제 15회 BOAZ 컨퍼런스

개미야 뭐하니?

: 투자자의 반응을 이용한 실시간 주식 등락 예측 (feat. 카프카)

16기 강지수 김서민 김윤기 문예진

개미야 뭐하니?

Contents

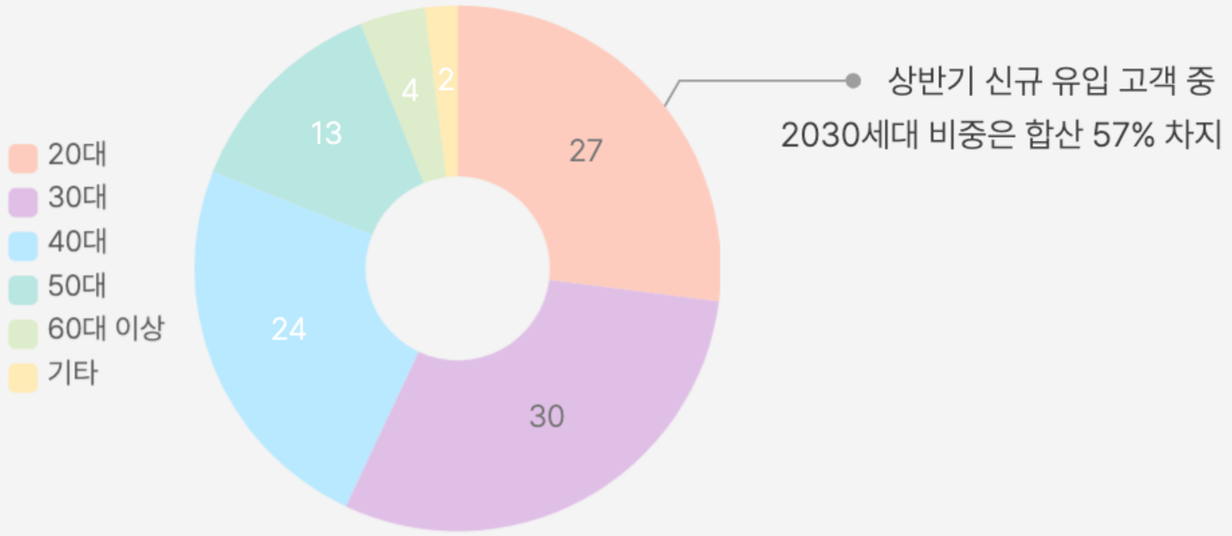
서론

- 01. 주제 선정 배경

커뮤니티에서의 투자자 반응을 이용한 실시간 주식 등락 예측

2030세대 주식 투자 급증

☑ 신규 주식 계좌 연령대별 비중 (2020년)



상반기 신규 유입 고객 중
2030세대 비중은 합산 57% 차지

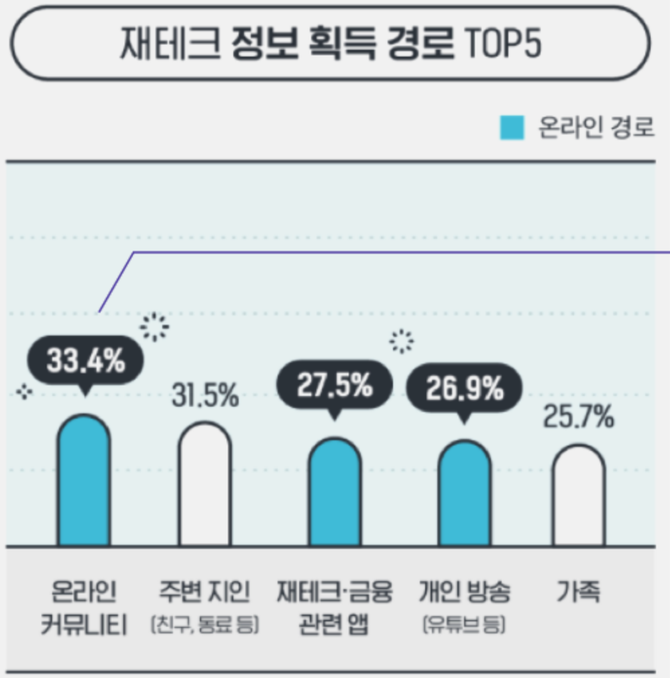
* 미래에셋 NH 한국 삼성 KB 키움 등
6개 증권사 기준 (단위 : %)

