Mingqing Teng

1) Which tasks have been completed?

A. Mingqing Teng (mt52): Installation of Anaconda including Numpy, Jupyter Notebook, Pandas.. etc.
B. Mingqing Teng (mt52): Finished tutorials of Jupyter NoteBook, Pandas
C. Mingqing Teng (mt52): Reading parquet file into pandas and investigating the data structure of the parquet file
D. Mingqing Teng (mt52): Focusing on column of users:
User-> Score, ranking by popularity, Top 50 or Top100.

2) Which tasks are pending?

E. Mingqing Teng (mt52): DateFrame of parquet file

3) Are you facing any challenges?

F. Mingqing Teng (mt52): How to use Pandas to do data analysis

Ben Chao

1) Which tasks have been completed?

G. Ben Chao (cwchao4): Installation of Numpy, Jupyter Notebook, Pandas.
H. Ben Chao (cwchao4): Finished tutorials of Jupyter NoteBook, Pandas, Numpy.
I. Ben Chao (cwchao4): Reading parquet file into pandas and investigating the data structure of the parquet file

2) Which tasks are pending?

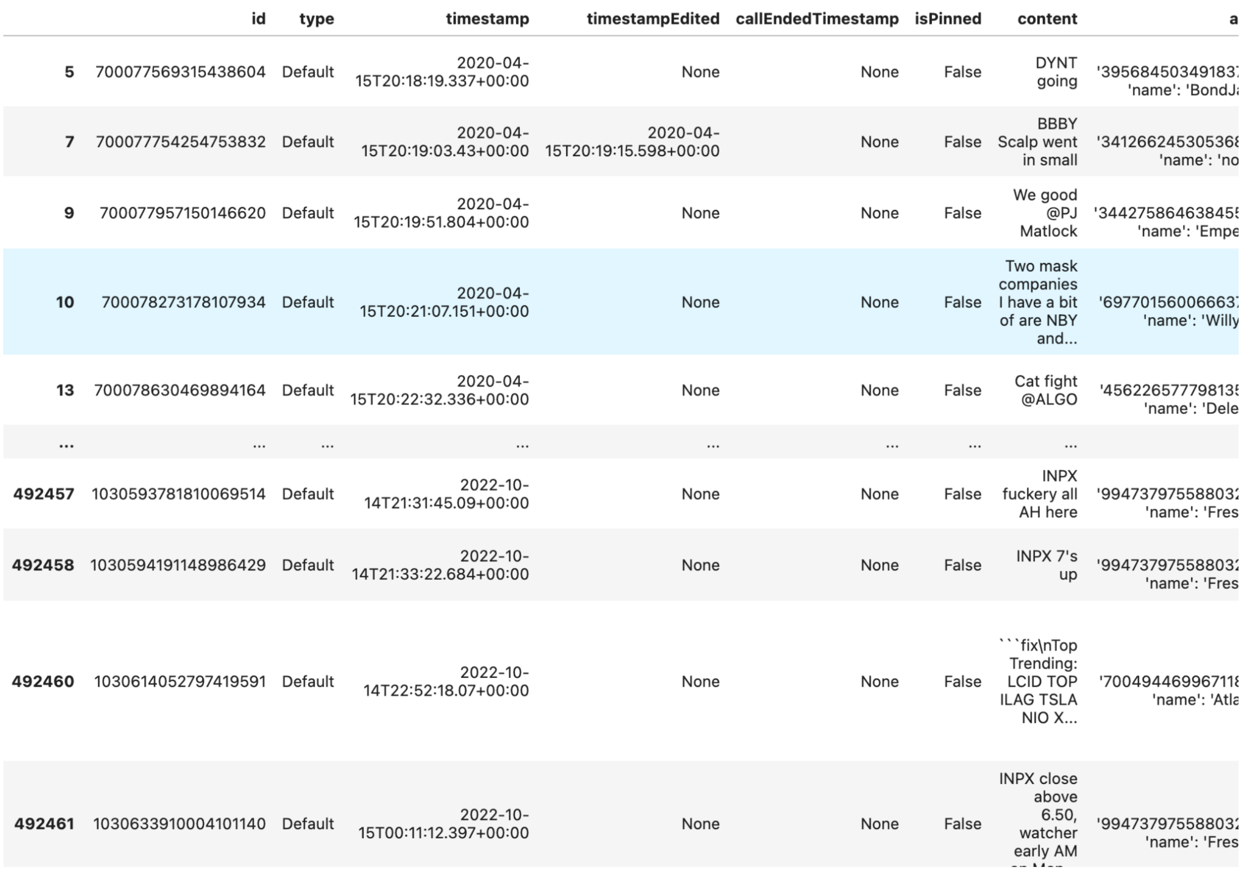
J. Ben Chao (cwchao4): Collect the top frequent Emoji and transfer Emoji into 1d array.

