기업 공시 데이터
CHATBOT

NLP mini Project

AntHouse

고준호 김다나 이선영 이지수

SeSAC 2

23.02.07 ~ 23.02.10

Contents

01
팀소개

AntHouse

개미집(AntHouse)

: 개미들을 위한 보금자리

고준호

김다나황금막내

이선영

이지수

02

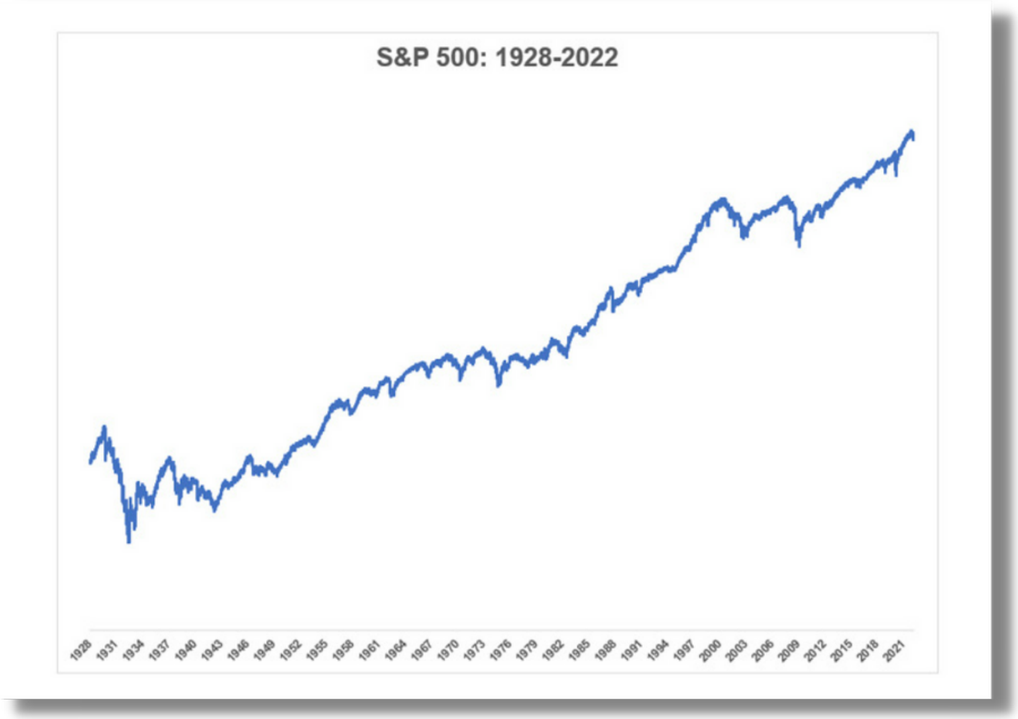
주제 설명

주제 선정 배경 | 프로젝트 목표

AntHouse

주제 선정 배경


주식거래 활동계좌 수 추이
자료: 금융투자협회
4000만
4000(만)
3500
3000
2500
2000
1500
1000
2007년 7월31일
2012년 5월17일
2020년 3월6일
2021년 3월19일