서론

- 01. 주제 선정 배경

커뮤니티에서의 투자자 반응을 이용한 실시간 주식 등락 예측

투자 정보 획득 경로



(Base: 3년 내 재테크 경험자, n=721, 복수 응답)

현 시대의 투자자들은
온라인 커뮤니티 및 온라인 경로를 통한
투자 정보 획득이 대다수

서론

- 02. 프로젝트 소개

커뮤니티에서의 투자자 반응을 이용한 실시간 주식 등락 예측

개인 투자자 반응 활용

전문가 투자자 반응 활용

실시간 주가 정보 활용

본론

- 01. 수집종목

친환경

종두산퓨얼셀 · 한화솔루션 · 에코프로 · 신성이엔지 · 유니슨

서비스업

카카오 · 네이버 · 엔씨소프트 · 넥슨지티 · SDS

반도체

삼성전자 · SK하이닉스 · SKC · 한솔케미칼 · DB하이텍

금융

KB금융 · 한국금융지주 · 신한지주 · 우리금융지주 · 하나금융지주

유통업

롯데쇼핑 · 신세계 · 호텔신라 · 삼성물산 · 동서

메타버스

한빛소프트 · 코세스 · 드래곤플라이 · 에이트원 · 엔텔스

본론

- 02. 데이터 소개

1) 실시간 커뮤니티 반응 수집

- DC 갤러리

하루에 올라오는 평균 개수 : 104.73
제목 당 평균 글자 수(공백포함) : 17.55

- 네이버 종목 토론방

하루에 올라오는 평균 개수 : 1608.98
제목 당 평균 글자 수(공백포함) : 12.48

2) 전문가 의견 : 신뢰도 높은 정보 수집 (pdf 형태)

