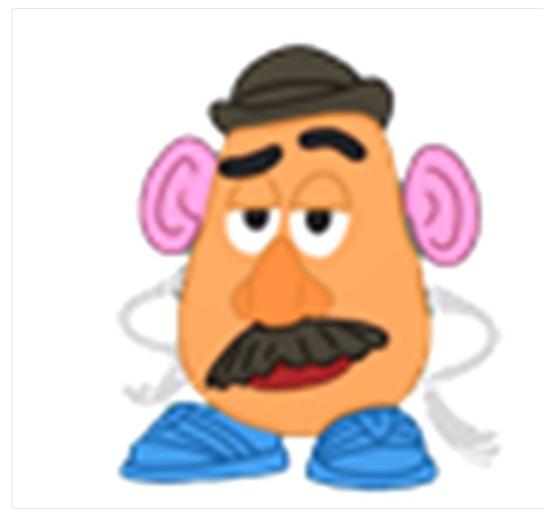
구성원 및 역할

Data Engineers



정길종

Data Scientists



인태우

김형림

윤보람

채길호

- 데이터 수집
- DB 설계 및 구축
- Django 웹 페이지 구축

- 데이터 전처리
- 데이터 및 시각화
- 감성사전 제작

- 머신러닝 분석 (ARIMA · FBProphet)
- 딥러닝 (LSTM) 및 강화 학습
- 앙상블 모델링

PART 1-2

데이터

- 데이터 수집
- 데이터 전처리
- 분석 모형

Sentiment analysis of news articles and YouTube texts and
prediction of stock price fluctuations using deep learning models

1-2. 데이터

데이터 수집 시가총액 상위 종목 삼성전자 · SK하이닉스 · LG화학 · 현대차 · 셀트리온

KOSPI 시가총액 상위 10종목(2021. 09. 기준) 중 그룹사 중복 및 네이버 · 카카오를 제외한 5개 종목 선정

Insert Web Page

This app allows you to insert secure web pages starting with https:// into the slide deck. Non-secure web pages are not supported for security reasons.

Please enter the URL below.

https:// public.tableau.com/app/profile/boram.yun/viz/stock-charts_ppt/sheet1?publish=yes

Note: Many popular websites allow secure access. Please click on the preview button to ensure the web page is accessible.

[Web Viewer Terms](#) | [Privacy & Cookies](#)

Preview

1-2. 데이터

데이터 수집 주식 · 재무지표



FinanceDataReader

Date	Open	High	Low	Close	Volume	Change
1997-08-20	1268	1329	1259	1301	218130	NaN
1997-08-21	1293	1313	1286	1287	100340	-0.010761
1997-08-22	1286	1301	1272	1296	100373	0.006993
1997-08-23	1313	1313	1297	1306	72280	0.007716
1997-08-25	1309	1313	1286	1301	72320	-0.003828
...
2021-08-26	76100	76200	74600	74600	16671494	-0.014531
2021-08-27	74300	75000	73800	74300	15172748	-0.004021
2021-08-30	75400	75500	74200	74600	12686999	0.004038
2021-08-31	74900	76700	74300	76700	24630370	0.028150
2021-09-01	76700	77100	75900	76800	16114775	0.001304

주식 거래 데이터

시 가 Open 종 가 Close
고 가 High 거래량 Volume
저 가 Low 변동률 Change



Date	PER	PBR	ROE	ROA
2018-01-02	8.403220	1.445614	0.172031	0.122212
2018-01-03	8.502043	1.462615	0.172031	0.122212
2018-01-04	8.413103	1.447314	0.172031	0.122212
2018-01-05	8.584395	1.476782	0.172031	0.122212
2018-01-08	8.567925	1.473949	0.172031	0.122212
...
2021-09-06	18.552538	1.671147	0.090076	0.066190
2021-09-07	18.264530	1.645204	0.090076	0.066190
2021-09-08	18.312532	1.649528	0.090076	0.066190
2021-09-09	18.072525	1.627909	0.090076	0.066190
2021-09-10	18.072525	1.627909	0.090076	0.066190

재무지표 데이터

주가수익비율 PER 자기자본이익률 ROE
주가순자산비율 PBR 총자산순이익률 ROA

1-2. 데이터

데이터 수집 경제 일간지 · 주식 콘텐츠 유튜브 채널

경제 일간지

매일경제 · MBN

아시아경제 □

사회 국제 부동산 증권 정치 IT·과학 문화

"어서와 이런 깜찍한 SUV는 처음이지?"...현대차, 캐스퍼 베일 벗었다

박소현 기자 | 입력 : 2021.09.01 09:01:54 | 수정 : 2021.09.01 14:23:32 | 11



△현대차, 신규 엔트리 SUV '캐스퍼' 외장 이미지 최초 공개 [사진제공=현대차]

현대자동차의 새로운 스포츠유틸리티차량(SUV) 캐스퍼가 엔트리 차급으로 SUV 라인업에 합류한다.

현대차는 올해 하반기 출시 예정인 엔트리 SUV '캐스퍼(CASPER)'의 외장 디자인을 1일 최초로 공개했다. 캐스퍼는 실용성과 개성을 추구하는 소비자의 취향과 라이프스타일을 반영해 기존에 없던 새로운 차급에서 처음으로 선보이는 모델이다.

관련뉴스

HYUNDAI MOTOR GROUP

"현대차 친환경 전환 속도, 세계 선두권"

제네시스 2030년부터 전기차만 생산

절대강자 없는 전기차...명품차 제네시스, 시장...

현대차-기아 8월 美 판매 1.3% 감소...반도체...

5년만에 제네시스 앞에 선 정의선...럭셔리에...

시민과 함께하는
스마트 행복도시 애호 Windo

주식 유튜브 채널

삼프로TV_경제의신과함께 ◉ 구독자 151만명

한국경제 TV ◉ 구독자 68.2만명

슈카월드 ◉ 구독자 169만명

기성세대인 40대, 50대가 위는 베이비부머 아래는 뭉뚱그려서 MZ세대라고 부르는거 아니나 X

구분	연도	연령	연령대	인구	중간값
베이비부머 세대	1946~1964년	55~75세	60대, 70대	1000만명	55년생
X세대	1965~1979년	43~54세	40대, 50대	1600만명	72년생
M세대(밀레니얼 세대)	1980~1995년	26~42세	10대, 20대, 30대	1800만명	87년생
Z세대	1996~2010년	12~25세			03년생
알파세대?	2010년 이후				

#X세대 #MZ세대 MZ세대의 요청 : M과 Z를 뒤지 말아주세요.

조회수 397,471회 · 2021. 8. 31.

7.3천 5.186 ▶ 공유 ⌂ 저장 ...

슈카월드
구독자 169만명

어렵고 딱딱한 경제/시사/금융 이야기를
쉽고 유쾌하게 들려내는
경제/시사/이슈/잡설 토크방송입니다.

스크립트

00:25 게임을 못 만들고 넥슨 안 만들고

00:27 넷마블은 외 만드냐 라고 얘길 했다

00:29 이게 그 nc는 목 기로에 서 있고

00:31 넥슨은 지금 종합 엔터테인먼트 회사로

00:34 가려고 넷마블은 주자 를 열심히

00:36 그러니까 이런 얘길 했는데 그거 이렇게

00:38 예약하시면 상당히 안됩니다 왜냐하면

00:39 x 게임을 안 만드는 게 아니죠 뒤

00:42 10시 만 들어요 저 수백 명의 아무

00:44 만들고 있을 거야 첨 머 이상이

00:46 게임을 만들고 있지 않을까 하는데

00:47 안에 있는 분들이야 게임에 다시

한국어 (자동 생성됨)

1-2. 데이터

데이터 수집 내역

삼성전자 기준 매일경제 · 아시아경제 일별 기사량 (2021.09.03)

삼성전자	005930	매일경제	2021090217	"D램 낸드가 전부 아니다"...삼성전자 세계 첫 개발 '2···	http://news.mk.co.kr/newsRead.php?no=85007...
삼성전자	005930	매일경제	2021090211	삼성전자, 최첨단 혁신 기능 담긴 '갤럭시 A52s 5G' 국···	http://news.mk.co.kr/newsRead.php?no=84804...
삼성전자	005930	매일경제	2021090211	"업계 최초 '2억 화소' 벽 뛰어넘었다" ...삼성전자 초격···	http://news.mk.co.kr/newsRead.php?no=84804...
삼성전자	005930	매일경제	2021090217	애니메이션으로 전자산업 배운다	http://news.mk.co.kr/newsRead.php?no=85007...
삼성전자	005930	매일경제	2021090217	출산계획까지 돋는 애플워치	http://news.mk.co.kr/newsRead.php?no=85012...
삼성전자	005930	매일경제	2021090217	美, 기술강국 韓·日·獨에 러브콜...5G장비등 한국기업...	http://news.mk.co.kr/newsRead.php?no=85016...
삼성전자	005930	매일경제	2021090204	[현장 둘보기] 반도체산업 경쟁력 키울 또 하나의 무기,...	http://news.mk.co.kr/newsRead.php?no=84680...
삼성전자	005930	매일경제	2021090220	"파월의 시간표 나왔다"...투자 고수 시선은 이 곳에	http://news.mk.co.kr/newsRead.php?no=85051...
삼성전자	005930	매일경제	2021090204	[Cover Story] '디지털 본사' 구축...언제 어디서든 세···	http://news.mk.co.kr/newsRead.php?no=84680...

삼성전자 기준 유튜브 채널 삼프로TV · 슈카월드

삼성전자	005930	삼프로TV_경제의신과함께	2021100200	[퇴근길 라이브&백브리핑] 현저한 악재 속 투자 아이디···	으 으으oo 아 아 으오투자는 사람들의 깊이 있는 대···
삼성전자	005930	삼프로TV_경제의신과함께	2021100100	[증시 셔터맨] 넷플릭스 후속작 대기중! 컨텐츠 업종 강···	4 애도 어른도에듀아르도아주 심플한 디자···
삼성전자	005930	삼프로TV_경제의신과함께	2021100100	[퇴근길 라이브&백브리핑] 최근 시장 조정 원인과 조정···	으 으seoul으오투자는 사람들의 깊이 있는데 와 상품···
삼성전자	005930	삼프로TV_경제의신과함께	2021100100	미 국제 금리 급등, 글로벌 경제에 미치는 영향은? _글···	[음악][음악]으으 으으 4 러버 안녕하세요 언제나처럼···
삼성전자	005930	삼프로TV_경제의신과함께	2021100100	[증시 염탐정] 하락 출발한 시장.. 미국장 영향에서 벗어···	2 베스트 40건의영불 이염 생활이란 이 모르겠습니다···
삼성전자	005930	삼프로TV_경제의신과함께	2021093000	[증시 셔터맨] 걱정되는 코스피, 심리는 이미 3천이 깨···	4 아예 온 제가 깜빡하고 많이 파토린 삼푸를 안 썼더니···
삼성전자	005930	삼프로TV_경제의신과함께	2021093000	[글로벌 이슈체크] 내년부터 메타버스 본게임?! 기업들···	디풀 iv 부 시작하도록 하겠습니다요즘은 장 방송 잠깐···
삼성전자	005930	삼프로TV_경제의신과함께	2021093000	[주인장백 브리핑] KB 이어 하나은행도 전세대출 한도···	4암튼 회장님의 어느 또 재미있는 소식갖고 오셨어요···
삼성전자	005930	삼프로TV_경제의신과함께	2021093000	[글로벌 개장시황] 불확실의 해소? 8월 보다 공포지수...	블롭 얼라이브 3부 시작하도록 하겠습니다어제는 빨리···
삼성전자	005930	삼프로TV_경제의신과함께	2021093000	[퇴근길 라이브&백브리핑] 원트전략, 올해보다 내년이...	[음악]으으 oo 아아 음[음악]으[음악]투자를 하···
삼성전자	005930	삼프로TV_경제의신과함께	2021093000	로블록스 말고~ 다양한 메타버스 관련 미국 주식 _글로···	으으 으으 글로벌 r 배우신 걸 환영합니다안녕하···
삼성전자	005930	삼프로TV_경제의신과함께	2021093000	[증시 염탐정] 혼조세의 장 초반.. 기관의 움직임이 중요···	자 그러면 저희가 진짜 오래 하고계신 우리ymb을 이염 생···
삼성전자	005930	삼프로TV_경제의신과함께	2021092900	[퇴근길 라이브&백브리핑] 중국 규제에도 불구하고 엔···	으으 으으 cole 음 으오투자를 하는 사람들의 깊이 있···
삼성전자	005930	삼프로TV_경제의신과함께	2021092900	[증시 염탐정] 출발 분위기 안 좋은 시장.. 수급 상황은?...	자 그러면에 우리 값이 하셨던ymb를 이 이벤트 투자증권···
삼성전자	005930	삼프로TV_경제의신과함께	2021092800	[심층 인터뷰] 전망 좋은 탄소배출권, 어떻게 투자해야...	아이언 터진 건가 함께하는 3%리그 실전투자대회 뭐···
삼성전자	005930	삼프로TV_경제의신과함께	2021092800	[퇴근길 라이브&백브리핑] 새로운 투자자산, 탄소배출...	[음악]으oo 아 아 음[음악]으[음악]박수투자를 하···
삼성전자	005930	삼프로TV_경제의신과함께	2021092800	밀레니얼, Z세대와 글로벌 주식투자 _글로벌 라이브...	[음악]으tobewon[음악]글로벌 라이브 에 오신걸 환영···

삼성전자	005930	아시아경제	2021090308	"코로나에 발목" 1위 삼성, 2분기 스마트폰 생산 크게...	https://view.asiae.co.kr/article/20210903080337...
삼성전자	005930	아시아경제	2021090310	코스피 상승 지속.. 외인 전기전자 순매수	https://view.asiae.co.kr/article/20210903110147...
삼성전자	005930	아시아경제	2021090313	애플, 2분기 '에어팟 프로' 1000만대 이상 팔았다	https://view.asiae.co.kr/article/20210903133936...
삼성전자	005930	아시아경제	2021090311	올 10대그룹 시총 80조 늘었다	https://view.asiae.co.kr/article/20210903095640...
삼성전자	005930	아시아경제	2021090311	국내 10대 기업, 코로나 딛고 해외서 일어섰지만...하반...	https://view.asiae.co.kr/article/20210902154245...
삼성전자	005930	아시아경제	2021090314	코스피 장중 3200 돌파	https://view.asiae.co.kr/article/20210903140842...
삼성전자	005930	아시아경제	2021090316	코스피 3201.06에 마감.. 돌아온 외인	https://view.asiae.co.kr/article/20210903160605...
삼성전자	005930	아시아경제	2021090309	코스피 상승 출발.. 외인 순매수 유입	https://view.asiae.co.kr/article/20210903093318...
삼성전자	005930	아시아경제	2021090314	10여년 만에 주상복합아파트 일반분양 용인 양지 '동문···	https://view.asiae.co.kr/article/20210224111318...
삼성전자	005930	아시아경제	2021090316	아파트 청약은 현금부자 뜻?대체재 주목받는 주거용 오···	https://view.asiae.co.kr/article/20210202115645...

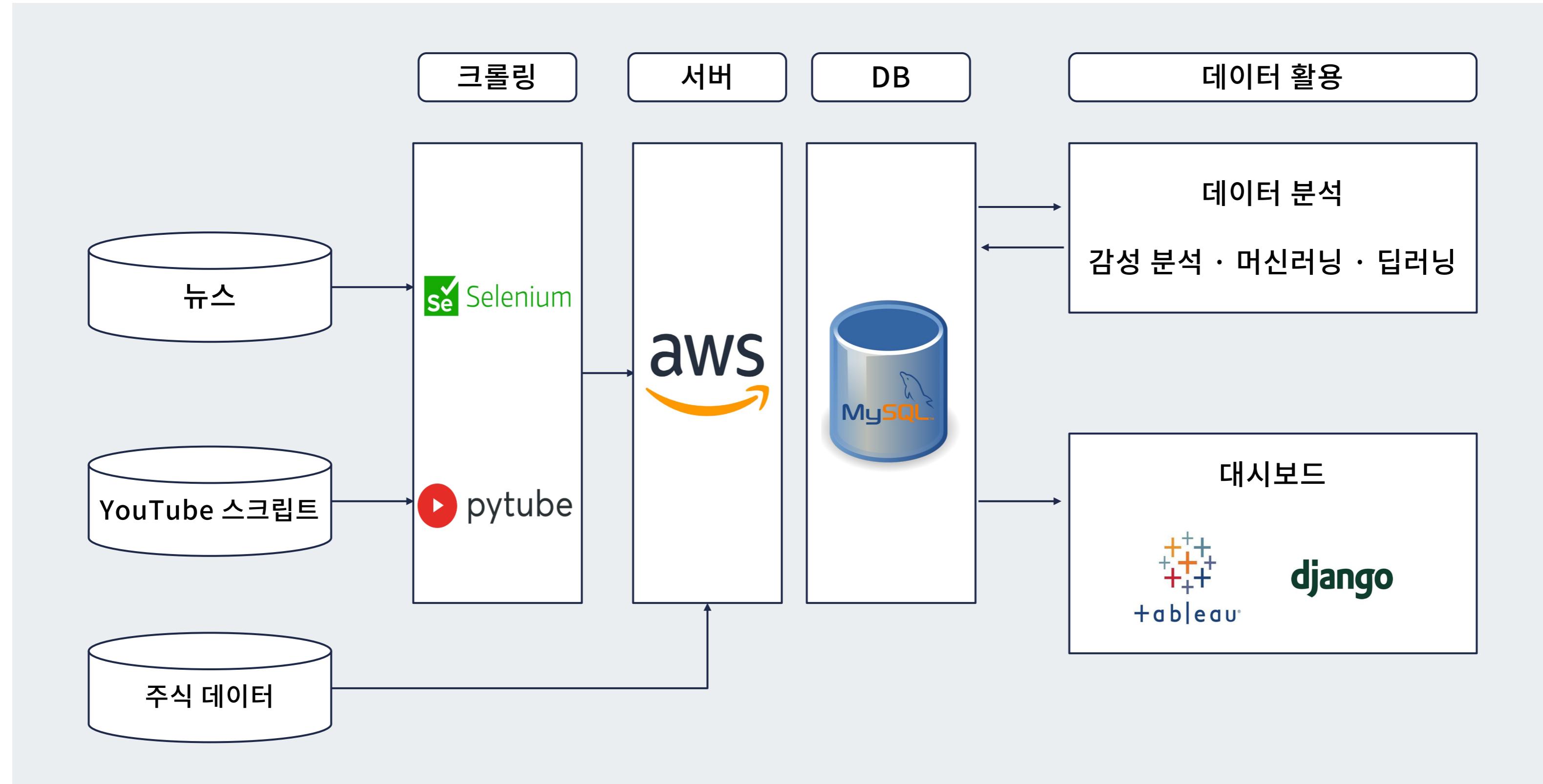
삼성전자	005930	슈카월드	2021091500	쇼크의 연속, 주린이에게 가혹한 2021년 주식시장	아[음악]wr요즘에 진짜 상황이 너무 너무힘들다고 느끼···
삼성전자	005930	슈카월드	2021091100	아직도 어려운 호칭과 높임말로 인한 오해들	아[음악]wr자 오늘도 그 반성하겠습니다 항상반성하는...
삼성전자	005930	슈카월드	2021091000	메타버스는 기존 온라인 게임이랑 무엇이 다른가	아[음악]wr제가 얼마전에 글로벌 라이브라는것을찍는···
삼성전자	005930	슈카월드	2021090900	수능이 붕괴된다? 벼랑 끝에 선 교육시장	아[음악]wr자재가 또 얼마전에 유튜브 s 도봤어요 진용...
삼성전자	005930	슈카월드	2021090800	전기차 충전기 기업은 대박났을까	아[음악]wr바이든 정부 행정 명령을 내렸죠미국에서 2...
삼성전자	005930	슈카월드	2021090700	초강력 문화 규제, 강력한 공산주의로 돌아가는 21세기...	아[음악]wr제가 얼마전에 중앙일보에서 이런기사가 났...
삼성전자	005930	슈카월드	2021090400	이슬람국단주의 IS는 왜 같은 이슬람인 탈레반을 공격···	아[음악]wr갑을 공황이 테러가 발생했다아프가니스탄...
삼성전자	005930	슈카월드	2021090300	내 경험에 자식에게 유전이 될까	으 르이[음악]오늘은 약간 재밌는 얘기를 하나 드릴텐...
삼성전자	005930	슈카월드	2021090200	정부는 돈을 뿐이고, 중앙은행은 돈을 흡수하는 기묘한...	아[음악]wr지난주에 어떤 일 있었나 면 대한민국이 기준...
삼성전자	005930	슈카월드	2021090100	MZ세대의 요청 : M과 Z를 묶지 말아주세요.	아[음악]wr여러분들 혹시 유튜버 나 스트림 하고싶으십...
삼성전자	005930	슈카월드	2021083100	NC소프트 충격적인 주식 손실, 분노의 주주들	아[음악]wr그 오해가 전 맨날 이제 방송하면일단 타과...
삼성전자	005930	슈카월드	2021082800	'공동부유', 시진핑 주석 3연임을 향한 새로운 이데올로...	아[음악]wr자우 자비에 여간 이런 얘기가있습니다 지난...
삼성전자	005930	슈카월드	2021082700	글로벌 기업들의 재생에너지 100% 사용선언과 우리의...	아[음악]wr키우 솔루션 한국에서의 비영리법인 데보다...
삼성전자	005930	슈카월드	2021082500	디지털 시대에 대응이란, 의외의 쪽박과 놀라운 대박	아[음악]wr자유의 보면 t 막힘 옛날엔 트로트였는데 티...
삼성전자	005930	슈카월드	2021082400	테슬라 AI 데이, 휴머노이드 '테슬라 로봇'의 등장	아[음악]wr자 우리의 그냥 팔림 중에 나의테슬라가 aid...
삼성전자	005930	슈카월드	2021082200	62년 카스트로 시대의 종언과 변화의 앞에 선 쿠바	아[음악]wr쿠바가 여러분들 잘 관심 없으시겠지만현재...
삼성전자	005930	슈카월드	2021082100	간절한 NC SOFT, 동상이몽 넥슨, 투자의 신 넷마블,...	아[음악]wr우리나라 대표 게임사 엔씨소프트가2분 비...

