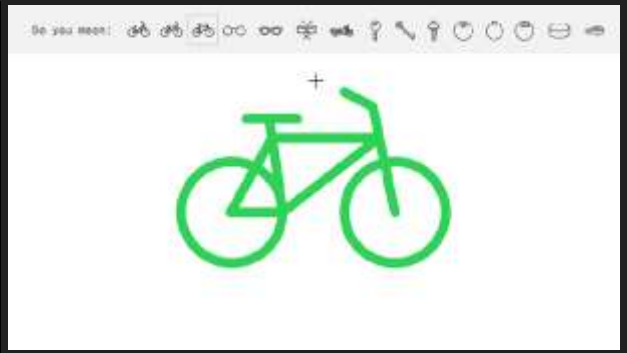
ML+주식 삽질기

phase 1

2017/07 신호철

요즘 딥러닝이 대세라던데~?

https://www.youtube.com/watch?v=V1eYniJ0Rnk



부푼 꿈을 안고, 일단 컴퓨터 구입

+

상품정보	판매가격	수량	주문금액
[ASUS] PRIME Z270-A STCOM (인텔Z270/ATX)	269,000원	1개	269,000원
[삼성전자] 삼성 DDR4 16GB PC4-17000	133,000원	1개	133,000원
[ADATA] SP920 New Series 256GB (가이드포함)[A/S 무상 5년] MLC	120,000원	1개	120,000원
[SEAGATE] 바라쿠다 2TB ST2000DM006 (3.5HDD/SATA3/7200rpm/64M)	88,000원	1개	88,000원
[GIGABYTE] GeForce GTX1070 UDV D5 8GB 윈드포스 ◆[구매 후 별도의 신청 시 2017 최고 기대작 게임 증정]	588,000원	1개	588,000원
[DEEPCOOL] GAMMAXX 400 For Intel BRAVOTEC	26,000원	1개	26,000원
[마이크로닉스] Classic II 600W +12V Single Rail 85+ (ATX/600W)	55,000원	1개	55,000원
[스카이디지탈] 유선 광마우스, NMOUSE C32 LED [블랙/USB]	4,800원	1개	4,800원
[마이크로소프트] Windows 10 Home K [한글/처음사용자용/패키지(FPP)/USB/32,64bit/멀티 랭귀지]	162,850원	1개	162,850원
[브리츠] [2채널스피커] BR-1000A Cuve 2 [블랙]	18,900원	1개	18,900원
[삼성전자] 삼성커브드모니터 C27F390F [무결점]	223,000원	1개	223,000원
[BRAVOTEC] 트레저 X6 630T 타이탄 글래스 블랙 (빅타워/ODD장착불가)	135,000원	1개	135,000원

환경설정

- 윈도우 사용
  - 증권사 연동때문.
- 아나콘다 설치
  - 헐... 윈도우 필수 뿐만 아니라, 증권사 연동 때문에 32비트 버전을 설치해야한다고 함.
    - https://wikidocs.net/2825
    - 이 당시는 python 3.6을 깔면 tensorflow-gpu가 설치 안 되므로 설치.
      - 4.2.0도 tensorflow-gpu는 설치 안됨…
      - 64비트 윈도우의 경우, anaconda x86 버전을 설치해도 t 다 설치 안됨.

가상화

- 가상화를 할 수 밖에 없으니...
  - 가상환경으로 32비트와 64비트를 나눠서 써야함.
    (32비트 환경에서만 돌리면 tensorflow가 제대로 동작 할리가...)
    - http://stackoverflow.com/questions/33709391/using-multiple-python-engines-32bit-64bit-and-2-7-3-5
      - >set CONDA_FORCE_32BIT=1
      - >conda create -n py35_32 python=3.5
      - >set CONDA_FORCE_32BIT=
      - >conda create -n py35_64 python=3.5
    - python 3.5 32bit 환경
      - >set CONDA_FORCE_32BIT=1
      - >activate py35_32

- 이건 마치...

tensorflow GPU 설치

- GTX1070을 샀는데 그냥 놀릴 수는 없지
- cuda 설치
  - http://jaejunyoo.blogspot.com/2017/02/start-tensorflow-gpu-window-10.html
  - https://www.tensorflow.org/install/install_windows
    - cuda 8.0.61_win10 설치
    - cudnn-8.0-windows10-x64-v6.0 설치
    - python 3.5 64bit 환경에서 설치
    - hello world 동작 체크

IDE

- pyCharm 설치
  - 이유: 그냥 이 기회에 pyCharm을 한번 써 보려고...
  - https://www.jetbrains.com/pycharm/download/#section=windows
  - 가상화 환경설정
    - https://www.jetbrains.com/help/pycharm/2016.3/conda-support-creating-conda-environment.html

증권 모듈

- 키움증권 설치 (2017/3/20)
  - 참고: https://wikidocs.net/4231
  - Open api 신청
  - Open api+ 모듈 설치
  - Koa studio 설치
  - 모의 투자 가입
    - 최대 3개월동안 모의계좌 사용 가능

목표 설정

- 일단위 데이터를 기반으로 분석

- 다음날 오를 것 같은 종목을 예측하는 것이 1차 목표임.