- 팍스넷

하루에 올라오는 평균 개수 : 2.40
게시글 당 글자 수 (공백 포함) : 12346.39

- 한경 컨센서스

하루에 올라오는 평균 개수 : 4.61
게시글 당 글자 수 (공백 포함) : 26125.21

3) 주식 정보 : label 정보 수집

네이버 금융에서 나타나는 종목별 분당 주가 정보를 크롤링하여 수집

본론

- 03. Kafka

Kafka 활용

주가는 분당 수십만건, 수십억 단위의 거래가 이루어지기 때문에
실시간 처리를 통한 의사 결정이 각광받는 추세

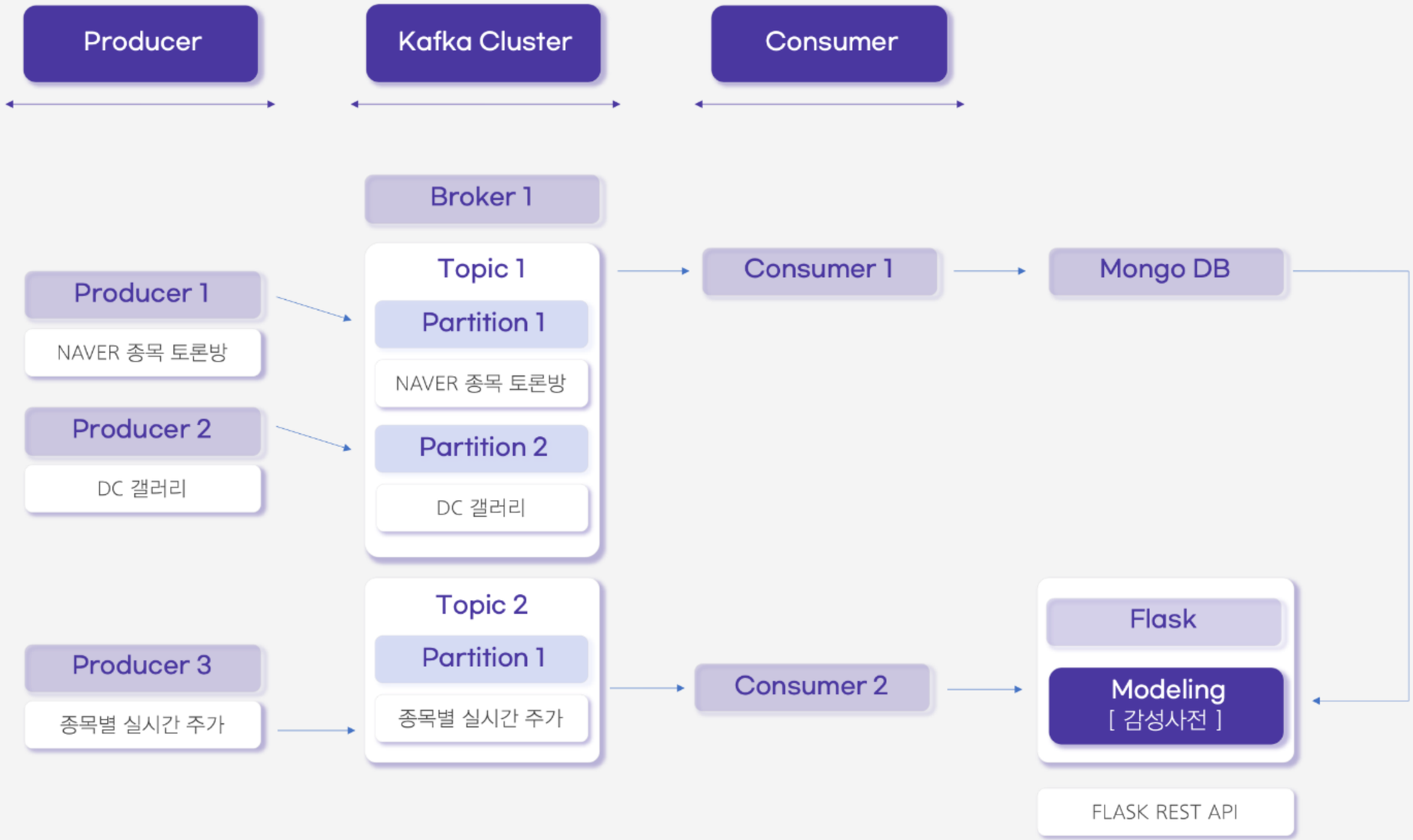
실시간 처리 → 주가 등락 예측

고가용성(High Availabilty)으로서 서버에 이슈가 생기길 경우
데이터 손실없이 복구가능

별도의 Processor를 통해 관리하기 때문에
낮은 지연(low latency)으로 대량의 데이터 실시간 처리 가능

본론

- 03. Kafka


Producer
Kafka Cluster
Consumer
Broker 1
Topic 1
Partition 1
NAVER 종목 토론방
Partition 2
DC 갤러리
Producer 1
NAVER 종목 토론방
Producer 2
DC 갤러리
Consumer 1
Mongo DB
Topic 2
Partition 1
종목별 실시간 주가
Producer 3
종목별 실시간 주가
Consumer 2
Flask
Modeling
[감성사전]
FLASK REST API