뉴스 기사 일일 데이터 수 평균 약 10개
주말에는 기사량이 적어 평균 이하로 수집

유튜브 채널의 경우 해당 채널마다
업로드 하는 영상의 일자가 불규칙적임

1-2. 데이터

데이터 수집 프로세스



1-2. 데이터

데이터 수집 프로세스

경제 일간지

매일경제 · MBN

아시아경제 ▶

-뉴스-
제목
내용

수집

-데이터 수집-
매일 아침
9시 30분

-설명-
전날 뉴스를 서버가
기동시간에 맞추어 수집

Selenium

주식 유튜브 채널



삼프로TV_경제의신과함께 ◉
구독자 151만명



한국경제TV ◉
구독자 68.2만명



슈카월드
구독자 169만명

-유튜브-
제목
스트립트

-데이터 수집-
매일 저녁
11시 30분

-설명-
유튜브 각 채널에서는
당일 업로드 선택 수집

저장

pytube



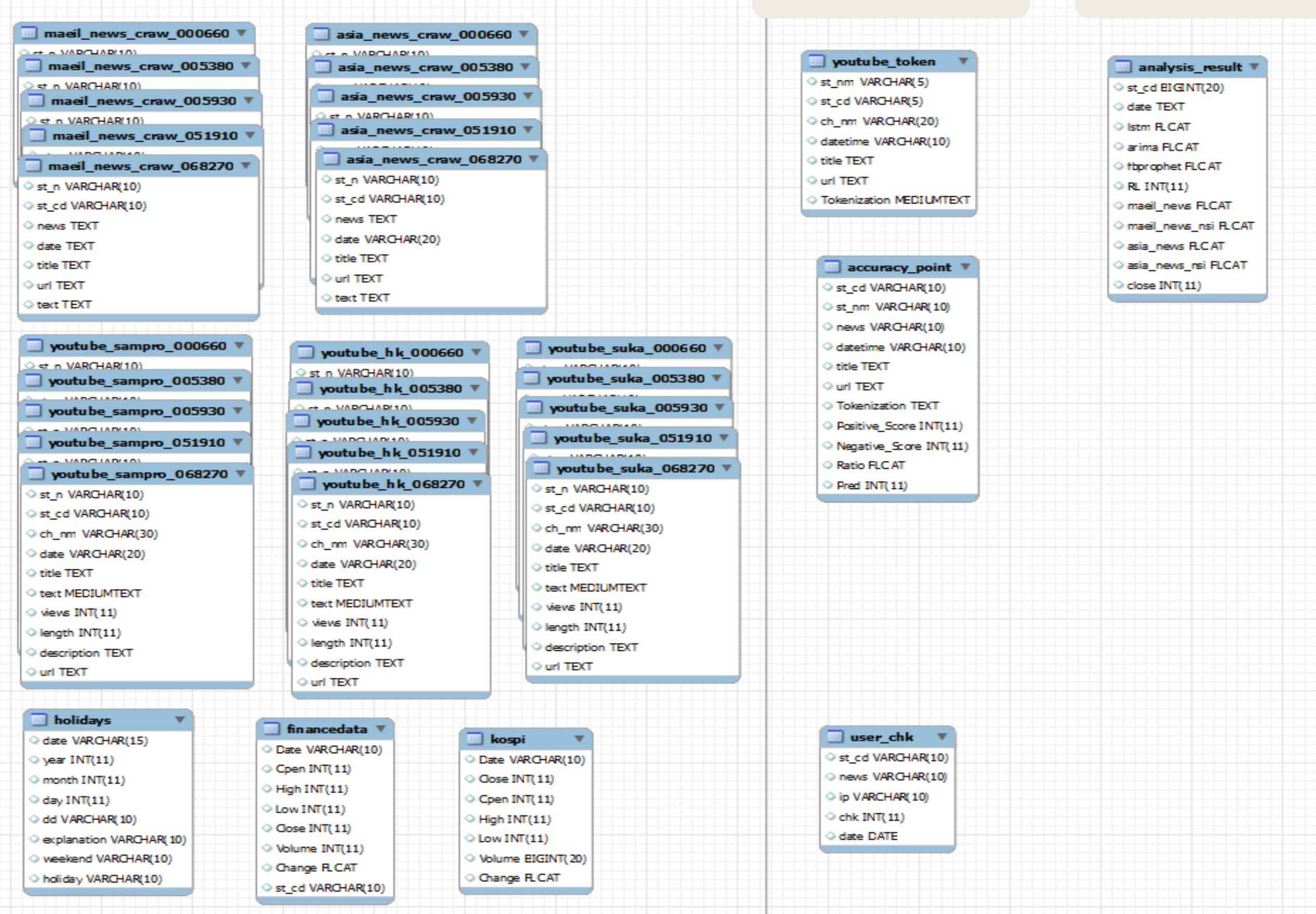
1-2. 데이터

데이터 수집 ERD

경제 일간지

유튜브

주식 · 재무

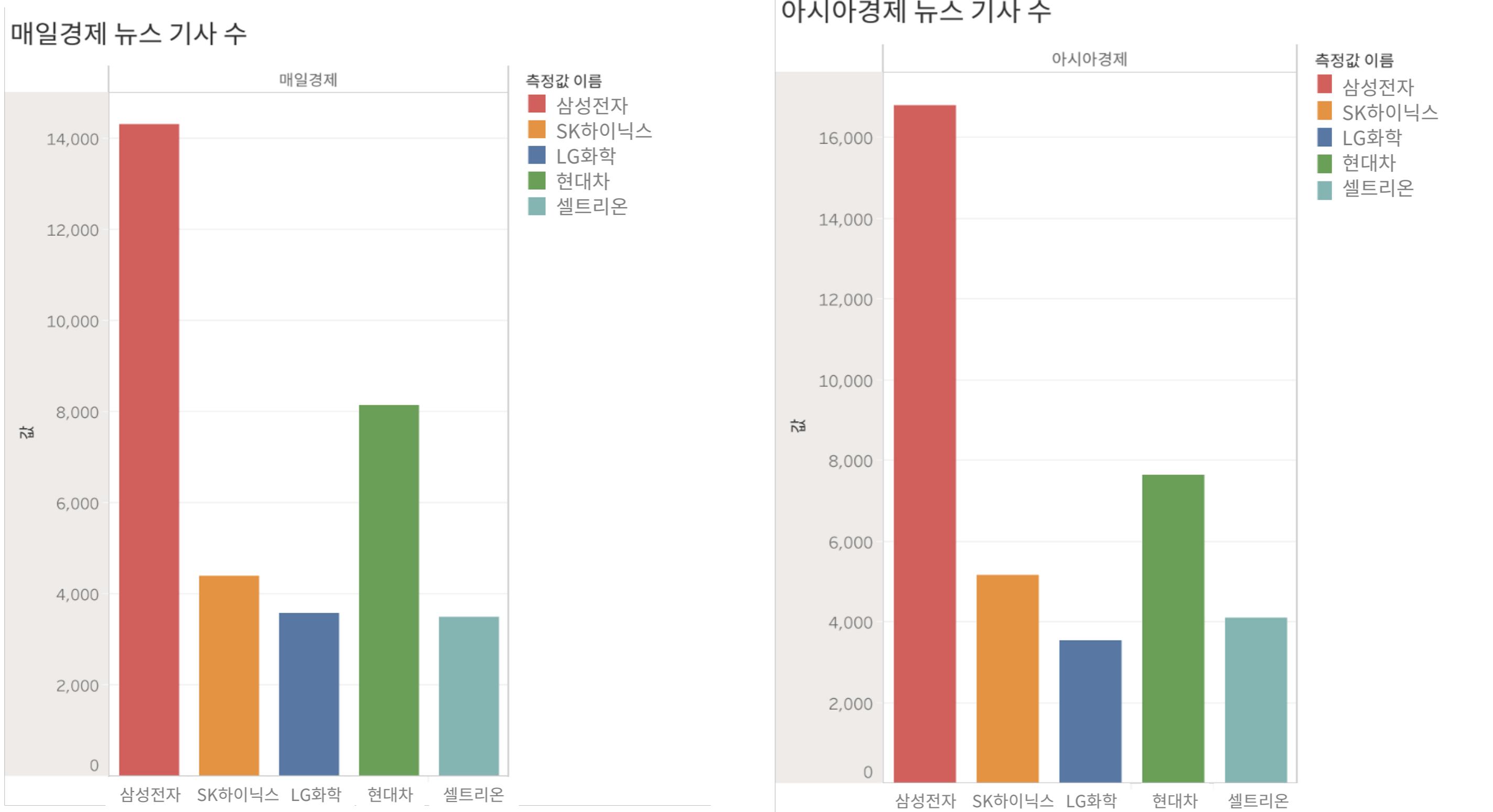


전처리

분석 결과

1-2. 데이터

데이터 수집 내역



1-2. 데이터

데이터 수집 총 데이터 수 2021. 10. 06. 기준

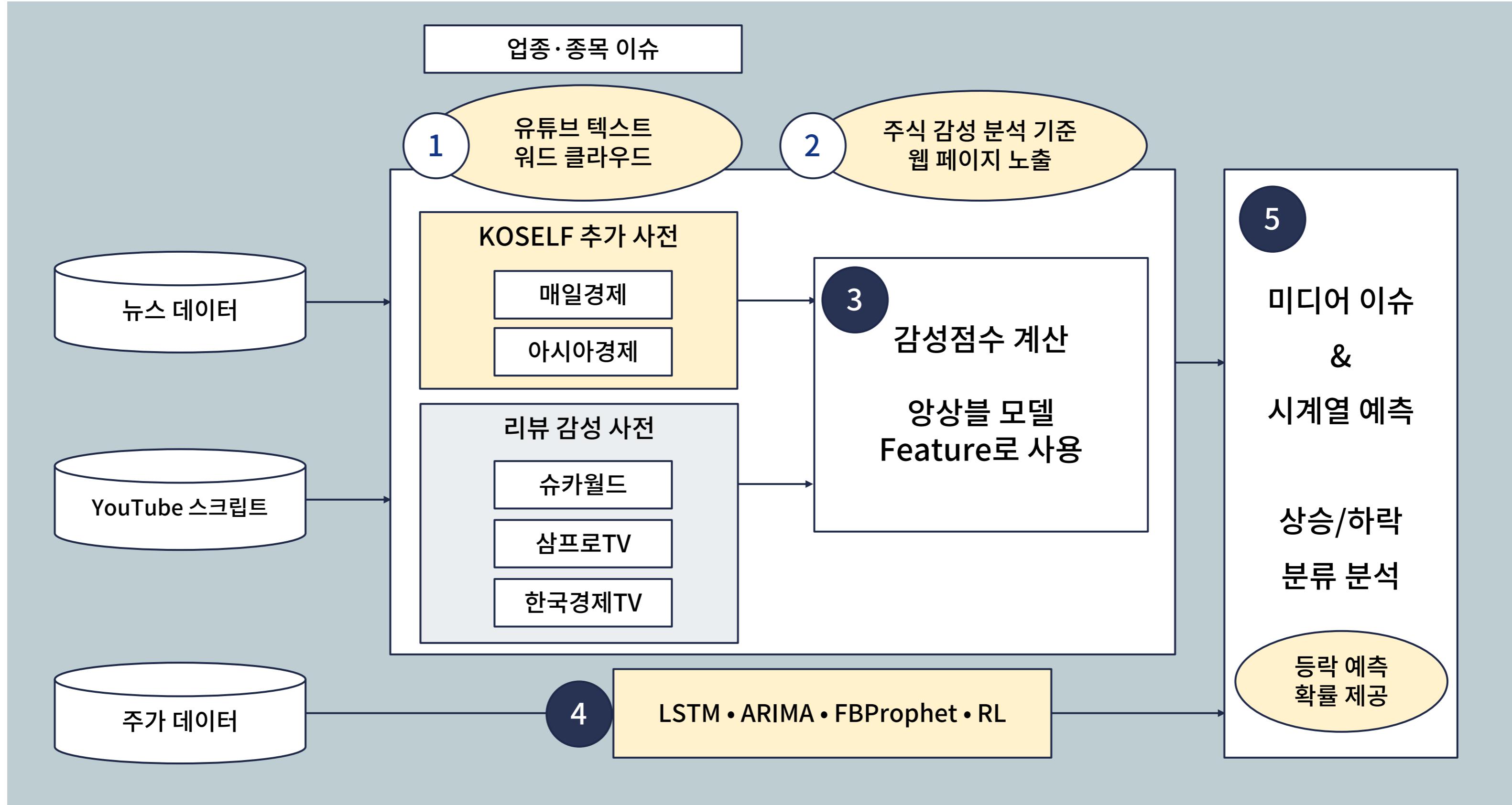
	삼성전자	SK하이닉스	LG화학	현대차	셀트리온	미디어별 합계
매일경제	14,304	4,383	3,572	8,150	3,505	33,914
아시아경제	16,788	5,175	3,530	7,627	4,100	37,220
슈카월드	448	90	64	403	197	1,202
삼프로TV	579	516	521	542	474	2,632
한국경제TV	577	512	507	544	486	2,626
종목별 합계	32,696	8,194	8,194	17,266	8,762	77,594

1-2. 데이터

데이터 전처리 뉴스 기사 및 유튜브 영상

전처리 항목		전처리 방법
휴장일 및 장 마감 시간	주말 및 공휴일 장 마감 (15:30)	<ul style="list-style-type: none">주말 및 공휴일 데이터 수집 후 리스트 생성휴장일 이후 데이터로 날짜 수정미디어 영향 반영 시점을 고려해 오후 3시 이후 업로드된 기사 및 동영상을 다음 거래일자로 변경
주가 변동률	상승 / 하락 / 보합	<ul style="list-style-type: none">전일 대비 종가 변동률이 양수이면 상승, 음수이면 하락, 0이면 보합으로 판단하여 데이터프레임 내 ud(updown) column 생성
텍스트 전처리	형태소 분석	<ul style="list-style-type: none">텍스트 정제 후 KoNLPy Okt 형태소 분석기 사용품사(POS) 태깅 하여 명사 및 형용사 추출불용어 처리 및 한 글자 단어 제거

1-2. 데이터 분석 모형



PART 2-1

감성 분석

- 감성사전
- 감성점수

Sentiment analysis of news articles and YouTube texts and
prediction of stock price fluctuations using deep learning models

2-1. 감성 분석

감성사전 KOSELF (KOrean SEntiment Lexicon for Finance) 기업 재무분석을 위한 한국어 감성사전

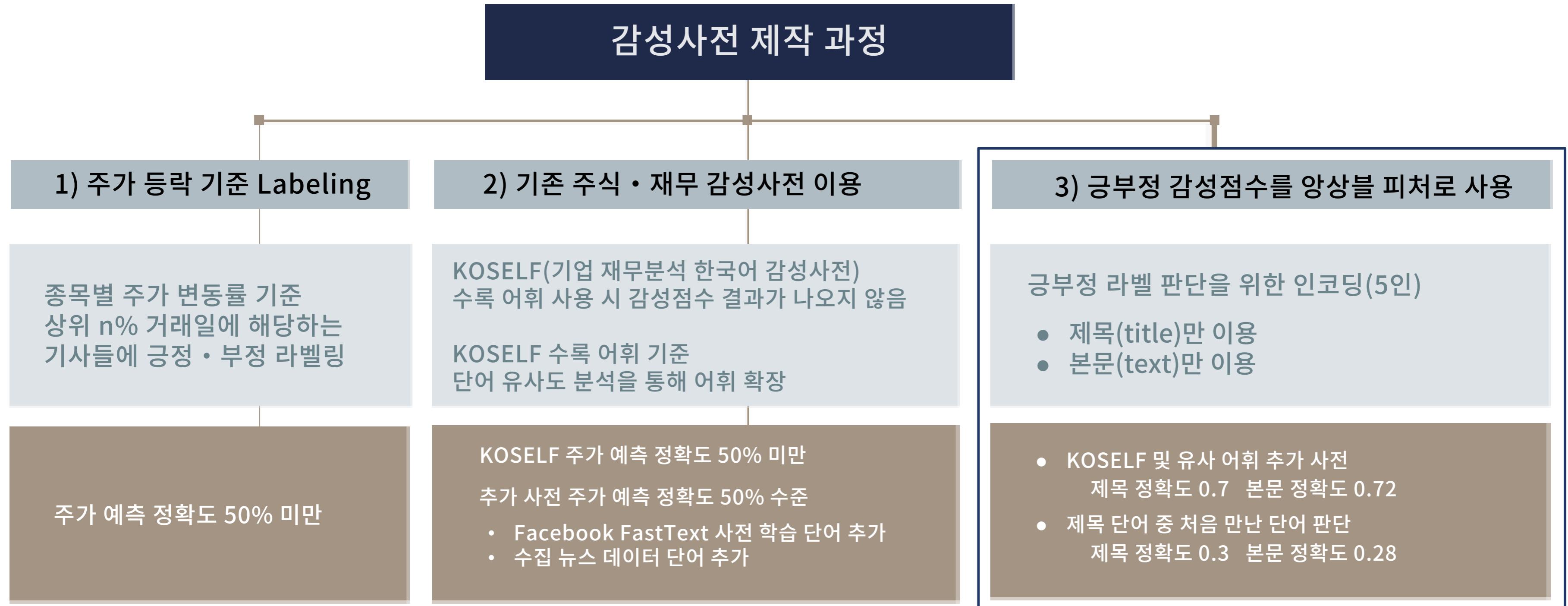


POSITIVE 긍정어 (48) 가치 가치 있는 강세 개선 개선된 개선되는 경신 경의 고급 기꺼이 더 좋은 도움이 되는 뛰어나 명성 바닥 본격적인 상승효과
상회한 성공 수혜 순조롭게 실현 안정된 완전한 우세 우월 우호적 의심의 여지가 없는 이끄는 이로운 인상적인 장점 적극 적극적으로 정성 중요한 증대 최고점 최선 추진 추천 충분히 충족 해결 호황 확보 활력 활성화

