Robo Adviser

# 자연어 처리 기반의 투자분석 및 예측시스템 개발

발표자 이 문 형, 강 민 재

Turnaround

00

# 프로젝트 개요 및 팀원 소개

| | |
|---|---|
| **Object** | 주가를 예측하는 로보어드바이저 개발 |
| **Mentor** | 정좌연 PE |
| **TeamMate** | 이지훈, 이문형, 강민재, 구병진, 김서정 |
| **Team** | TurnAround |

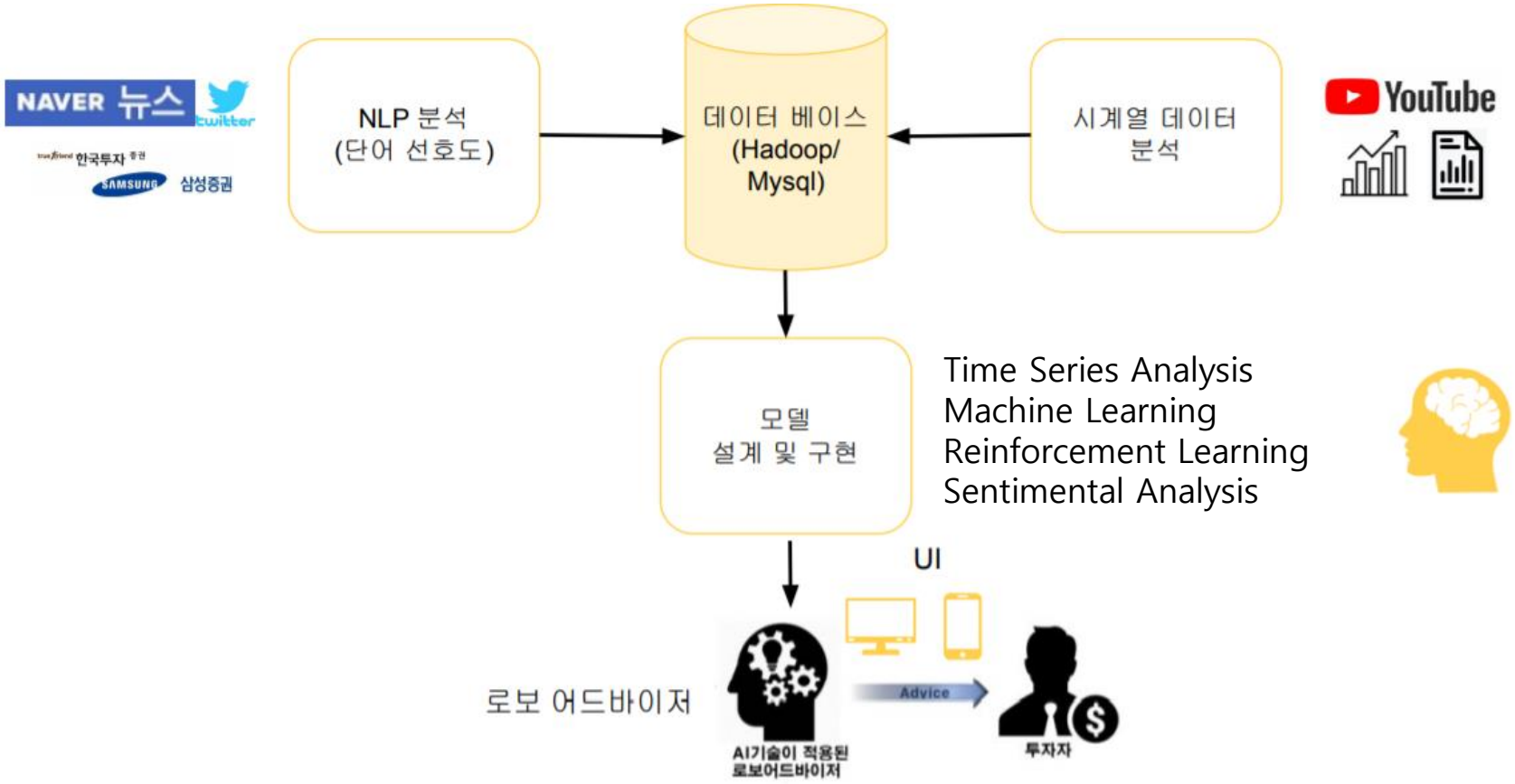
# CONTENTS



Turnaround

01

# 프로젝트개요 및 UI



01

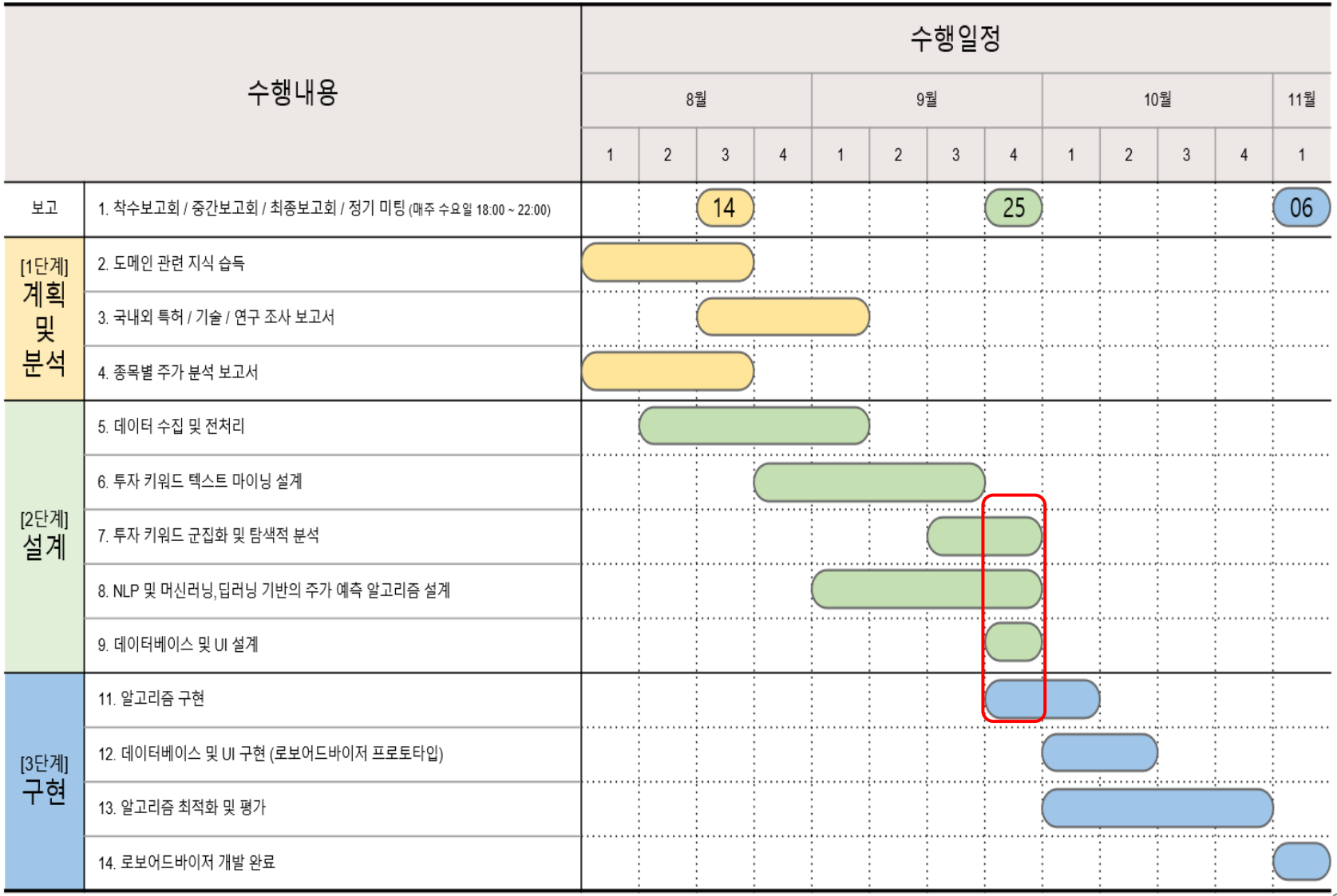
# 프로젝트개요 및 UI


1. 종목 선택

1. 주가 분석
2. 주가 키워드 분석
3. 주가 긍정/부정 단어 분석

1. 기간 입력
2. 매매 시점 예측
3. 투자손익률 예측

# Gantt Chart


수행내용
수행일정
8월
9월
10월
11월
1
2
3
4
1
2
3
4
1
2
3
4
1
보고
1. 착수보고회 / 중간보고회 / 최종보고회 / 정기 미팅 (매주 수요일 18:00 ~ 22:00)
14
25
06
[1단계] 계획 및 분석
2. 도메인 관련 지식 습득
3. 국내외 특허 / 기술 / 연구 조사 보고서
4. 종목별 주가 분석 보고서
[2단계] 설계
5. 데이터 수집 및 전처리
6. 투자 키워드 텍스트 마이닝 설계
7. 투자 키워드 군집화 및 탐색적 분석
8. NLP 및 머신러닝,딥러닝 기반의 주가 예측 알고리즘 설계
9. 데이터베이스 및 UI 설계
[3단계] 구현
11. 알고리즘 구현
12. 데이터베이스 및 UI 구현 (로보어드바이저 프로토타입)
13. 알고리즘 최적화 및 평가
14. 로보어드바이저 개발 완료

02

# 데이터 수집 및 전처리

## 개요

| 정형 데이터 | 비정형 데이터 (추후 수집 계획) |
| --- | --- |
| • 주가 지수 (KOSPI 거래실적, 지수 추이, KOSPI200, 시리즈, 산업군)<br>• 공시 자료 (공시, 분기/반기/사업보고서)<br>• 종목 정보 (시간대별/일자별/월별 시세, 재무분석, 주식대용가)<br>• 투자 지표 (PER/PBR, 가치분석, 투자분석, 투자자별/ 회원사별 거래실적)<br>• 시장 지표 (환율 USD/JPY/EUR) | • 뉴스 (네이버증권 종목별 뉴스, 스낵, 더벨, 비즈니스 워치)<br>• 유튜브<br>• NICE BIZ INFO (연관 기업 빅데이터 분석)<br>• 증권사 리포트 (한경 컨센서스, 네이버증권 리포트) |