데이터 수집

- 일데이터 수집
- 본래는 pandas datareader를 활용하려고 했음.
  - 이렇게 잘 되면 좋을텐데… 보통 웹에 떠 돌아다니는 예제들은 pandas datareader를 사용함.
    - https://github.com/dspshin/stock-analyzer/blob/master/noti.py 이건 뭐, 시작부터 난관이야.

- 문제점
  - 코스닥 지원 안됨.
  - 지연이 심함. 이틀 전 데이터가 최신 데이터임. --;
  - 결과적으로 증권사 api를 사용할 수 밖에 없음

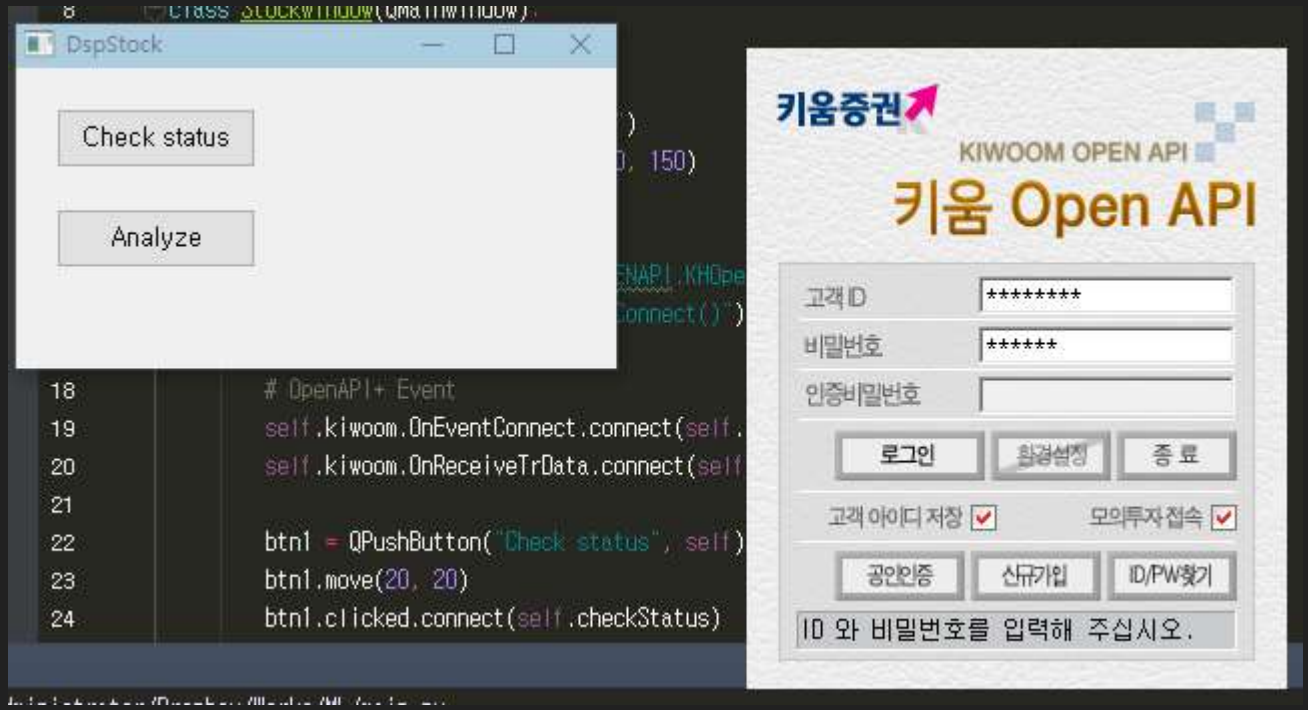
키움 로그인

main.py

StockWindow __init__()

```python
import sys
from PyQt5.QtWidgets import *
from PyQt5.QtGui import *
from PyQt5.QAxContainer import *

STOCK_COLUMNS = ["등락률", "거래량", "종가", "외인순매수", "기관순매수", "개인순매수"]

class StockWindow(QMainWindow):
    def __init__(self):
        super().__init__()
        self.setWindowTitle("DspStock")
        self.setGeometry(300, 300, 300, 150)

        # login
        self.kiwoom = QAxWidget("KHOPENAPI.KHOpenAPICtrl.1")
        self.kiwoom.dynamicCall("CommConnect()")

        # OpenAPI+ Event
        self.kiwoom.OnEventConnect.connect(self.eventConnect)
        self.kiwoom.OnReceiveTrData.connect(self.OnReceiveTrData)