NEGATIVE 부정어(47) 결국 결함 공허한 과적 극심한 둔화 마이너스 몰수 미완성 배상 부적합한 부정적 부정적인 부주의 불리한 소란 스캔들 실패
악화 약점 약화 여파 연기 우발적 의심 의혹 잘못 저품질 저해 정체 조치 주의 지나친 지독한 지연 질타 차질 초라한 충격 침체
투자 회수 파산 평가절하 하락 해체 혼란 훼손

2-1. 감성 분석

감성사전 제작



2-1. 감성 분석

감성사전 기준 사전 이용 감성사전의 어휘와 뉴스 기사 토큰화 단어들 간의 유사도 확인

```
pos_lexicon_2018 = pd.DataFrame(columns=['date', 'news_num', 'KOSELF_pos_word', 'news_word',  
  
# KOSELF_pos와의 Cosine Similarity 계산  
for x in range(len(lexicon_2018['Tokenization'])):  
    news_num = x+1  
    for y in range(len(positive)):  
        for z in range(len(list(set(lexicon_2018['Tokenization'][x].split())))):  
            if (ko_model.wv.similarity(positive[y], list(set(lexicon_2018['Tokenization'][x].split()))[z])):  
                freq = 0  
                for w in range(len(lexicon_2018['Tokenization'][x].split())):  
                    if lexicon_2018['Tokenization'][x].split()[w] == list(set(lexicon_2018['Tokenization'][x].split()))[z]:  
                        freq += 1  
                data = {  
                    'date': lexicon_2018['date'][x],  
                    'news_num': news_num,  
                    'KOSELF_pos_word': positive[y],  
                    'news_word': list(set(lexicon_2018['Tokenization'][x].split()))[z],  
                    'cosine_similarity': ko_model.wv.similarity(positive[y], list(set(lexicon_2018['Tokenization'][x].split()))[z]),  
                    'frequency': freq  
                }  
                pos_lexicon_2018 = pos_lexicon_2018.append(data, ignore_index=True)  
print("{{0}} Cosine Similarity between <{1}> & <{2}> : {3}".format(lexicon_2018['Tokenization'][0], lexicon_2018['Tokenization'][1], lexicon_2018['Tokenization'][2], lexicon_2018['Tokenization'][3]))
```

```
***2018-01-02 Cosine Similarity between <경신> & <기록> : 0.522990882396698  
***2018-01-02 Cosine Similarity between <장점> & <특성> : 0.5244326591491699  
***2018-01-02 Cosine Similarity between <경신> & <기록> : 0.522990882396698  
***2018-01-02 Cosine Similarity between <의심의 여지가 없는> & <자공시> : 0.5188294053077698  
***2018-01-02 Cosine Similarity between <증대> & <증가> : 0.6946426630020142  
***2018-01-02 Cosine Similarity between <추진> & <증인> : 0.539746880531311  
***2018-01-02 Cosine Similarity between <호황> & <훈풍> : 0.5522076487541199  
***2018-01-02 Cosine Similarity between <추진> & <사업> : 0.5213737487792969  
***2018-01-02 Cosine Similarity between <추진> & <계획> : 0.5205424427986145  
***2018-01-02 Cosine Similarity between <감세> & <약세> : 0.7028064727783203  
***2018-01-02 Cosine Similarity between <개선> & <방안> : 0.5635414123535156  
***2018-01-02 Cosine Similarity between <실현> & <차익> : 0.531851053237915
```

	date	news_num	KOSELF_pos_word	news_word	cosine_similarity	frequency
0	2018-01-02	1	경신	기록	0.522991	1
1	2018-01-02	1	장점	특성	0.524433	8
2	2018-01-02	2	경신	기록	0.522991	13
3	2018-01-02	2	의심의 여지가 없는	자공시	0.518829	1
4	2018-01-02	2	증대	증가	0.694643	2
...
17939	2018-12-12	4144	의심의 여지가 없는	핸디레이	0.577344	3
17940	2018-12-12	4144	증가	가장	0.537125	17
17941	2018-12-12	4144	추진	사업	0.521374	1
17942	2018-12-12	4144	추진	계획	0.520542	1
17943	2018-12-12	4144	추진	진행	0.535857	1

1. FastText 한글 모델 단어 중 KOSELF 단어와 유사도 비교
→ 높은 단어 상위 각 10개를 추가 감성사전 ver1.0 구축

2. 크롤링 뉴스 데이터 안의 단어와 KOSELF 단어의 유사도
→ 유사도>0.5인 단어를 추가 종목별 감성사전 ver1.0 구축

2-1. 감성 분석

감성사전 감성점수 계산 후 양상블 피처 사용

감성점수 계산의 정확도 확인을 위해 팀원 5인 개별 감성 라벨 인코딩(제목, 본문) 후 다수결 방식으로 최종 라벨링

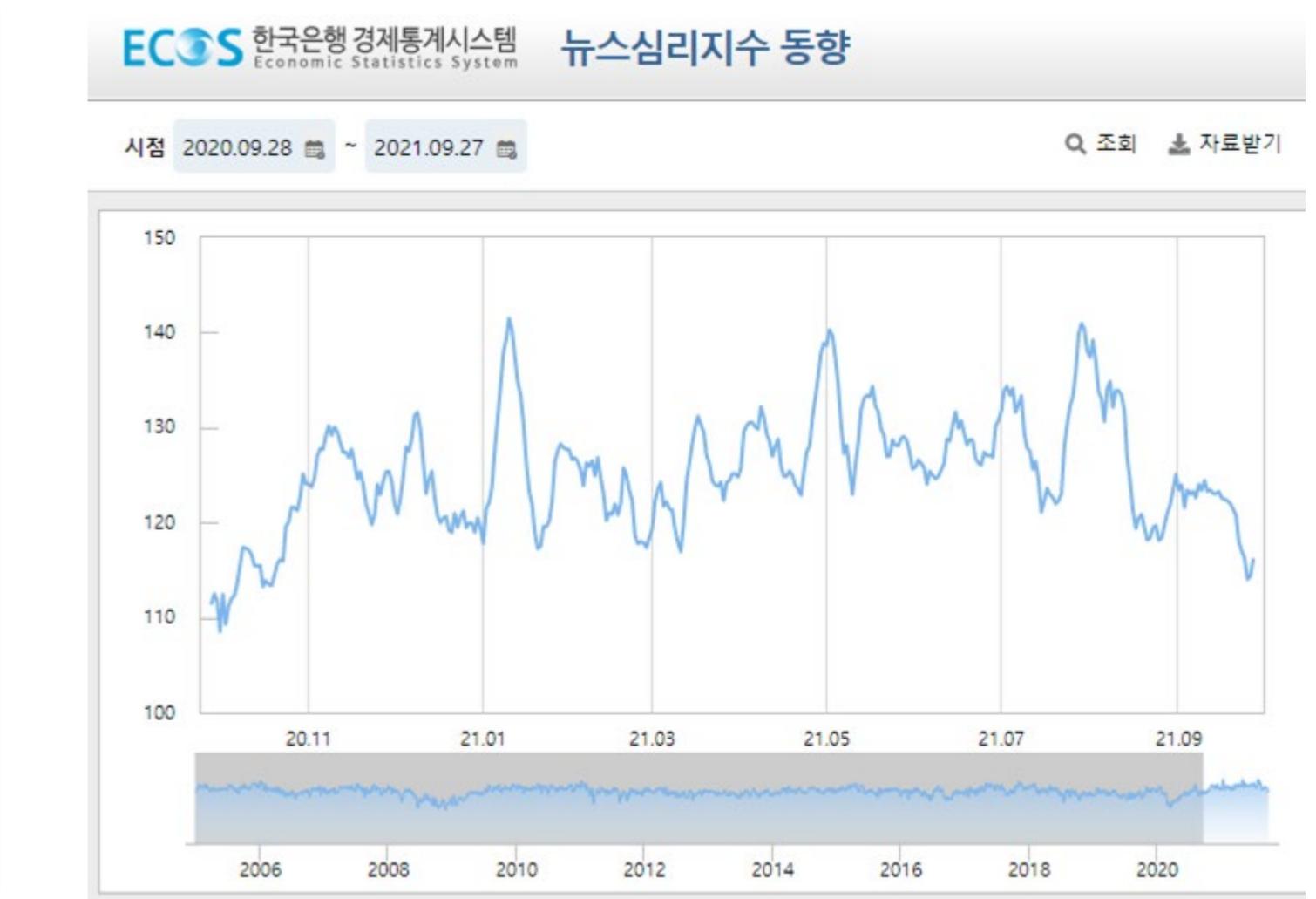
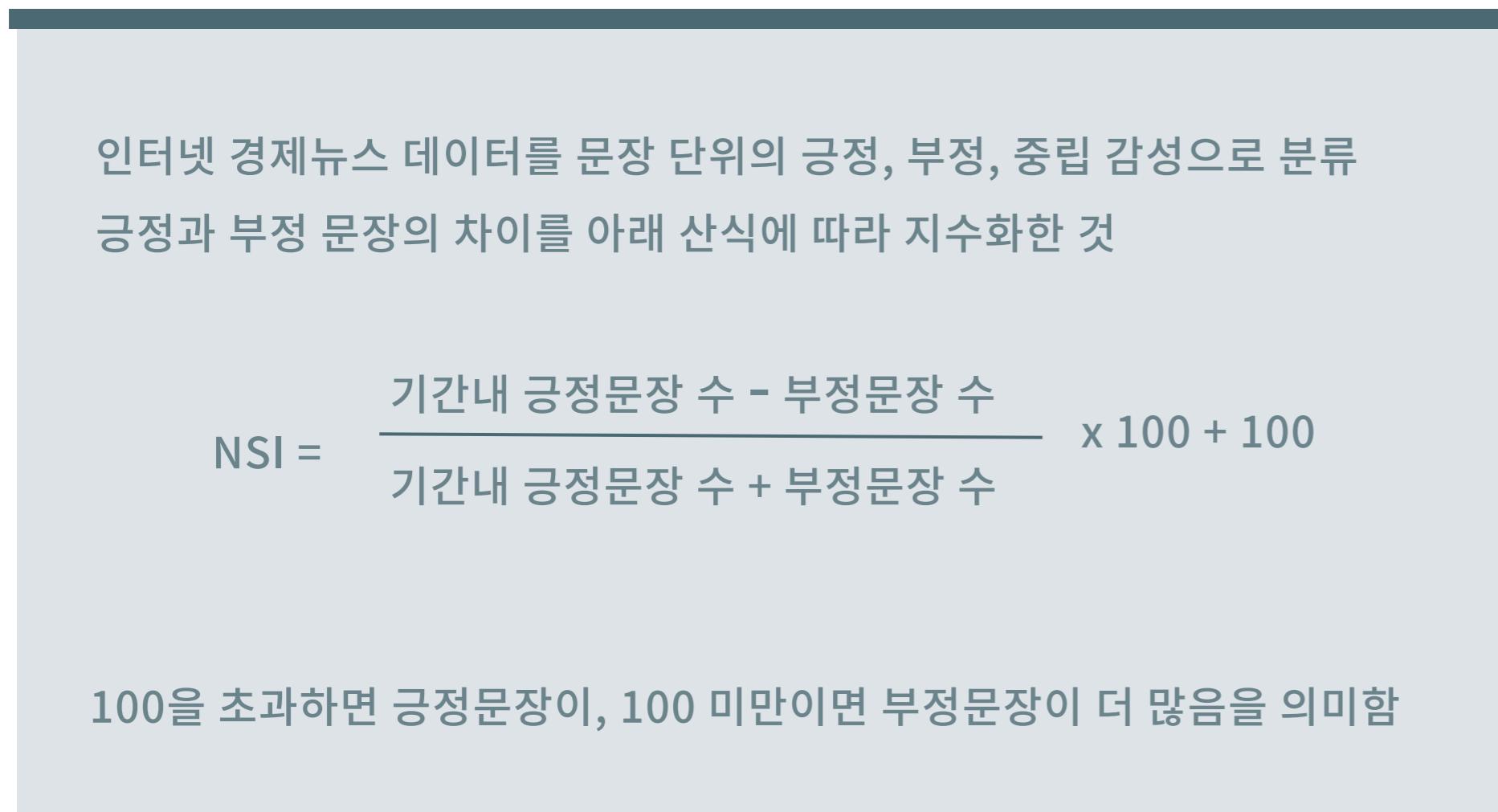
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
1	index	sample_in	st_n	st_cd	news	date	title	url	text	label_titl	label_te	label_te	label_te	label_te	label_te	label_text						
2	0	1277	sk하이닉스	660	아시아경제	2018030509	코스피, 상승 출발 후 하락 반전...2400선 붕괴(종합)	http://view.asiae.co.kr		-1	-1	-1	-1	-1		1	1	-1	-1	-1		
3	1	4010	sk하이닉스	660	아시아경제	2021032108	반도체 부품 수요 확대 추세...원익QnC 살아날까	https://view.asiae.co.kr	미국 한파에 오스	1	1	1	1	1			1	1	1	1	1	
4	2	2215	sk하이닉스	660	아시아경제	2019072909	코스피, FOMC까지 관망심리 커져...2050선까지 후퇴	https://view.asiae.co.kr	[아시아경제 박훈]	-1	-1	-1	-1	-1			-1	-1	-1	-1	-1	
5	3	2217	sk하이닉스	660	아시아경제	2019071916	미국 금리인하 기대감...코스피 2090선 '회복'	https://view.asiae.co.kr	[아시아경제 유현]	1	1	1	1	1			1	1	1	1	1	
6	4	3825	sk하이닉스	660	아시아경제	2021050311	[속수무책 기술유출]美 첨단산업 경제에 궁지 몰린 中...	https://view.asiae.co.kr	반도체·2차전지	-1	-1	-1	-1	-1			-1	-1	-1	-1	-1	
7	5	642	현대차	5380	아시아경제	2018041213	[포토] 현대차 '넥쏘' 내부 공개	http://view.asiae.co.kr	12일 서울 강남구	0	1	1	1	1			0	1	1	1	1	
8	6	689	현대차	5380	아시아경제	2018121112	이광국 부사장 "팰리세이드, 폭발적 반응에 판매 목표	http://view.asiae.co.kr	팰리세이드, 사전	1	1	1	1	1			1	1	1	1	1	
9	7	7228	현대차	5380	아시아경제	2021032416	끌내 3000 밀으로...外人·기관 매도세에 내려 앉은 코스닥은 0.8%↑	https://view.asiae.co.kr	코스닥은 0.8%↑	-1	-1	-1	-1	-1			-1	-1	-1	-1	-1	
10	8	106	현대차	5380	아시아경제	2018101815	현대차 정몽구 재단, '산의 날' 유공 대통령 표창 수상	http://view.asiae.co.kr		1	1	1	1	1			1	1	1	0	1	
11	9	2186	현대차	5380	아시아경제	2019041515	"유럽 수소상용차 시장 공략"...현대차, 스위스 수소에	http://view.asiae.co.kr	[아시아경제 우수]	1	1	1	1	1			1	1	1	1	1	
12	10	12472	삼성전자	5930	아시아경제	2021022109	외국인, 3주만에 '팔자' 전환...SK그룹株는 사들여	https://view.asiae.co.kr	[아시아경제 송호]	0	-1	1	1	-1		0	-1	-1	1	1	-1	
13	11	4777	삼성전자	5930	아시아경제	2019050111	[택트체크] 青"문 대통령과 이재용 7번 만났다고 하는	http://view.asiae.co.kr	문재인 대통령이	0	1	-1	-1	0		0	0	1	-1	0	0	0
14	12	13152	삼성전자	5930	아시아경제	2021051109	국내 증시, 고점 부담·美기술주 하락 여파에 동반 약세	https://view.asiae.co.kr	[이미지출처=연합뉴스]	-1	-1	-1	-1	-1			-1	-1	-1	-1	-1	
15	13	6100	삼성전자	5930	아시아경제	2019072508	전자랜드 서산점, '파워센터'로 리뉴얼 오픈	https://view.asiae.co.kr	[아시아경제 성기]	1	1	0	0	1			1	1	0	0	1	
16	14	3469	삼성전자	5930	아시아경제	2018091815	[평양회담]특별수행단, 北김영남 접견...경제인, 리용호	http://view.asiae.co.kr	정당 대표단, 北朝	0	0	1	1	1			1	0	1	1	1	
17	15	2807	LG화학	51910	아시아경제	2021042012	오후 코스피, 외국인·기관 순매수 속 3220선 등락	https://view.asiae.co.kr	20일 서울 종구	0	1	1	1	1			1	1	1	1	1	
18	16	1696	LG화학	51910	아시아경제	2021082410	동반성장위원회-LG화학, '협력사 ESG 지원사업' 협약	https://view.asiae.co.kr	교육·역량진단·컨	1	1	1	1	1			1	1	1	1	1	
19	17	2722	LG화학	51910	아시아경제	2021051815	기관매수 지속...코스피 1%대 상승 유지	https://view.asiae.co.kr	코스닥도 970선	1	1	1	1	1			1	1	1	1	1	
20	18	2053	LG화학	51910	아시아경제	2021041915	[단독]에이프로, LG엔솔-GM 배터리공장에 1천억 규모	https://view.asiae.co.kr	합작 배터리 1공	1	1	1	1	1			1	1	1	1	1	
21	19	1286	LG화학	51910	아시아경제	2018081109	[증권사 주간 추천 종목]SK증권	http://view.asiae.co.kr		1	-1	0	0	1		1	1	1	0	1	1	
22	20	3247	셀트리온	68270	아시아경제	2019092210	코스피, 11일 연속 상승...2100 넘어설까	https://view.asiae.co.kr	샘蹂면낫湲 늘쳤	1	1	1	1	0			1	1	1	1	-1	
23	21	4050	셀트리온	68270	아시아경제	2018012006	다음 주 코스피 2490~2550 전망...주요국 통화정책 회	http://www.asiae.co.kr	지난해 4Q 실적	1	1	1	0	1			1	1	1	1	1	
24	22	2691	셀트리온	68270	아시아경제	2020081909	코스피, 개인 순매수 나서며 상승 출발 ... 코스닥 2%↑	https://view.asiae.co.kr	[이미지출처=연합뉴스]	1	1	1	1	1			1	1	1	1	1	
25	23	2901	셀트리온	68270	아시아경제	2020032515	코스피, 글로벌 경기 부양 기대감에 1700선 회복... 코스닥 2%↑	https://view.asiae.co.kr	[이미지출처=연합뉴스]	1	1	1	1	1			-1	-1	0	-1	1	
26	24	626	셀트리온	68270	아시아경제	2021052409	[클릭 e종목]"마이크로디지탈, 코로나 백신 생산 필수품	https://view.asiae.co.kr	[아시아경제 유현]	1	1	1	1	1			1	1	1	1	1	
27	25	1252	sk하이닉스	660	매일경제	2019090615	코스피, 미중 무역협상 재개 기대감에 사흘째 오름세	http://news.mk.co.kr	미국과 중국이 두	1	1	1	1	1			-1	-1	0	-1	1	
28	26	1211	sk하이닉스	660	매일경제	2019012017	"SK 혁신기술력 널리 전파"...최재원 수석부회장의 주	http://news.mk.co.kr		1	1	1	1	1			1	1	1	1	1	
29	27	3342	sk하이닉스	660	매일경제	2021061007	매경이 전하는 세상의 지식 (매-세-지, 6월 10일)	http://premium.mk.co.kr	매경이 전하는 서	0	0	0	0	0		0	1	1	0	-1	0	0
30	28	306	sk하이닉스	660	매일경제	2018020513	[컨콜]SKT 배당 SK하이닉스 배당과 연계하기 어	http://news.mk.co.kr	유영상 SK텔레콤	-1	0	-1	-1	0		-1	1	1	0	-1	0	1