1. 텍스트 데이터 (뉴스 등)
2. 시계열 데이터 (차트 데이터, 투자 지표, 주가 지수, 보조 지표 등)

02

# 데이터 수집 및 전처리

## 개요서 작성

| 데이터 분류 | 데이터 이름 | 데이터 유형 | 데이터 수집 대상 | 데이터 수집 방식 | 데이터 수집 내용 | 데이터 수집 주기 | 데이터 수집 건수 |
|---|---|---|---|---|---|---|---|
| 주가 지수 | KOSPI 거래실적 (거래대금) | 정형 / 연속형 | KOSPI | 한국 거래소, csv 추출 | 월별 거래대금(정규매매, 시간외 매매, 온라인, HTS 등) | 매달 1회 | 2010/01 ~ 2020/07 |
| | KOSPI 거래실적 (거래량) | 정형 / 이산형 | KOSPI | 한국 거래소, csv 추출 | 월별 거래량(정규매매, 시간외 매매, 온라인, HTS 등) | 매달 1회 | 2010/01 ~ 2020/07 |
| | KOSPI 지수 추이 | 정형 / 연속형 | KOSPI | 한국 거래소, csv 추출 | 일별 주가(종가, 대비, 등락률, 시가, 고가, 저가, 거래량, 거래대금, 시가총액 등) | 매일 1회 | 2018/0818 ~ 2020/0819 |
| | KOSPI200 거래실적 (거래대금) | 정형 / 연속형 | KOSPI200 | 한국 거래소, csv 추출 | 월별 거래대금(정규매매, 시간외 매매, 온라인, HTS 등) | 매달 1회 | 2010/01 ~ 2020/07 |
| | KOSPI200 거래실적 (거래량) | 정형 / 이산형 | KOSPI200 | 한국 거래소, csv 추출 | 월별 거래량(정규매매, 시간외 매매, 온라인, HTS 등) | 매달 1회 | 2010/01 ~ 2020/07 |
| | KOSPI시리즈 서비스업 | 정형 / 연속형 | KOSPI 서비스업 | 한국 거래소, xls 추출 | 일별 주가(종가, 대비, 등락률, 시가, 고가, 저가, 거래량, 거래대금, 시가총액 등) | 매일 1회 | 2010/01 ~ 2020/07 |

