뉴스와 어텐션 메커니즘을 활용한
주가 등락 예측

팀 1: 한경훈, 허정욱, 이형석

1. 배경 및 동기

- **주제 선정배경 및 이유**
  - 팀원 모두의 **공통적인 관심사**에 해당
  - 지금까지 공부해온 내용을 바탕으로 직접 구현이 가능한 미니 프로젝트라 판단

- **구현의 목적 및 기대효과**
  - 공부한 NLP메커니즘 중 적합한 것을 선별하고 적용할 수 있는 능력 배양
  - **학습데이터를 직접 공수**하고 이를 **전처리** 하는 과정을 통해 실력 향상
  - 예측 정확도를 **55%~60%** 정도로 예상하는데, 이를 끌어올리는 과정을 통해 심화학습 수행

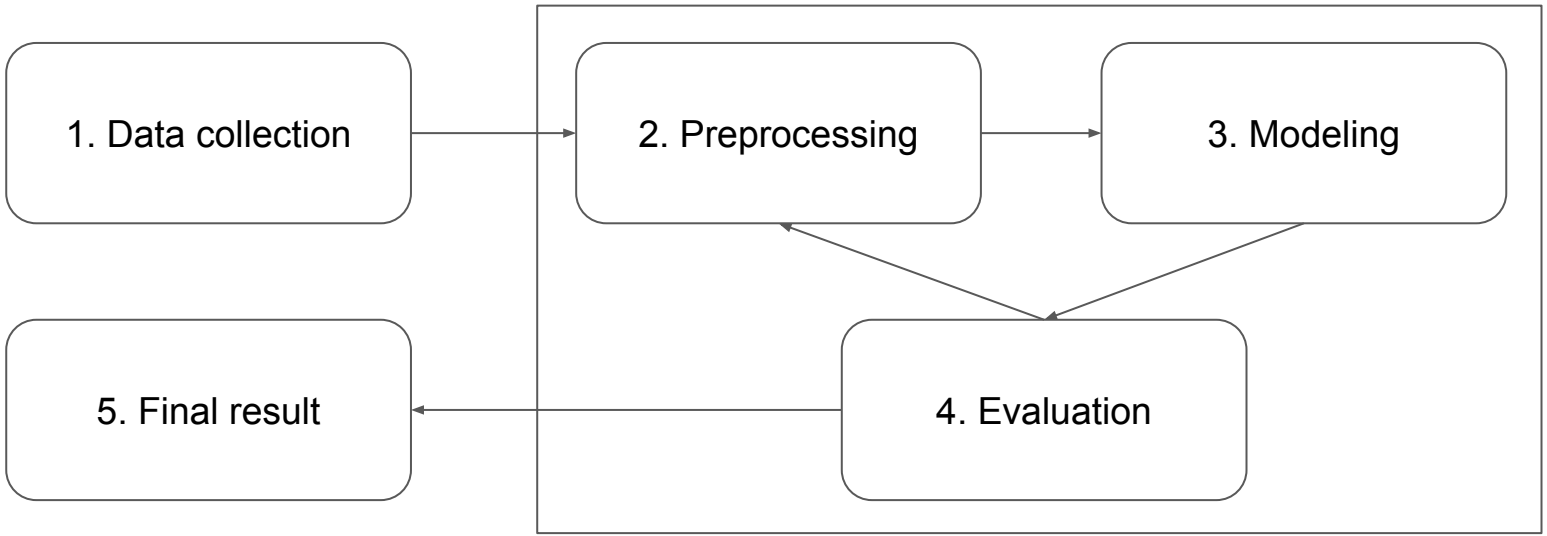
2. 실행 계획

- **Data Science 관점에서의 문제 정의**
  - **기존 모델들은 정형데이터(주가, 환율, 각종 차트 모델 등) 중심**으로 학습함
  - **이 프로젝트에선 비정형데이터인 뉴스기사만을 가지고 익일의 주가를 예측**하고자 함