2-1. 감성 분석

감성점수 뉴스심리지수 NSI

한국은행 경제통계국에서 개발한 뉴스심리지수를 추가적인 감성점수 계산에 사용

뉴스심리지수(NSI, News Sentiment Index)



2-1. 감성 분석

감성점수 유튜브 스크립트 유튜브에서 제공되는 STT(Speech-To-Text) 스크립트 전처리

STT 구현이 완벽하지 않아서 잘못 입력된 단어들이 많이 분포

- 자동 생성된 스크립트의 의성어, 특수문자, 반복되는 의미없는 단어 등 삭제
- 기존 뉴스데이터에서 사용하던 불용어 사전을 통해 불용어를 삭제
토큰화를 진행한 후에 단어 길이가 1 이하인 단어 삭제

스크립트	스크립트
00:05 [음악]	00:25 게임을 못 만들고 네슨 안만들고
00:15 으	00:27 넷마블은 외 만드냐 라고 얘길 했다
00:17 으 으	00:29 이게 그 nc 는 목 기로에 서 있고
00:20 cool	00:31 네슨은 지금 종합 엔터테인먼트 회사로
00:21 으	00:34 가려고 넷마블은 추자 를 열심히
00:22 [음악]	00:36 하니까 이런 얘길 했는데 그거 이렇게
00:27 으	00:38 예약하시면 상당히 안됩니다 왜냐하면
00:34 으	00:39 x 게임을 안 만드는 게 아니죠 뒤
00:39 [음악]	00:42 10시 만들어요 저 수백 명의 아무
00:58 [음악]	00:44 만들고 있을 거야 첨 며 이상이
01:08 에	00:46 게임을 만들고 있지 않을까 하는데
01:08 오랜만에 팽 달 알고	00:47 안에 있는 분들이야 게임에 다시
한국어 (자동 생성됨)	한국어 (자동 생성됨) 맞춤법이 맞지 않는 단어와 [음악]과 같은 의성어가 많음

1

수작업으로 의미 없는 단어 삭제

```
youtube_df.text = youtube_df.text.str.replace('[으*]', '')
youtube_df.text = youtube_df.text.str.replace('[₩₩]', '')
youtube_df.text = youtube_df.text.str.replace('음악', '')
youtube_df.text = youtube_df.text.str.replace('자', '')
youtube_df.text = youtube_df.text.str.replace('[a-zA-Z]', '')
youtube_df.text = youtube_df.text.str.replace('[구독*]', '')
youtube_df.text = youtube_df.text.str.replace('[좋아요]', '')
youtube_df.text = youtube_df.text.str.replace('[₩s]', '')
youtube_df.text = youtube_df.text.str.strip() # 앞뒤 공백 제거
```

2

토큰화 진행 후 불용어 사전 이용 및 단어 길이 1 이하 단어 삭제

```
# 토큰화
for i in range(start, len(youtube_df)):
    temp = []

    temp = okt.nouns(youtube_df['text'].iloc[i])
    temp = [word for word in temp if not word in stopwords] # 불용어 제거
    temp = [word for word in temp if len(word) > 1] # 불용어 제거
    # print(i, '번 토큰화, 불용어 제거 완료')

    temp = list(set(temp))
```

2-1. 감성 분석

감성점수 유튜브 스크립트 라벨링 작업을 위한 모델 선정

1. 기존의 감성 사전 패키지를 통해 라벨링을 진행

- VADER 감성사전 ⇒ 거의 모든 자료를 중립 라벨링
- rhinoMorph 감성사전 ⇒ 거의 모든 자료를 긍정 라벨링

▶ 맞춤법에 맞지 않거나 신조어 인식을 못하는 경우가 있고,
여러 사람이 대화를 하면서 방송을 진행하는 등의 이유로
라벨링이 제대로 되지 않는 것을 확인함

```
sent_i_score = sent_i_analyzer.polarity_scores('꿀잼')
print(sent_i_score)

[nltk_data] Downloading package vader_lexicon to /root/nltk_data...
[nltk_data] Package vader_lexicon is already up-to-date!
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

```
filepath: /usr/local/lib/python3.7/dist-packages
classpath: /usr/local/lib/python3.7/dist-packages/rhinoMorph/lib/rhino.jar
RHINO started!
rn
<java class='rhino.RHINO'>
0.0
['0']
<class 'numpy.ndarray'>
긍정적인 글입니다
```

```
rn = rhinoMorph.startRhino()
print('rn', rn)
new_input = '재미 더럽 없'
```

2. 라벨링 된 리뷰를 학습 데이터로 사용하여 텍스트 임베딩 모델로 라벨링을 진행

- 구매 리뷰 데이터
⇒ 긍정 데이터가 많아 결과가 긍정적으로 편향
- 네이버 영화 댓글 데이터
⇒ 데이터에 긍정과 부정이 적절하게 섞여 있음

▶ 실제 사람들의 글을 통해서 훈련시켜 실제 수작업으로 라벨링 한
부분과의 정확도가 1번보다 높게 나옴

```
labeling.label.value_counts()
```

```
1 5816
0 547
.. ..
```

0.76

```
youtube_df.naver.value_
1 4716
-1 1647
```

0.8

PART 2-2

데이터 분석

- ◆ 시계열 분석 및 강화 학습
- ◆ 앙상블

Sentiment analysis of news articles and YouTube texts and
prediction of stock price fluctuations using deep learning models

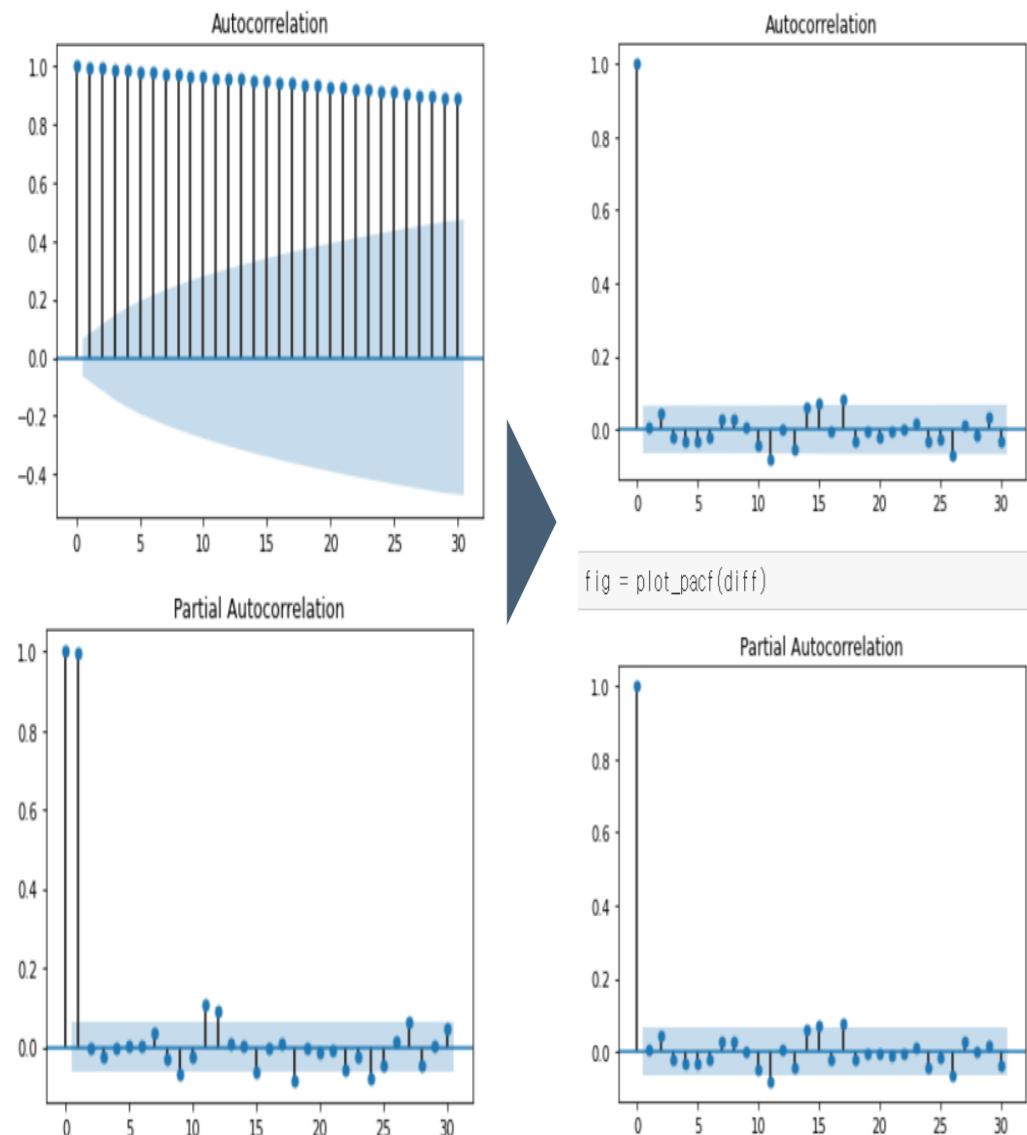
2-2. 데이터 분석

시계열 분석 ARIMA 파라미터 조정

데이터의 정상성을 확인하기 위해
종목별로 종가 그래프, ACF그래프,
PACF 그래프를 확인

Augmented Dickey-Fuller Test를 통해
적절한 차분 수를 선정하고 시계열 정상성 검정

AR, MA 차수를 변경하면서 적절한 모델 선택
삼성전자의 경우
const를 제외한 AR : 2, 차분 : 1, MA : 2일 때 유의한 결과



로그 변환과 차분 진행



ADF statistic: -0.5569406214310959
p-value: 0.880417913118139

ADF statistic: -30.033008581502763
p-value: 0.0

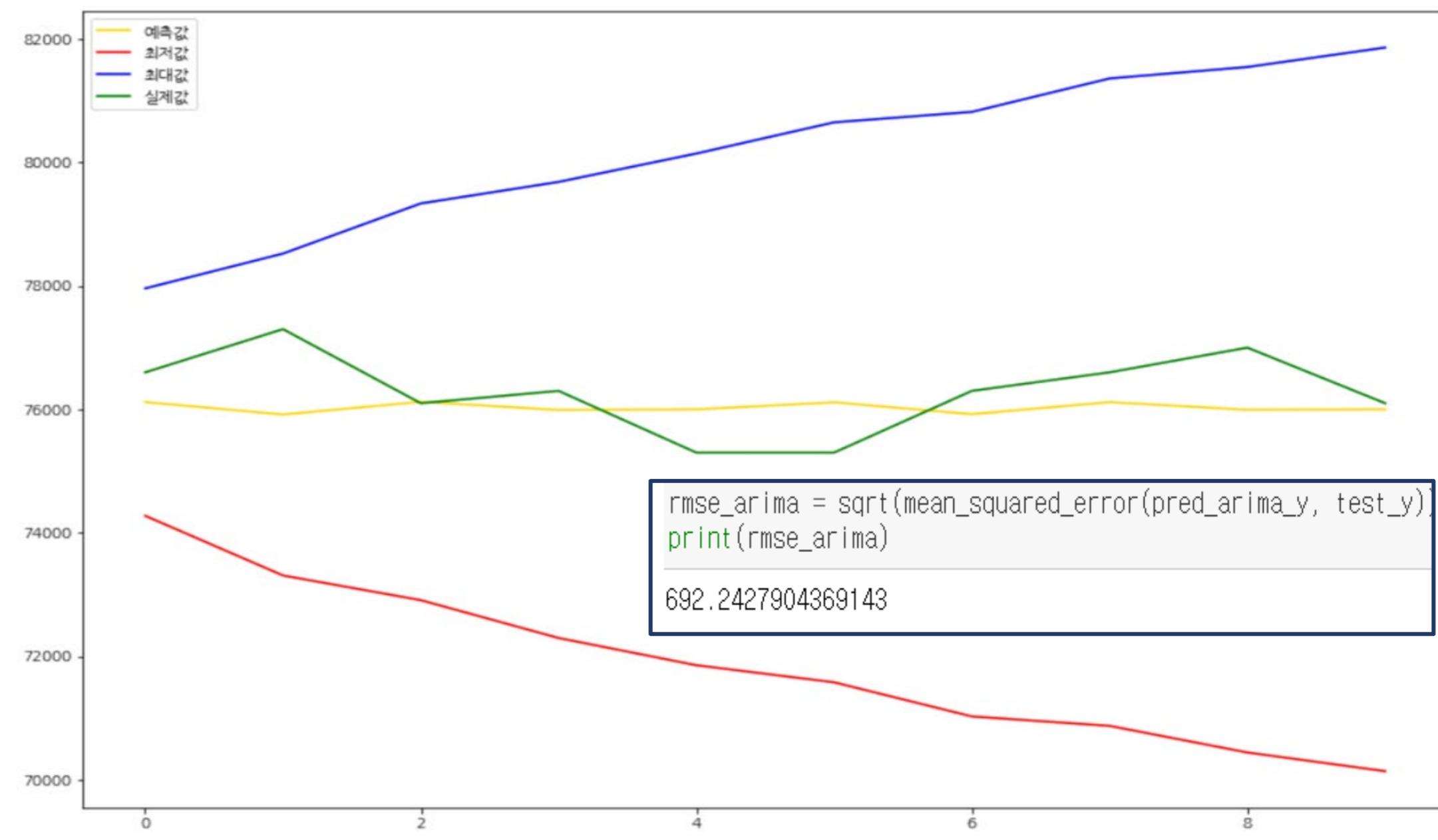
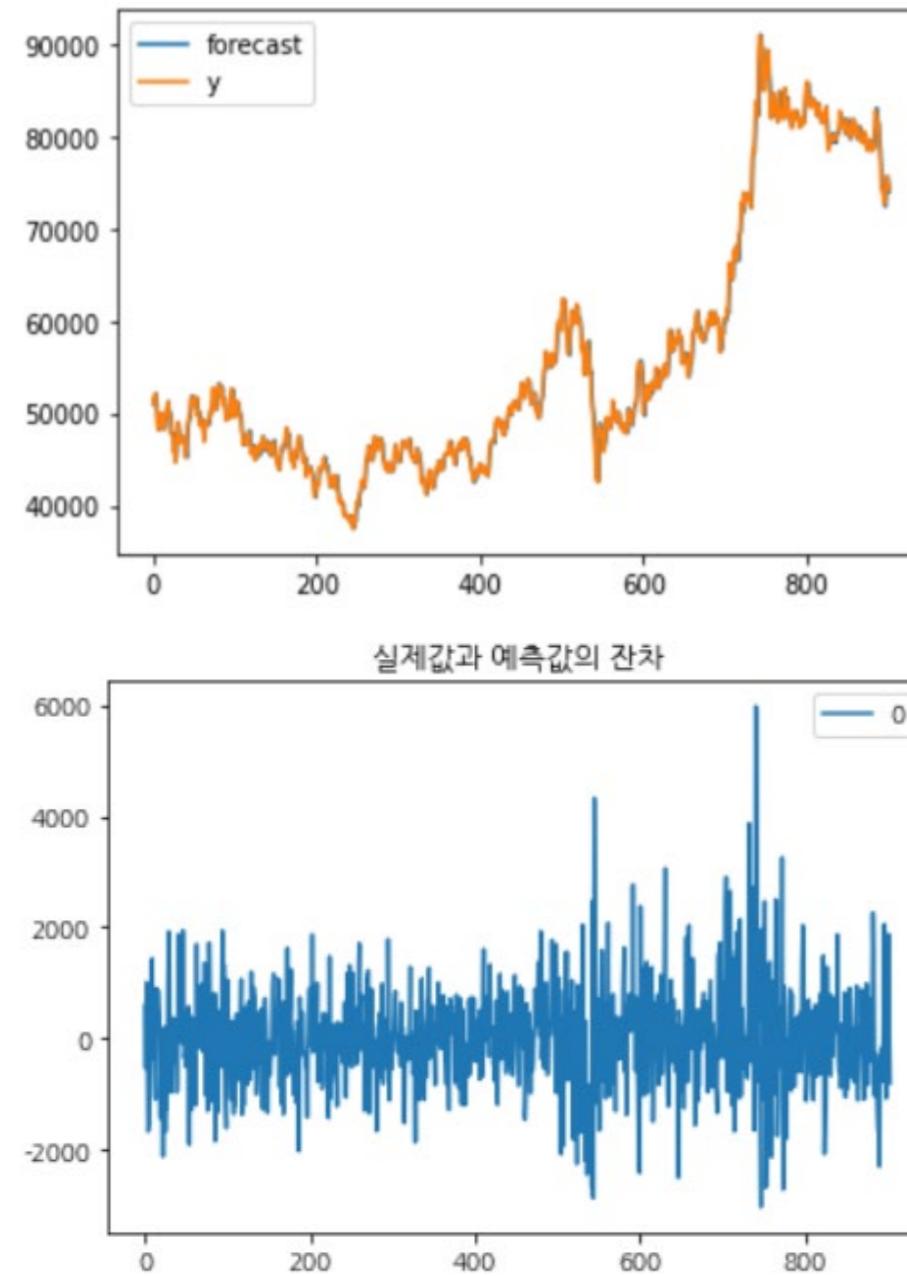
ARIMA Model Results							ARIMA Model Results							
Dep. Variable:	D.y	No. Observations:	912	Dep. Variable:	D.y	No. Observations:	912	Dep. Variable:	D.y	No. Observations:	912			
Model:	ARIMA(2, 1, 2)	Log Likelihood	-7538.858	Model:	ARIMA(2, 1, 2)	Log Likelihood	-7539.154	Model:	ARIMA(2, 1, 2)	Log Likelihood	-7539.154			
Method:	css-mle	S.D. of innovations	938.892	Method:	css-mle	S.D. of innovations	939.200	Method:	css-mle	S.D. of innovations	939.200			
Date:	Mon, 13 Sep 2021	AIC	15089.716	Date:	Mon, 13 Sep 2021	AIC	15088.309	Date:	Mon, 13 Sep 2021	AIC	15088.309			
Time:	08:45:47	BIC	15118.609	Time:	08:45:53	BIC	15112.387	Time:	08:45:53	BIC	15112.387			
Sample:	1	HQIC	15100.746	Sample:	1	HQIC	15097.501	Sample:	1	HQIC	15097.501			
coef std err z P> z [0.025 0.975]							coef std err z P> z [0.025 0.975]							
const	13.7394	31.285	0.439	0.661	-47.577	75.056	const	-1.5978	0.011	-139.518	0.000	-1.620	-1.575	
ar.L1.D.y	-1.5979	0.012	-138.725	0.000	-1.620	-1.575	ar.L1.D.y	-0.9801	0.009	-113.423	0.000	-0.997	-0.963	
ar.L2.D.y	-0.9802	0.009	-113.114	0.000	-0.997	-0.963	ar.L2.D.y	1.5997	0.008	208.092	0.000	1.585	1.615	
ma.L1.D.y	1.5998	0.008	207.644	0.000	1.585	1.615	ma.L1.D.y	0.9999	0.008	132.236	0.000	0.985	1.015	
ma.L2.D.y	0.9999	0.008	132.669	0.000	0.985	1.015	ma.L2.D.y	Roots						
Real Imaginary Modulus Frequency							Real Imaginary Modulus Frequency							
AR.1	-0.8151	-0.5965j	1.0101	-0.3995	AR.1	-0.8151	-0.5965j	1.0101	-0.3995	AR.2	-0.8151	+0.5965j	1.0101	0.3995
AR.2	-0.8151	+0.5965j	1.0101	0.3995	AR.2	-0.8151	+0.5965j	1.0101	0.3995	MA.1	-0.8000	-0.6001j	1.0001	-0.3976
MA.1	-0.8000	-0.6001j	1.0001	-0.3976	MA.1	-0.8000	-0.6001j	1.0001	-0.3976	MA.2	-0.8000	+0.6001j	1.0001	0.3976

2-2. 데이터 분석

시계열 분석 ARIMA 학습 결과

마지막 10일을 예측하기 위한 Train-Test Set 분할 및 학습 진행 결과
5, 10, 20일 중에 10일 구간이 RMSE가 가장 작음(5일 : 약 1,100, 20일 : 약 2,200)

구성한 모델을 통해 10일 예측 결과 비교 및 RMSE

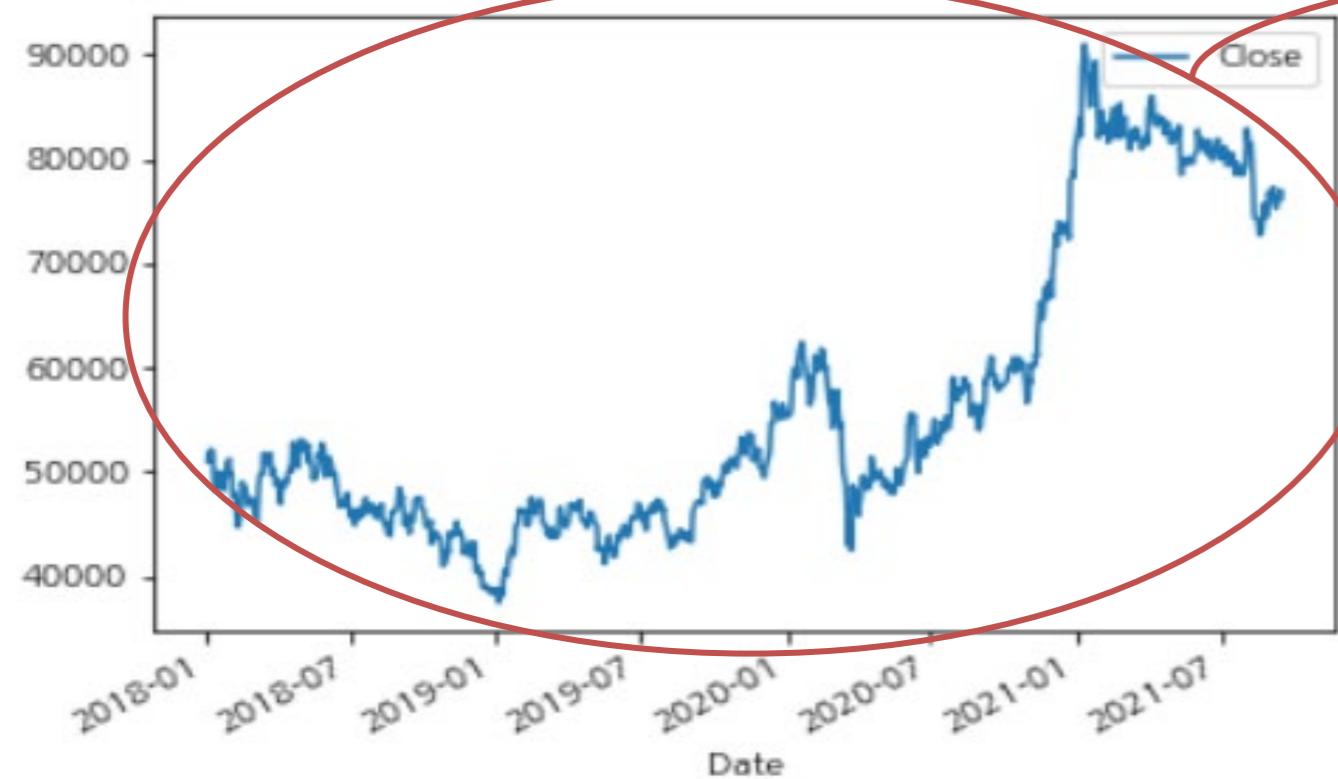


2-2. 데이터 분석

시계열 분석 FBProphet 모델 선정 이유 및 파라미터 조정

선형 회귀 방법인 ARIMA 모델의 단점을 극복하기 위해서 Additive 모델을 기반한 Prophet 시계열 모델을 사용

- ◆ FBProphet은 시계열 데이터의 트렌드성(연간/월간/일간)을 예측하는 것에 초점이 맞추어져 있음
- ◆ FBProphet 모델을 사용하기 위해 데이터를 ds - y 형태로 전처리 시행



