1주차 ( 9/9 ~ )

**<주제 선정>**

고려사항 : 실현가능성, 구체성

1) 네이버 금융

* 사이트 규모가 적어보임
* 테이블이 명시적임

2) airbnb, 야놀자

* 크롤링에 적대적일것같음
* 특정지역의 모든 집을 다 긁어온다는게 쉽지는 않을것같음
* 동적으로 변하는 사이트는 난이도가 높다

3) 네이버 쇼핑, 쿠팡, 11번가

* 긁어올 정보가 명확히 정의될 필요가 있을것같음.
* 각 사이트에서 주는 정보가 다를텐데 차라리 사이트를 하나로 정하는게 나을것같음.
* 물건의 범주를 정해야될거같음 (ex. 신발,의자..)
* 상품의 특성에도 명확한 정의가 필요할거같음. (ex, 썸네일 이미지, 본문의 설명일수도있는데 텍스트로 안 올라와 있고 jpg로 올라와있음, 데이터로 피처 추출할 때 유사성이 잘 나올지 모르겠고, 이미지를 사용하면 deep learning 사용해야 할 거 같은데 한 학기 내에 하기에 무리가 있어 보임)

**< 보고서의 가이드라인 >**

- 섹션 5개

` 이 주제 선정 이유

` 데이터 수집과정 (크롤링과정)

` 적용한 방법 (eda 해보니까 데이터셋 어떻게 만들어졌는지, 어떤 모델사용할지)