- **프로젝트 프로세스**
  - **가설**
    - 뉴스 **제목**으로부터 호재인지 악재인지 구별할 수 있다고 가정
    - 뉴스가 공개되고 **하루 지난 시점**에 주가에 반영된다고 가정
    - 종목 선정은 **삼성전자**로 함
      - 국내 시가총액 1위의 삼성전자가 관련 뉴스가 가장 많고 다양하며, 또한 오랜 기간 데이터가 쌓임
      - 주식 시장의 예상치 못한 변수로 인한 주가 변동성이 낮다고 봄
  - **데이터셋 소개**
    - 네이버 웹스크래핑을 통해 삼성전자 관련 기사의 **날짜, 언론사, 제목만 추출**하여 저장
    - 기간: 1998.05.20 ~ 현재(7월말 예정)
  - **데이터셋 활용 방안**
    - 일자를 기준으로 **뉴스 타이틀**을 concat하여 **일자별 문서**처럼 만들고자 함
    - 그 후, 전처리를 통해 컴퓨터가 인식할 수 있는 숫자로 변경하고, **attention을 포함한 다양한** seq2seq 메커니즘을 사용해 학습 후 주가의 등락을 예측해 보고자 함
  - **이후 방향**
    - 성능지표를 확인 후 정형데이터를 붙여보면서 성능이 올라가는지 확인해 보려고 함

- Random Walk: 무작위 걸음, 무작위 행보

3. 진행 과정


1. Data collection
2. Preprocessing
3. Modeling
4. Evaluation
5. Final result

3-1. 진행 과정_Data collection

셀러니움(*)을 사용해 '삼성전자' 네이버 뉴스검색 결과를 동적 웹스크래핑

3대의 PC, 3일간의 웹스크래핑, 약 400만건의 '삼성전자' 관련 기사 수집

	언론사	날짜	제목
0	연합뉴스	1998.05.20.	주가 사흘째 상승세
1	연합뉴스	1998.05.20.	아남텔레콤, PDA 데이터 전송서비스 개시
2	연합뉴스	1998.05.20.	<회전목마> 삼성전자, 사이버 사외보 제작
3	연합뉴스	1998.05.20.	<주식시황> 사흘째 상승, 3백70선 다가서
4	연합뉴스	1998.05.20.	-춘계승마- 신창무,싼졸쥬마장마술 우승
...	...	...	...
3851965	한국경제	2022.07.21.	디지털이미지뱅크, 한투등 3개사서 20억원 투자 유치
3851966	한국경제	2022.07.21.	[코스닥공시] 세보기계
3851967	한국경제	2022.07.21.	전자.통신업체 신용등급도 급상승세
3851968	한국경제	2022.07.21.	종합주가 폭락세, 916.01(-37.21) 전장마감
3851969	한국경제	2022.07.21.	산자부, 수치제어장치 산.학.연 공동연구 개발에 성공

3851970 rows × 3 columns

- 셀러니움: 웹사이트 테스트를 위한 도구로 브라우저 동작을 자동화 할 수 있음

3-2. 진행 과정_Preprocessing

1. 데이터 레이블링
2. 데이터 정제하기
3. 토큰화
4. 정수 인코딩
5. 패딩

```
# 삼성전자 주가 정보 불러올 수 있도록 Library
import
```

```
import FinanceDataReader as fdr
# 삼성전자(005930) 전체 (1998-05-20 ~
2022-07-21)
samsung = fdr.DataReader('005930')
# 주가가 상승하였을 경우 1, 하락하였을 경우 0으로 변환
```

```
def change_of_next_day_of_stock_market(today):
    return_change = -1
    today = datetime.date.isoformat(datetime.date.fromisoformat(today)
            + datetime.timedelta(days=1)) # 하루 미룸
    while return_change == -1:
        try:
            if stock_df.loc[today]["Change"] > 0:
                return_change = 1
            else:
                return_change = 0
        except:
            today = datetime.date.isoformat(datetime.date.
                    fromisoformat(today) + datetime.timedelta(days=1))
            # While문을 통해 찾아질 때까지 뒤로 하루씩 미룸
    return return_change
```

3-2. 진행 과정_Preprocessing

1. 데이터 로딩 및 분할하기
2. 데이터 정제하기
3. 토큰화
4. 정수 인코딩
5. 패딩

```
# 중복 열 제거
train_data.drop_duplicates(subset=['document'],
inplace=True)
# Null이 있나 확인 후 있으면 제거
train_data.loc[train_data.document.isnull()]
train_data = train_data.dropna(how = 'any') # Null 값이
존재하는 행 제거
```

```
# 날짜 별 기사를 모아 일자 별 문서 만들기 (*이후 중요*)
sentences = []
for i in dates:
    titles = df[df["Date"] == i]["Title"].values
    sentence = ""
    for title in titles:
        sentence = sentence + title + "\n"
    sentences.append(sentence)
```

```
# 한글과 공백을 제외하고 모두 제거
train_data['document'] =
train_data['document'].str.replace("[^ㄱ-ㅎㅏ-ㅣ가-힣
]","")
```