본론

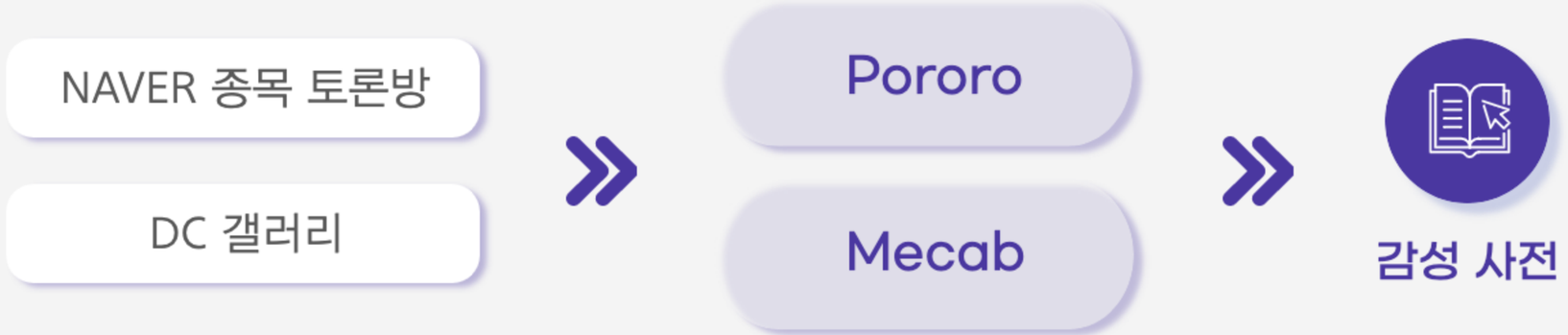
- 04. 주식 등락 예측

  1) 감성 사전

목표

NAVER 종목 토론방과 DC 갤러리에 올라오는
일반 투자자들의 반응을 감성 분석하여 이를 기반으로 5분 뒤 등락 예측

방법 : 감성사전 구축



본론

- 04. 주식 등락 예측

  1) 감성 사전

[1차 감성 사전 만들기]

1. 카카오브레인 Pororo의 감성분석 모델을 활용하여, 문장의 감성 점수 도출
2. Mecab으로 토큰화, 토큰이 나타난 게시글의 조회수를 가중치로 설정하여 감성 사전 토큰별로 감성 점수 도출

C `{'넥슨': 0.29874059815645493}`

code	date	title	mecab	view	positive
008770	2021-10-02 18:54:00	솔직히	[]	5	0.030165
008770	2021-10-02 14:44:00	애 호텔주 아니라니까	[(애, NNG), (호텔, NNG), (주, NNG), (아니, VCN)]	116	0.057826
008770	2021-10-02 13:04:00	솔브레인 화요일 시초에 담아라	[(솔, NNP), (브레인, NNP), (화요일, NNG), (시초, NNG), ...	204	0.856793
008770	2021-10-02 12:04:00	기대가 크면 실망이 크다	[(기대, NNG), (크, VA), (실망, NNG), (크, VA)]	272	0.011657
008770	2021-10-02 10:20:00	기가막히게 쌍바닥이네	[(기, NNG), (막히, VV), (쌍, NNG), (바닥, NNG), (이, ...	209	0.022041
...	...	...	...	...	...
041140	2022-01-03 10:33:37	저번주에 넥슨지티 만지작 햇는데	[(저번, NNG), (주, NNG), (넥슨, NNP), (지티, NNP), (햇...	76	0.242665
041140	2022-01-03 10:13:53	넥슨지티감사합니다	[(넥슨, NNP), (지티, NNP), (감사, NNG)]	66	0.502039
041140	2022-01-03 10:13:36	넥슨지티 물린거 개같이 탈출	[(넥슨, NNP), (지티, NNP), (물린, VV+ETM), (거, NNB),...	169	0.074371
041140	2022-01-02 14:19:26	넷게임즈 넥슨지티중에 뭐사야됨	[(넷, IC), (게임즈, NNP), (넥슨, NNP), (지티, NNP), (중...	123	0.563175
030350	2022-01-03 12:59:04	게임 해보신분	[(게임, NNG), (보, VX), (분, NNB)]	46	0.441391

A

B

```
[(170, 0.35563676059246063),
 (120, 0.3136954978108406),
 (803, 0.11383883841335773),
 (419, 0.13565491791814566),
 (234, 0.1426195427775383),
 (428, 0.1957569718360901),
 (113, 0.9835244119167328),
 (187, 0.08162579871714115),
 (371, 0.026991701684892178),
 (781, 0.050540365278720856),
 (111, 0.09827875532209873),
 (358, 0.4869778901338577),
 (449, 0.2586163282394409),
 (131, 0.17834227345883846),
 (455, 0.12103454349562526),
...
 (113, 0.5057604908943176),
 (193, 0.05699286446906626),
 (76, 0.24266507476568222),
 (66, 0.50203937292099),
 (169, 0.07437133695930243),
 (123, 0.5631746351718903)]
```

A 토큰 = 넥슨

B '넥슨'에 대한 감성점수 리스트

C 조회수를 가중치로 사용하여 도출한 '넥슨'의 감성 점수

본론

- 04. 주식 등락 예측

  1) 감성 사전

[1차 감성 사전 만들기] 조회수 가중합이란?

$$\frac{v_1}{v_1+\cdots v_n} * positive_1 + \frac{v_2}{v_1+\cdots v_n} * positive_2 + \cdots + \frac{v_n}{v_1+\cdots v_n} * positive_n = POSITIVE_{token}$$