```
prophet = Prophet(seasonality_mode = 'multiplicative',
                   yearly_seasonality=True,
                   weekly_seasonality=True,
                   daily_seasonality=True,
                   changepoint_prior_scale=0.6)
```

```
prophet.fit(train)
```

0.5, 0.6, 0.7로 시행한 결과 0.6 이상이 넘어가면 과적합이 발생 → 0.6으로 선택

- **seasonality_mode** : 연간, 월간, 주간, 일간 등의 트렌드성을 반영하는 것을 의미하는 파라미터
- Additive는 데이터의 진폭이 일정함을 의미하고, Multiplicative는 데이터의 진폭이 점점 증가하거나 감소하는 것을 뜻함
- **changepoint_prior_scale** : 트렌드가 변경되는 문맥을 반영하는 파라미터로, 수치가 높을수록 모델은 과적합에 가까워짐

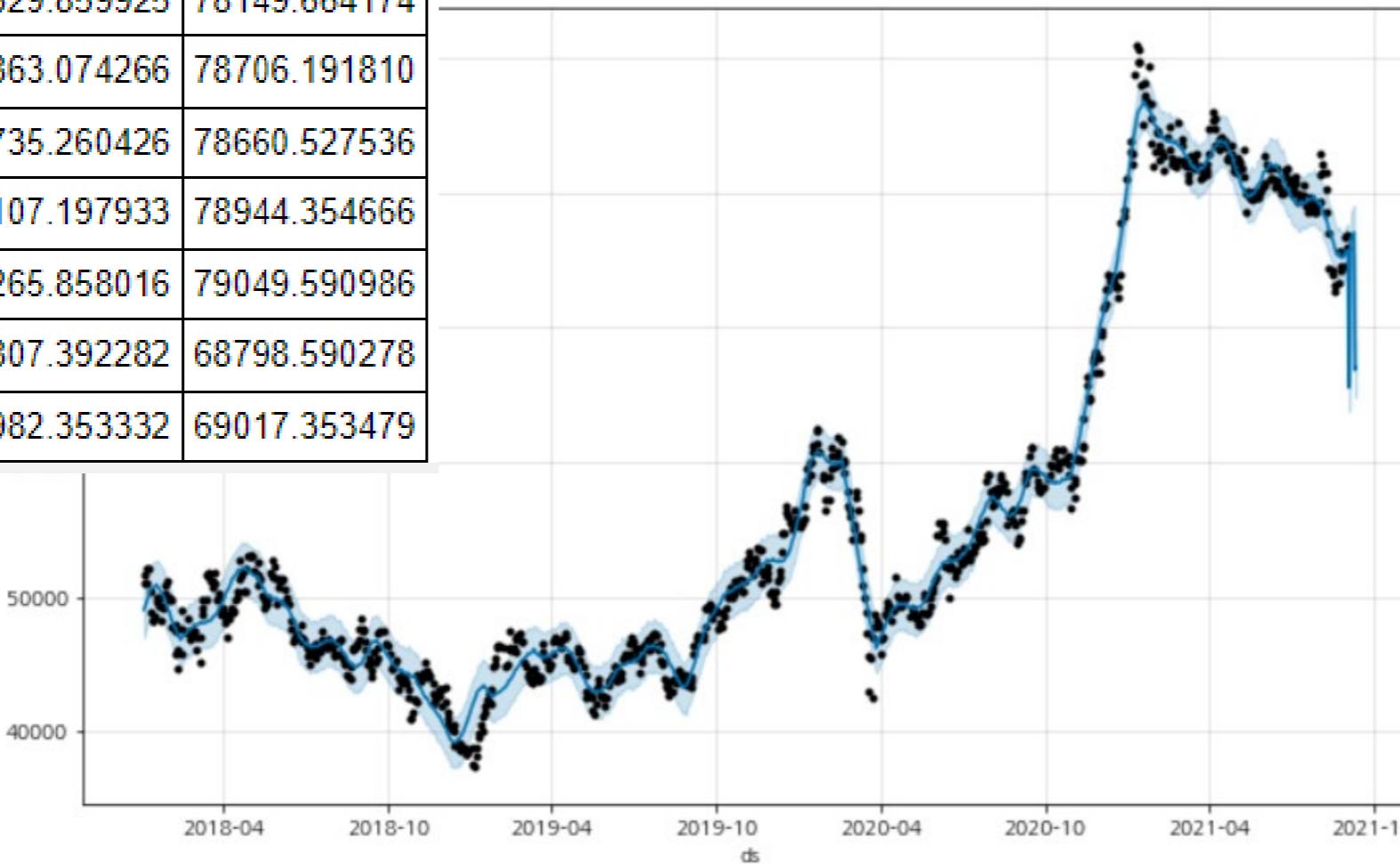
2-2. 데이터 분석

시계열 분석 FBProphet 마지막 10일을 예측하기 위한 Train-Test Set 분할 및 학습 진행 결과

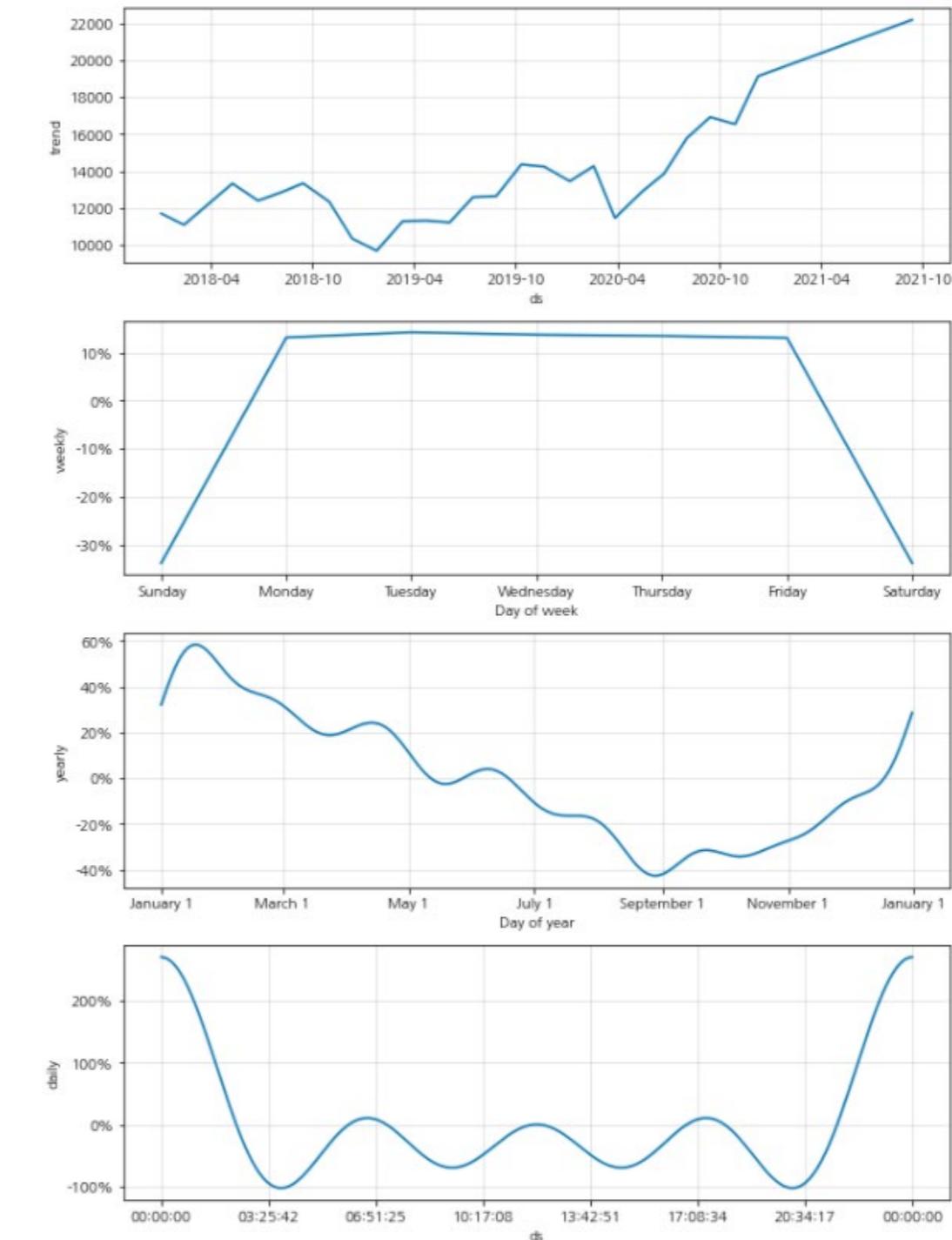
구성한 모델을 통해 10일 예측값 및 신뢰구간

	ds	yhat	yhat_lower	yhat_upper
906	2021-09-03	75735.972526	73785.030477	77689.028833
907	2021-09-04	65527.745085	63776.890039	67336.098026
908	2021-09-05	65696.858848	63776.494852	67609.315805
909	2021-09-06	76275.115353	74529.859925	78149.664174
910	2021-09-07	76709.484111	74863.074266	78706.191810
911	2021-09-08	76790.502887	74735.260426	78660.527536
912	2021-09-09	76933.875584	75107.197933	78944.354666
913	2021-09-10	77041.958459	75265.858016	79049.590986
914	2021-09-11	66830.439298	64807.392282	68798.590278
915	2021-09-12	67013.161507	64982.353332	69017.353479

모델 적합 결과



트렌드 곡선

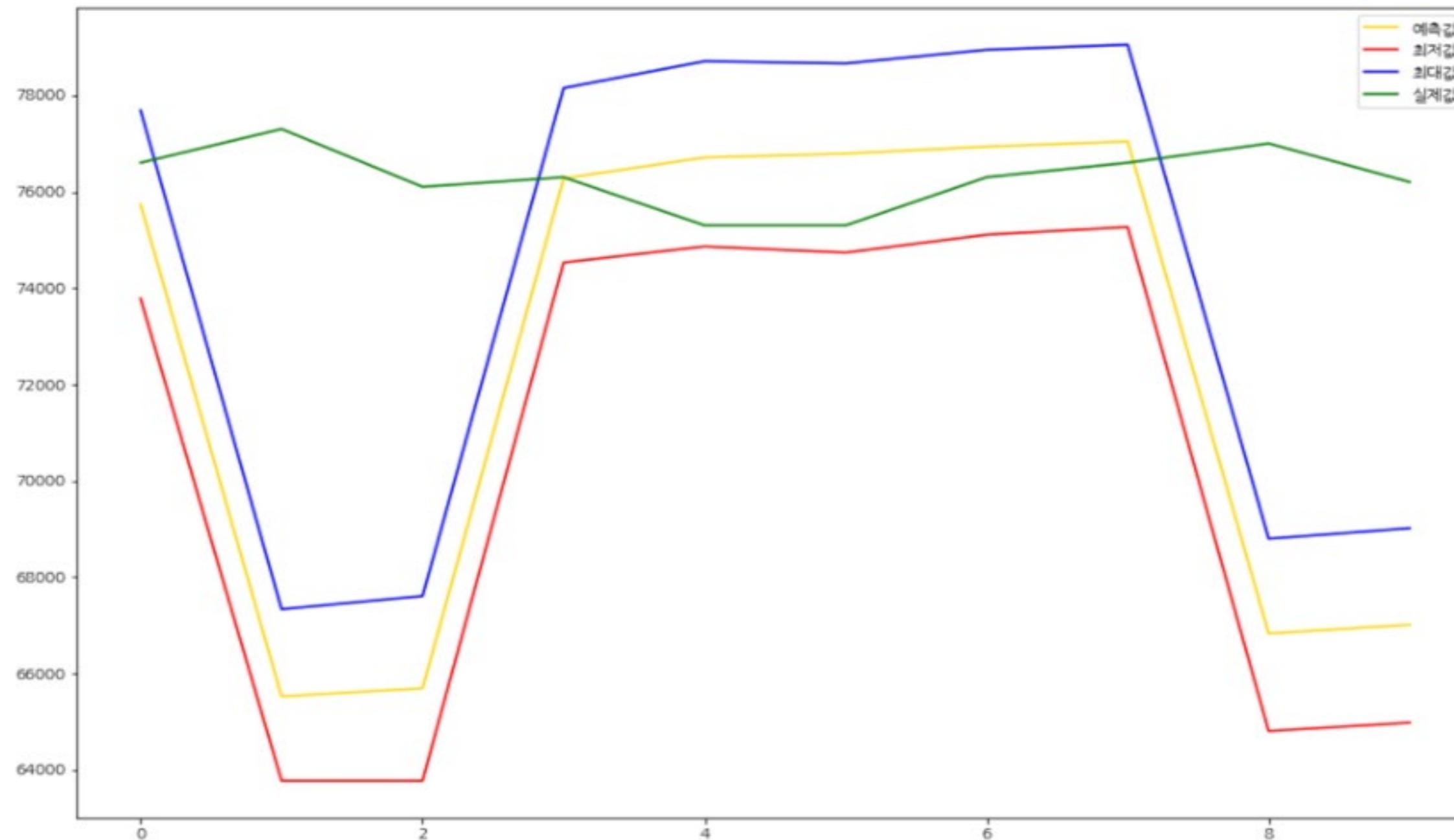


2-2. 데이터 분석

시계열 분석 FBProphet 마지막 10일을 예측하기 위한 Train-Test Set 분할 및 학습 진행 결과

```
prophet_arima = sqrt(mean_squared_error(pred_fbprophet_y, test_y))  
print(prophet_arima)
```