' 결과+해석

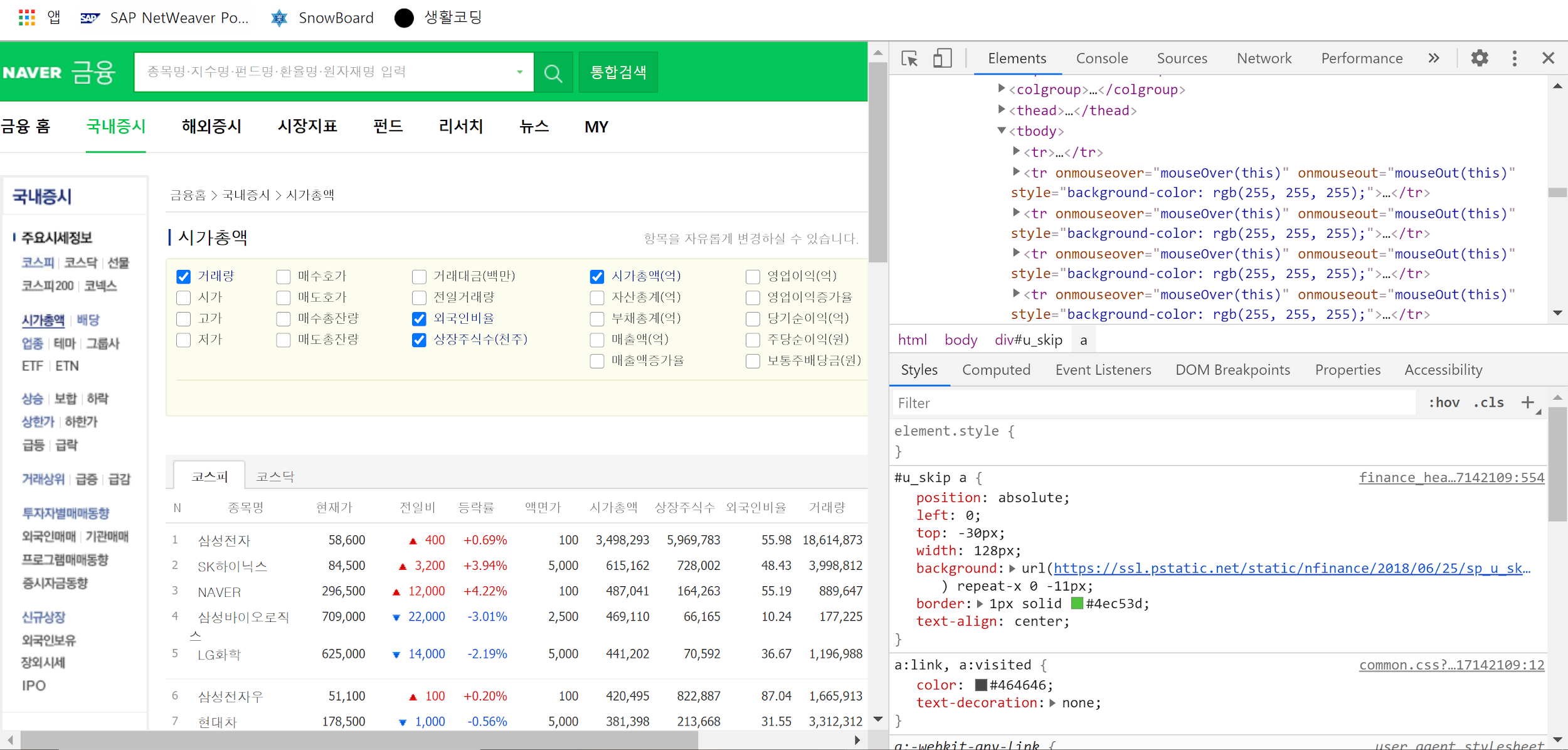
' 결론

2주차 ( 9/16 ~ )

**<참고 사이트>**

* 네이버 금융 : <https://finance.naver.com/>
* <https://namkisec.tistory.com/entry/1%EC%8B%9C%EA%B0%84%EC%95%88%EC%97%90-%EA%B8%88%EC%9C%B5%EC%A3%BC%EA%B0%80-%ED%81%AC%EB%A1%A4%EB%A7%81-%EC%BD%94%EB%94%A910%EB%B6%84-%ED%8F%AC%EC%8A%A4%ED%8C%85-45%EB%B6%84>
* <https://infinite-learning.tistory.com/entry/%ED%8C%8C%EC%9D%B4%EC%8D%AC-%ED%81%AC%EB%A1%A4%EB%A7%81-%EB%84%A4%EC%9D%B4%EB%B2%84-%EA%B8%88%EC%9C%B5-%EC%BD%94%EC%8A%A4%ED%94%BC-10%EB%85%84%EC%B9%98-%EC%A2%85%EA%B0%80-%EC%A0%95%EB%B3%B4-%EC%88%98%EC%A7%91%EC%9D%BC%EB%B4%89>

**<분석 방향>**



* 2019년의 주식 정보를 학습자료로 사용하여 2020년을 예측하고자 함.
* 주식의 시가, 종가, 고가, 저가 등을 크롤링하여 초기 자본금을 10,000,000원으로 설정해 주식투자 시뮬레이션을 하고자함.
* 에포크를 1000으로 지정해두고, 몇번의 매수와 매도를 하였을지 그리고 관망의 횟수를 파악하고 최종 포트폴리오 가치를 파악해보기.
* 시가총액이 비교적 높은 기업과 낮은 기업들에 똑같은 금액을 투자했을 때와 차이를 두고 투자했을 때의 시뮬레이션 결과를 살펴본다.

**<역할분담>**

* 주제 제안자 -> 안희재
* 회의록 작성 -> 이혜원
* 크롤링 -> 이혜원 김서영
* eda -> 김서영
* modeling -> 안희재 이혜원
* 결과 + 해석 -> 안희재 김서영
* 보고서 작성 -> 이혜원 김서영 안희재

3주차 ( 9/23 ~ )

**<역할분담>**

* 주제 제안자 : 안희재
* 회의록 작성 : 이혜원
* 크롤링 : 전원
* eda : 김서영
* modeling : 전원
* 발표자 : 전원
* 보고서 총 정리 : 안희재
* 포스터 총 정리 : 김서영
* 코드 총 정리 : 이혜원

**<기업 선정>**

* 삼성전자 (005930)
* SK하이닉스 (000660)
* LG화학 (051910)

**<크롤링 할 변수들>**

* 날짜
* 시가
* 종가
* 고가
* 저가
* 전일비
* 거래량
* ...