S&P 500: 1928-2022

주제 선정 배경

코로나 장기화 인한 개인 투자자들의 대거 유입으로 주식시장 과열

주식 투자자들이 챗봇을 통해 기업 공시 데이터에 접근할 수 있도록 하고자 함

증권거래소에 상장된 기업중 시가총액 200위까지인 KOSPI200을 데이터 수집 대상으로 지정

KOPSI200 기업 공시 데이터 챗봇을 주제로 선정

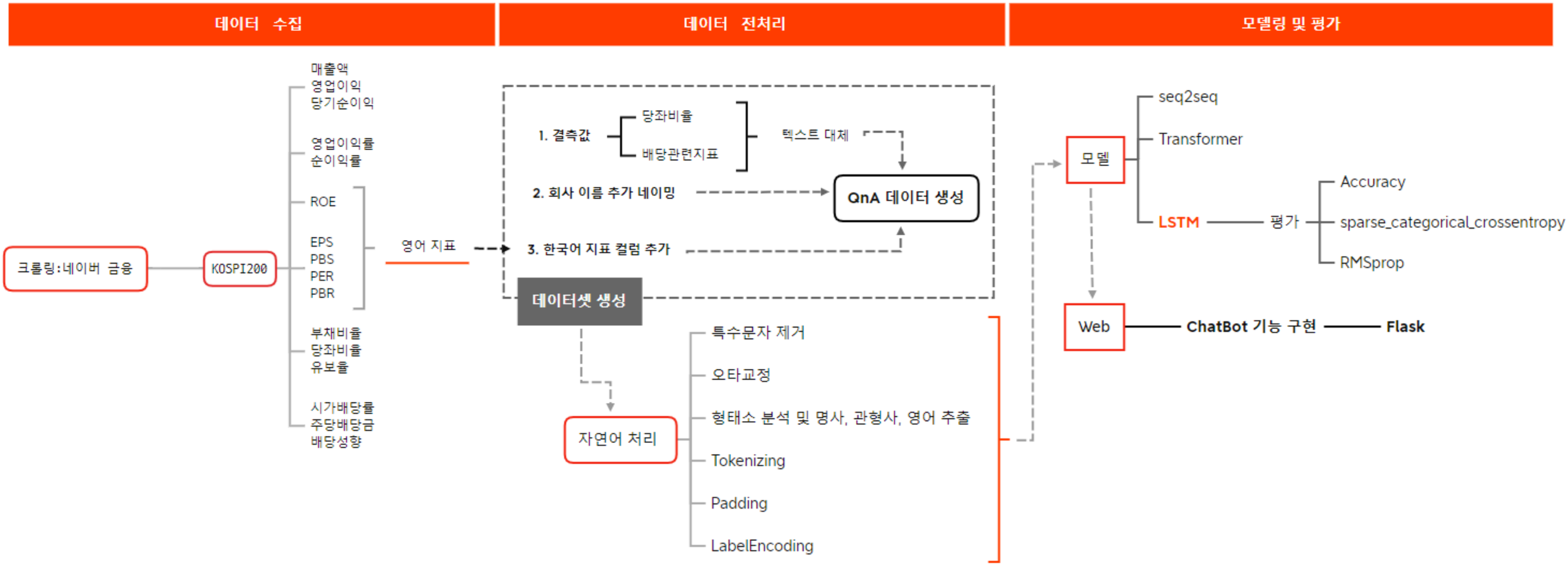
프로젝트 목표

"상장기업의 공시 정보를 챗봇으로 편리하게 제공"

프로젝트 일정

SUN	MON	TUE	WED	THU	FRI	SAT
					27	28
					미니 주제 확정	KOSPI 200
29	30	31	1	2	3	4
당기순이익 txt	데이터 통합	1차 모델링 및 피드백			FLASK WEB 구현	
	전처리					데이터 확장
5	6	7	8	9	10	
전처리		최종 모델링		PPT	발표	
데이터 확장		FLASK WEB 구현				

Flow Chart


데이터 수집
데이터 전처리
모델링 및 평가
크롤링:네이버 금융
KOSPI200
매출액
영업이익
당기순이익
영업이익률
순이익률
ROE
EPS
PBS
PER
PBR
부채비율
당좌비율
유보율
시가배당률
주당배당금
배당성향
영어 지표
1. 결측값
당좌비율
배당관련지표
텍스트 대체
2. 회사 이름 추가 네이밍
QnA 데이터 생성
3. 한국어 지표 컬럼 추가
데이터셋 생성
자연어 처리
특수문자 제거
오타교정
형태소 분석 및 명사, 관형사, 영어 추출
Tokenizing
Padding
LabelEncoding
모델
seq2seq
Transformer
LSTM
평가
Accuracy
sparse_categorical_crossentropy
RMSprop
Web
ChatBot 기능 구현
Flask

03
데이터

데이터 설명 | 전처리 | 데이터셋

AntHouse

데이터 설명

수집 대상 : KOSPI 200

수집 데이터 : 기업의 경영 실적 데이터

*가장 최근년도인 2022년 데이터.
*아직 공시 안됐을 경우 2021년 데이터

수집 방법 : 네이버 금융 페이지 크롤링

전체 데이터 크기 :