토큰 → 리뷰의 감성점수

token	positive	view
바닥	0.022041	209
바닥	0.006177	382
바닥	0.006177	382
바닥	0.120078	337
바닥	0.096808	317
...	...	...
바닥	0.003337	86
바닥	0.173715	72
바닥	0.018028	120
바닥	0.069639	126
바닥	0.865479	124

조회수 → 조회수 가중치

token	positive	view	weight
바닥	0.022041	209	0.000684
바닥	0.006177	382	0.001251
바닥	0.006177	382	0.001251
바닥	0.120078	337	0.001103
바닥	0.096808	317	0.001038
...	...	...	...
바닥	0.003337	86	0.000282
바닥	0.173715	72	0.000236
바닥	0.018028	120	0.000393
바닥	0.069639	126	0.000413
바닥	0.865479	124	0.000406

토큰의 감성점수

token	positive
바닥	0.000327

본론

- 04. 주식 등락 예측

1) 감성 사전

[주가 등락 반영 감성 사전 업데이트]

1. 5분간 데이터 수집 후, 등락을 반영하여 감성 점수 업데이트
2. 종목별로 리뷰의 전체 word counting 진행 후, 이를 기반으로 감성 점수 업데이트
(rank에 따라 30% 30% 40% 삼 분위로 구분)

$$x_{t+1} = x_t + n * x_t * max(\text{ 해당 분위 단어들의 감성 점수 })$$

$$\min(\max(x_{t+1}, 0), 1)$$

- ☑ x_t t 시점의 토큰 감성 점수
- ☑ n $n = \dfrac{p_{t+1} - p_t}{p_t}$

본론

- 04. 주식 등락 예측

  1) 감성 사전

[주가 등락 반영 감성 사전 업데이트] 예시 캡쳐본

code	view	positive	mecab	date_time	price_time	price	time_pred	price_pred	change
035720	151	0.088650	[(거머리, 0.0), (기생충, 0.0), (카카오, 0.0), (망해라, 0.0...	2021-09-29 14:49:00	2021-09-29 14:50:00	116000	2021-09-29 14:55:00	116000	0.0
035720	264	0.097173	[(카카오, 0.0), (물린, 0.0), (흑우, 0.0), (쪼다, 0.0), ...	2021-09-29 14:47:00	2021-09-29 14:50:00	116000	2021-09-29 14:55:00	116000	0.0
035720	167	0.574085	[(오늘, 0.0), (외국인, 0.0), (개관, 0.0), (형, 0.0), (...	2021-09-29 14:45:00	2021-09-29 14:50:00	116000	2021-09-29 14:55:00	116000	0.0
035720	354	0.580836	[(나, 0.0), (평단, 0.0), (높, 0.0), (사람, 0.0), (있,...	2021-09-29 14:45:00	2021-09-29 14:50:00	116000	2021-09-29 14:55:00	116000	0.0
035720	177	0.130612	[(카카오, 0.0), (평단, 0.0), (나, 0.0), (높, 0.0), (사...	2021-09-29 14:48:23	2021-09-29 14:50:00	116000	2021-09-29 14:55:00	116000	0.0
035720	116	0.022100	[(카카오, 0.0), (게임, 0.0), (원래, 0.0), (운영, 0.0), ...	2021-09-29 14:46:20	2021-09-29 14:50:00	116000	2021-09-29 14:55:00	116000	0.0

token	word_count	rank
카카오	5	1
있	2	1
나	2	1
평단	2	1
높	2	1
사람	2	1
거머리	1	1
기생충	1	1
망해라	1	1

정부	1	2
차원	1	2
망하	1	2
해라	1	2
물린	1	2
흑우	1	2
쪼다	1	2
도	1	2
오늘	1	2

외국인	1	3
개관	1	3
형	1	3
들	1	3
충실	1	3
게임	1	3
원래	1	3
운영	1	3
뭐	1	3
해서	1	3
이미지	1	3
였	1	3

rank

word count

본론

- 04. 주식 등락 예측

  1) 감성 사전

[종목별 등락 확률 예측]