3) Are you facing any challenges?

Ji Ma (jima2)

1) Which tasks have been completed?

- Scraping trader’s chat data from discord and condensed it to a parquet data frame consists of information like content, reactions, timestamp, author, etc.

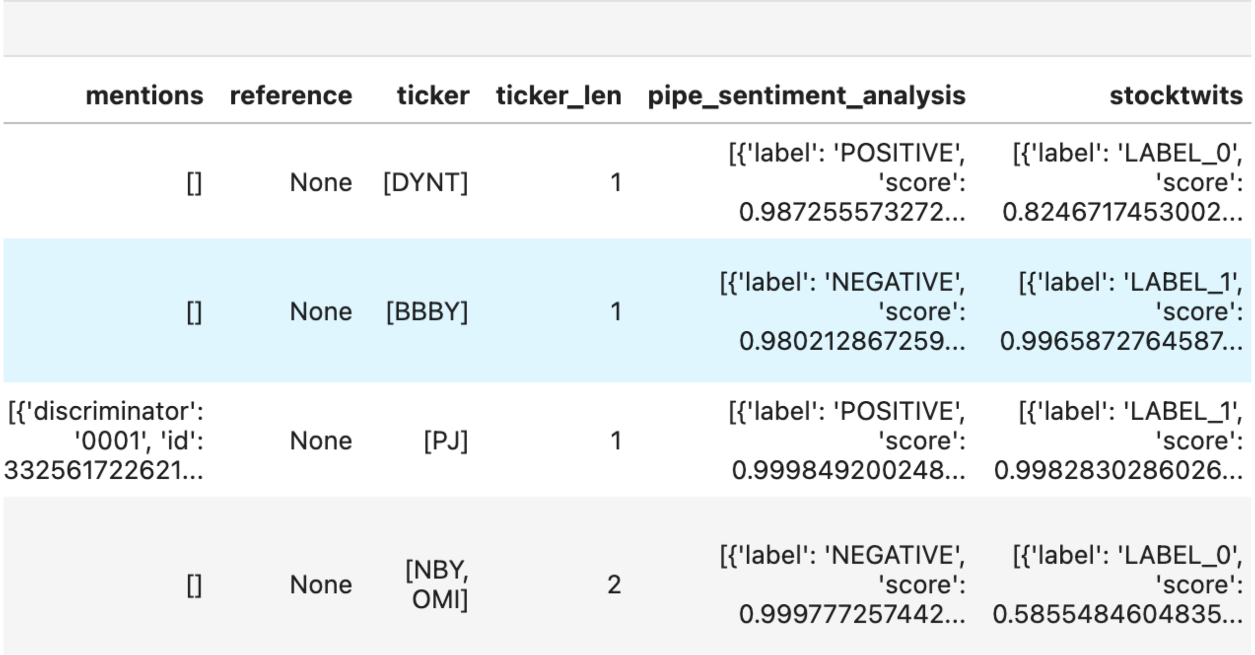


	id	type	timestamp	timestampEdited	callEndedTimestamp	isPinned	content	a
5	700077569315438604	Default	2020-04-15T20:18:19.337+00:00	None	None	False	DYNT going	'39568450349183 'name': 'BondJ
7	700077754254753832	Default	2020-04-15T20:19:03.43+00:00	2020-04-15T20:19:15.598+00:00	None	False	BBBY Scalp went in small	'34126624530536 'name': 'no
9	700077957150146620	Default	2020-04-15T20:19:51.804+00:00	None	None	False	We good @PJ Matlock	'34427586463845 'name': 'Empe
10	700078273178107934	Default	2020-04-15T20:21:07.151+00:00	None	None	False	Two mask companies I have a bit of are NBY and...	'69770156006663 'name': 'Willy
13	700078630469894164	Default	2020-04-15T20:22:32.336+00:00	None	None	False	Cat fight @ALGO	'45622657779813 'name': 'Dele
...	...	...	...	...	...	...	...	
492457	1030593781810069514	Default	2022-10-14T21:31:45.09+00:00	None	None	False	INPX fuckery all AH here	'99473797558803 'name': 'Fres
492458	1030594191148986429	Default	2022-10-14T21:33:22.684+00:00	None	None	False	INPX 7's up	'99473797558803 'name': 'Fres
492460	1030614052797419591	Default	2022-10-14T22:52:18.07+00:00	None	None	False	```fix\nTop Trending: LCID TOP ILAG TSLA NIO X...	'70049446996711 'name': 'Atla
492461	1030633910004101140	Default	2022-10-15T00:11:12.397+00:00	None	None	False	INPX close above 6.50, watcher early AM	'99473797558803 'name': 'Fres