02

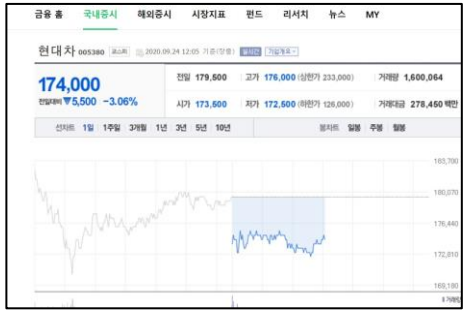
# 텍스트 데이터 수집

## 뉴스기사 분석

# “뉴스기사 분석”

**주식의 상하한가 키워드 학습**

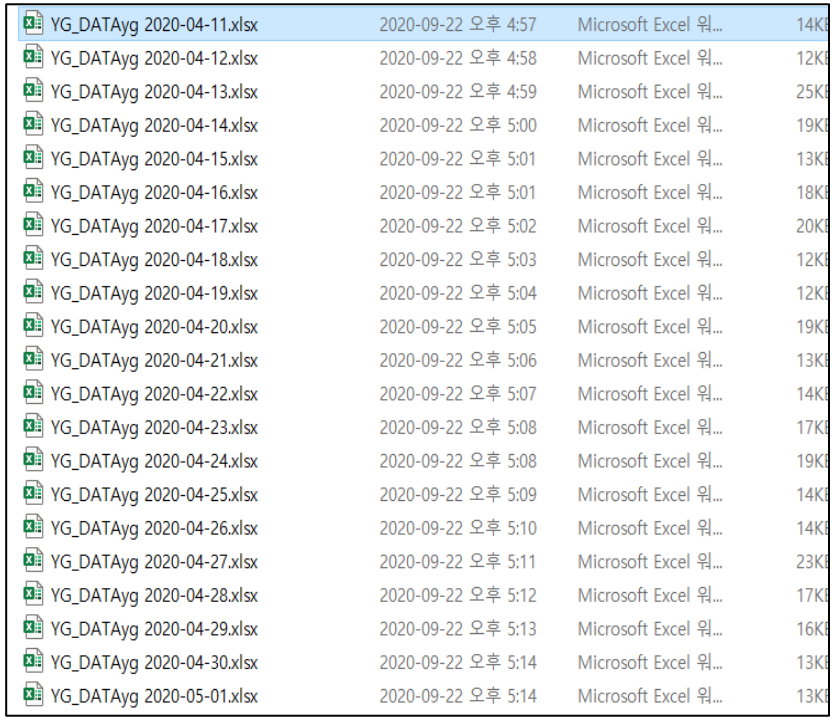
**주식의 예측과 관련한 논문 자료를 사전 리서치.**

02

# 텍스트 데이터 수집

## NAVER 뉴스 크롤링 (관련도 순 검색)



| | | | |
|---|---|---|---|
| YG_DATAyg 2020-04-11.xlsx | 2020-09-22 오후 4:57 | Microsoft Excel 워... | 14KB |
| YG_DATAyg 2020-04-12.xlsx | 2020-09-22 오후 4:58 | Microsoft Excel 워... | 12KB |
| YG_DATAyg 2020-04-13.xlsx | 2020-09-22 오후 4:59 | Microsoft Excel 워... | 25KB |
| YG_DATAyg 2020-04-14.xlsx | 2020-09-22 오후 5:00 | Microsoft Excel 워... | 19KB |
| YG_DATAyg 2020-04-15.xlsx | 2020-09-22 오후 5:01 | Microsoft Excel 워... | 13KB |
| YG_DATAyg 2020-04-16.xlsx | 2020-09-22 오후 5:01 | Microsoft Excel 워... | 18KB |
| YG_DATAyg 2020-04-17.xlsx | 2020-09-22 오후 5:02 | Microsoft Excel 워... | 20KB |
| YG_DATAyg 2020-04-18.xlsx | 2020-09-22 오후 5:03 | Microsoft Excel 워... | 12KB |
| YG_DATAyg 2020-04-19.xlsx | 2020-09-22 오후 5:04 | Microsoft Excel 워... | 12KB |
| YG_DATAyg 2020-04-20.xlsx | 2020-09-22 오후 5:05 | Microsoft Excel 워... | 19KB |
| YG_DATAyg 2020-04-21.xlsx | 2020-09-22 오후 5:06 | Microsoft Excel 워... | 13KB |
| YG_DATAyg 2020-04-22.xlsx | 2020-09-22 오후 5:07 | Microsoft Excel 워... | 14KB |
| YG_DATAyg 2020-04-23.xlsx | 2020-09-22 오후 5:08 | Microsoft Excel 워... | 17KB |
| YG_DATAyg 2020-04-24.xlsx | 2020-09-22 오후 5:08 | Microsoft Excel 워... | 19KB |
| YG_DATAyg 2020-04-25.xlsx | 2020-09-22 오후 5:09 | Microsoft Excel 워... | 14KB |
| YG_DATAyg 2020-04-26.xlsx | 2020-09-22 오후 5:10 | Microsoft Excel 워... | 14KB |
| YG_DATAyg 2020-04-27.xlsx | 2020-09-22 오후 5:11 | Microsoft Excel 워... | 23KB |
| YG_DATAyg 2020-04-28.xlsx | 2020-09-22 오후 5:12 | Microsoft Excel 워... | 17KB |
| YG_DATAyg 2020-04-29.xlsx | 2020-09-22 오후 5:13 | Microsoft Excel 워... | 16KB |
| YG_DATAyg 2020-04-30.xlsx | 2020-09-22 오후 5:14 | Microsoft Excel 워... | 13KB |
| YG_DATAyg 2020-05-01.xlsx | 2020-09-22 오후 5:14 | Microsoft Excel 워... | 13KB |

| | date | title | source | contents | link |
|---|---|---|---|---|---|
| 0 | 2020.04.11 | [단독] 'K팝 | TV리포트 | 'K팝스타3' | https://www.tvreport.c |
| 1 | 2020.04.11 | 블랙핑크, | 스포츠서울 | 블랙핑크는 | http://www.sportsseou |
| 2 | 2020.04.11 | 위너, 'Rem | MK스포츠 | YG엔터테인 | http://mksports.co.kr/v |
| 3 | 2020.04.11 | Inside the | The New Y | Shaquille O | https://www.nytimes.c |
| 4 | 2020.04.11 | "BTS·기생충 | 헤럴드경제 | YG엔터테인 | http://news.heraldcorp |
| 5 | 2020.04.11 | [연계소문] | 한국경제 | 2013년 SM | https://www.hankyung |
| 6 | 2020.04.11 | Quibi: 9 Sh | The New Y | This docur | https://www.nytimes.c |
| 7 | 2020.04.11 | 블랙핑크, | 아시아뉴스 | 블랙핑크 ( | http://www.anewsa.co |
| 8 | 2020.04.11 | [코스피 주 | 블록체인빌 | YG PLUS ( | http://www.fintechpos |
| 9 | 2020.04.11 | [단독] 서원 | TV리포트 | 서원진 프로 | https://www.tvreport.c |
| 10 | 2020.04.11 | 위너 'Rem | 뉴스엔 | YG엔터테인 | https://www.newsen.c |
| 11 | 2020.04.11 | 블랙핑크, | 뉴스엔 | 초청하기 대 | https://www.newsen.c |
| 12 | 2020.04.11 | [단독] 서원 | TV리포트 | 서원진 프로 | https://www.tvreport.c |
| 13 | 2020.04.11 | [연계소문] | 한국경제 | 2013년 SM | https://www.hankyung |
| 14 | 2020.04.11 | 위너, 'Rem | 비즈엔터 | YG엔터테인 | http://enter.etoday.co. |