200개 기업 * 16개 지표 = 3200개

<그림1 : 삼성전자 크롤링 페이지 >

NAVER 증권 | 삼성전자 | 통합검색

기업실적분석

주요재무정보	최근 연간 실적				최근 분기 실적				
	2019.12	2020.12	2021.12	2022.12 (E)	2021.12	2022.03	2022.06	2022.09	2022.12 (E)
	IFRS 연결	IFRS 연결	IFRS 연결	IFRS 연결	IFRS 연결	IFRS 연결	IFRS 연결	IFRS 연결	IFRS 연결
매출액(억원)	2,304,009	2,368,070	2,796,048	3,054,876	765,655	777,815	772,036	767,817	735,244
영업이익(억원)	277,685	359,939	516,339	463,279	138,667	141,214	140,970	108,520	72,102
당기순이익(억원)	217,389	264,078	399,074	371,012	108,379	113,246	110,988	93,892	62,429
영업이익률(%)	12.05	15.20	18.47	15.17	18.11	18.15	18.26	14.13	9.81
순이익률(%)	9.44	11.15	14.27	12.14	14.16	14.56	14.38	12.23	8.49
ROE(%)	8.69	9.98	13.92	11.67	13.92	15.13	15.10	13.42	
부채비율(%)	34.12	37.07	39.92		39.92	39.34	36.64	36.35	
당좌비율(%)	233.57	214.82	196.75		196.75	202.26	219.39	226.19	
유보율(%)	28,856.02	30,692.79	33,143.62		33,143.62	34,110.56	35,054.68	35,798.23	
EPS(원)	3,166	3,841	5,777	5,374	1,567	1,638	1,613	1,346	788
PER(배)	17.63	21.09	13.55	10.29	13.55	10.92	8.65	8.61	70.19
BPS(원)	37,528	39,406	43,611	48,462	43,611	45,106	46,937	49,387	48,462
PBR(배)	1.49	2.06	1.80	1.14	1.80	1.54	1.21	1.08	1.14
주당배당금(원)	1,416	2,994	1,444	1,521					
시가배당률(%)	2.54	3.70	1.84						
배당성향(%)	44.73	77.95	25.00						

데이터 전처리

결측값

당좌비율(당좌자산 / 유동부채)	배당관련지표 (주당배당금, 시가배당률, 배당성향)
• 결측 이유 : 금융주의 경우 예금이 부채로 잡히는 특성 상 유동부채가 의미가 없음 • 결측값 대체 : '금융주의 경우 당좌비율이 확인되지 않습니다' • 결측값 개수 : 21개	• 결측 이유 : 최근 배당 관련 공시가 안된 기업들 • 결측값 대체 : '최근 배당내역이 없습니다' • 결측값 개수 : 35개

데이터 전처리

데이터 추가

영어 지표 추가

ROE → 자기자본이익률

EPS → 주당순이익

PER → 주가수익비율

BPS → 주당순자산가치

PBR → 주가순자산비율

컬럼 5개 추가 생성

회사이름 추가

1. 줄이기

   삼성전자 → 삼전

2. 영-한,한-영 변환

   코웨이 → COWAY

3. 띄어쓰기

   카카오뱅크 → 카카오 뱅크

기업 당 2개씩 추가 네이밍

데이터 셋

회사명 네이밍 파일

기업명(띄어쓰기, 줄이기, 영-한 변환) 2개씩 네이밍
총 600개 기업이름 네이밍

삼전(삼성전자)
삼성 전자(삼성전자)
엘지에너지(LG에너지솔루션)
엘지에너지솔루션(LG에너지솔루션)
에스케이하이닉스(SK하이닉스)
에스케이하이(SK하이닉스)
삼성바이오(삼성바이오로직스)
삼성 바이오로직스(삼성바이오로직스)
엘지화학(LG화학)
엘지 화학(LG화학)
삼성에스디아이(삼성SDI)
삼성 에스디아이(삼성SDI)
현대자동차(현대차)
현대(현대차)
네이버(NAVER)
네이 버(NAVER)
KAKAO(카카오)
카톡(카카오)