```

로그인 실행화면


DspStock
Check status
Analyze
키움증권
KIWOOM OPEN API
키움 Open API
고객ID
****
비밀번호
******
인증비밀번호
로그인
환경설정
종 료
고객 아이디 저장
모의투자 접속
공인인증
신규가입
ID/PW찾기
ID 와 비밀번호를 입력해 주십시오.
18 # OpenAPI+ Event
19 self.kiwoom.OnEventConnect.connect(self.
20 self.kiwoom.OnReceiveTrData.connect(self
21
22 btn1 = QPushButton("Check status", self)
23 btn1.move(20, 20)
24 btn1.clicked.connect(self.checkStatus)

종목 정보 가져 오기

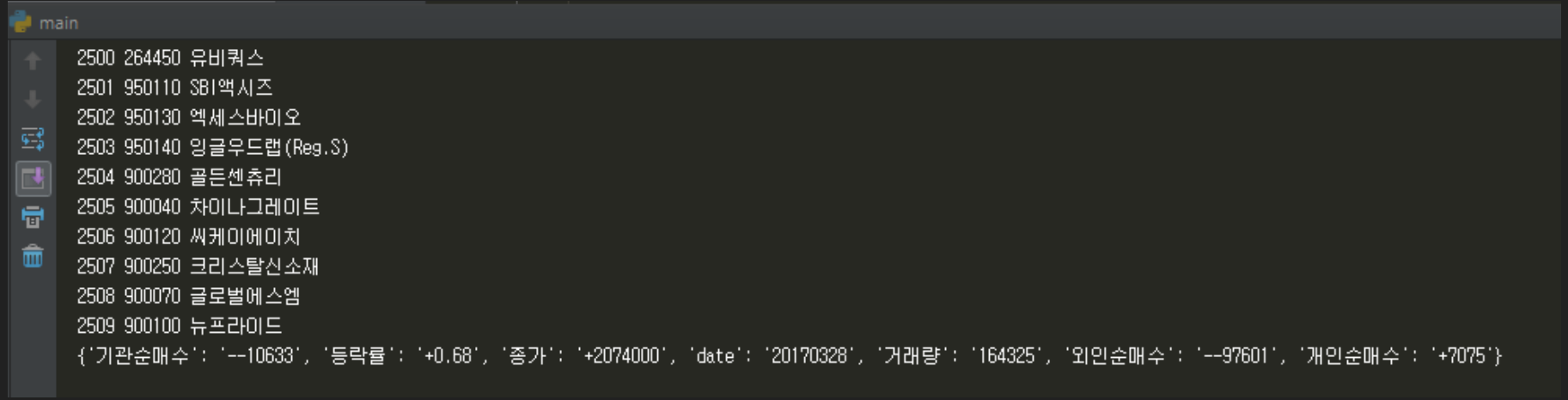
```
44      def analyze(self):
45          kospi_codes = self.kiwoom.dynamicCall("GetCodeListByMarket(QString)", ["0"]).split(';')
46
47          kosdaq_codes = self.kiwoom.dynamicCall("GetCodeListByMarket(QString)", ["10"]).split(';')
48
49          all_codes = kospi_codes + kosdaq_codes;
50
51          #kospi_code_name_list = []
52          i=0
53          for c in all_codes:
54              if c:
55                  name = self.kiwoom.dynamicCall("GetMasterCodeName(QString)", [c])
56                  #kospi_code_name_list.append(x + " : " + name)
57                  print(i, c, name)
58                  i+=1
```

일별주가정보 가져오기

```
60          # 일별주가정보 조회 테스트
61          self.set("종목코드", "005930")
62          self.set("조회일자", "20170328")
63          self.set('표시구분', '0')
64
65          self.kiwoom.dynamicCall("CommRqData(QString, QString, int, QString)", "005930_daily", "opt10086", 0, "0000")
66
67
68      def OnReceiveTrData(self, ScrNo, RQName, TrCode, RecordName, PrevNext, DataLength, ErrorCode, Message, SplmMsg):
69          if RQName == "005930_daily":
70              date = self.kiwoom.dynamicCall("CommGetData(QString, QString, QString, int, QString)", TrCode, "",
71                                             RQName, 0, "날짜").strip()
72              data = {
73                  'date':date
74              }
75              for col in STOCK_COLUMNS:
76                  data[col] = self.kiwoom.dynamicCall("CommGetData(QString, QString, QString, int, QString)", TrCode, "",
77                                                 RQName, 0, col).strip()
78              print(data)
```