**<모델 선정>**

* LSTM

4주차 ( 9/30 ~ )

**<크롤링>**

크롤링진행 - 주피터 확인



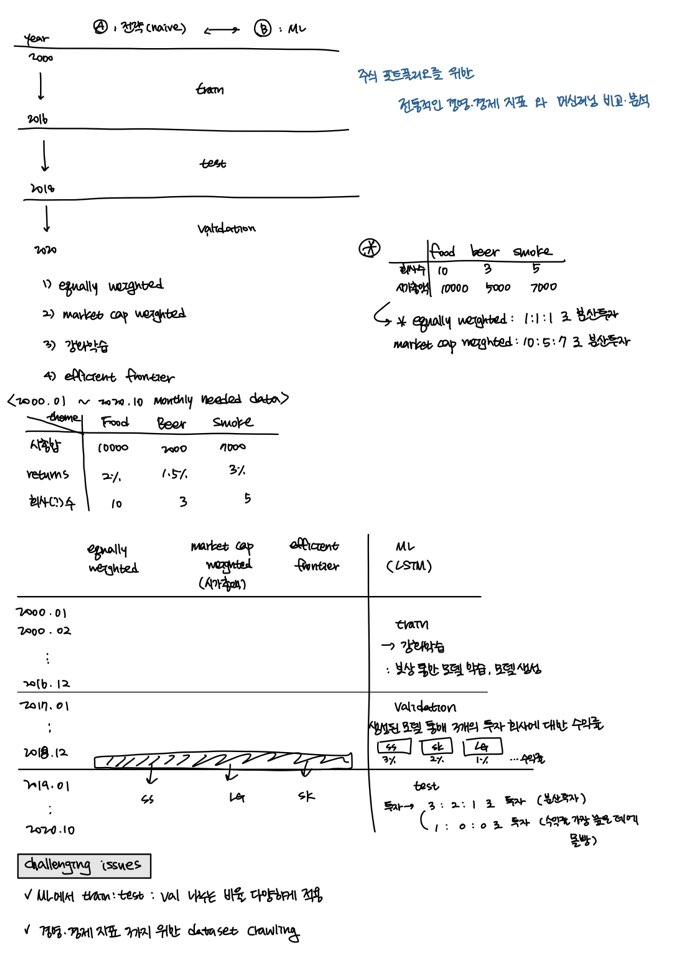
5주차 ( 10/7 ~ )

* 포트폴리오에 필요한 정보 찾아보기
* 금융분석과 관려된 책들 찾아보면서 주식 기본 지식 이해하기
* 트레이딩 전략에 따라 포트폴리오를 3가지 정도 만들자.

6주차 ( 10/14 ~ )

* 포트폴리오 구성 각자 계속 구성해보기
* 월요일에 교수님 면담 후 포스터 & 보고서 작업 시작
* 테마별로 크롤링 새로 하기

7주차 ( 10/19~ )



8주차 ( 10/28~ )

**<EDA>**

* 새로 15개 크롤링 새로하기
* 크롤링해온 데이터 셋 비어있는 부분 NULL값 처리
* 2000년 ~ 2011년 training-set
* 2012년 ~ 2016년 validation-set
* 2017년 ~ 2020년 test-set
* 2017년꺼부터 효율적투자선+샤프지수 그리기
* 서영이 파이썬 downgrade 계속 시도,,

**<피드백>**

* 피드백 받은 거 읽어보면서 보고서나 포스터에 추가할 내용이 뭐가 있을지 생각하기
* 참고란에 경제용어 설명란을 추가하는 것이 좋을 것 같음
* 강화학습에 대한 설명 추가적으로 하는 것이 좋을 듯
* 코로나로 우리의 데이터에 영향을 주는 부분이 있는지 생각해봐야할듯

9,10주차( 11/4~)