챗봇 학습시킬 QA 파일

600개 기업이름 * 지표 21개 * 질문 8개 = **100,800**개

Q	A
삼성전자 매출액	삼성전자의 2022년도 매출액은 3,054,876(억원)입니다.
삼성전자의 최근 매출액	삼성전자의 2022년도 매출액은 3,054,876(억원)입니다.
삼성전자 매출액 알려줘	삼성전자의 2022년도 매출액은 3,054,876(억원)입니다.
삼성전자 매출액 궁금해	삼성전자의 2022년도 매출액은 3,054,876(억원)입니다.
삼성전자 매출액 얼마야	삼성전자의 2022년도 매출액은 3,054,876(억원)입니다.
삼성전자 매출액 어때	삼성전자의 2022년도 매출액은 3,054,876(억원)입니다.
삼성전자 매출액 말해봐	삼성전자의 2022년도 매출액은 3,054,876(억원)입니다.
삼성전자 매출액 괜찮아	삼성전자의 2022년도 매출액은 3,054,876(억원)입니다.
삼성전자 영업이익	삼성전자의 2022년도 영업이익은 463,279(억원)입니다.
삼성전자의 최근 영업이익	삼성전자의 2022년도 영업이익은 463,279(억원)입니다.
삼성전자 영업이익 알려줘	삼성전자의 2022년도 영업이익은 463,279(억원)입니다.
삼성전자 영업이익 궁금해	삼성전자의 2022년도 영업이익은 463,279(억원)입니다.
삼성전자 영업이익 얼마야	삼성전자의 2022년도 영업이익은 463,279(억원)입니다.
삼성전자 영업이익 어때	삼성전자의 2022년도 영업이익은 463,279(억원)입니다.
삼성전자 영업이익 말해봐	삼성전자의 2022년도 영업이익은 463,279(억원)입니다.
삼성전자 영업이익 괜찮아	삼성전자의 2022년도 영업이익은 463,279(억원)입니다.
삼성전자 당기순이익	삼성전자의 2022년도 당기순이익은 371,012(억원)입니다.
삼성전자의 최근 당기순이익	삼성전자의 2022년도 당기순이익은 371,012(억원)입니다.
삼성전자 당기순이익 알려줘	삼성전자의 2022년도 당기순이익은 371,012(억원)입니다.

04

모델링

전처리 | 모델 | 결과(WEB)

AntHouse

전처리

1 ***CHANGE_FILTER*** **"([~!?\"':;&)(])"**

: 숫자 표현에 필요한 . 과 , 는 제거하지 않음

```
import re
CHANGE_FILTER = re.compile("([~!?\"':;&)(])")
```

2 ***hanspell***을 사용해 조사 **'은/는', '을/를'** 교정

```
from hanspell import spell_checker
question = '삼성전자은 당기순이익를 궁금해'
ok = spell_checker.check(question)
question = ok.checked
print(question)
```

```
삼성전자는 당기순이익을 궁금해
```

3 ***okt. pos*** 를 통해 형태소 구분
명사,관형사, 영어만 추출하여 문장 재구성

```
from konlpy.tag import Okt
okt = Okt()
clean_words = []
question = '삼성 전자는 당기순이익을 궁금해'
words = okt.pos(question)
for word in words:
    if word[1] in ['Noun', 'Modifier', 'Alpha']:
        clean_words.append(word[0])
    question = ' '.join(clean_words)
print(question)
```

```
삼성 전자 당 기 순이익
```

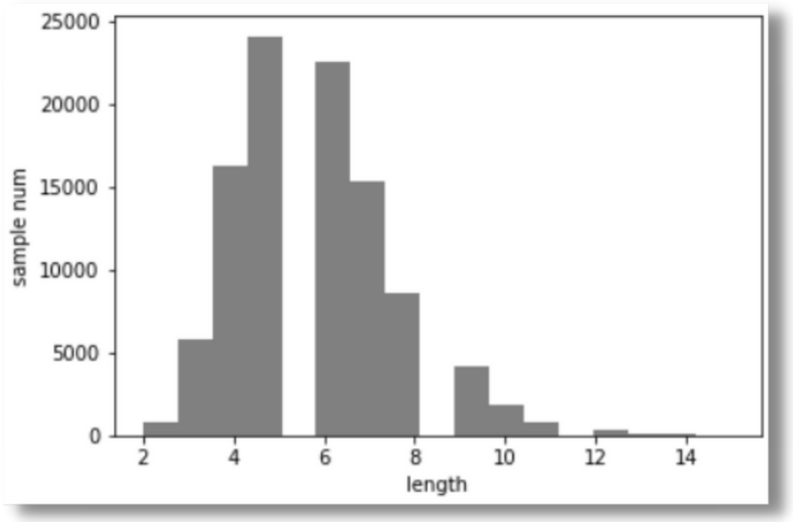
전처리

4 **Tokenizing**

```
[[1], [2], [], [3], [4], [], [5], [], [6], [], [7], [8], [9]]
```