# 시계열 데이터 수집

## 차트 데이터 및 투자지표 수집

※ 차트 데이터

| 지표명 | 표현 |
|---|---|
| 년/월/일 | date |
| 시가 | open |
| 고가 | high |
| 저가 | low |
| 종가 | close |
| 거래량 | volume |
| 거래대금 | value |
| 시가총액 | stock_value |
| 상장주식수 | stock_volume |

※ 투자지표

| 지표명 | 표현 |
|---|---|
| 주당순이익 | eps |
| 주가수익비율 | per |
| 주당 순자산가치 | bps |
| 주가순자산비율 | pbr |
| 주당배당금 | dividend_per_stock |
| 배당수익률 | dividend_yeild_ratio |
| 기관_매수량 | volume_inst_buy |
| 기관_매도량 | volume_inst_sell |
| 기관_순매수량 | volume_inst_pure_buy |
| 외국인_매수량 | volume_fore_buy |
| 외국인_매도량 | volume_fore_sell |
| 외국인_순매수량 | volume_fore_pure_buy |
| 기관_매수대금 | value_inst_buy |
| 기관_매도대금 | value_inst_sell |
| 기관_순매수대금 | value_inst_pure_buy |
| 외국인_매수대금 | value_fore_buy |
| 외국인_매도대금 | value_fore_sell |
| 외국인_순매수대금 | value_fore_pure_buy |

# 시계열 데이터 수집

## 주가지수 수집 및 보조지표 생성

※ 주가지수 및 주가지수 관련 투자지표

| 구분 | 지수명(지표명) | 표현 |
|---|---|---|
| 종합지수 | 코스닥 | kosdaq_xxx |
| | (종가) | kosdaq_close |
| | (시가) | kosdaq_open |
| | (고가) | kosdaq_high |
| | (저가) | kosdaq_low |
| | (거래량) | kosdaq_volume |
| | (거래대금) | kosdaq_value |
| | (시가총액) | kosdaq_stock_value |
| | (배당수익률) | kosdaq_dividend_yield_ratio |
| | (주가수익비율) | kosdaq_per |
| | (주가순자산비율) | kosdaq_pbr |
| 대표지수 | 코스닥 150 | kosdaq150_xxx, (이하 동일) |
| 섹터지수 | 코스닥 150 커뮤니케이션 | kosdaq150_comm_xxx, (이하 동일) |
| 산업별지수 | 오락,문화 | kosdaq_enter_xxx, (이하 동일) |
| 시가총액 규모별 지수 | 코스닥 대형주 | kosdaq_large_xxx, (이하 동일) |
| 소속부 지수 | 코스닥 우량기업부 | kosdaq_super_xxx, (이하 동일) |

※ 보조지표

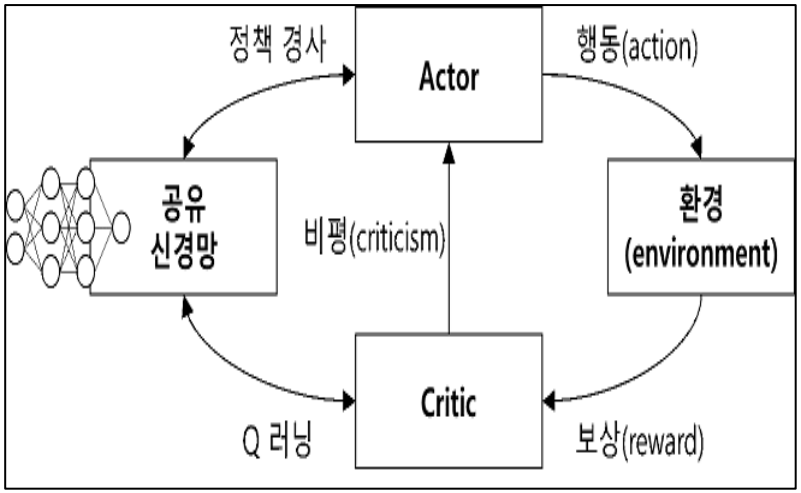
| 지표명 | 파라미터 값 | 표현 |
|---|---|---|
| 이평선 (이동평균, 지수평균, 가중평균) | 5, 10, 20, 30, 60, 120 | 종가, 거래량, 기관 및 외국인 거래량에 적용<br>ma_xxx, ema_xxx, wma_xxx |
| 볼린저밴드 | 20, 2 | ubb, mbb, lbb |
| MACD (이동평균수렴확산) | 12, 26 | macd, macdsignal9, macdhist |
| RSI (상대강도지수) | 14 | rsi |
| 스토캐스틱 | 5, 3<br>14, 5 | slowk, slowd, fastk, fastd<br>fastk_rsi, fastd_rsi |
| CCI (Commodity Channel Index) | 14 | cci |
| Williams'%R | 14 | willR |
| parabolic SAR | | sar |
| ADX (Average Directional Movement Index) | 14 | adx |
| plusDI (Plus Directional Indicator) | 14 | plus_di |
| plusDM (Plus Directional Movement) | 14 | plus_dm |
| ATR (Average True Range) | 14 | atr |
| OBV (On Balance Volume) | | obv |
| Variance | 5, 1 | var |
| Three Line Strike | | line_str |
| Three Black Crows | | blk_crw |
| Evening Star | | evn_star |
| Abandoned Baby | | abn_baby |