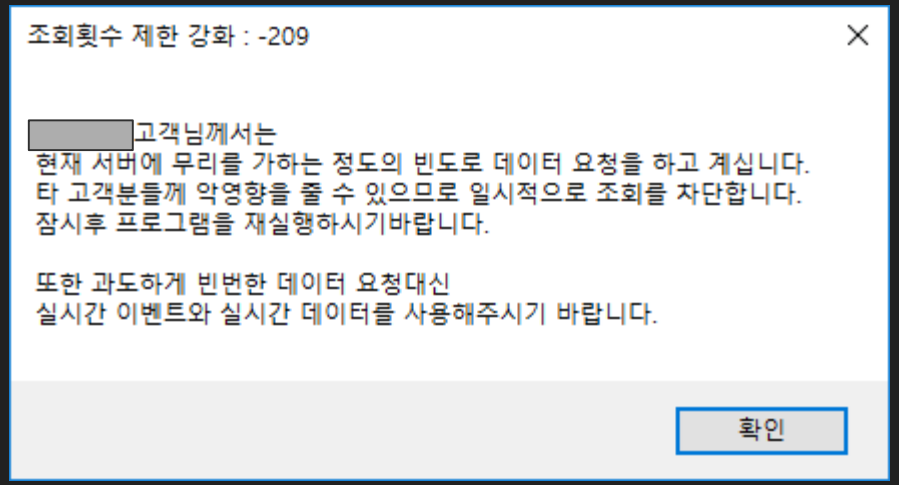
```
def set(self, *args):
    return self.kiwoom.dynamicCall("SetInputValue(QString, QString)", *args)

STOCK_COLUMNS = ["등락률", "거래량", "종가", "외인순매수", "기관순매수", "개인순매수"]
```

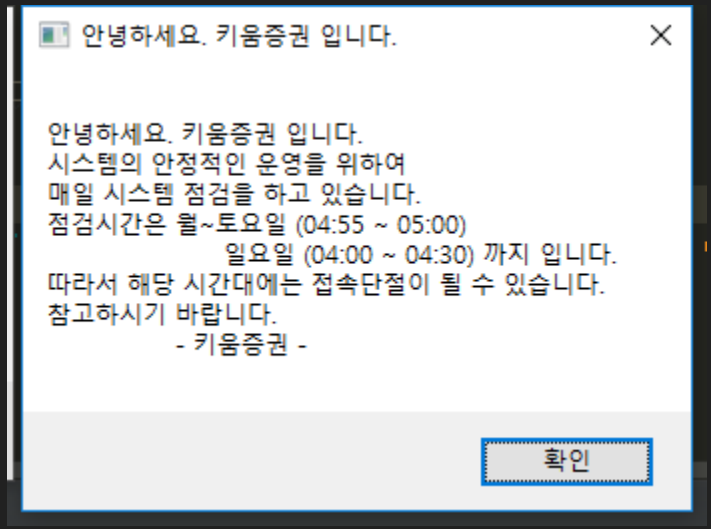
종목 정보 및 일별주가정보 실행화면



일별 주가정보 수집 및 DB에 저장...

- 안전하게 1초 delay로 호출했는데...





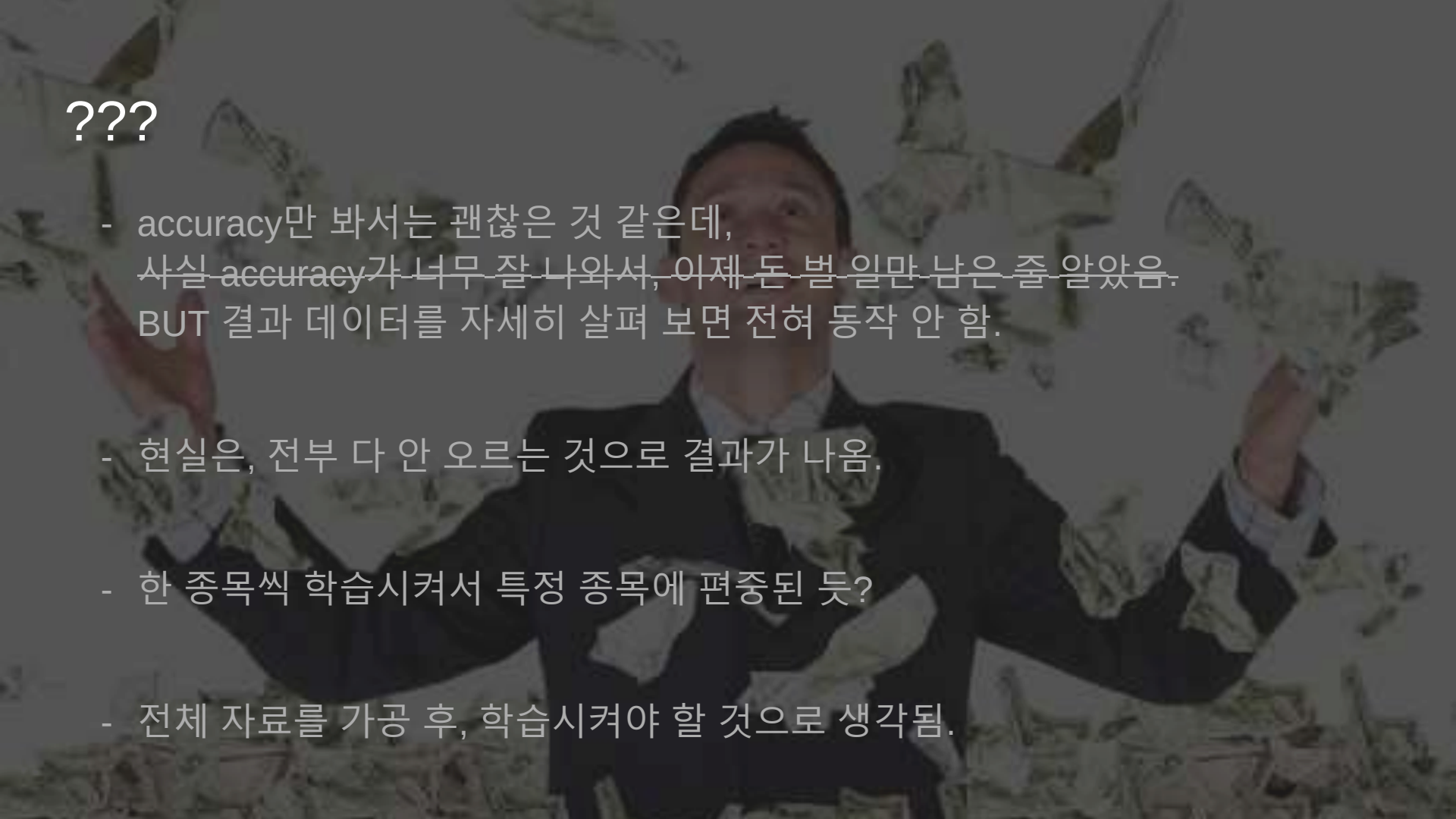
- 데이터를 1개 가져오려고 해도 중간중간 연결이 많이 필요해서 그런듯.
- 게다가 점검시간도 있음.