1 리뷰 내 토큰들의 감성 점수를 평균내어 리뷰의 감성 점수로 사용

2 각 리뷰의 조회수 가중치를 기반으로 리뷰의 감성 점수들을 가중합하여 종목별 등락 확률 도출

$$\frac{\text{해당 리뷰의 view 수}}{\text{종목별 리뷰의 전체 조회 수}}$$

code	date	mecab	view	time_map	positive
035720	2022-01-19 09:19:02	[(카카오, 0.23746586374235407), (윤석열, 0.274815488...	201	4	0.230787
035720	2022-01-19 09:17:52	[(카카오, 0.23746586374235407), (설거지, 0.167072099...	100	4	0.166251
035720	2022-01-19 09:16:04	[(카카오, 0.23746586374235407), (착한, 0.4950597159...	138	4	0.275566

code	time_now	time_pred	sentiment
035720	2022-01-19 09:20:00	2022-01-19 09:25:00	0.230163

본론

- 04. 주식 등락 예측
  2) PDF Modeling

목표

신뢰도 있는 정보를 반영하여 주가 등락 예측의 정확도를 높이고자
각 증권사의 전문가가 분석한 리포트와 투자의견이 담긴 pdf 게시글 모델링 진행

방법 : 딥러닝 기반

게시글의 길이가 길고
전문적으로 작성

게시글의 수가 적으므로
실시간 예측이 불가능

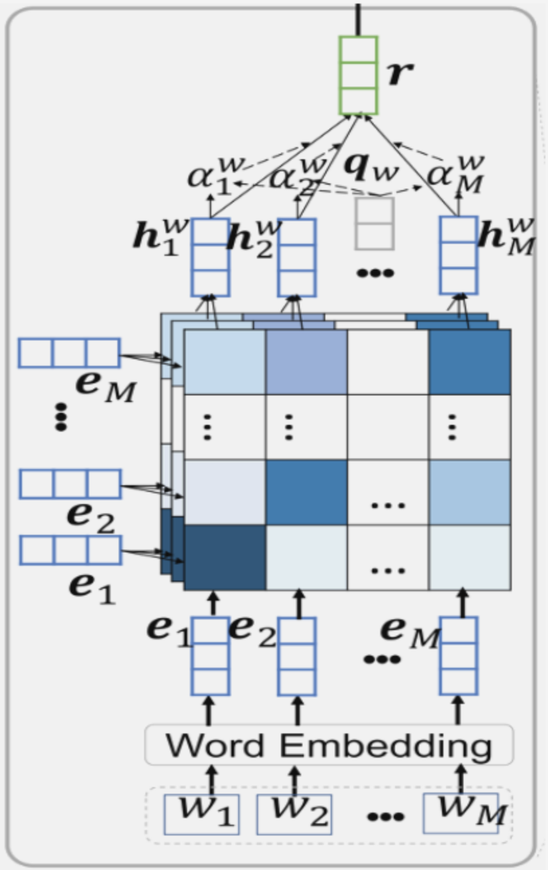
01 매일 주식장 마감 후,
당일에 업로드 된 게시글을 바탕으로 예측

02 다음 날 개장 후, 주가 등락을 예측하는 동안
예측의 신뢰도를 높이는 부가 요소로 작용

본론

- 04. 주식 등락 예측
  2) PDF Modeling

PDF Modeling



4 Classifier

문장 벡터를 이용하여 등락 확률 도출

3 Attention

classification을 위해
각 단어 벡터를 합쳐 하나의 문장 벡터 생성

2 Self Attention

context 반영

1 Word Embedding

사전적인 의미 반영

본론

- 04. 주식 등락 예측

3) 등락 예측 - 적용

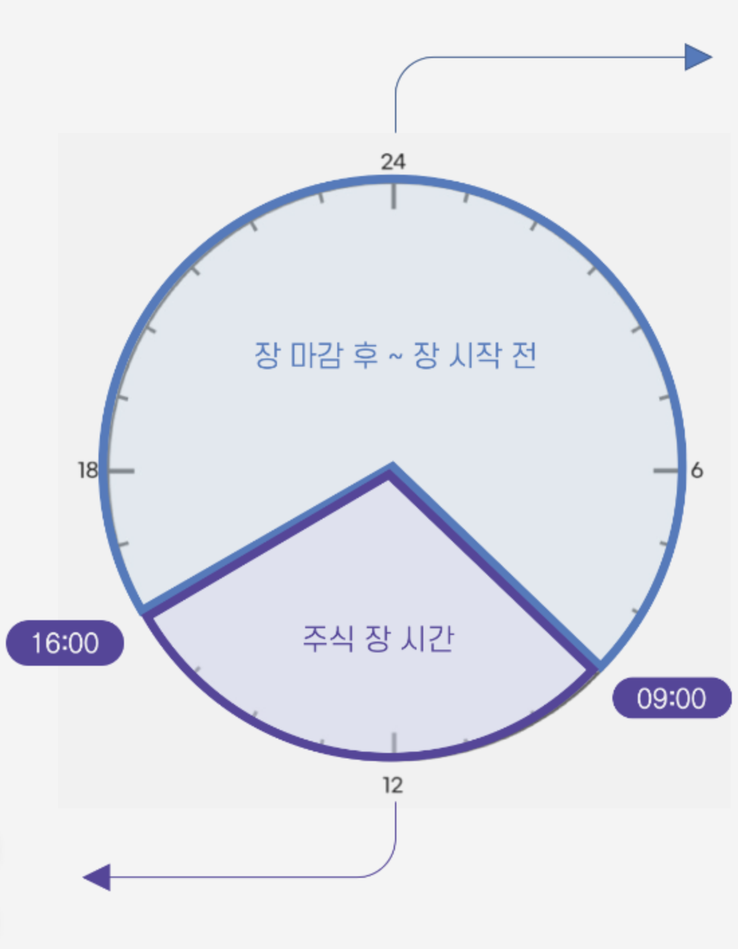
주식 등락 예측 : (1 – α) * 감성 점수 + α *(PDF 모델링 값)

☑ PDF 모델링

다음 날 하루종일 영향을 미치기 때문에
하루 단위로 학습이 이루어짐

☑ 종목당 PDF가 여러개일 경우

