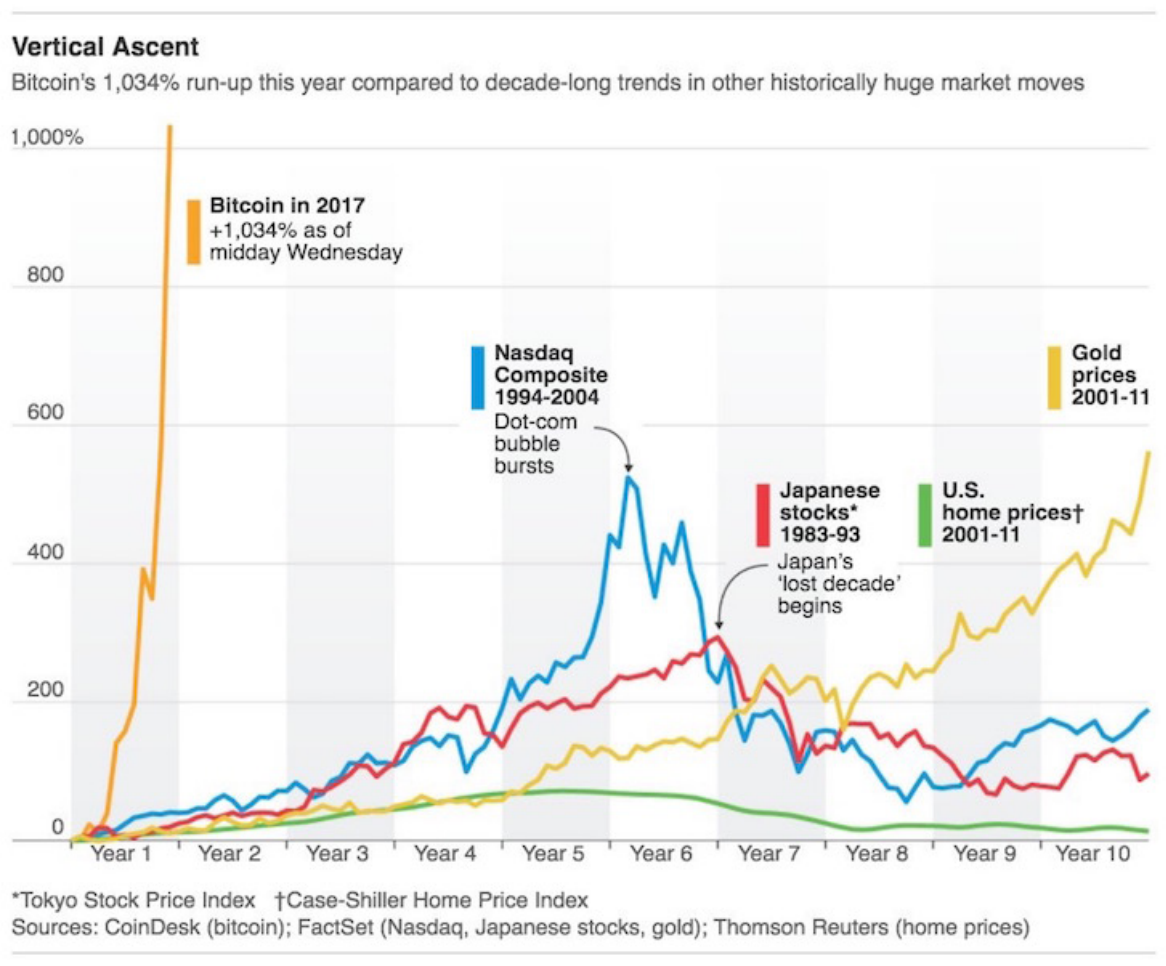
The crowd and the coin

—using sentiment to predict bitcoin price

Summer Yuan

Bitcoin is on fire…

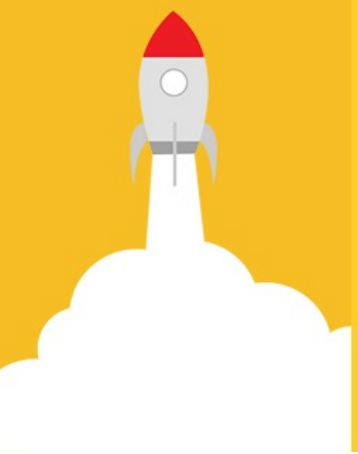


1. As of 2017, bitcoin offers over 1000% return rate

2. Outperforms real estate, nasdaq, equities, commodities

Content

- Data
- EDA
- NLP analysis
- Price prediction (ARIMA vs LSTM)