일데이터 분석을 위한 준비들

- sqlite3 to pandas
    - http://stackoverflow.com/questions/36028759/how-to-open-and-convert-sqlite-database-to-pandas-dataframe
- pandas tutorial
    - https://github.com/jvns/pandas-cookbook/tree/master/cookbook
- 주식 지표 계산
    - 하나하나 ~~계산할 수는 없고~~ 귀찮고, stockstats 모듈을 활용
        - https://github.com/jealous/stockstats
    - 노트북에서 일단 어떤 지표를 쓸까 체크 및 연습

Test log #1 - 뭐가뭔지 모르겠고 일단 돌리고 결과를 적어 놓자.

- kospi + 34, 100, 20, 2 + 10 epoch + 0.001 learning rate =
  - 2297 sec, 0.971548 accuracy
- kospi + 34, 100, 100, 20, 2 + 10 epoch + 0.001 learning rate =
  - 2286 sec, 0.9713 accuracy
- kospi + 34, 200, 100, 20, 2 + 10 epoch + 0.001 learning rate =
  - 2347 sec, 0.9699 accuracy
- kospi+kosdaq + 34, 200, 100, 20, 2 + 10 epoch + 0.001 learning rate =
  - 5586 sec, 0.95725 accuracy



???

- accuracy만 봐서는 괜찮은 것 같은데,
  ~~사실 accuracy가 너무 잘 나와서, 이제 돈 벌 일만 남은 줄 알았음.~~
  BUT 결과 데이터를 자세히 살펴 보면 전혀 동작 안 함.

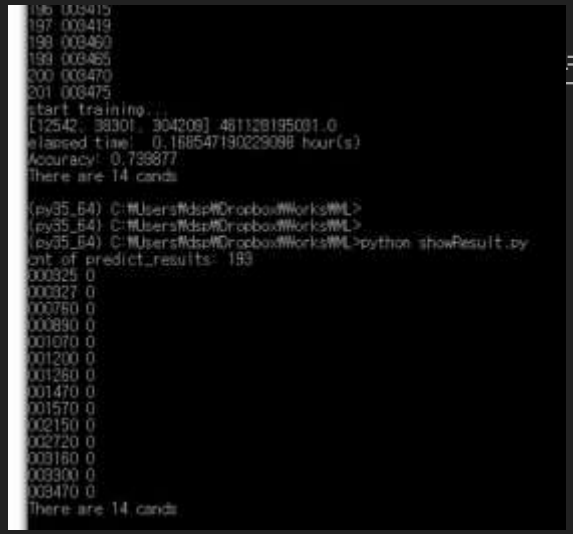
- 현실은, 전부 다 안 오르는 것으로 결과가 나옴.

- 한 종목씩 학습시켜서 특정 종목에 편중된 듯?

- 전체 자료를 가공 후, 학습시켜야 할 것으로 생각됨.

Test log #2

- 초기값에 편중되지 않도록 전체세트를 먼저 구하고 훈련.
- 200개 종목에 대해서만 우선 실험한 결과.
  - network : 35, 200, 3



한 후, 0504결과를 예측한 결과.