3-2. 진행 과정_Preprocessing - 날짜 별 기사 하나로 모으기 유무 비교

	Press	Date	Title	Change
0	연합뉴스	1998-05-20	주가 사흘째 상승세	0
1	연합뉴스	1998-05-20	아남텔레콤, PDA 데이터 전송서비스 개시	0
2	연합뉴스	1998-05-20	<회전목마> 삼성전자, 사이버 사외보 제작	0
3	연합뉴스	1998-05-20	<주식시황> 사흘째 상승, 3백70선 다가서	0
4	연합뉴스	1998-05-20	-춘계승마- 신창무,싼졸쥬마장마술 우승	0
...	...	...	...	...
3851965	한국경제	2022-07-21	디지털이미지뱅크, 한투등 3개사서 20억원 투자 유치	1
3851966	한국경제	2022-07-21	[코스닥공시] 세보기계	1
3851967	한국경제	2022-07-21	전자.통신업체 신용등급도 급상승세	1
3851968	한국경제	2022-07-21	종합주가 폭락세, 916.01(-37.21) 전장마감	1
3851969	한국경제	2022-07-21	산자부, 수치제어장치 산.학.연 공동연구 개발에 성공	1

기사 제목별 1개 Data

	Date	Change	Title
0	1998-05-28	1	주가 상승세 삼성전자, 하나더 판촉 확대실시 <주식시황> 주가 상승 3백20선 회복...
1	1998-05-29	1	< 6.4선거 이후의 과제 > ...(下) 대기업그룹별 현안 전자업계 수출드라이브,...
2	1998-05-30	0	정보통신업계, `실직자를 잡아라' 주가 하락세 반전 삼성전자, 세계 공용 DVD 개...
3	1998-06-01	0	<회전목마> 삼성전자, 인터넷잡지 창간 지난해 상장기업 수출 28.8% 증가 현대전...
4	1998-06-02	0	공장도價 과다 책정...소비자불신 초래 <약수터> 환경부 고위공직자 근무지이탈 말썽...
...	...	...	...
8316	2022-07-15	1	발명의 날 기념 행사 개최 朴통산장관,무역수지 개선위한 업계 간담회 최우수 상장기업...
8317	2022-07-16	1	회사채 수익률 다시 상승 12월결산법인 올해 순익 56%이상 감소 추정 삼성전자, ...
8318	2022-07-17	1	安통산장관, 무역업계와 간담회 삼성전자, 인력재배치 작업 본격화 삼성전자,2백40만...
8319	2022-07-18	1	불경기에도 자동차.화장품 광고는 활발 <주식시황> 주가, 큰 폭 상승 삼성전자 정보...
8320	2022-07-19	0	반도체 업체들, 보따리장사로 골머리 주가 이틀째 급상승 三星電子, 세계 최경량 PC...

기사 날짜별 1개 Data

3-2. 진행 과정_Preprocessing

1. 데이터 로딩 및 분할하기
2. 데이터 정제하기
3. 토큰화
4. 정수 인코딩
5. 패딩

```
# Okt는 KoNLPy Library에서 제공하는 형태소 분석기임

한국어을 토큰화할 때는 영어처럼 띄어쓰기 기준으로 토큰화를 하는 것이 아니라,
주로 형태소 분석기를 사용함 ('이런'이 '이렇다'로 변환되고,
'만드는'이 '만들다'로 변환됨)

okt = Okt()

stopwords =
['의','가','이','은','들','는','좀','잘','걍','과',
'도','를','으로','자','에','와','한','하다']

X_train = []
for sentence in tqdm(train_data['document']):
    tokenized_sentence = okt.morphs(sentence) # 토큰화
    stopwords_removed_sentence = [word for word in
tokenized_sentence if not word in stopwords] # 불용어 제거
    X_train.append(stopwords_removed_sentence)
```