```
6634.620038351937
```



2-2. 데이터 분석

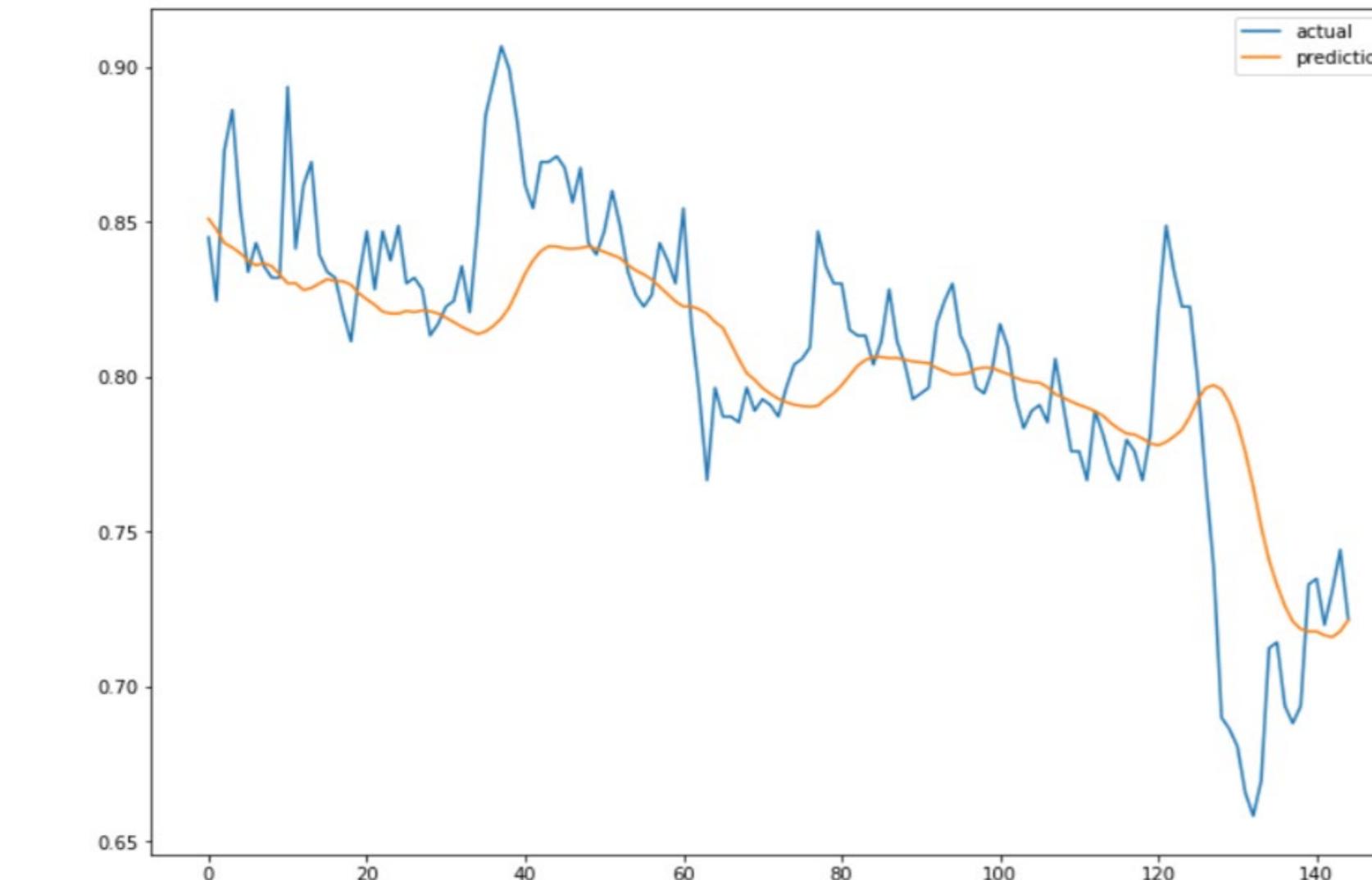
시계열 분석 LSTM

날짜 기반으로 종가만 예측하는 모델과 재무지표를 포함한 데이터로 종가를 예측하는 모델을 시행 해보고 결과를 비교

확인 결과 재무지표를 포함한 데이터가 더 정밀하게 예측하는 것으로 보임

- window size \Rightarrow 10일 기준으로 shift 진행
- feature = 12개(주식 정보 + 재무지표) * 10일 = 120개

주식 정보 LSTM



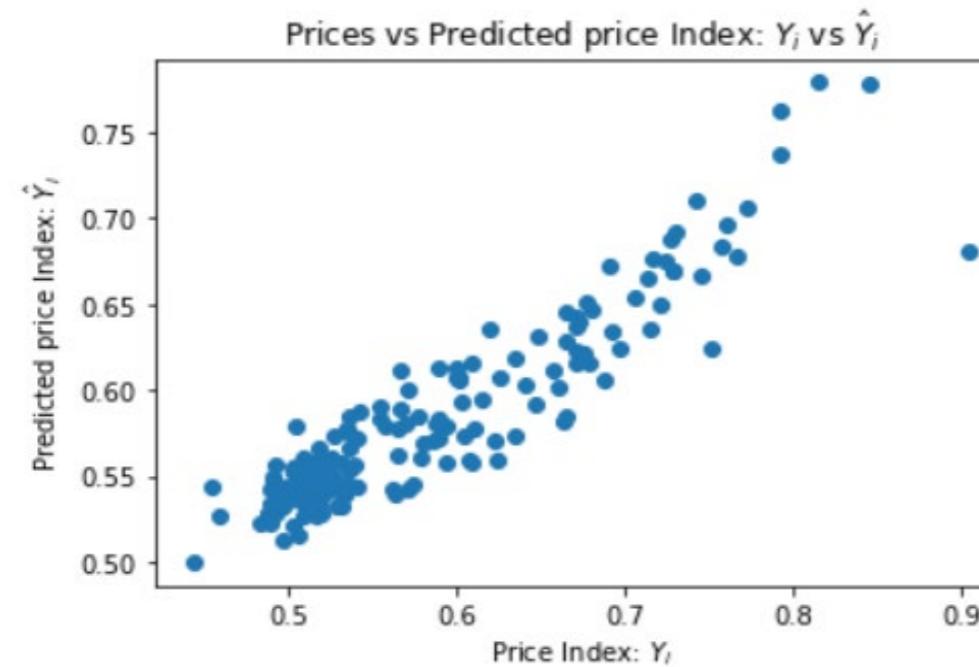
주식 정보 + 재무지표 LSTM



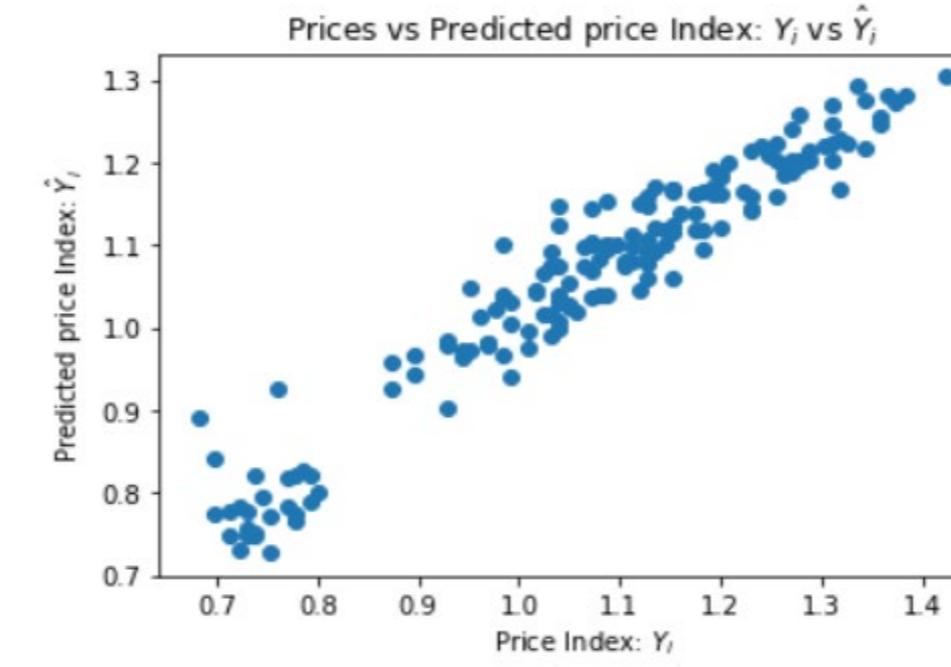
2-2. 데이터 분석

시계열 분석 LSTM 재무 정보를 추가한 데이터로 5개의 업종 LSTM 실행 후 RMSE 결과 및 산점도

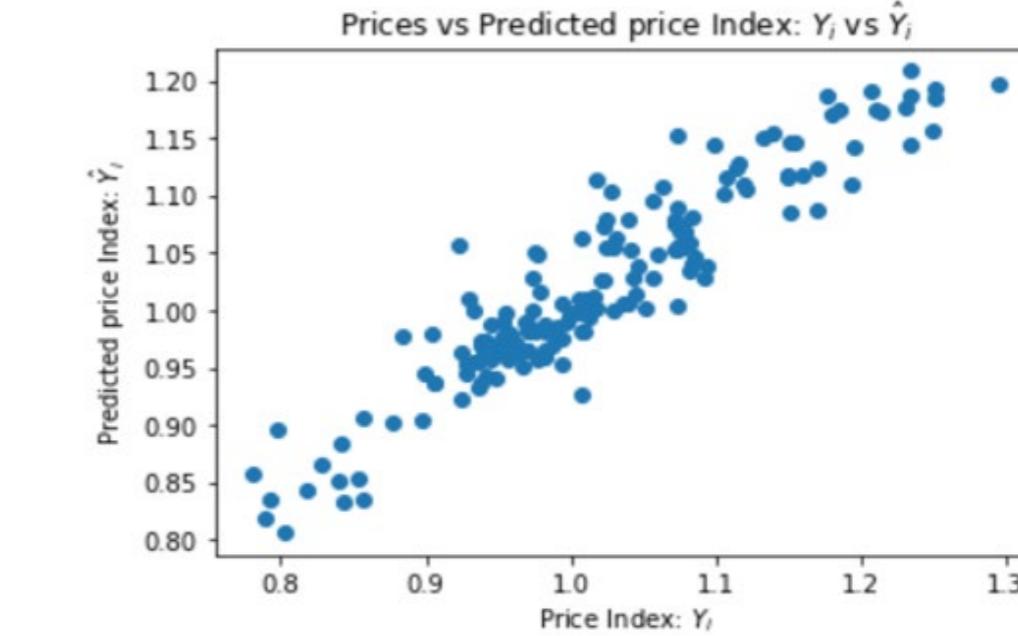
산점도가 선형에 가까울수록 예측 정확도 높음 → x : 실제값, y : 예측값



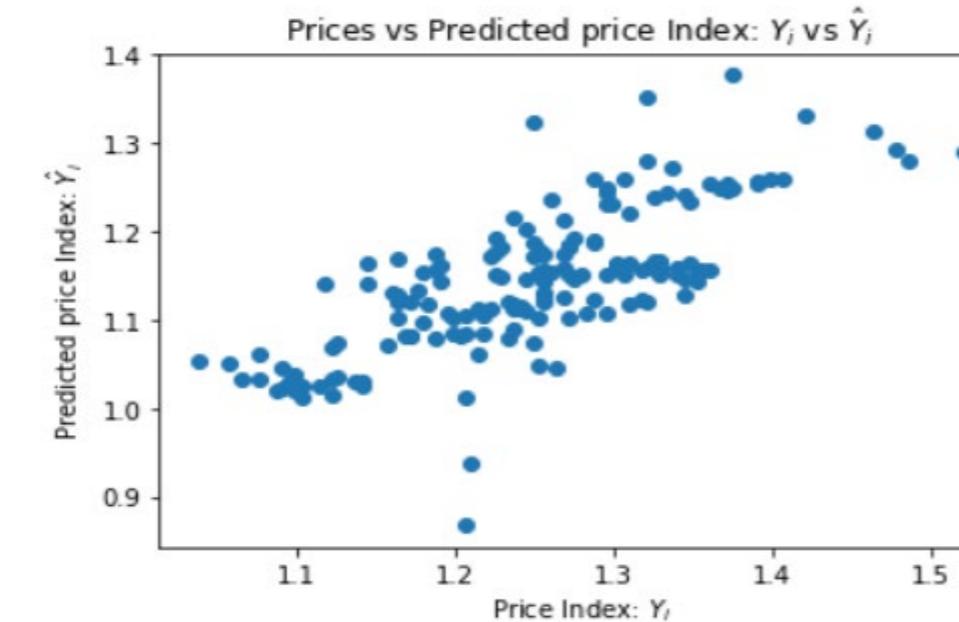
삼성 lstm rmse : 9614



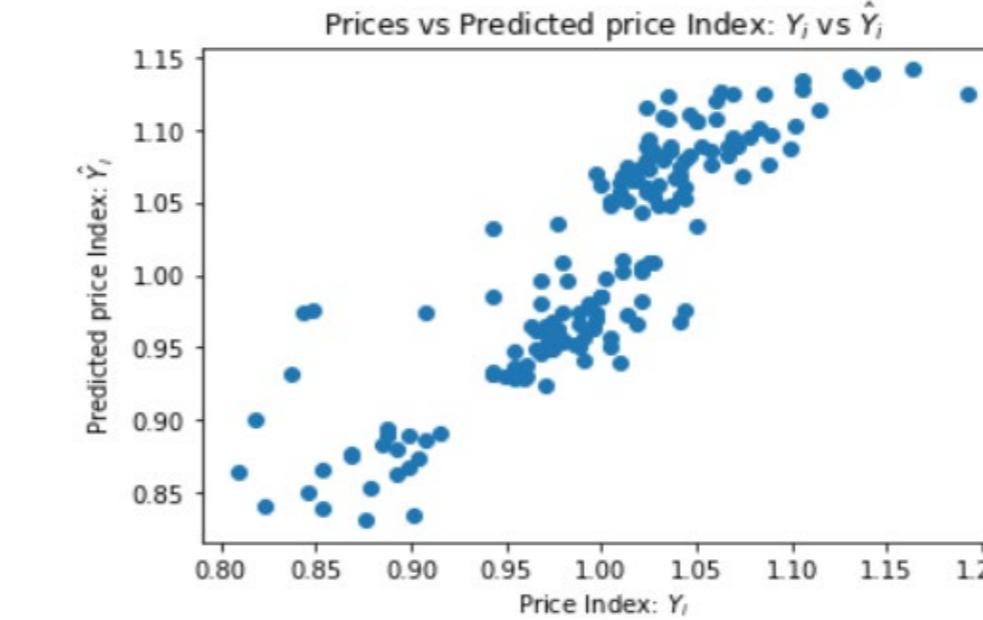
lstm 하이닉스 결과 rmse : 27238



lstm lg화학결과 rmse : 177384



lstm 현대차 결과 rmse : 46457



lstm 셀트리온 결과 rmse : 41497

2-2. 데이터 분석

강화 학습 RL Train-Test 데이터 분할 및 학습 진행

```
reward = 0

# action
# 0: idle
# 1: 매수
# 2: 매도
if act == 1: # 매수
    self.positions.append(self.data.iloc[self.t, :]['Close'])
elif act == 2: # 매도
    if len(self.positions) == 0:
        reward = -1
    else:
        profits = 0
        for p in self.positions:
            profits += (self.data.iloc[self.t, :]['Close'] - p)
        reward += profits
        self.profits += profits
        self.positions = []

# set next time
self.t += 1

self.position_value = 0
for p in self.positions:
    self.position_value += (self.data.iloc[self.t, :]['Close'] - p)
self.history.pop(0)
self.history.append(self.data.iloc[self.t, :]['Close'] - self.data.iloc[(self.t-1), :]['Close'])
if (self.t==len(self.data)-1):
    self.done=True
# clipping reward
if reward > 0:
    reward = 1
elif reward < 0:
    reward = -1
#print ("t=%d, done=%s"%(self.t,self.done))
return [self.position_value] + self.history, reward, self.done # obs, reward, done
```

- 데이터 분할

- Train : 0.8, Test : 0.2

- 클래스 라벨링

- 1 : 매수 시점(상승 예측)
- 2 : 매도 시점(하락 예측)
- 0 : 유지(보합 예측)

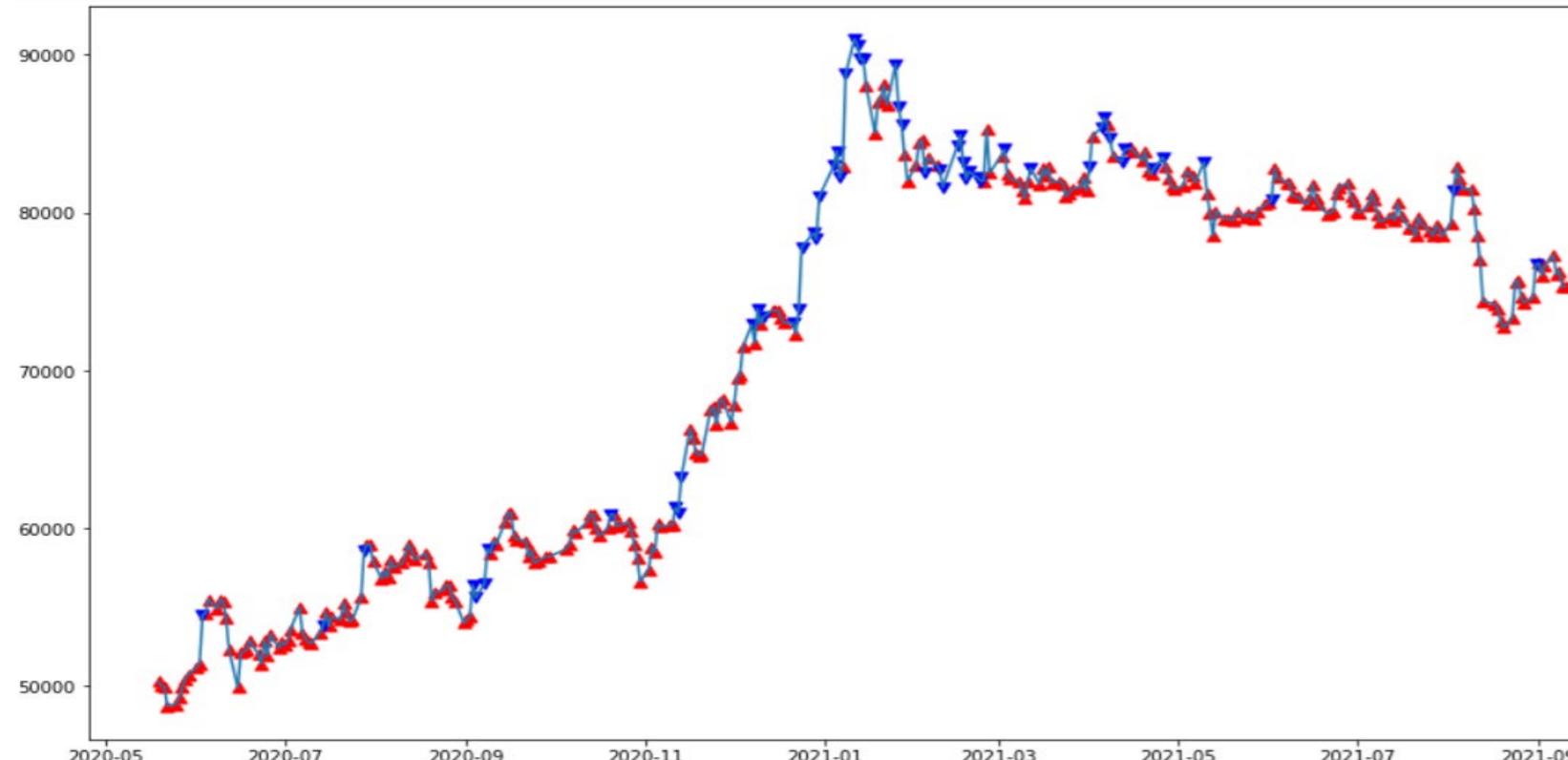
layer_summary(layer 정보)