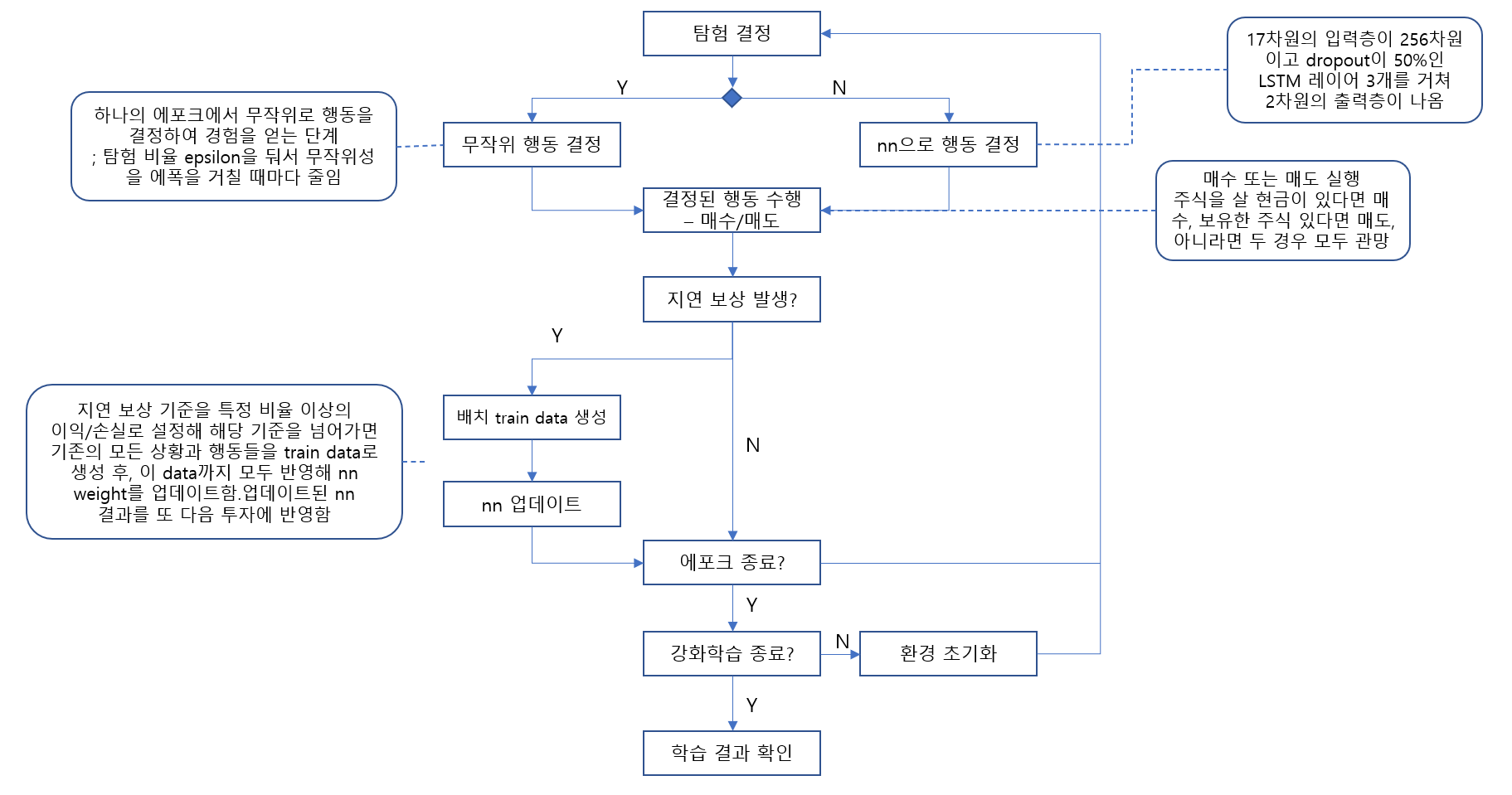
**<방향>**

* 피드백-코로나관련 : 따로 언급하지 않기로 함.
* 강화학습 시도는 이번까지만
* 실패 시 다른 방법으로 구현하는 것이 바람직함
* 다른 지표들도 만들어보기
* ex) 볼린저밴드, 위험균형, ...
* 볼린저밴드 - 기간 설정 논의할 필요 있어보임
* 너무 보통 경제흐름을 10일(2주)로 많이하는데 우리는 데이터가 기간이 길기 때문에 15일(3주)로 해보는 것은 어떠할지,,

**<강화학습 진행 단계>**

각 날의 시종고가, 거래량으로 이루어진 데이터는 주가와 거래량의 단위가 다르므로 학습 데이터를 비율값으로 범위를 같게 만들어 종목에 의존하지 않도록 만든다.