3-2. 진행 과정_Preprocessing

1. 데이터 로딩 및 분할하기
2. 데이터 정제하기
3. 토큰화
4. 정수 인코딩
5. 패딩

```
# 훈련데이터에 대해서 단어집합(Vocaburary) 만들기
tokenizer = Tokenizer(vocab_size)
tokenizer.fit_on_texts(X_train)

# 단어 집합 크기를 반영해서 정수 배정
X_train = tokenizer.texts_to_sequences(X_train)
X_test = tokenizer.texts_to_sequences(X_test)
```

적용 예시

[['barber', 'person'], ['barber', 'good', 'person'], ['barber', 'huge', 'person'] …]

↓ 단어집합 만들기

{'barber': 1, 'good': 2, 'person': 3}

↓ 정수 인코딩

[[1, 5], [1, 5], [1, 3, 5], …]

3-2. 진행 과정_Preprocessing

1. 데이터 로딩 및 분할하기
2. 데이터 정제하기
3. 토큰화
4. 정수 인코딩
5. 패딩

자연어 처리를 하다보면 각 문장이 서로 길이가 다를 수 있음
그러나 기계는 길이가 같은 문장들에 대해서는,한꺼번에 묶어서 처리할 수 있음
즉, 병렬연산을 위해서 여러 문장의 길이를 동일하게 맞춰주는 작업이 필요함

```
X_train = pad_sequences(X_train, maxlen=MAX_LEN)
X_test = pad_sequences(X_test, maxlen=MAX_LEN)
```

적용 예시

[[1, 5],
[7, 7, 3, 2, 10, 1, 11],
[1, 12, 3, 13]]

패딩(열 길이를 맞춤)

[[1, 5, 0, 0, 0, 0, 0],
[7, 7, 3, 2, 10, 1, 11],
[1, 12, 3, 13, 0, 0, 0]]

3-3. 진행 과정_Modeling

모델	LSTM (Long Short-Term Memory)	양방향 LSTM with Attention	Transformer
상세설명	• 바닐라 RNN 모델은 오래된 기억을 제대로 활용하지 못하는 장기 의존성 문제를 지님 • LSTM은 은닉층의 메모리 셀에 입력, 망각, 출력 게이트를 추가하여 불필요한 기억을 지우고, 기억해야 할 것을 정해 장기 의존성 문제를 어느정도 해결함	• RNN에 기반한 모델(LSTM)은 고정된 크기의 벡터에 모든 정보를 압축하려고 해서 정보 손실이 발생하고, 기울기 소실 문제가 존재함 • Attention은 디코더에서 출력을 예측하는 매 시점마다, 인코더에서 예측해야할 단어와 연관이 있는 단어부분을 좀 더 집중해서 봄으로써 이러한 문제를 해결함	• RNN에 기반한 Seq2Seq 모델이 가지고 있는 단점을 Attention을 통해 보정하는 것 방식이 아니라 Attention만으로 인코더와 디코더를 만든 모델 • 기존 RNN을 사용하지 않고도 더 우수한 성능을 보임

출처 : 딥 러닝을 이용한 자연어 처리 입문

4. 결과

Model	Optimizer	Dropout	Train data (행) 수	Max sample 길이	Accuracy
LSTM	SGD	-	2,830 / 6,656 / 2,012,775 /	4,000 / 8,000 / 15 /	52.68% / 50.21% / 52.74% /
	ADAM	-	2,830 / 6,656 / 2,012,775 /	4,000 / 8,000 / 15 /	48.87% / 51.83% / 58.08% /
양방향 LSTM with Attention	SGD	0.5	2,830 / 2,012,775	4,000 / 15	50.28% / 50.50%
	ADAM	0.5	2,830 / 2,012,775	4,000 / 15	50.28% / 57.29%
Transformer	SGD	0.1	2,012,775	15	50.50%
	ADAM	0.1	2,012,775	15	50.50%

5. 모델 활용 분석 - 8월 3일 삼성전자 관련 뉴스 제목 두 개 분석

1) 입력받음: "삼성전자 비중 확대"…한국운용, 삼성그룹펀드 전략 재편
한글만: 삼성전자비중확대한국운용삼성그룹펀드전략재편
형태소 정제: ['삼', '성', '전', '자비', '중', '확대', '한국', '운용', '삼성', '그룹', '펀드', '전략', '재편']
불용어 제거: ['삼', '성', '전', '자비', '중', '확대', '한국', '운용', '삼성', '그룹', '펀드', '전략', '재편']
숫자로: [[16, 10, 30, 9419, 38, 100, 32, 1740, 1, 71, 301, 120, 1970]]
패들링: [[0 0 16 10 30 9419 38 100 32 1740 1 71 301 120 1970]]
스코어: 0.4960644841194153
49.61% 확률로 다운!