# Reinforcement Learning

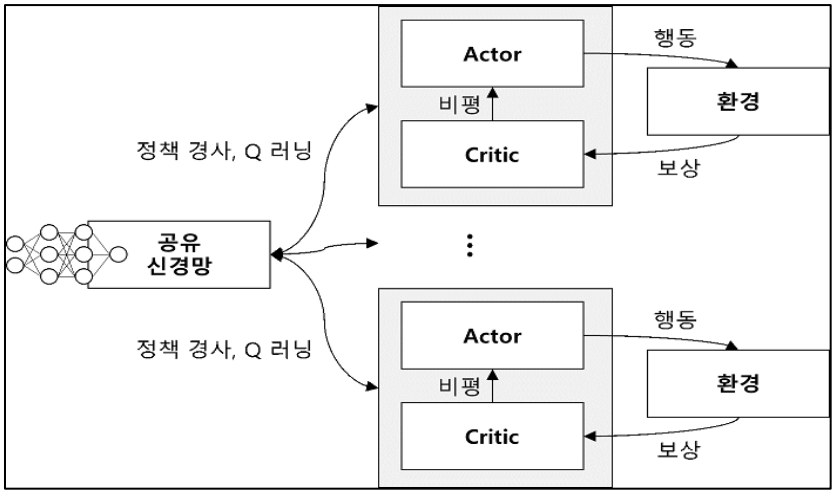
A2C, A3C, LSTM, Machine Learning 등

## 2) 알고리즘 절차

데이터 및 파라미터 입력 → Feature Selection Model (DT, RF, lgbm 회귀/분류 모델)
→ 정책/행동 신경망(LSTM) 가중치 학습 → A2C, A3C 강화학습 수행 → 출력값 도출




A2C Structure




A3C Structure

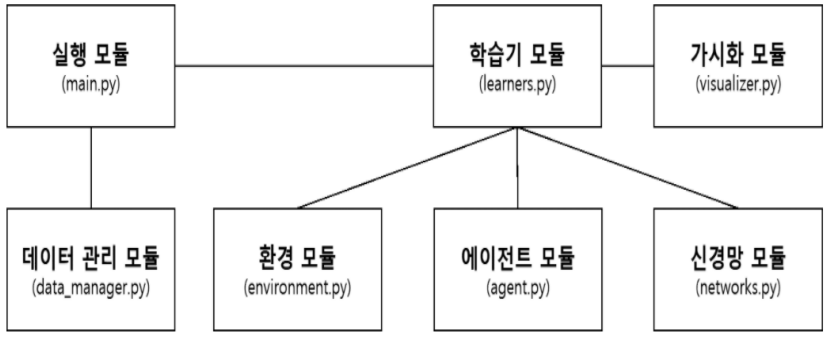
# Reinforcement Learning

Baseline Code : https://github.com/quantylab/rltrader

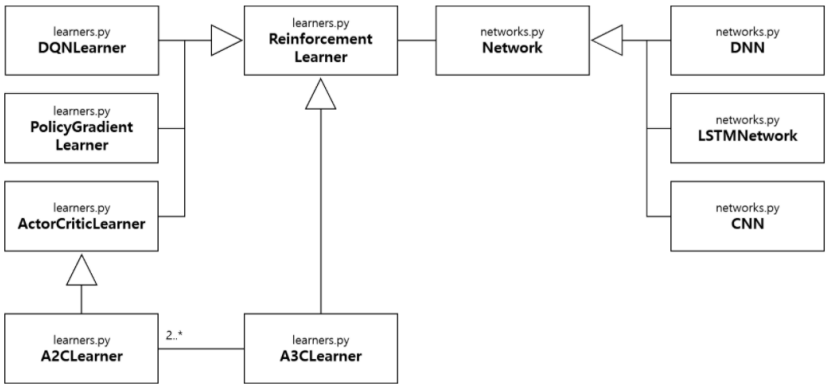
## 2) 알고리즘 구현 진행사항

- 데이터 전처리, 프로토타입 구현 및 In-Out Data 문서화 작업 완료
- Feature Selection Model, 알고리즘 커스터마이징 및 최적화 작업 진행 중
- 추후 시계열 예측 모델, 자연어 처리 모델과 앙상블 계획

모듈 구조




클래스 다이어그램



# Reinforcement Learning

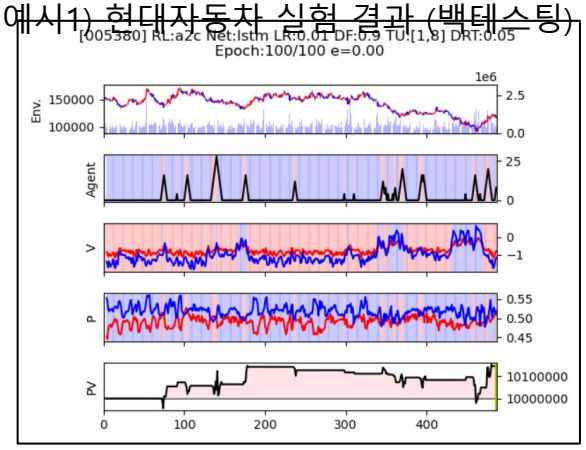
Baseline Code : https://github.com/quantylab/rltrader

## 3) 알고리즘 실행 결과

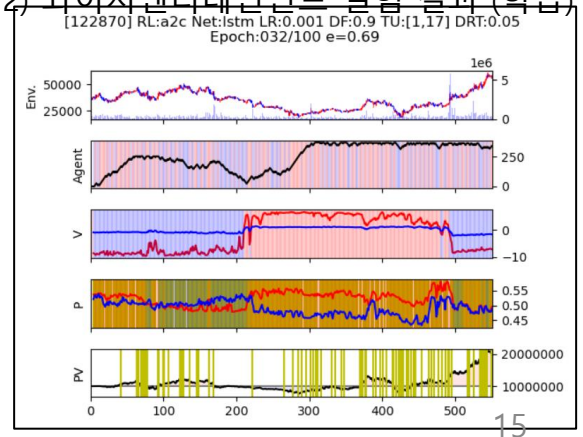
```
[122870][Epoch 093/100] Epsilon:0.0707 #Expl.:41/547 #Buy:112 #Sell:89 #Hold:346 #Stocks:297 PV:17,267,428 LC:72 Loss:67.768916 ET:11.435
[122870][Epoch 094/100] Epsilon:0.0606 #Expl.:30/547 #Buy:66 #Sell:40 #Hold:441 #Stocks:272 PV:15,795,025 LC:101 Loss:60.675441 ET:13.362
[122870][Epoch 095/100] Epsilon:0.0505 #Expl.:23/547 #Buy:67 #Sell:41 #Hold:439 #Stocks:235 PV:15,181,302 LC:99 Loss:57.752592 ET:13.7077
[122870][Epoch 096/100] Epsilon:0.0404 #Expl.:22/547 #Buy:72 #Sell:45 #Hold:430 #Stocks:225 PV:14,618,703 LC:100 Loss:53.745067 ET:13.283
[122870][Epoch 097/100] Epsilon:0.0303 #Expl.:17/547 #Buy:48 #Sell:51 #Hold:448 #Stocks:0 PV:10,830,778 LC:93 Loss:37.350014 ET:13.2926
[122870][Epoch 098/100] Epsilon:0.0202 #Expl.:9/547 #Buy:118 #Sell:123 #Hold:306 #Stocks:0 PV:9,906,256 LC:78 Loss:33.564916 ET:12.3166
[122870][Epoch 099/100] Epsilon:0.0101 #Expl.:4/547 #Buy:58 #Sell:67 #Hold:422 #Stocks:0 PV:15,059,591 LC:54 Loss:51.690092 ET:10.4394
[122870][Epoch 100/100] Epsilon:0.0000 #Expl.:0/547 #Buy:46 #Sell:53 #Hold:448 #Stocks:0 PV:15,454,129 LC:54 Loss:49.348068 ET:10.6300
[122870] Elapsed Time:1563.1269 Max PV:20,837,626 #Win:99
```