```
(fc_val): Sequential(
    (0): Linear(in_features=91, out_features=100, bias=True)
    (1): ReLU()
    (2): Linear(in_features=100, out_features=100, bias=True)
    (3): ReLU()
    (4): Linear(in_features=100, out_features=3, bias=True)
)
```

2-2. 데이터 분석

강화 학습 RL 학습 진행 후 결과 확인

셀트리온 종목을 제외한 나머지 4가지 종목은 재무 정보를 포함한 결과의 정확도가 더 높게 나옴



Date	Open	High	Low	Close	Volume	Change	BPS	PER	PBR	EPS	DIV	DPS	Action
2020-12-22	72500	73200	72100	72300	16304910	-0.009589	37528	22.84	1.93	3166	1.96	1416	1
2020-12-23	72400	74000	72300	73900	19411326	0.022130	37528	23.34	1.97	3166	1.92	1416	2
2020-12-24	74100	78800	74000	77800	32502870	0.052774	37528	24.57	2.07	3166	1.82	1416	2
2020-12-28	79000	80100	78200	78700	40085044	0.011568	37528	24.86	2.10	3166	1.80	1416	2
2020-12-29	78800	78900	77300	78300	30339449	-0.005083	37528	24.73	2.09	3166	1.81	1416	2
...
2021-09-09	76400	76600	75000	75300	17600770	-0.013106	39406	19.60	1.91	3841	3.98	2994	1
2021-09-10	75300	75600	74800	75300	10103212	0.000000	39406	19.60	1.91	3841	3.98	2994	0
2021-09-13	75200	76300	75100	76300	11397775	0.013280	39406	19.86	1.94	3841	3.92	2994	2
2021-09-14	77100	77700	76600	76600	18167057	0.003932	39406	19.94	1.94	3841	3.91	2994	1
2021-09-15	77400	77400	76400	77000	12500473	0.005222	39406	20.05	1.95	3841	3.89	2994	2

종목별 정확도

lg화학 강화학습 결과_재무정보 57%

삼성전자 강화학습_재무정보 추가 정확도 62%

셀트리온 강화학습 결과 54%

하이닉스 강화학습 결과 64%

현대차 강화학습 결과_재무정보 56%

셀트리온_RL.csv

셀트리온 강화학습 결과 64%

셀트리온_RL_재무.csv

셀트리온 강화학습 결과 54%

2-2. 데이터 분석

양상블 모델 텍스트 감성점수 · 종가 예측

업종별 모델 결과 및 신문사 뉴스 긍정지수, 뉴스심리지수(NSI) 계산 결과

매일 뉴스 크롤링 후에 업데이트 가능하도록 코드 작성

code	date	lstm	arima	fbprophet	RL	mail_news	mail_news_nsi	asia_news	asia_news_nsi	sampro_youtube	hk_youtube	suka_youtube	close
005930	2021-09-15	75635.2426 7578125	76491.08497 472802	76623.23586 24819	-1	0.782352941 1764706	156.7647058823 5293	0.507620851 3708514	101.8575036075 036	0.37133069170845 88	0.512560352 6830673	0.50873059034 34753	77000
005930	2021-09-16	75637.3295 8984375	76936.87973 354082	76644.81668 759834	0	0.547222222 2222222	109.8611111111 111	0.589542483 6601307	118.0849673202 6143	0.37133069170845 88	0.563625961 5421295	0.50873059034 34753	76100
005930	2021-09-17	76096.1528 3203125	76320.41718 345125	76621.52012 548641	-1	0.772108843 537415	154.4217687074 8298	0.548018648 018648	109.8764568764 569	0.24351451794306 436	0.535122096 5385437	0.50873059034 34753	77200
005930	2021-09-23	76369.5266 1132812	76895.10958 142759	76429.80359 181295	-1	0.659796725 0141163	132.3071710897 7976	0.564516116 3859226	113.1137495929 7396	0.29997785687446 593	0.580723367 6314354	0.50873059034 34753	77400
005930	2021-09-24	76418.1506 3476562	77680.41069 898686	76287.74944 202403	-1	0.438888888 88888883	87.944444444444 446	0.645292207 7922078	129.0584415584 416	0.30108477175235 75	0.484348393 9766884	0.50873059034 34753	77300
005930	2021-09-27	76359.4067 3828125	77162.59166 319207	75936.43952 160457	-1	0.465548340 5483406	93.38239538239 54	0.683407528 1443702	136.8657161551 8983	0.18567574024200 44	0.218094944 95391846	0.50873059034 34753	77700
005930	2021-09-28	76422.0	77641.18038 252764	75967.82153 906283	0	0.571807359 3073593	114.6796536796 5367	0.247715247 71524772	49.54304954304 9534	0.39847799284117 563	0.324798623 72080487	0.50873059034 34753	76300
005930	2021-09-29	76854.0910 6445312	76541.42479 116977	75715.30224 376015	1	0.509740259 7402598	102.0389610389 6105	0.454241746 0226949	91.04065689684 666	0.28415226936340 33	0.442739695 31059265	0.50873059034 34753	74100
005930	2021-09-30	76312.1123 046875	73766.22464 82903	75549.74877 666075	-1	0.508241758 2417582	101.9560439560 4394	0.282424242 42424236	56.68484848484 849	0.41668616873877 39	0.604040116 071701	0.50873059034 34753	74100
005930	2021-10-01	75350.9401 8554688	74356.54373 860663	75355.39555 011716	-1	0.508241758 2417582	101.9560439560 4394	0.282424242 42424236	56.68484848484 849	0.66808604200681 05	0.965788066 3871765	0.50873059034 34753	73200

2-2. 데이터 분석

양상을 모델 텍스트 감성점수 · 종가 예측

1. 하드보팅 방법을 통해서 상승, 하락, 보합을 예측

- 1, 1, 1, 1, 0, 1, 0, 0 → 1

2. 로지스틱 분류 모델을 이용하여 분류

- 분류 결과 및 확률 도출

예측 정확도 비교

최종 웹 페이지 분석 결과 제공

2-2. 데이터 분석

양상블 모델 텍스트 감성점수 · 종가 예측

1) 하드보팅 방법을 통해서 상승, 하락, 보합을 예측

code	date	lstm	arima	fbprophet	RL	mail_news	mail_news_nsi	asia_news	asia_news_nsi	sampro_youtube	hk_youtube	suka_youtube	change
5930	2021-09-15	75550.37	76491.08	76623.24	-1	1	1	1	1	-1	1	1	1
5930	2021-09-16	1.0	-1.0	1.0	0	1	1	1	1	-1	1	1	-1
5930	2021-09-17	1.0	1.0	-1.0	-1	1	1	1	1	-1	1	1	1
5930	2021-09-23	1.0	-1.0	-1.0	-1	1	1	1	1	-1	1	1	0
5930	2021-09-24	1.0	1.0	-1.0	-1	-1	-1	1	1	-1	-1	1	0
5930	2021-09-27	-1.0	-1.0	-1.0	-1	-1	-1	1	1	-1	-1	1	1
5930	2021-09-28	1.0	-1.0	1.0	0	1	1	-1	-1	-1	-1	1	-1
5930	2021-09-29	1.0	1.0	-1.0	1	1	1	-1	-1	-1	-1	1	-1
5930	2021-09-30	-1.0	-1.0	-1.0	-1	1	1	-1	-1	-1	1	1	0
5930	2021-10-01	-1.0	1.0	-1.0	-1	1	1	-1	-1	-1	1	1	-1

상승(1) 텁하락(-1) 개수가 더 많은 결과 ⇒ 41% 정확도

2-2. 데이터 분석

양상블 모델 텍스트 감성점수 • 종가 예측

2) 로지스틱 분류 모델을 이용하여 분류

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score, roc_auc_score, roc_curve
import statsmodels.api as sm
import matplotlib.pyplot as plt
import time

logit = pd.read_csv("./total_table.csv")
logit.dropna(inplace=True)
logit.drop('close',1,inplace=True)
logit['change'] = np.where(logit['change']>0,1,
                           np.where(logit['change']<0,-1,0))
```

```
# 회귀모델에서 b0를 위한 상수항 추가

logit = sm.add_constant(logit,has_constant = "add")
a=logit[logit['change'] == 0].index
logit = logit.drop(a)
logit = logit.reset_index().drop(['index','code','date'],1)
logit.head()
```

```
def acc(cfmat):
    return (cfmat[0,0] + cfmat[1,1])/(cfmat[0,0] + cfmat[1,1] + cfmat[0,1] + cfmat[1,0])
acc(cfmat)

0.6291390728476821
```

모든 데이터의 결과를 피처로 사용해

로지스틱 분류 결과 ⇒ **약 63% 정확도**

2-2. 데이터 분석

양상블 모델 텍스트 감성점수 · 종가 예측 최종 결과

로지스틱을 통한 분류모델로 긍정 확률 도출

18년도부터의 모든 데이터를 사용하여 Train-Test Set을 test_size = 0.2으로 나누어서 모델에 적합

const	code	date	lstm	arima	fbprophet	RL	mail_news	mail_news_nsi	asia_news	asia_news_nsi	sampro_youtube	hk_youtube	suka_youtube	change
1.0	660	2019-07-01	69690.480298	69334.439383	68097.004960	-1	0.359347	71.869489	0.341326	68.348494	0.34345	0.884782	0.027747	1
1.0	660	2019-07-02	70907.199246	69969.206936	68680.444525	-1	0.342177	68.435374	0.306667	61.333333	0.34345	0.884782	0.027747	1
1.0	660	2019-07-03	71219.647998	71587.378666	69012.552825	-1	0.362879	72.775758	0.208333	42.000000	0.34345	0.884782	0.027747	-1
1.0	660	2019-07-04	69626.333362	68893.380793	69538.926271	-1	0.425714	85.342857	0.554945	110.989011	0.34345	0.884782	0.027747	1
1.0	660	2019-07-05	70574.226451	70282.182786	69951.558884	-1	0.380159	76.031746	0.246032	49.206349	0.34345	0.884782	0.027747	-1



Train-Test Set 분할
및 데이터 수 확인

```
print(train_x.shape, test_x.shape,  
(373, 12) (160, 12) (373,) (160,)
```

```
pred_y = results.predict_proba(test_x)  
pred_y  
array([[ 0.51959642,  0.48040358],  
       [ 0.37235797,  0.62764203],  
       [ 0.55167075,  0.44832925],  
       [ 0.35007614,  0.64992386],  
       [ 0.43035047,  0.56964953],  
       [ 0.28998814,  0.71001186],  
       [ 0.4712168 ,  0.5287832 ],  
       [ 0.3681827 ,  0.6318173 ],  
       [ 0.67204294,  0.32795706],  
       [ 0.28628736,  0.71371264],
```

threshold	performance	
	ACC	
0.0	0.48125	0.5
0.1	0.48125	0.6
0.2	0.51875	0.7
0.3	0.56250	0.8
0.4	0.64375	0.9

예측 결과
하락
확률 51%



predict_probability 확인

threshold 값에 따른 정확도 확인

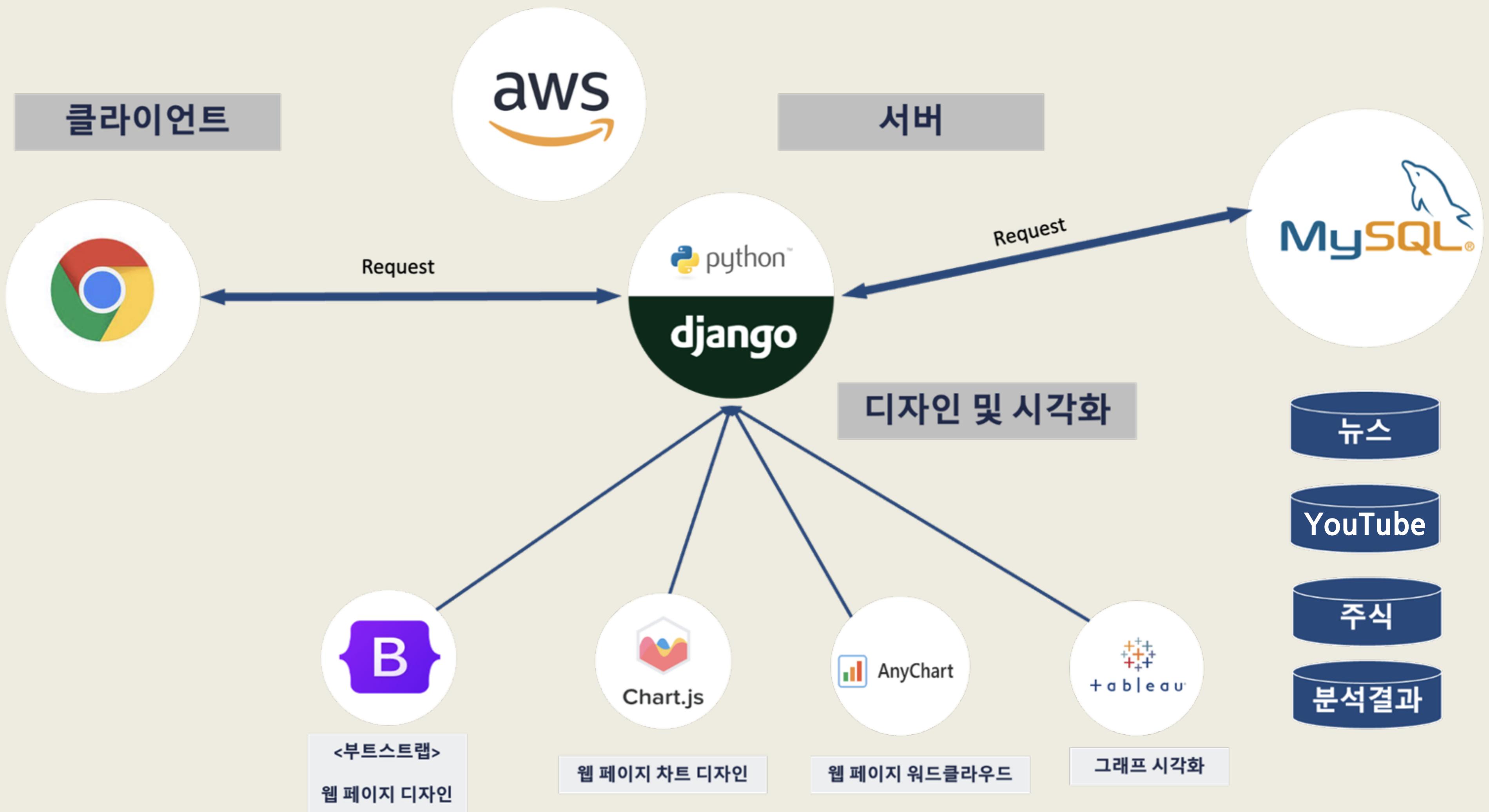
확률 및 결과 표현

PART 3.

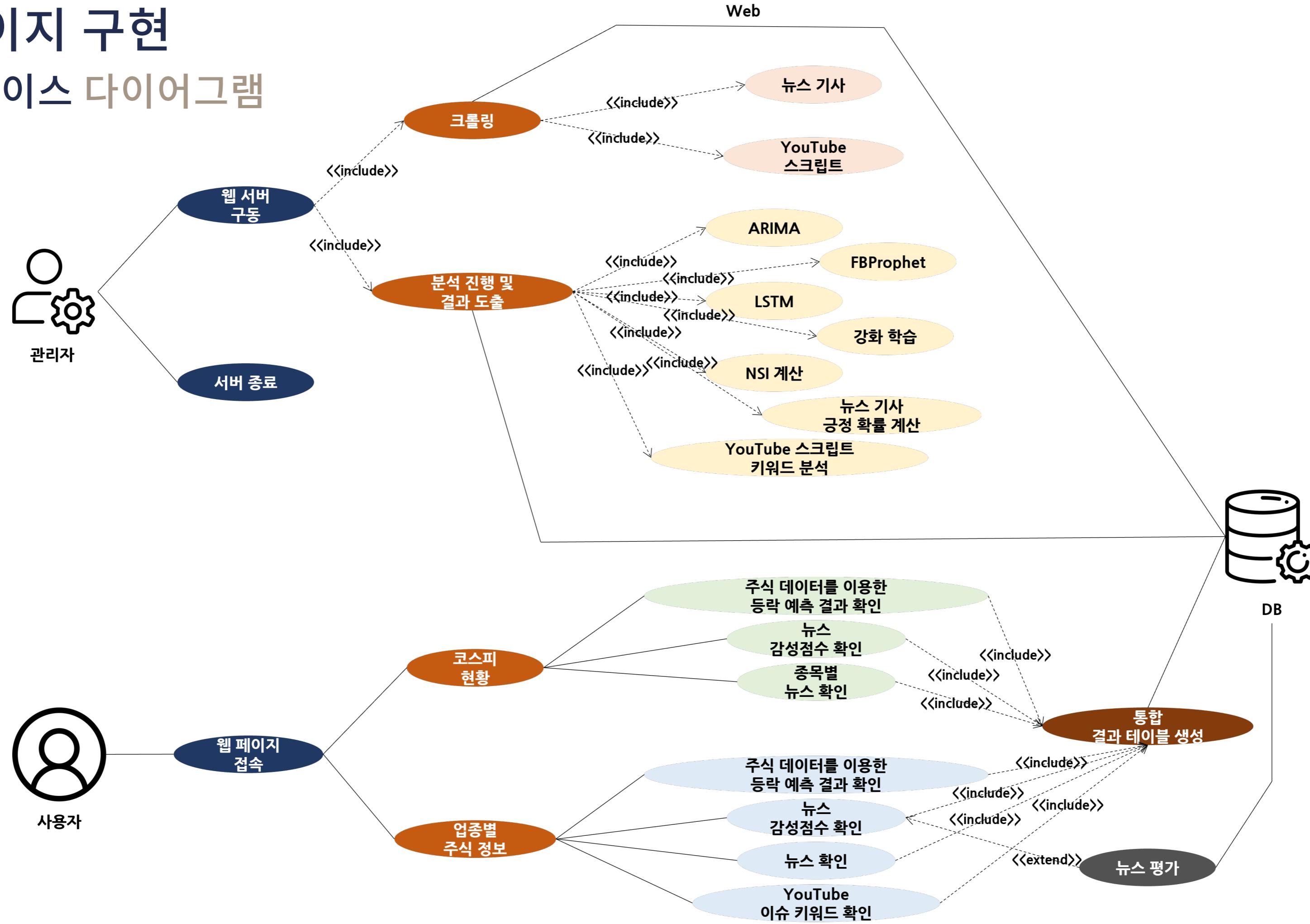
웹 페이지 구현

- 웹 페이지 프로세스
- 웹 페이지 화면

Sentiment analysis of news articles and YouTube texts and
prediction of stock price fluctuations using deep learning models



3. 웹 페이지 구현 유스 케이스 다이어그램



Demonstration

Sentiment analysis of news articles and YouTube texts and
prediction of stock price fluctuations using deep learning models

3. 웹 페이지 구현 메인 페이지

STOCK PREDICTION¹

HOME

INTERFACE

삼성전자

SK하이닉스

LG화학

현대차

셀트리온

ADDONS

Pages

Charts

Tables

더 많은 주식 정보는 [Premium features, components, and more!](#)입니다.

네이버 금융!

1

삼성전자 2021-09-28 뉴스 바로가기→ "얼마나 잘 팔리길래...스마트폰 대기가 4주?" 갤Z폴드3 플립3 배송지연에 소비자 친다

하이닉스 2021-09-28 뉴스 바로가기→ 대기업 장애인 고용률 2.38%...대우건설 등 0%대

현대차 2021-09-28 뉴스 바로가기→ 현대차 수소트럭 국내 누빈다...11월 '엑시언트' 실증 투입

LG화학 2021-09-27 뉴스 바로가기→ 中 전력 사용 제한 조치...현지 한국기업 비상(종합)

셀트리온 2021-09-27 뉴스 바로가기→ 코스피 외국인 기관 쌍끌이 매수...8.40포인트↑, 3133.64 장 마감

2

코스피 (KOSPI) 2018~현재 2021 2020 2019 2018

3

언론사별 긍정 확률

언론사	긍정 확률 (%)
매일경제	78.2 %
아시아경제	50.8 %
Youtube 슈카월드	50.9 %
Youtube 한국경제TV	51.3 %
Youtube 삼프로	37.1 %

4

오늘의 이슈 30선

1. 종목 기사

삼성전자, SK하이닉스, 현대차, LG화학, 셀트리온 순으로 긍정/부정점수가 가장 큰 기사 표출

2. 코스피 지수

2018~2021년까지의 월별 코스피 지수 평균

연도별 코스피 지수 데이터

3. 언론사별 긍정/부정 판단

5개의 뉴스 매체 및 YouTube 채널 하루 평균 긍정/부정 판단 시각화

4. 오늘의 이슈 30선 워드 클라우드

5개의 언론사 및 YouTube에서 언급된 상위 30개 단어 워드 클라우드

3. 웹 페이지 구현 메인 페이지

STOCK PREDICTION¹

HOME

INTERFACE

삼성전자

SK하이닉스

LG화학

현대차

셀트리온

ADDONS

Pages

Charts

Tables

더 많은 주식 정보는 [Premium features, components, and more!](#)로 제공됩니다.