2) 입력받음: 로봇주, 삼성전자 '무인공장' 도입 보도에 일제히 상승
한글만: 로봇주삼성전자무인공장도입보도에일제히상승
형태소 정제: ['로봇', '주', '삼', '성전', '자', '무인', '공장', '도입', '보도', '에', '일제', '히', '상승']
불용어 제거: ['로봇', '주', '삼', '성전', '무인', '공장', '도입', '보도', '일제', '히', '상승']
숫자로: [[628, 11, 16, 385, 5913, 158, 539, 1870, 1157, 1191, 25]]
패들링: [[0 0 0 0 628 11 16 385 5913 158 539 1870 1157 1191 25]]
스코어: 0.911342978477478
91.13% 확률로 업!

6. 모델 활용 예시 - 8월 5일 10개 기사 제목 분석

입력받음: 연세대 의대 졸업→복지부 공보의→삼성전자 ======> 10.51% 확률로 다운!

입력받음: [ET시선]미룰 수 없는 이재용 부회장 사면 ======> 26.18% 확률로 다운!

입력받음: 오전증시 상승유지…'제약·2차 전지' 투심 집중 ======> 53.24% 확률로 업!

입력받음: 델리오, 웹3.0 디지털자산 지갑 자체 개발 ======> 51.15% 확률로 업!

입력받음: "이번엔 다를까"…번번이 '특사' 실패한 이재용, 긍정 여론 속 삼성 '침묵' ======> 76.04% 확률로 업!

입력받음: 페이 포인트에 명품까지···이통3사, 신작 갤럭시 사전알람 혜택 경쟁 ======> 70.50% 확률로 업!

입력받음: 삼성 언팩 D-5…Z폴드4·플립4 물들여 '보라' ======> 26.91% 확률로 다운!

입력받음: KB증권, 8월 첫째주 삼성SDI 등 8종목 매수 추천 ======> 52.90% 확률로 업!

입력받음: 경제단체, 이재용·신동빈 광복절 특별사면 건의 추진 ======> 85.65% 확률로 업!

입력받음: 삼성·LG전자 앞다퉈…가전, 예술을 입다 ======> 34.61% 확률로 다운!

6. 모델 활용 예시 - 8월 5일 10개 기사 제목 분석

10.51%	53.24%	76.04%	26.91%	85.65%
26.18%	51.15%	70.50%	52.90%	34.61%
평균 48.0% -> 오늘 장 마감 전에 파는 것을 추천!				

7. 추가 연구

- 다양한 형태소 분석기를 활용할 필요성이 있다
  - 일관되지 못 함 예) ['삼', '성', '전', '자비', '중', '확대', '한국', '운용', '삼성', '그룹', '펀드', '전략', '재편']
  - 대안

```
okt = Okt()
kkma = Kkma()
han = Hannanum()
```

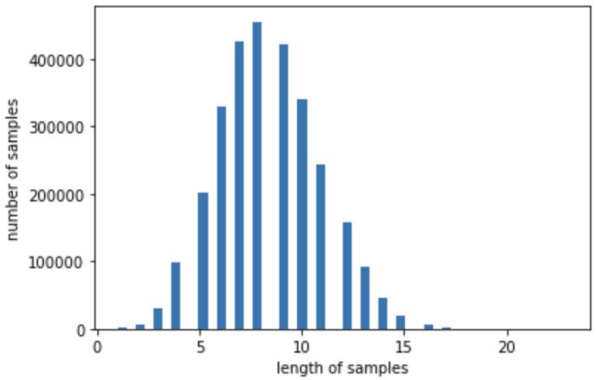
- 형태소 분석기에서 걸러지지 못한 단어들은 불용어 처리를 통해 거를 필요가 있다
  - 삼성전자를 불용어로 넣기

```
print(tokenizer.word_index)
```