**<강화학습이 진행되는 과정 도식표>**



학습 초기에는 탐험 비율을 크게 하여 무작위로 투자를 많이 진행하고, 이러한 탐험을 통해 특정 상황에서 좋은 행동이 무엇인지 행동을 결정하기 위한 경험을 쌓는다. 이 탐험 비율을 에포크를 거쳐 반복 학습을 통한 경험이 쌓일수록 줄어들면서 신경망 모델이 포트폴리오 가치를 높이는 방향으로 갱신된다. 무작위로 매수/매도를 결정하는 것과 반대로 정책 신경망을 통한 행동 결정은 train data를 input으로 넣어 3개의 LSTM 레이어를 거치고 매수, 매도에 대한 확률을 출력한다. 이 매수, 매도의 확률을 비교해 행동을 결정한다. 매수/매도를 한 후 발생한 이익과 손실에 대해 지연 보상 기준을 정해 이 기준보다 높은 이익이나 손실이 발생한 경우 이 전까지의 상황을 train data로 함께 고려하여 모델을 업데이트한다. 업데이트된 결과가 다음 투자에 또 반영되어 모델이 계속 포트폴리오 가치를 최대화하는 방향으로 학습이 진행된다. 포트폴리오 가치는 주식잔고\*현재주가+현금잔고로 계산하고 현재 투자자의 주식 잔고가 많을 때 주가가 상승하면 포트폴리오 가치가 크게 증가할 수 있으므로 주가가 상승될 것 같을 때 매수를 해 주식 잔고를 늘리고 반대의 경우 매도한다. 학습을 위해 필요한 역할과 과정을 분담해 투자자의 상태 정보, 투자할 차트 데이터, 투자자의 행동을 결정하는 부분을 클래스로 만들어 학습을 진행한다.

**< 계획 >**

* 강화학습 무리라고 판단 -> 연구 방향 바꾸기
* 크롤링한 데이터를 사용하여 2020년의 주식을 예측
* 미리 크롤링 한 2020년의 데이터와 비교
* 미래의 데이터로 포트폴리오 구성하기

11주차 ( 11/12 ~ )

**< 교수님 피드백 >**

dataset – 2015년부터.

2000년부터 2020년까지는 있음. (앞부분 데이터는 null이 많음.)

- 예측 태스크 있긴 있으니까 training set은 많을 수 있는 방향으로.

- [a] 2015년-2019의 데이터셋으로 2020 예측 -> 포트폴리오

- [b] 2000년-2019의 데이터셋으로 2020 예측 -> 포트폴리오

사람들이 궁금할만한 것

- [a] vs. [b] 2020년 예측 성능

- [a] vs. [b] 포트폴리오 성적

예측할 때 모델 기법들

- Deep 계열 - (1) LSTM만 했다 하면 털림

- (제대로 할거면 hyperparemter optimizaton 빡세게 해야 할 듯)

- Tradition ML 계열 - (2) 안전하게 가려면 다른 traditonal machine learning도 집어넣을 것. 어렵게 가지 말고 h2o 정도로 가볍게.

- 간단 계열 - (3) moving average 같은 간단한 기법들

action points (-11/30 4:00p)

- 위에 써있는거 한번은 건드려보겠다.