네이버 금융!

STOCK PREDICTION¹

삼성전자
2021-09-28
뉴스 바로가기-
"얼마나 잘 팔리길래...스마트폰 대기가 4주?" 갤Z폴드3 플립3 배송지연에 소비자 친다

하이닉스
2021-09-28
뉴스 바로가기-
대기업 장애인 고용률 2.38%...대우건설 등 0%대

현대차
2021-09-28
뉴스 바로가기-
현대차 수소트럭 국내 누빈다...11월 '엑시언트' 실증 투입

LG화학
2021-09-27
뉴스 바로가기-
中 전력 사용 제한 조치...현지 한국기업 비상(종합)

셀트리온
2021-09-27
뉴스 바로가기-
코스피 외국인 기관 쌍끌이 매수...8.40포인트↑, 3133.64 장 마감

코스피 (KOSPI) 2018~현재 2021 2020 2019 2018

언론사 별 긍정 확률

언론사	긍정 확률 (%)
매일경제	78.2 %
아시아경제	50.8 %
Youtube 슈퍼월드	50.9 %
Youtube 한국경제TV	51.3 %
Youtube 삼프로	37.1 %

오늘의 이슈 30선

일단 사실 오늘 경우 정도 가격 때문 부분 우리 얘기 시작 투자 생각 시장 상황 랜드 계속 말씀 종목 저희 가지 미국 한반 조급 업체 탄소배출권

AnyChart Trial Version

The screenshot displays a stock prediction application's main interface. On the left, a sidebar lists navigation items like HOME, INTERFACE, and various company names. The main area features five news cards for Samsung Electronics, SK Hynix, Hyundai Motor, LG Chem, and Celltrion, each with a date, a link, and a brief summary. Below the news is a line chart for the KOSPI index from January 2021 to August 2021, showing a general upward trend with some fluctuations. To the right of the chart is a bar chart titled '언론사 별 긍정 확률' (Positive Probability by News Source) comparing five media outlets. At the bottom is a section titled '오늘의 이슈 30선' (Top 30 News of the Day) with a grid of Korean words representing different topics.

5. 코스피 2021년 클릭 화면

3. 웹 페이지 구현 종목별 페이지



1. 종목 전날 주가 데이터

주가 : 전날 삼성전자 주가

거래량 : 전날 삼성전자 거래량

전일대비 : 전날 구가 등락률

2. 예측값

양상별 분석을 통한 주식 가격 상승/하락 예측

3. 주가

종목 2018년~현재 주가 데이터 그래프

2018/2019/2020/2021년 연도별 그래프 표출

4. 언론사별 긍정/부정 판단

각 수집 언론사별 감성 분석을 통해 얻은 기사 및 YouTube 영상의 긍정/부정 판단

3. 웹 페이지 구현 종목별 페이지



5. 주가

2018년 주가 그래프 화면

3. 웹 페이지 구현 종목별 페이지



언론사 별 이슈 정리

매일경제 2021-09-28
"얼마나 잘 팔리길래...스마트폰 대기가 4주?" 갤Z플드3
플립3 배송지연에 소비자 지친다
뉴스 바로가기→

6

7

금정 중립 부정

금정 중립 부정



6. 언론사별 이슈 기사

매일경제, 아시아경제 각각 당일 기사 중
긍정 단어의 비율이 가장 큰 기사 표출

(말풍선) 색을 통해 긍정/중립/부정 기사 표현
말풍선 색 파랑 – 긍정
말풍선 색 회색 – 보합
말풍선 색 빨강 – 부정

7. (버튼) 사용자가 기사의 긍정/중립/부정 판단

사용자가 기사를 읽고 해당 내용이
긍정/중립/부정인지를 직접 판단하여 투표

8. 분석 결과

ARIMA, FBProphet, LSTM 순으로 분석 결과 표출

9. 분석 결과 그래프

ARIMA, FBProphet, LSTM 분석 결과 표출

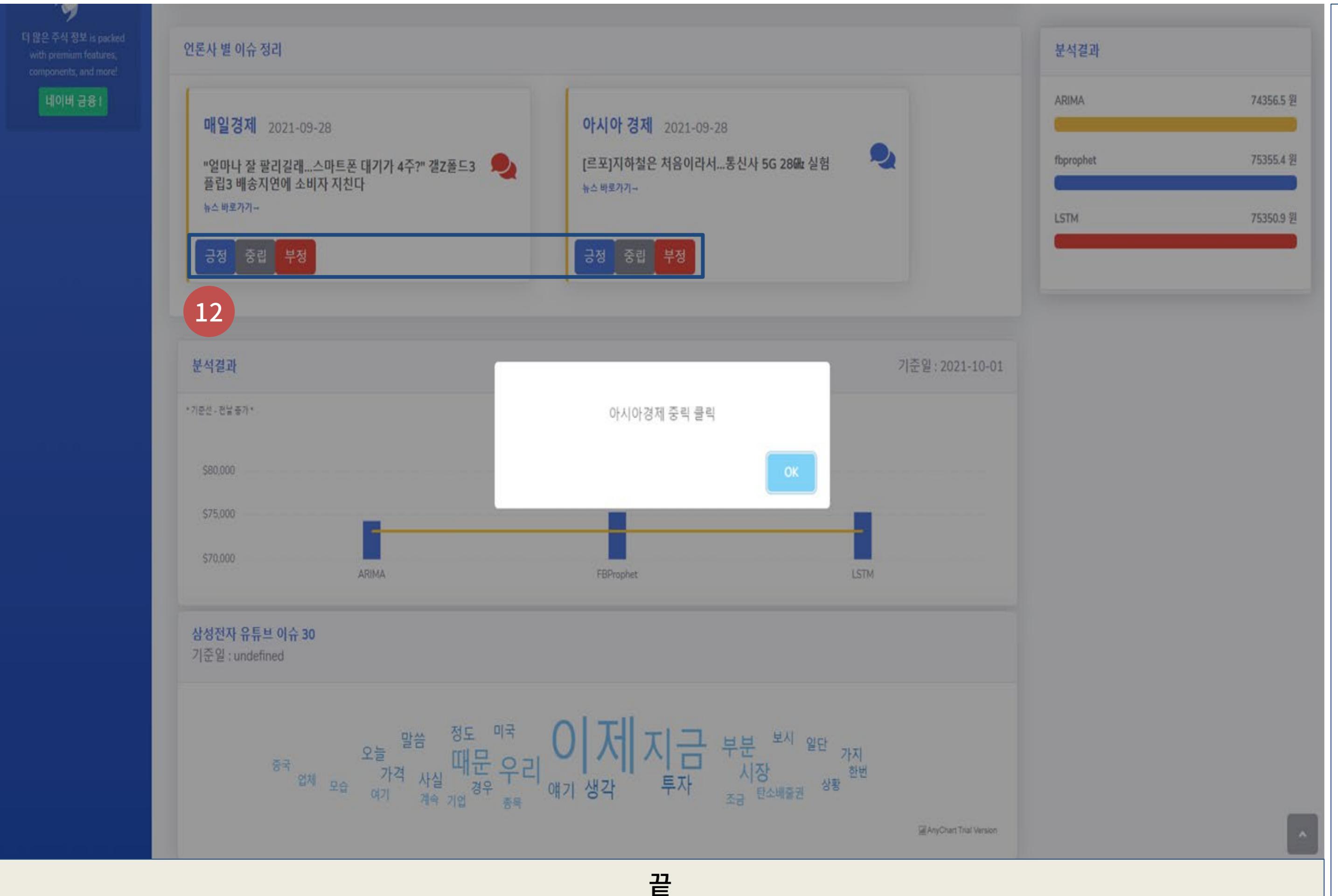
(노란색 선) 선으로 현재 가격과
예측 결과의 차이를 보여줌

10. 종목별 워드 클라우드 표출

종목 관련 YouTube 키워드 30개 표출
마우스 오버레이 시 키워드 개수 및 비율 표출

11. (버튼) 종목 페이지 – 상단 번으로 이동

3. 웹 페이지 구현 종목별 페이지



12. 아시아경제-중립-버튼 클릭 화면

PART 4.

발전방향 및 느낀점

- ◆ 한계점 및 발전방향
- ◆ 느낀점

Sentiment analysis of news articles and YouTube texts and
prediction of stock price fluctuations using deep learning models

4. 발전방향 및 느낀점

한계점

- ◆ 유튜브 STT(Speech-To-Text) 스크립트의 한계로 기사와 동일한 방법으로 감성점수 산정 불가
- ◆ 한정된 일간지 · 유튜브 채널 미디어의 데이터 수집 및 5개 종목만을 대상으로 함
- ◆ 긍정 및 부정 라벨 인코딩 데이터 수의 부족

발전방향

- ◆ STT 기술 발전에 따라 서비스(감성 분석 및 워드 클라우드) 동반 발전 가능
- ◆ KOSPI 10대 종목 중 5개 종목 이외의 여러 종목으로 확대 가능
- ◆ 데이터 수집 경제 일간지 및 유튜브 채널의 확대로 다양한 정보 제공
- ◆ 종목별 분석 모델에서 범용적인 모델로 통합하여 구축 가능
- ◆ 서비스 이용자 긍부정 의견을 반영하여 객관성 확보된 라벨링 데이터로 모델 성능 개선
- ◆ 개인 맞춤 종목 분석 결과 서비스 제공

4. 발전방향 및 느낀점

느낀점

정길종

크롤링에 대해 배워보고 싶었는데, 프로젝트를 통해서 다양한 방법과 여러 사이트 크롤링을 할 수 있는 좋은 기회였습니다. 서버 측면에서는 Django, SQL문을 배웠고, 이를 AWS클라우드 환경에서 진행한 점이 많은 도움이 되었습니다. 또한, 분석 부분을 직접 수행 하지는 않았지만, 조원분들이 정리한 내용으로 시계열 분석, 감성사전 등 많은 내용을 배울 수 있었습니다.

김형림

무엇보다도 텍스트 분석을 경험할 수 있어서 좋았습니다. 지난 번 프로젝트 때보다도 코드 구현 실력도 한층 업그레이드 된 것 같아 뿌듯합니다. 하지만 한편으로는 감성사전 구축이라는 하나의 주제에 너무 몰두한 나머지 다른 분석 기법들에 대해서는 비교적 많이 공부하지 못한 점이 아쉽기도 합니다. 프로젝트가 끝난 이후에는 이번에 사용한 모델들에 대해 조금 더 깊게 공부해보고, 더 발전시킬 수 있었으면 좋겠습니다. 또 하나 빠질 수 없는 얘기는, 팀원들 모두 무슨 능력자들인 것 같이 마법처럼 맑은 일을 해오시는 걸 보면서 나도 질 수는 없지! 라고 생각하며 좋은 시너지 효과를 낼 수 있었습니다.
모든 팀원들께 감사합니다.

채길호

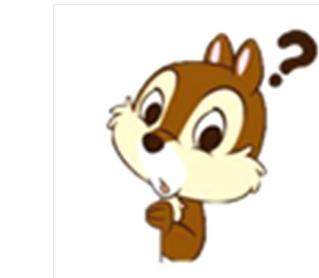
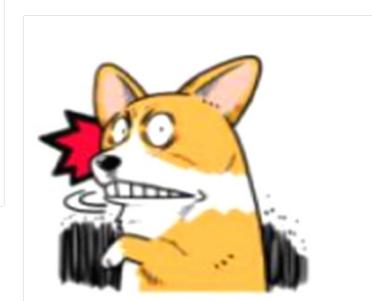
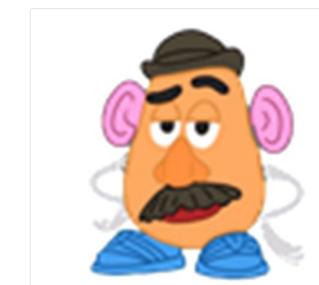
다양한 분석을 진행함으로써 공부가 많이 되었고, 특히 텍스트 분석 부분에서 감성사전을 구축해보는 새로운 경험을 해보아서 유익했습니다. 또한 팀원 간 소통이 잘 되고, 각자 맡은 업무를 잘 끝내주어서 좋았습니다. 분석부터 서비스 구현까지의 프로젝트가 완성도 있게 진행된 것 같아서 추후에 실무에서도 도움이 될 것 같습니다.

인태우

이번 최종 프로젝트를 통해 4개월 동안 배운 내용을 종합적으로 사용하며 복습하는 계기가 되었고, 해보지 않았던 서비스 구현을 통해 많은 것을 배울수 있는 기회였다고 생각합니다. 아직 부족한 데이터 분석쪽에서도 팀원분들이 잘 이끌어 주시고 많은 도움을 받아 프로젝트를 무사히 마친것 같습니다. 여러가지 데이터 수집 방법과 분석 방법, 서비스 구현까지 실무에서 직접 해볼 법한 일을 경험한게 좋았고 프로젝트를 통해 더 성장했다고 생각합니다. 끝까지 함께 해주신 팀원분들께 감사하고 강사님께도 그동안 고생하셨다고 전해드리고 싶습니다.

윤보람

데이터 수집과 분석, 그리고 서비스 구현까지 모든 과정을 경험해볼 수 있어서 굉장히 유익한 시간이었습니다. 지난 5개월 동안 배운 것들을 되돌아보며 더 발전할 수 있는 계기가 된 것 같습니다. 특히 이번 프로젝트에서 BI, 데이터 시각화 부분을 배우고 실습할 수 있어서 앞으로 이 분야에 대한 업무를 하게 되면 큰 도움이 될 것 같습니다. 훌륭한 팀원분들 덕분에 옆에서 많이 배울 수 있었고, 다들 열심히 하시는 모습이 정말 인상적이었습니다. 제가 부족한 부분을 깨닫고 앞으로 보완해갈 수 있는 좋은 길잡이가 되어 주셔서 고생하셨다는 말과 함께 감사의 인사를 전하고 싶습니다. 길면서 짧았던 시간 동안 열성적으로 지도해주신 강사님들께도 진심으로 감사드립니다. 그리고 멀티캠퍼스 매니저님, 초반에 많은 도움 주셨던 FT님들께도 감사드립니다. 마지막으로 다들 원하는 곳에서 일할 수 있는 좋은 결과 있으시길 바랍니다. 수고하셨습니다.



4. 발전방향 및 느낀점

참고 자료

데이터 자료 참고

KOSELF 감성사전 <https://sites.google.com/view/cheolwon-yang/koself?authuser=0>

금부정 라벨링 훈련 데이터 자료 <https://github.com/e9t/nsmc/>

모델링 자료 참고

자연어 처리 <https://wikidocs.net/21698>

VADER 사용법 <https://nicola-ml.tistory.com/45>

ARIMA 모형 정상성 확인 <https://dinonotes.com/archives/2476>, <https://skyeong.net/285>

유사도 파악 <https://inahjeon.github.io/fasttext/>

논문 및 연구 자료

2021, 서울대학교, 강유 외 3인, 「Accurate Multivariate Stock Movement Prediction via Data-Axis Transformer with Multi-Level Contexts」

2017, 숭실대학교 대학원, 김기준, 「뉴스 감성 분석과 시계열 예측 기반의 주가 등락 예측」

2021, 한국증권학회지, 조수지 외 2인, 「기업 재무분석을 위한 한국어 감성사전(KOSELF) 구축」

2021, 한국은행, 통계기획팀, 「뉴스심리지수(NSI) 개요 및 특성」

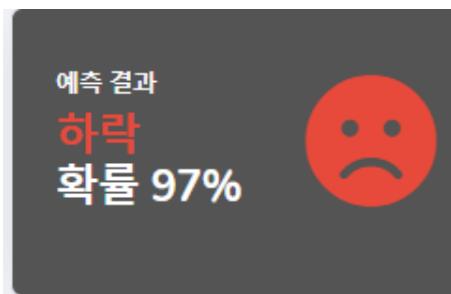
2020, 서울시립대학교, 전종준 외 3인, 「경제용어 감성사전 구축방안 연구」

2018, 한국빅데이터학회지, 김다예 외 1인, 「Word2Vec을 활용한 뉴스 기반 주가지수 방향성 예측용 감성사전 구축」

Appendix.

2021. 10. 07. 실제 주가 등락과 예측 확률 비교

삼성전자



삼성전자

71,600 KRW
+300 (0.42%) ↑ 오늘

10월 7일 오후 3:30 GMT+9 · 면책조항

SK하이닉스

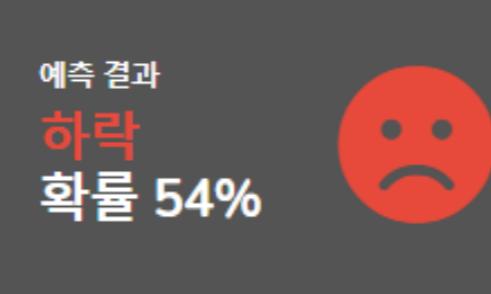


SK하이닉스

95,700 KRW
-800 (0.83%) ↓ 오늘

10월 7일 오후 3:30 GMT+9 · 면책조항

LG화학

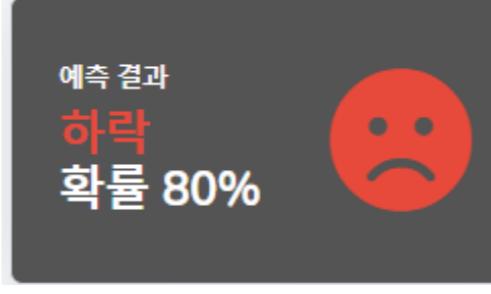


LG화학

737,000 KRW
-6,000 (0.81%) ↓ 오늘

10월 7일 오후 3:30 GMT+9 · 면책조항

현대차

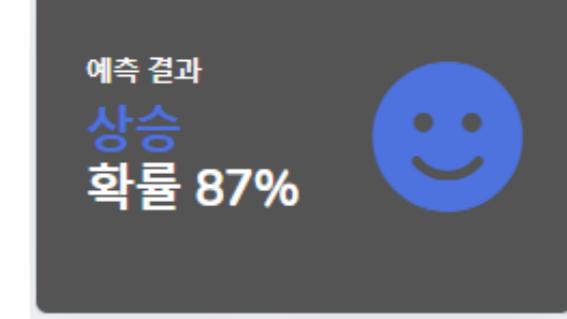


현대차

201,000 KRW
+7,000 (3.61%) ↑ 오늘

10월 7일 오후 3:30 GMT+9 · 면책조항

셀트리온



셀트리온

213,000 KRW
+1,000 (0.47%) ↑ 오늘

10월 7일 오후 3:30 GMT+9 · 면책조항

실제 예측 결과 →

5개 종목 중 3개 종목 예측
60% 정확도

Thank you

Sentiment analysis of news articles and YouTube texts and
prediction of stock price fluctuations using deep learning models