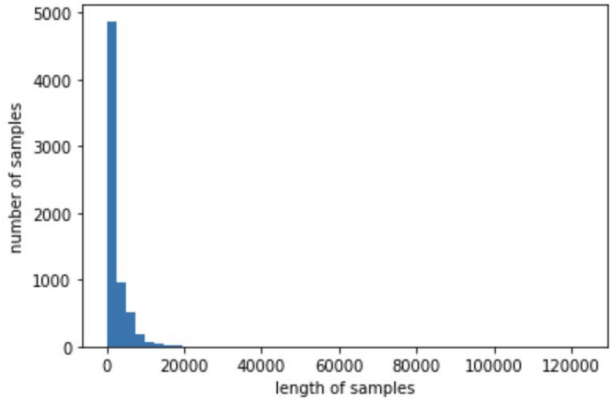
```
{'삼성': 1, '전자': 2, '코스피': 3, '갤럭시': 4, '이재용': 5, '일': 6, '선': 7, '시': 8, '반도체': 9, '성': 10, '주': 11, '출시': 12, '년': 13, '기업': 14, '위': 15, '삼': 16, '시장': 17, '마감': 18, '서': 19, '황': 20, '투자': 21, '분기': 22, '등': 23, '장': 24, '상승': 25, '스마트폰': 26, '하락': 27, '실적': 28, '노트': 29, '전': 30, '조': 31, '한국': 32, '월': 33, '외국인': 34, '만원': 35, '대': 36, '기관': 37, '중': 38, '애플': 39, '되다': 40, '만': 41, '기술': 42, '부회장': 43, '보다': 44, '증시': 45, '세계': 46, '억': 47, '최고': 48, '공개': 49, '종합': 50, '회장': 51, '사업': 52, '매수': 53, '국내': 54, '만에': 55, '증권': 56, '대다': 57, '사': 58, '폰': 59, '적': 60, '기': 61, '원': 62, '순': 63, '첫': 64, '글로벌': 65, '최대': 66, '판매': 67, '영업': 68, '포토': 69, '세': 70, '그룹': 71, '개다': 72, '코스닥': 73, '을': 74, '회복': 75, '사장': 76, '까지': 77, '주가': 78, '제': 79, '화': 80, '없다': 81, '종목': 82, '스마트': 83, '명': 84, '에서': 85, '형': 86, '수': 87, '지': 88, '상': 89, '오늘': 90,
```

8. 의문점

제목의 최대 길이 : 23
제목의 평균 길이 : 8.381338133604693



Title의 최대 길이 : 123238
Title의 평균 길이 : 2095.9888822115386



전체 샘플 중 길이가 15 이하인 샘플의 비율: 99.72

전체 샘플 중 길이가 8000 이하인 샘플의 비율: 91.2109375

패딩 과정에서 **불필요한 정보**가 많이 들어간 것이 성능 하락의 원인으로 추측

9. 결론

- 성능이 50% 초중반 수준으로 찍는 수준이라 이를 개선하기 위해 필요한 사항
  - 정형데이터(주가, 금리, 환율 등) 추가해서 feature vector 수를 늘리고자 함
  - 컴퓨터 성능 문제로 데이터를 축소한 경향이 있어 Data 양을 늘려 보고자 함
    - 데이터가 부족했던 것이 Adam optimizer를 적용했을 때 성능 저하를 가져왔을 것으로 보임

- 같은 labeling, 다른 결과
  - **기사를 일자별로** 합하여 학습 시켰을 때와 **기사 제목 하나를 한 데이터**로 합하여 학습 시켰을 때, **같은 labeling 값**을 갖고 있지만 **다른 성능**을 보임
    - 일자별로 더 많은 데이터로 학습했던게 더 좋은 성능을 낸 것이 아닐까 추측
  - 기사를 일자별로 합했을 때 편차가 큰 상황 발생
    - 패딩 과정에서 **불필요한 정보**가 많이 들어간 것이 성능 하락의 원인으로 추측
  - 모델의 옵티마이저를 바꿀 때마다 큰 성능 차이 발생
    - 다양한 옵티마이저를 최대한 활용할 필요성이 있다

8. 출처

코드 참조

- https://wikidocs.net/book/2155 – **딥 러닝을 이용한 자연어 처리 입문**

사용 라이브러리

- numpy, pandas, matplotlib, konlpy, tensorflow, FinanceDataReader, Datetime, pickle

구현 환경

- Jupyter-notebook, Google Colab, Visual Studio Code

Q & A

- 끝까지 들어주셔서 감사합니다!