5 **Padding**

max_len(=15) 을 기준으로 데이터 길이 맞춰주기



6 **LabelEncoding**

분류에 사용하기 위해 Answer 값 인코딩

```
{0: 'BGF리테일의 2021년도 당좌비율은 68.85(%)입니다.',
 1: 'BGF리테일의 2021년도 배당성향은 35.10(%)입니다.',
 2: 'BGF리테일의 2021년도 부채비율은 220.81(%)입니다.',
 3: 'BGF리테일의 2021년도 시가배당율은 2.06(%)입니다.',
 4: 'BGF리테일의 2021년도 유보율은 4,439.53(%)입니다.',
 5: 'BGF리테일의 2022년도 BPS은 55,268(원)입니다.',
 6: 'BGF리테일의 2022년도 EPS은 11,332(원)입니다.',
 7: 'BGF리테일의 2022년도 PBR은 3.45(배)입니다.',
 8: 'BGF리테일의 2022년도 PER은 16.85(배)입니다.',
 9: 'BGF리테일의 2022년도 ROE은 22.21(%)입니다.',
 10: 'BGF리테일의 2022년도 당기순이익은 1,963(억원)입니다.',
 11: 'BGF리테일의 2022년도 매출액은 76,042(억원)입니다.',
 12: 'BGF리테일의 2022년도 순이익율은 2.58(%)입니다.',
 13: 'BGF리테일의 2022년도 영업이익율은 3.45(%)입니다.',
 14: 'BGF리테일의 2022년도 영업이익은 2,626(억원)입니다.',
 15: 'BGF리테일의 2022년도 자기자본이익율은 22.21(%)입니다.',
 16: 'BGF리테일의 2022년도 주가수익비율은 11,332(배)입니다.',
 17: 'BGF리테일의 2022년도 주가순자산비율은 3.45(배)입니다.',
 18: 'BGF리테일의 2022년도 주당배당금은 3,307(원)입니다.',
 19: 'BGF리테일의 2022년도 주당순이익은 11,332(원)입니다.',
 20: 'BGF리테일의 2022년도 주당순자산가치은 55,268(원)입니다.',
```