-
- Did a quick analysis on the top frequent words

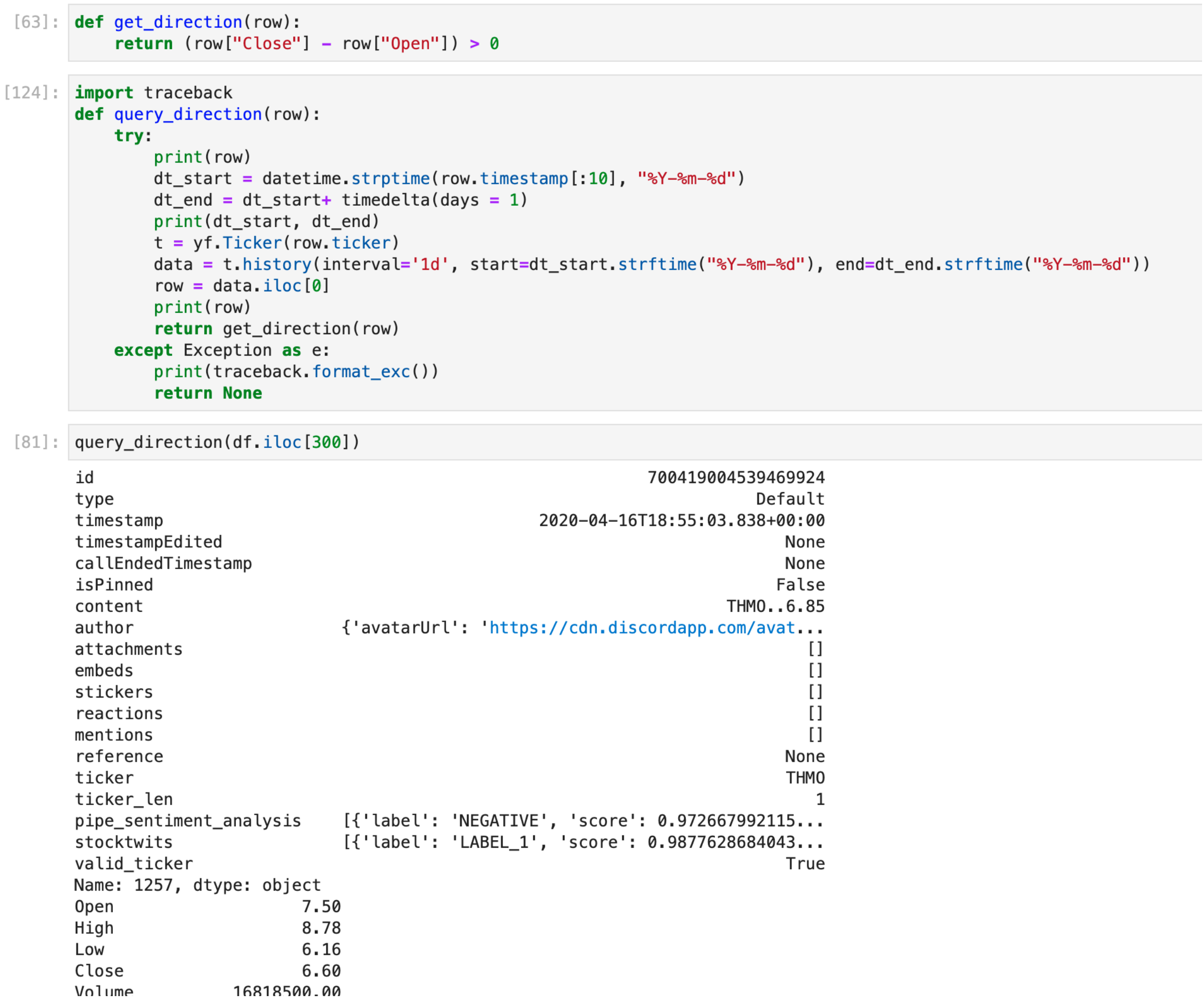
volatility halted
top trending
fix top
vwap
new
break
watch
volume
watching
will
trying
play
ww
today
nhod
now
day
see
added
deleted user
open
dip
going
lol
good
bounce
think
hit
nice
moving
still
trogdor
took
look
looking
move
back
pop
close
well
halt
mara
zack morris
market
share
spy
might
curl
pt
holding
hold
gap
know
entry
getting
tsla
eye
breakout
daily
knife
vol
buy
btc
ah
got
big
level
guy
next
high
gnus
chart
long
nakd
trade
need
pm
swing
want
morning
another
may
run
short
u
keep
taking
ocgn
sndl
right
test
go
offering
one
hod
support
addyesterday
time
amc
thank
yeah
make
though
imo
lot
s
made
first
tickerpossible
really
er
uone
come
far
coming
new high
fcel
pm high
n
g
p
call
htz
push
drop
set
hl
area
spi
tos
around
trading
sava
wish
low
min
something
great
work
f
dpw
cei
stock
twitter zack
sold
maybe
position
gme
wow
pullback
order
said
scalp
sunw
morris mrzackmorris
goes
nola
sell
damn
small
adding
little
vxrt
wwr
take