종목당 여러 개의 예측 등락 확률을
구하여 평균값을 취함



본론

- 05. 정확도 확인

정확도 확인

2022년 1월 20일

code	time_now	price_now	sentiment	time_pred	price_pred	pdf
008770	2022-01-20 11:00:00	73600	0.320514	2022-01-20 11:05:00	73600	0.0
008770	2022-01-20 10:35:00	73200	0.243312	2022-01-20 10:40:00	73300	0.0
008770	2022-01-20 10:15:00	73500	0.235418	2022-01-20 10:20:00	73200	0.0
008770	2022-01-20 09:50:00	73400	0.195416	2022-01-20 09:55:00	73300	0.0
008770	2022-01-20 09:40:00	73600	0.222513	2022-01-20 09:45:00	73600	0.0
...	...	...	...	...	...	...
047080	2022-01-20 09:30:00	5120	0.322713	2022-01-20 09:35:00	5110	0.0
047080	2022-01-20 09:20:00	5160	0.224654	2022-01-20 09:25:00	5140	0.0
047080	2022-01-20 09:05:00	5190	0.329614	2022-01-20 09:10:00	5150	0.0
047080	2022-01-20 09:00:00	5200	0.230660	2022-01-20 09:05:00	5190	0.0
069410	2022-01-20 09:00:00	9040	0.385917	2022-01-20 09:05:00	9120	0.0

총 383개의 주가 정보

2개 종목에 대한 값

α 에 따른 정확도

	정확도
0.2	66.84
0.25	66.84
0.3	66.84
0.5	66.84

결론

- 01. 의의 및 개선사항

의의

- Kafka 를 활용하여 투자자들의 커뮤니티 반응을 실시간으로 수집
- 주식 거래가 이루어지는 동안 5분 단위로 주가 등락 예측
- 데이터 수집부터 모델링까지 아우르는 전체적인 데이터 파이프라인 구축

개선사항

- 정확한 등락 예측을 위한 추가적인 투자자 반응 수집 필요
- 광고성 게시글을 필터링할 수 있는 방안 필요
- Kafka Streaming을 활용하여 파이프라인 고도화

개미야 뭐하니?

: 투자자의 반응을 이용한
실시간 주식 등락 예측
(feat. 카프카)

https://github.com/jayleenym/AYOA

Q & A