y 데이터의 idx_label

모델

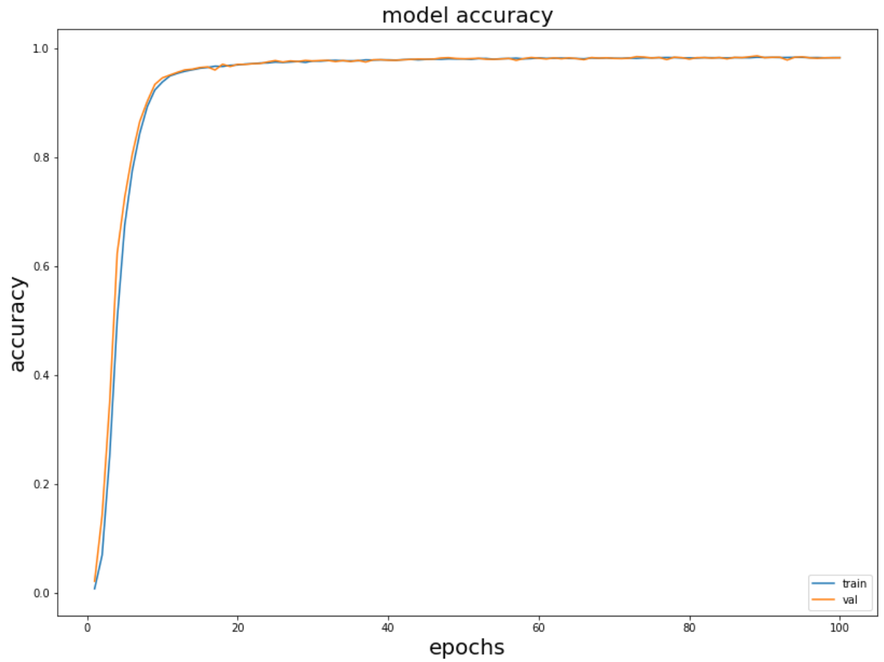
LSTM을 활용한 다중분류 모델

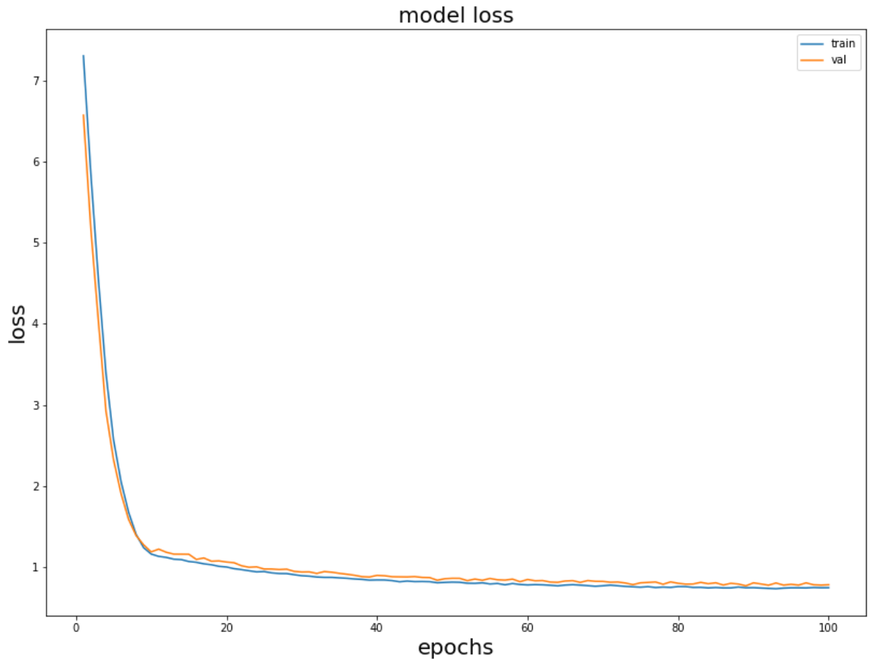
```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_2 (Embedding)     (None, None, 64)          37888

 lstm_2 (LSTM)               (None, 256)               328704

 dense_5 (Dense)             (None, 3886)              998702

=================================================================
Total params: 1,365,294
Trainable params: 1,365,294
Non-trainable params: 0
_________________________________________________________________
```

Optimizer	**RMSProp**
Loss	**다중 분류** **sparse categorical crossentropy**
Metrics	**Accuracy**