| | |
|---|---|
| 종목의 일봉 차트 | |
| 보유 주식 수 및 에이전트 행동 | 매수(빨간색), 매도(파란색) |
| 가치 신경망 출력 | 매수 가치(빨간색), 매도 가치(파란색) |
| 정책 신경망 출력 및 탐험 | 매수 확률(빨간색), 매도 확률(파란색), 탐험(노란색) |
| 포트폴리오 가치 및 학습 지점 | 수익(빨간색), 손실(파란색), 학습 지점(노란색) |

예시1) 현대자동차 실험 결과 (백테스팅)




예시2) 와이지엔터테인먼트 실험 결과 (학습)



# ARIMA

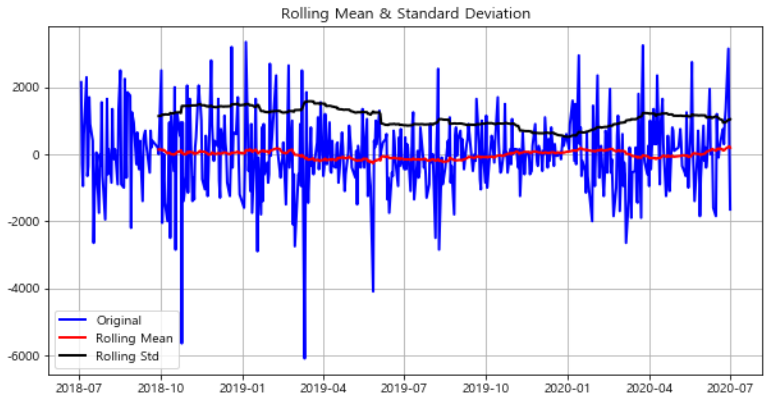
과거의 관측 값과 오차를 사용해서 현재의 시계열 값을 설명하는
ARMA(Auto-regressive Moving Average) 모델을 일반화 한 것

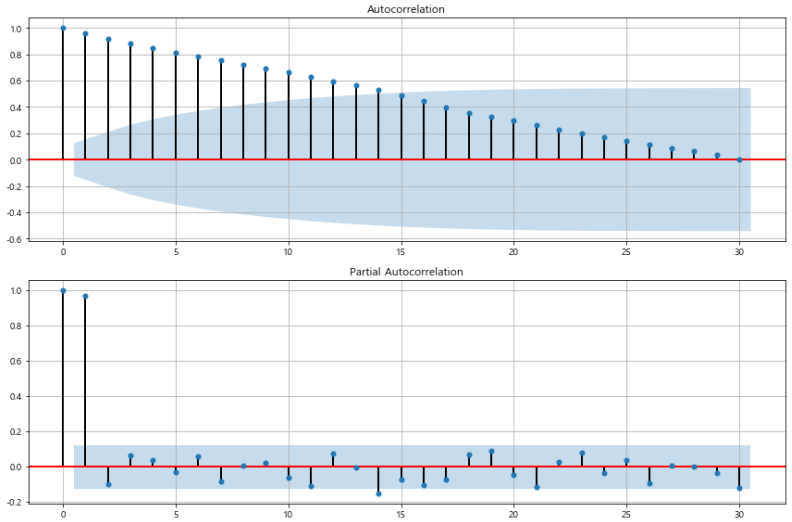
비교적 덜 안정(Non-Stationary)된 시계열 데이터에도 적용가능한 모델

차분(Difference)을 통해 데이터 whitening(분산과 평균을 일정하게)시키는 과정을 포함

**2)** ARIMA(p,d,q) -> (p,d,q) : 파라미터

최적의 파라미터를 찾기위해 ACF, PACF그래프, AIC,BIC 등 활용



# ARIMA

**3)** 여러 파라미터로부터 나온 값중, AIC가 최소인 파라미터 선정

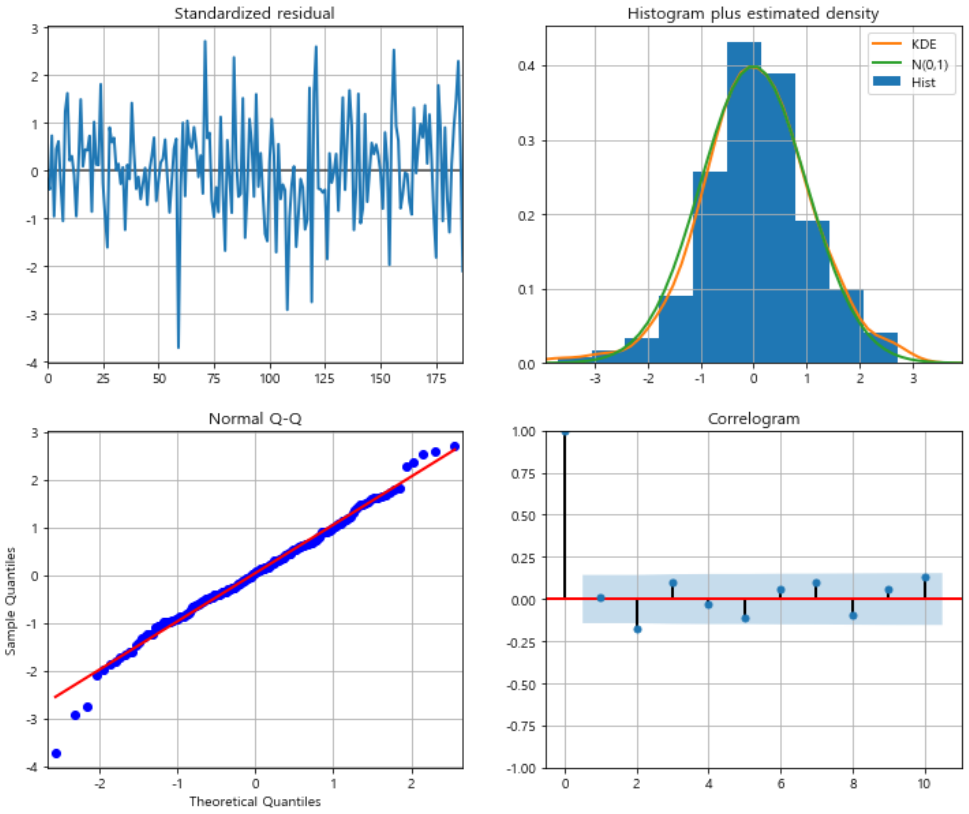
*AIC : 주어진 데이터 셋에 대한 통계 모델의 상대적인 품질을 평가한 지표 (가장 최소의 정보 손실 -> 낮을수록 좋은 모델)

```
Best model:  ARIMA(0,1,0)(0,1,1)[60]
Total fit time: 383.295 seconds
```