much
pr
gevo
halted
kodk
cbat
re
ayro
sl
uav
bit
soon
rip
aht
love
tomorrow
starter
thing
let
room
company
stop
broke
reversal
week
say
tkat
srne
curling bbig
wall
way
earlier
potential

-
- Added sentiment analysis with roberta model



mentions	reference	ticker	ticker_len	pipe_sentiment_analysis	stocktwits
[]	None	[DYNT]	1	[{'label': 'POSITIVE', 'score': 0.987255573272...	[{'label': 'LABEL_0', 'score': 0.8246717453002...
[]	None	[BBBY]	1	[{'label': 'NEGATIVE', 'score': 0.980212867259...	[{'label': 'LABEL_1', 'score': 0.9965872764587...
[{'discriminator': '0001', 'id': 332561722621...	None	[PJ]	1	[{'label': 'POSITIVE', 'score': 0.999849200248...	[{'label': 'LABEL_1', 'score': 0.9982830286026...
[]	None	[NBY, OMI]	2	[{'label': 'NEGATIVE', 'score': 0.999777257442...	[{'label': 'LABEL_0', 'score': 0.5855484604835...

- Extracted ticker/stock information and validated those ticker information. So that each row would have one corresponding ticker
- Generated label based on one day price movement of a given row’s stock/ticker.