모델

epoch = 100 batchsize = 64

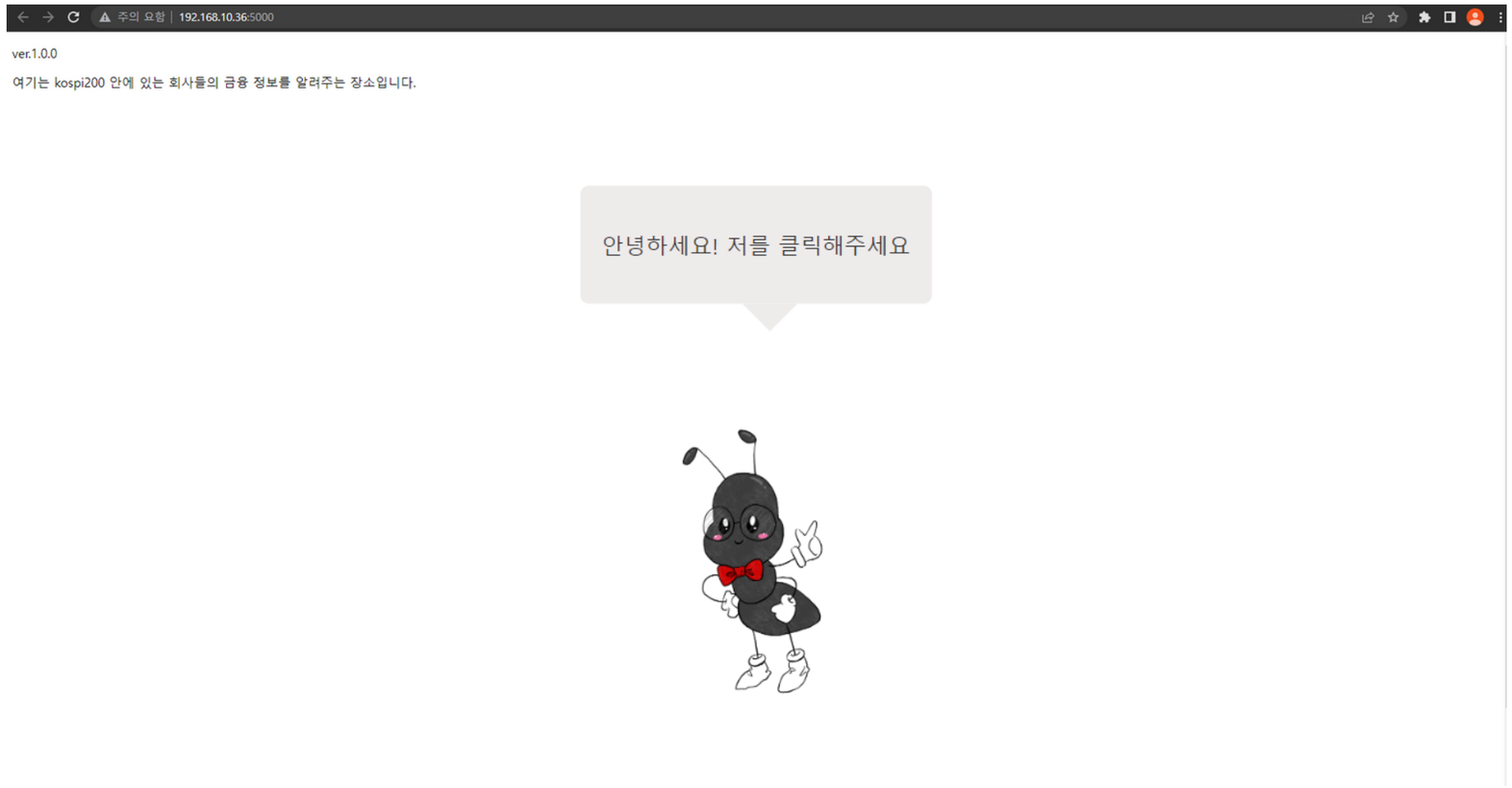




train_acc: 0.9834
val_acc: 0.9831

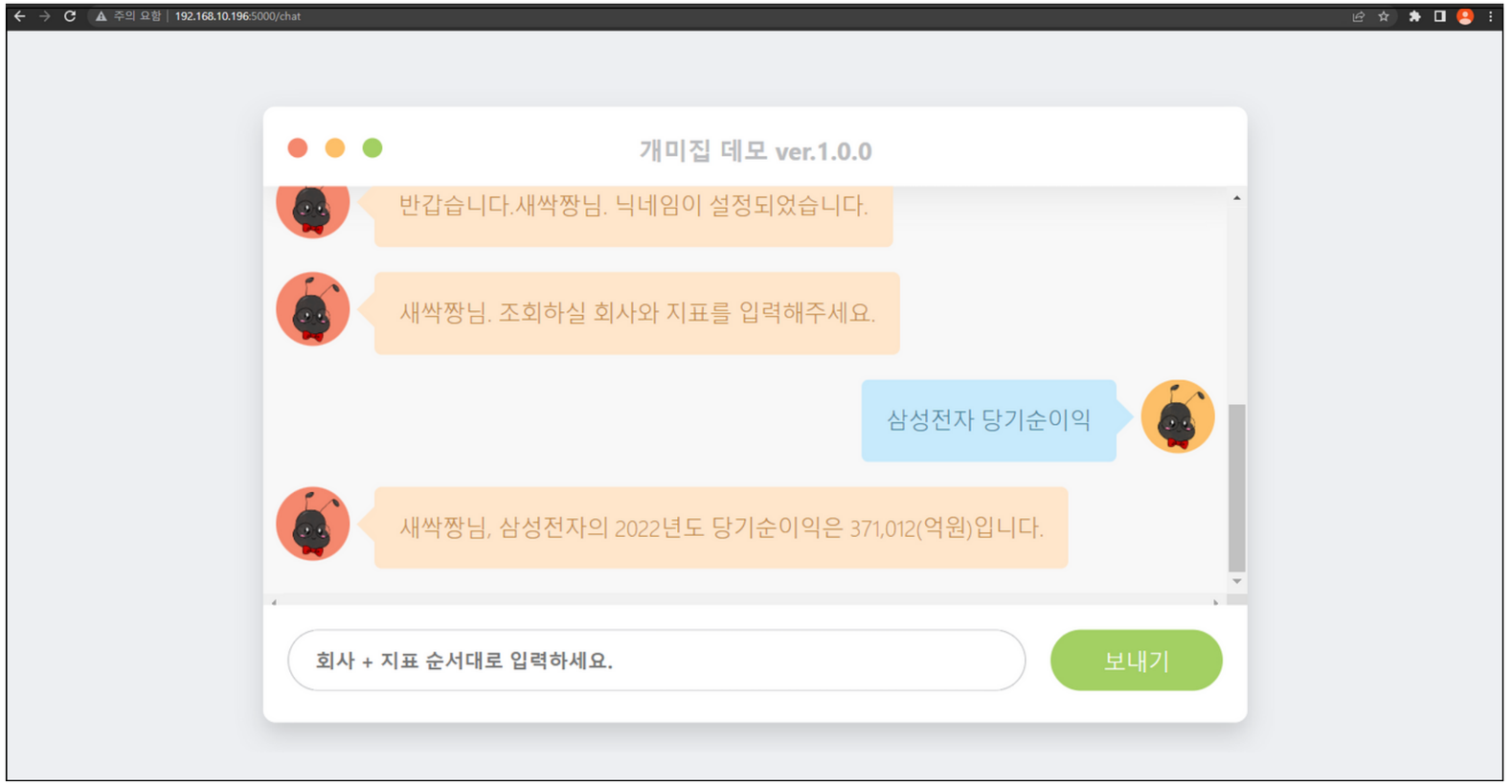
train_loss: 0.7449
val_loss: 0.7807

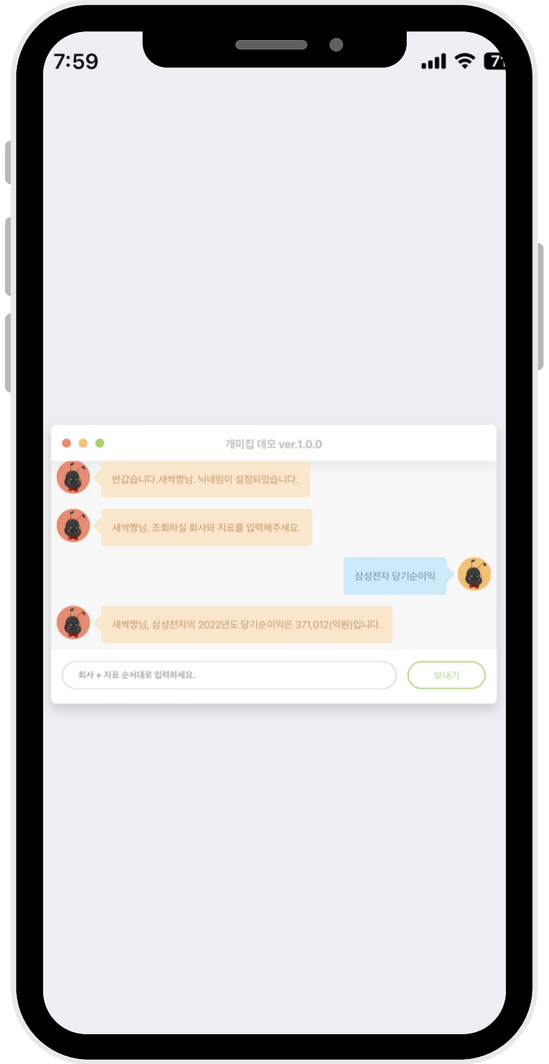
WEBPAGE 구현


주의 요함 | 192.168.10.36:5000
ver.1.0.0
여기는 kospi200 안에 있는 회사들의 금융 정보를 알려주는 장소입니다.
안녕하세요! 저를 클릭해주세요


7:57
1.0.0
는 kospi200 안에 있는 회사들의 금융 정보를 알려주는 장
다.
안녕하세요! 저를 클릭해주
세요

WEBPAGE 구현


주의 요함 | 192.168.10.196:5000/chat
개미집 데모 ver.1.0.0
반갑습니다.새싹짱님. 닉네임이 설정되었습니다.
새싹짱님. 조회하실 회사와 지표를 입력해주세요.
삼성전자 당기순이익
새싹짱님, 삼성전자의 2022년도 당기순이익은 371,012(억원)입니다.
회사 + 지표 순서대로 입력하세요.
보내기


7:59