전체를 훈련시키기 위한 개선

- 전체를 모두 메모리에 올릴 수가 없는 문제 발생
  - OOM 이슈
- Batch로 나눠서 실행할 필요가 있음.
  - 이전까지는 왜 batch가 필요한지 모르고 있었음. --;

해 놓았다가 다시 꺼내서 훈련함.

Test log #3

- ~2017/05/04까지의 데이터를 가지고, 0508를 예측
  - all + 34, 512, 1024, 512, 64, 3 + 0.0001 learning rate + 500 epoch
    - 2.5 hours, accuracy 0.789907, 112 cands(all 1)
  - all + 34, 512, 512, 512, 64, 3 + 0.0001 + 500
    - 2.2 hours, accuracy 0.839576, 1 cands
  - all + 34, 1024, 1024, 1024, 64, 3 + 0.0001 + 100
    - 2.0 hours, accuracy 0.835347, 19 cands
      - 001470 0
      - 005740 0
      - 005745 0
      - 009275 0
      - 023350 0
      - 049770 0
      - 051900 0

...

- all + 34, 1024, 1024, 1024, 1024, 1024, 64, 3 + 0.0001 + 50
  - 0.818191, 2+52 cands.
- …

- 더 deep + wide 하게 가도 의미가 없는 것 같아

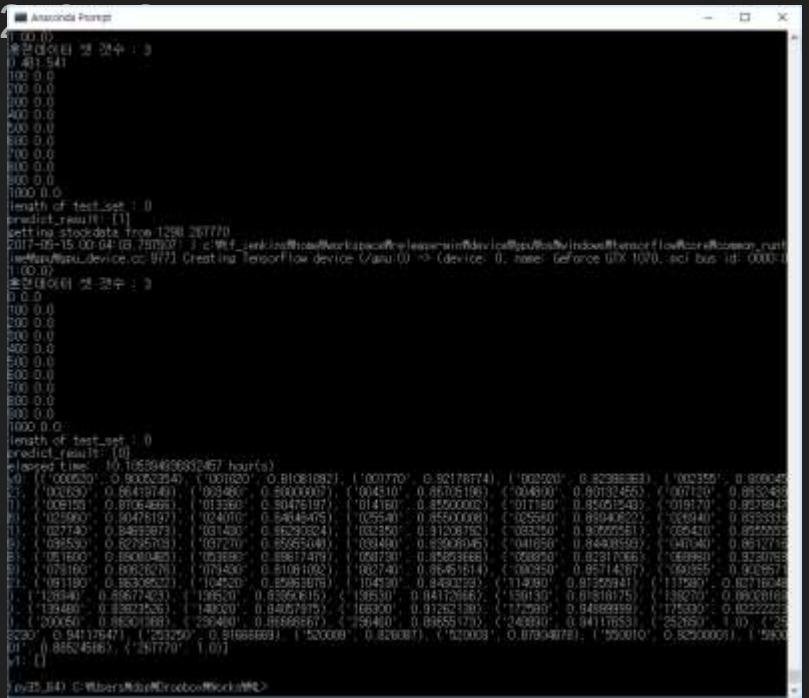
- 플라랩 동료의견 :
  - 한 종목씩 따로 훈련해 보는건 어떨까?

입력에 변화를 줘 보자

- 입력 데이터를 오늘 데이터만 넣으면 충분하다고 생각했으나, 결과가 안 좋으니 입력데이터로 어제+그제 데이터를 넣어볼까?
    - 35 → 35 x 3 days
    - 35 → 35 x 5 days
- Normalization을 안하고 있었네?
- 의견수렴: 한 종목에 대해서만 훈련

출력 변화

- 다음날 결과만을 예측 → 3일 연속 오르는지 체크 (binary classification)
- layer: 35*5 , 512, 512, 512, 512, 512
- 0.0001 learning rate, 1000 epoch
- 20150504 기준날로 예측한 결과