- 기존 포트폴리오 기법

- 예측 + 기존 포트폴리오 기법

- (1)-(3) 중에 두 가지 정도

**< 데이터 셋 나누기 >**

* train : 2013.1.1 ~ 2016.12.31
* validation : 2017.1.1 ~ 2018.12.31
* test : 2019.1.1 ~ 2019.12.31

**< 방향 >**

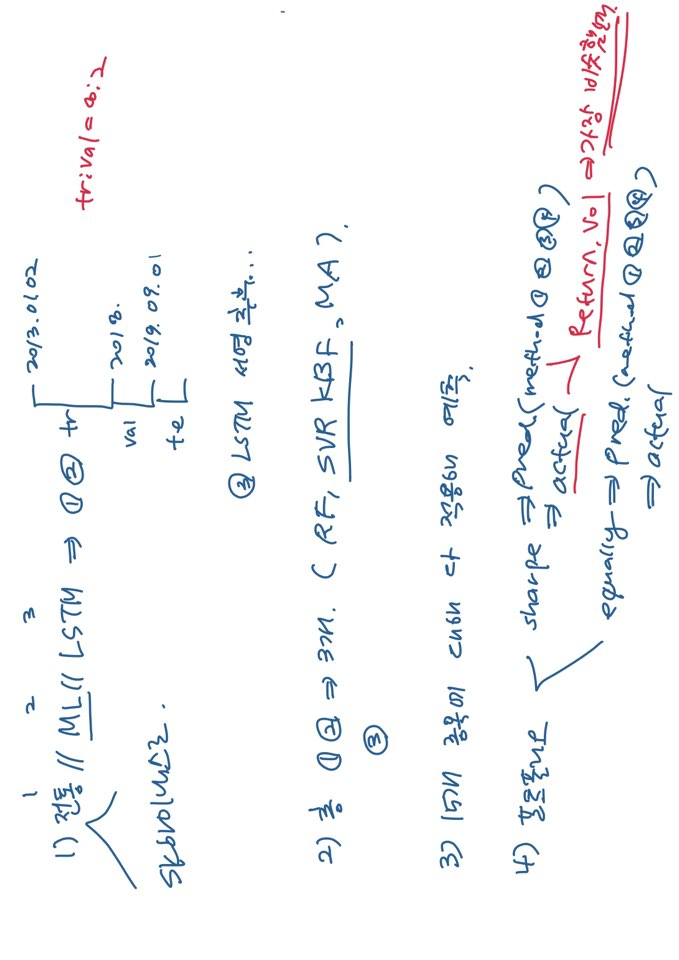
* sk하이닉스로 다양한 모델 평가해보기 ( traditional, ml, dl, h2o )
* ex) LM, MA, KNN, ... , h2o, LSTM
* 평가지표 3가지정도로 고려하여 제일 좋은 모델 2~3개 선정
* ex) RMSE, MAPE,
* 선택된 모델로 다른 종목(2-3개)에 적용하여 최종 모델 선택
* 15개의 종목의 종가를 예측하여 포트폴리오를 만들기.
* 포트폴리오에는 최소분산투자, …
* 1년 예측 말고 분기 예측은 어떨까?

**< 다음 미팅까지 할 일 >**

* 혜원 : MA,LM, KNN, … 등으로 주가 예측
* 희재 : h2o으로 주가 예측 -> prophet
* 서영 : LSTM으로 주가 예측

11/30 ~

* 머신러닝 & LSTM 성능 발전시키기
* 모델 3-4개 선정
* 15개 종목 test error 구해보기



12/7 ~

<피드백>

1, 머신러닝으로 피팅이 안된이유적기

- iid가 아니여서

2, 모델 추가(?)

- 최신 데이터에서 trend만 뽑는것

-> 최근것만 가지고, linear로 해보기 ( 기존 linear에서 기간만 다시 하면될듯)

-> train 기간 다른게 아님, 81일 이전에 weight를 0으로 주면되는것.

3, 데이터를 설명하는게 중요함 ( EDA 부분이 적어도됨 )

-> 대신, 맥락을 이해시키는데 중점을 두기

(=> text로만 쓰면 이해하기엔 무리가 있을수도 있음 )

->"액면분할, 거래정지" 한거 시계열데이터만 그림 넣고 이게 무엇인지 설명하기

4, 분량

-> 포트폴리오 설명은 1~2장으로 해야됨.

-> 방법론보단 사람들 모르는 분야쪽을 설명하는게 맞음