개미집 데모 ver.1.0.0
반갑습니다.새싹짱님. 닉네임이 설정되었습니다.
새싹짱님. 조회하실 회사와 지표를 입력해주세요.
삼성전자 당기순이익
새싹짱님, 삼성전자의 2022년도 당기순이익은 371,012(억원)입니다.
회사 + 지표 순서대로 입력하세요.
보내기

결과_WEB

Flask 를 이용한 챗봇 서버 구현

사용언어 : HTML, Python, CSS , JS

http://192.168.10.196:5000

웹으로 확인해보세요

05

마무리

의의 | 한계 | 확장

AntHouse

의의

데이터 수집부터 라벨링까지 모든 단계를 경험

기업명 네이밍 자동화를 통해 데이터셋 구축

transformer, seq2seq모델 성능평가 비교 후 keras모델 사용

HTML, Python, CSS , JS으로 웹구현

한계

dart openapi를 사용했지만 네이버 금융 페이지에 의존

커스텀 모델 대신 keras의 sequential 모델을 사용하여 복잡한 대화는 처리 불가능

Final을 향해

aws 서버 오픈 및 카카오톡 채널 오픈 예정

상장된 기업 뉴스기사 감정분류

새로운 지표 생성 및 머신러닝, 딥러닝을 통한 예측 결과 제공

Thank you!

We are AntHouse