| | | | |
|---|---|---|---|
| **Dep. Variable:** | y | **No. Observations:** | 249 |
| **Model:** | SARIMAX(0, 1, 0)x(0, 1, [1], 60) | **Log Likelihood** | -1603.628 |
| **Date:** | Tue, 22 Sep 2020 | **AIC** | 3211.257 |
| **Time:** | 14:52:28 | **BIC** | 3217.730 |
| **Sample:** | 0 | **HQIC** | 3213.879 |
| | - 249 | | |
| **Covariance Type:** | opg | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **ma.S.L60** | -0.3072 | 0.029 | -10.464 | 0.000 | -0.365 | -0.250 |
| **sigma2** | 1.44e+06 | 1.47e+05 | 9.803 | 0.000 | 1.15e+06 | 1.73e+06 |

| | | | |
|---|---|---|---|
| **Ljung-Box (Q):** | 55.42 | **Jarque-Bera (JB):** | 7.45 |
| **Prob(Q):** | 0.05 | **Prob(JB):** | 0.02 |
| **Heteroskedasticity (H):** | 1.51 | **Skew:** | -0.20 |
| **Prob(H) (two-sided):** | 0.11 | **Kurtosis:** | 3.89 |



# ProPhet

페이스북이 만든 시계열 예측 라이브러리

**1)** 시간에 종속적이지 않고 Curve Fitting으로 문제를 해결

- 추세가 변경되는 지점(changing point)을 자동으로 감지해 추세를 예측 (감지하는 것을 사용자가 조절할 수도 있음)

- 휴일이나 모델에 반영하고 싶은 이벤트가 있으면 Dataframe을 생성해 반영할 수 있음

**2)** 직관적 파라미터를 통해 모형을 조정

- growth, changepoints, n_changepoints, daily_seasonality, yearly_seasonality etc

| ds | trend | cap | floor | yhat_lower | yhat_upper |
|---|---|---|---|---|---|
| 2018-07-02 | 35943.982486 | 43140.0 | 28760.0 | 34341.116875 | 37536.899705 |
| 2018-07-03 | 35942.705043 | 43140.0 | 28760.0 | 34688.470216 | 37903.913127 |
| 2018-07-04 | 38090.914926 | 45720.0 | 30480.0 | 36581.759105 | 39786.682550 |
| 2018-07-05 | 37789.643282 | 45360.0 | 30240.0 | 36531.727689 | 39596.614499 |
| 2018-07-06 | 36838.594149 | 44220.0 | 29480.0 | 35578.567626 | 38841.614998 |

# Komoran

KOrean MORphological ANalyzer

Java로 구현한 한국어 형태소 분석기.

# Komoran

KOrean MORphological ANalyzer

Java로 구현한 한국어 형태소 분석기.

**단어 사전** 단어경계점수 ⇨ 단어 확인

```
[('블랙핑크', 1135, 1.0),
 ('YG', 676, 0.76),
 ('공개', 443, 1.0),
 ('신곡', 381, 0.75),
 ('트레저', 335, 1.0),
 ('데뷔', 273, 1.0),
 ('컴백', 267, 1.0),
 ("That'", 264, 1.0),
 ('포스터', 258, 0.7142857142857143),
 ('돌파', 241, 1.0),
 ('MV', 233, 0.8461538461538461),
 ('뮤비', 211, 1.0),
 ('1위', 203, 1.0),
 ('신인', 200, 1.0),
 ('양현석', 197, 1.0),
 ('K팝', 170, 1.0),
 ('고메즈', 168, 1.0),
 ('세계', 152, 1.0),
 ('영상', 146, 0.8),
 ('개인', 142, 1.0),
 ("Cream'", 141, 1.0),
```

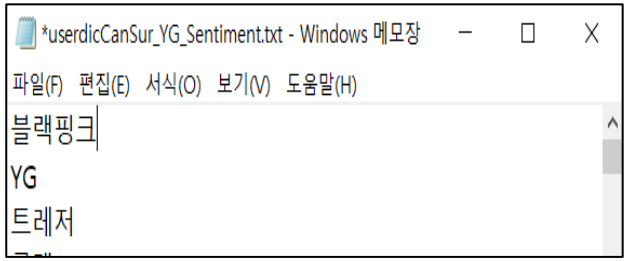
# Komoran

KOrean MORphological ANalyzer

Java로 구현한 한국어 형태소 분석기.