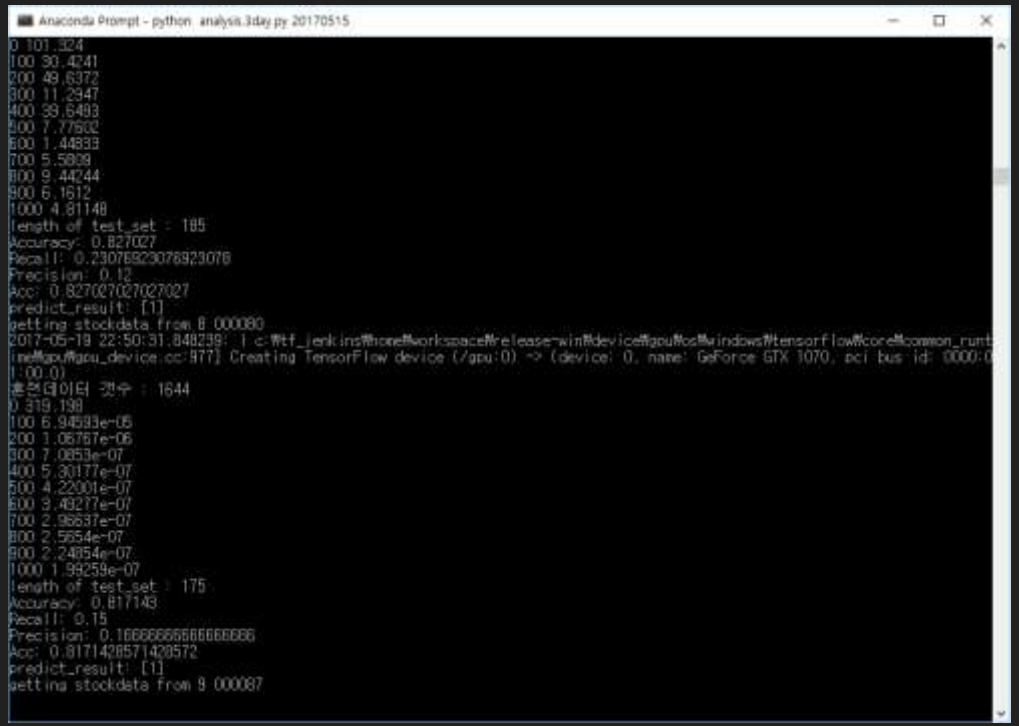
Accuracy 만으로는 부족

- Accuracy 만으로는 제대로 예측이 되었는지 알 수가 없음.
- recall, precision도 측정 필요.
    - recall = tp / (tp+fn)
        - 재현율이란, 실제 True인것 중에서 True로 예측한 것의 비율
    - precision = tp / (tp+fp)
        - 즉, 정확도는 True로 예측한 것 중에서 실제 True인 것의 비율

	실제결과	
	Positive	Negative
예측결과 positive	True Positive	False Positive
예측결과 negative	False Negative	True Negative

결과 update

- 결과는 더욱 처참해짐.



CNN ?

- 어차피 stockstat에서 특징을 뽑아내었으므로 크게 의미가 없다고 생각했으나 …
  - 일단 지금은 패스
  - 나중에 한번 실험해 봐야 겠음.

RNN

- 사실 이미 과거의 데이터를 사용하고 있긴 하지만...
- LS

```
81      # init tensors
82      X = tf.placeholder(tf.float32, [None, WINDOW_SIZE, len(columns)])
83      Y = tf.placeholder(tf.float32, [None, 1])
84
85      def lstm_cell():
86          cell = tf.contrib.rnn.BasicLSTMCell(num_units=hidden_dim, state_is_tuple=True, activation=tf.tanh, forget_bias=1.0)
87          return cell
88      multi_cells = tf.contrib.rnn.MultiRNNCell([lstm_cell() for _ in range(MULTI_CELLS)], state_is_tuple=True)
89
90      outputs, _states = tf.nn.dynamic_rnn(multi_cells, X, dtype=tf.float32)
91      Y_pred = tf.contrib.layers.fully_connected(
92          outputs[:, -1], output_dim, activation_fn=None)  # We use the last cell's output
93
94      # cost/loss
95      loss = tf.reduce_sum(tf.square(Y_pred - Y))  # sum of the squares
96      # optimizer
97      optimizer = tf.train.AdamOptimizer(LEARNING_RATE)
98      train = optimizer.minimize(loss)
99
100     # RMSE
101     targets = tf.placeholder(tf.float32, [None, 1])
102     predictions = tf.placeholder(tf.float32, [None, 1])
103     rmse = tf.sqrt(tf.reduce_mean(tf.square(targets - predictions)))
```

LSTM 결과

- 그러나 결과는 역시 안습
  - precision이 영 나아지지 않음.

- facebook tensorflow KR 커뮤니티에도 도움 요청을 했으나…
  - https://www.facebook.com/groups/TensorFlowKR/?multi_permalinks=473913729616357
  - 댓글들이 거의 다 좌절스러움.

Re-try LSTM

- facebook 댓글 중, 3년 어치 데이터에 대해 모든 종목 값들을 training하면 어떨까란 댓글이 그럴듯해서 적용.

```
[step: 2400] loss: 77185.1698404948
[step: 2500] loss: 75676.47521972656
[step: 2600] loss: 73147.57641601562
[step: 2700] loss: 72684.11617024739
[step: 2800] loss: 76013.92810058594
[step: 2900] loss: 71640.28544108073
[step: 3000] loss: 68808.03889973958
length of test_set : 59052
RMSE: 3.2340409755706787
Recall: 0.5517436380772855
Precision: 0.5032380079087627
Accuracy: 0.4648106753369911
predict_result: [[ 0.40456176]
 [ 2.58484292]
 [ 0.22989161]
 ...,
 [-0.06882797]
 [ 0.14535643]
 [ 0.36506465]]
elapsed time:  8.33378691944811 hour(s)

(py35_64) C:₩Users₩dsp₩Dropbox₩Works₩ML>
```

- 후... 그러나 역시 결과는...