- 

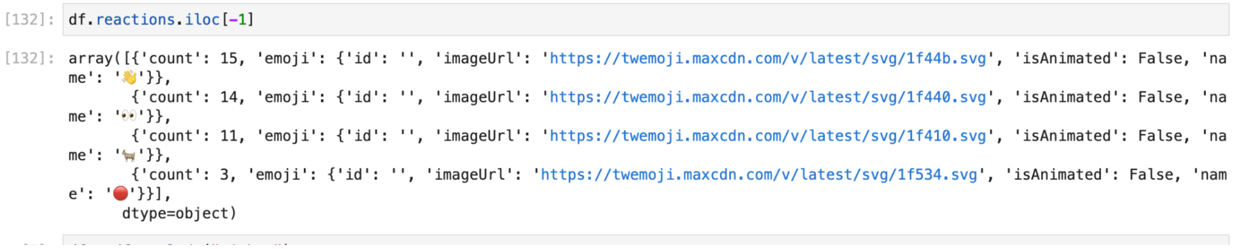
```
[63]: def get_direction(row):
          return (row["Close"] - row["Open"]) > 0
```

```
[124]: import traceback
       def query_direction(row):
           try:
               print(row)
               dt_start = datetime.strptime(row.timestamp[:10], "%Y-%m-%d")
               dt_end = dt_start+ timedelta(days = 1)
               print(dt_start, dt_end)
               t = yf.Ticker(row.ticker)
               data = t.history(interval='1d', start=dt_start.strftime("%Y-%m-%d"), end=dt_end.strftime("%Y-%m-%d"))
               row = data.iloc[0]
               print(row)
               return get_direction(row)
           except Exception as e:
               print(traceback.format_exc())
               return None
```

```
[81]: query_direction(df.iloc[300])
```

```
id                                                     700419004539469924
type                                                              Default
timestamp                                    2020-04-16T18:55:03.838+00:00
timestampEdited                                                      None
callEndedTimestamp                                                   None
isPinned                                                            False
content                                                       THMO..6.85
author                     {'avatarUrl': 'https://cdn.discordapp.com/avat...
attachments                                                            []
embeds                                                                 []
stickers                                                               []
reactions                                                              []
mentions                                                               []
reference                                                            None
ticker                                                               THMO
ticker_len                                                              1
pipe_sentiment_analysis    [{'label': 'NEGATIVE', 'score': 0.972667992115...
stocktwits                 [{'label': 'LABEL_1', 'score': 0.9877628684043...
valid_ticker                                                         True
Name: 1257, dtype: object
Open                    7.50
High                    8.78
Low                     6.16
Close                   6.60
Volume           16818500.00
```

2) Which tasks are pending?

- Further sanity check the data in terms of label
- Add more labels besides 1 day price movement, we can consider 7 day or 1 hour price movement as well as prediction a task
- More feature engineering such as one hot encoding for reactions, and time of day, day of week.
- 

```
[132]: df.reactions.iloc[-1]
```

```
[132]: array([{'count': 15, 'emoji': {'id': '', 'imageUrl': 'https://twemoji.maxcdn.com/v/latest/svg/1f44b.svg', 'isAnimated': False, 'na
me': '👋'}},
              {'count': 14, 'emoji': {'id': '', 'imageUrl': 'https://twemoji.maxcdn.com/v/latest/svg/1f440.svg', 'isAnimated': False, 'na
me': '👀'}},
              {'count': 11, 'emoji': {'id': '', 'imageUrl': 'https://twemoji.maxcdn.com/v/latest/svg/1f410.svg', 'isAnimated': False, 'na
me': '🐐'}},
              {'count': 3, 'emoji': {'id': '', 'imageUrl': 'https://twemoji.maxcdn.com/v/latest/svg/1f534.svg', 'isAnimated': False, 'nam
e': '🔴'}}],
       dtype=object)
```

- Build a model for prediction.

3) Are you facing any challenges?

- Have to deal with rate limiting when pulling data from discord and yahoo finance.

- The data from yahoo finance’s coverage is very low, a lot of None ended up with labels. Need to find an alternative data source.