**빈도수 확인**

| | |
|---|---|
| 블랙핑크 | 2923 |
| 트레저 | 1242 |
| YG | 1072 |
| 공개 | 845 |
| 돌파 | 543 |
| 신곡 | 466 |
| 컴백 | 458 |
| 데뷔 | 439 |
| K팝 | 405 |
| How You Like That | 401 |




```
sent='YG승리 블랙핑크 MV가 100만뷰를 돌파'
print(f'코모란 형태소 분리 : {komoran.nouns(sent)}')
##사용자 사전 추가하기 & 사용자사전이 잘 되는지 테스트
komoran_userdic=Komoran(userdic='./userdicCanSur_YG_Sentiment.txt')
# 형태소분석 결과 보기 komoran.pos(sent)
print(f'사용자 정의사전 추가 : {komoran_userdic.nouns(sent)}')
```

```
코모란 형태소 분리 : ['승리', '블랙', '핑크', '돌파']
사용자 정의사전 추가 : ['YG', '승리', '블랙핑크', 'MV', '돌파']
```

Soynlp의 단어 경계 분석을 이용해
사용자 사전 구성 및 추가

04

# Sentimental Analysis

## Labeling

### 1) Data processing (0: 부정, 1: 중립, 2: 긍정)

| Title | Labeling | Contents | Link |
|---|---|---|---|
| '도박 혐의' 양현석, 첫 공판 9월 9일로 연기 | 0 | 도박 혐의를 받고 있는 양현석 전 YG | https://www.newsen.com/news_v |
| '독보적 행보' 트레저, 데뷔 앨범 초동 판매량 16만장 돌파 | 2 | YG 신인 트레저(TREASURE)가 음원 | http://www.wowtv.co.kr/NewsCen |
| '독특 헤어스타일' 악뮤 이찬혁 근황…"회사 차리면 이하이 왔으면" | 1 | 영상이 공개된 후 이찬혁과 YG엔터테 | http://sports.khan.co.kr/news/sk_ |
| '따상' BTS, 증시에서 '그래미상'거머쥘까 | 1 | 이는 전일 종가 기준 '3대 기획사'로 일 | https://www.etoday.co.kr/news/vie |
| '맛남의 광장' 블랙핑크 지수, "여기 나오고 싶어서 YG에 얘기해" 솔직 | 1 | 블랙핑크 지수가 출연을 결심한 이유 | https://news.mtn.co.kr/newscente |
| '무표정으로 걸어오는 양현석 전 YG 대표' [포토엔HD] | 0 | 해외 원정 도박 혐의 양현석 전 YG 엔 | https://www.newsen.com/news_v |
| '방탄소년단 테마주' 모두 하락, 엔터테인먼트3사는 YG만 올라 | 2 | 02%(800원) 떨어진 1만9100원에, SM | http://www.businesspost.co.kr/BP |

개수: 7,116

# Sentimental Analysis

## Labeling

### 2) Code

```
import re
def message_cleaning(docs):
    docs = [str(doc) for doc in docs] # series의 object를 str로 변경.
    #사진이나 이모티콘제거
    pattern1 = re.compile("Photo|Emoticon")
    docs = [pattern1.sub("", doc) for doc in docs]

    # 자음이나 모음만 존재하는 표현 제거, 예: ㅠㅠ, ㅋㅋㅋ
    pattern2 = re.compile("[ㄱ-ㅎ]*[ㅏ-ㅣ]*")
    docs = [pattern2.sub("", doc) for doc in docs]

    # http://로 시작하는 하이퍼링크 제거
    pattern3 = re.compile(r"₩b(https?:₩/₩/)?([₩w.]+){1,2}(₩.[₩w]{2,4}){1,2}(.*)")
    docs = [pattern3.sub("", doc) for doc in docs]

    # 특수문제 제거
    pattern4 = re.compile("[₩{₩}₩[₩]₩/?.,::|₩)*~`!^₩-_+<>@₩#$%&₩₩₩=₩(₩'₩"]")
    docs = [pattern4.sub("", doc) for doc in docs]

    return docs
```

```
def text_cleaning(docs): # 한글만 남기는 함수
    for doc in docs:
        doc = re.sub("[^ㄱ-ㅎㅏ-ㅣ가-힣 ]", "", doc)
    return docs

def text_tokenizing(corpus, tokenizer):
    token_corpus = []
    if tokenizer == "noun":
        for n in tqdm_notebook(range(len(corpus)), desc="Preprocessing"):
            token_text = komoran_userdic.nouns(corpus[n])
            token_text = [word for word in token_text if word not in SW and len(word) > 1]
            token_corpus.append(token_text)

    elif tokenized == "morph":
        for n in tqdm_notebook(range(len(corpus)), desc="Preprocessing"):
            token_text = komoran_userdic.nouns(corpus[n])
            token_text = [word for word in token_text if word not in SW and len(word) > 1]
            token_corpus.append(token_text)

    elif tokenizer == "word":
        for n in tqdm_notebook(range(len(corpus)), desc="Preprocessing"):
            token_text = corpus[n].split()
            token_text = [word for word in token_text if word not in SW and len(word) > 1]
            token_corpus.append(token_text)
```

# Sentimental Analysis

## Labeling

### 3) Prediction

```
Topic ID: 0
        트레저   0.059326134622097015
        블랙핑크          0.040209293365478516
        YG      0.03763829916715622
        공개    0.028189143165946007
        데뷔    0.017605207860469818
        신인    0.01743585616350174
        컴백    0.013787531293928623
        신곡    0.012346041388809681
        포스터  0.008313209749758244
        1위     0.00823524035513401
        글로벌  0.007483654655516148
        공식    0.00738660711795091 6
        BTS     0.007181905675679445
        돌파    0.00717254774644970 9
        BOY     0.007064988370984793
        차트    0.00702896201983094 2
        고메즈  0.006647278554737568
        앨범    0.006628982257097959 5
        셀레나  0.006385885644704103 5
        빅히트  0.006099694408476353
        사랑해  0.00586106814444065 1
        유튜브  0.005399358458817005
        K팝     0.005058085545897484
        만장    0.00477723404765129 1
        엔터테인먼트      0.004612428601831198
        확정    0.004486884921789169
        싱글    0.00442185997962951 7
        선주문  0.004404489416629076
        예고    0.004385652951896191
        MV      0.00438411440700292 6
```

```
Topic ID: 1
        블랙핑크          0.030400115996599197
        YG      0.028986113145947456
        양현석  0.012726676650345325
        이하이  0.010981475934386253
        AOMG    0.009560856968164444
        돌파    0.009084714576601982
        도박    0.009079246781766415
        협의    0.008512331172823906
        트레저  0.008481076918542385
        오늘    0.008245328441262245
        K팝     0.00816620048135519
        공식    0.007991990074515343
        원정    0.007586676627397537
        How You Like That       0.007528911344707012
        한서희  0.006375005003064871
        아이콘  0.006311759818345308
        기록    0.00607891334220767
        마약    0.005731870885938406
        음원    0.005613033194094896
        공개    0.005137780681252479 6
        MV      0.005119821522384882
        인정    0.005078026559203863
        구준회  0.00487259728834033
        김진환  0.004834835417568684
        데뷔    0.004470687359571457
        뮤비    0.00445277802646160 1
        출석    0.004375442396849394
        신곡    0.004362788051366806
        계약    0.004265055526047945
        기대    0.004197795875370502 5


-6.9058221825705495
```

# THANK YOU

Turnaround