News

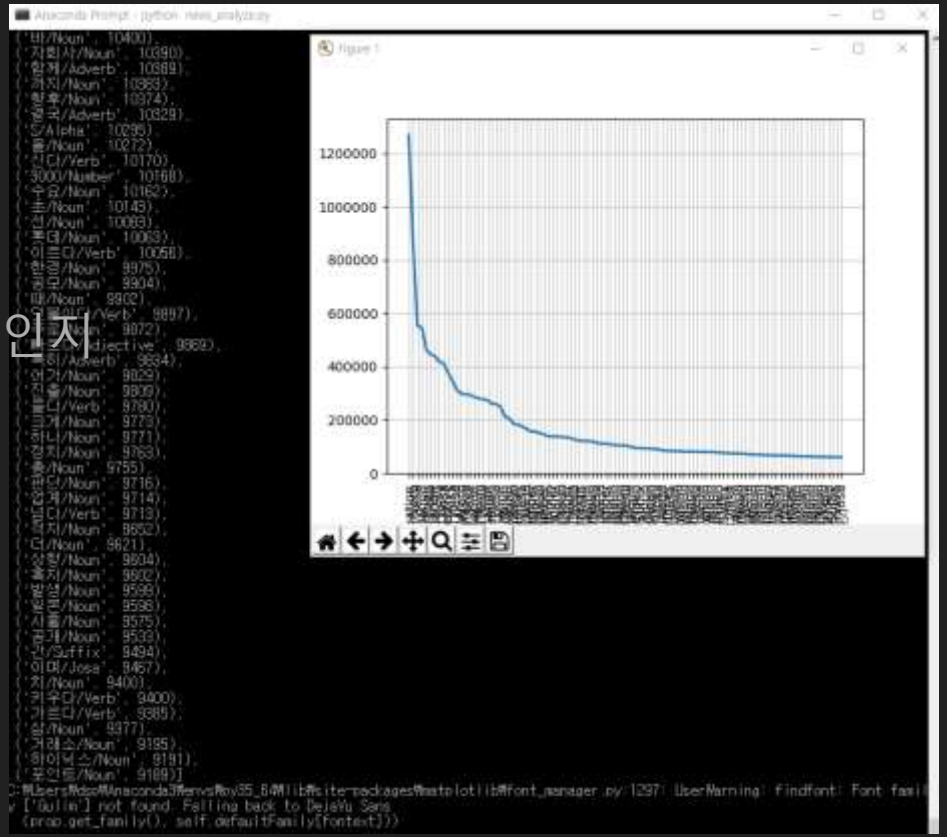
- 여기서 좌절할 수 없지. 뉴스를 수집해서 버무려볼까?
- 뉴스를 수집해서 분석하려면 한국어 NLP가 필요.
- konlpy 사용
  -
```
>>> from konlpy.tag import Hannanum
>>> text='이낙연 국무총리 후보자에 대한 국회 인준을 두고 여야가 대치하는 가운데 남화토건(091590)이 급등하고 있다. 앞서 남화토건은 장미 대선이 끝난 직후 이낙연 국무총리 후보자 관련주로 꼽히며 한차례 급등을 연출했던 적이 있다.'
>>> h.nouns(text)
['이낙연', '국무총리', '후보자', '국회', '인준', '여야', '대치', '가운데', '남화토', '091590', '급등', '남화토건', '장미', '대선', '직후', '이낙연', '국무총리', '후보', '관련주', '한차례', '급등', '연출', '적']
```
- 처리로직
  - 문단별로 긍/부정 카운트하여 scoring해 보자꾸나.

Crawling News

- 일단 데이터가 있어야 하니, 뉴스들을 긁어모아 보자~
- 네이버 금융 장중특징주 crawling
  - 시드 : http://finance.naver.com/news/market_special.nhn?&page=1
  - field : datetime, title, url, content, …
  - logic
    - page=1 에서 한번 크롤한 결과가 나올때, 혹은 article이 없는 마지막 페이지가 나올 때 까지
      - selecting articles : $(".boardList2 tbody tr")
        - url이 이미 저장된 것인지 새것인지 확인.

뉴스 분석

- 형태소별로 histogram 뽑아보고,
  - 빈도별 긍부정 단어를 체크해서,
  - +/- 점수를 주면 종목별로 어떤 뉴스인지 판단이 가능할 것으로 예상됨.



To be continued...

- 약 3개월간 주말(+평일 퇴근 후 밤)에 한 삽질은 여기까지입니다.

- 조금 지치지만, 계속 삽질 중입니다.
  - 요즘 술만 먹고 있음 ㅠㅜ;

- 언젠가는 컴퓨터 투자비를 뽑을 수 있겠죠?
  - ~~빨리 게임이나 열심히 해야겠다.~~

The End