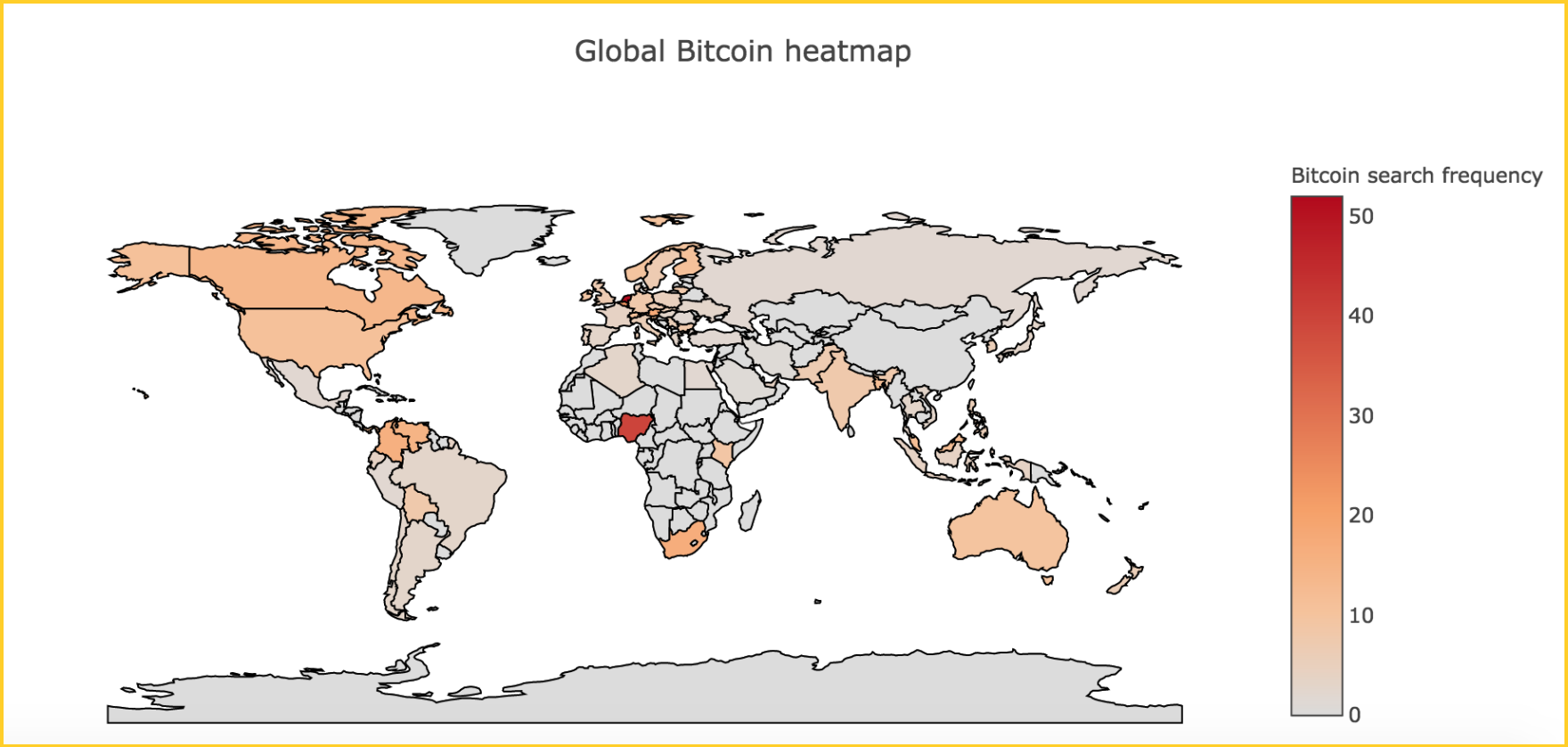


Data

- historical bitcoin price(Quandl API)
- ‘bitcoin’ google search frequency (pytrends API)
- tweets about bitcoin

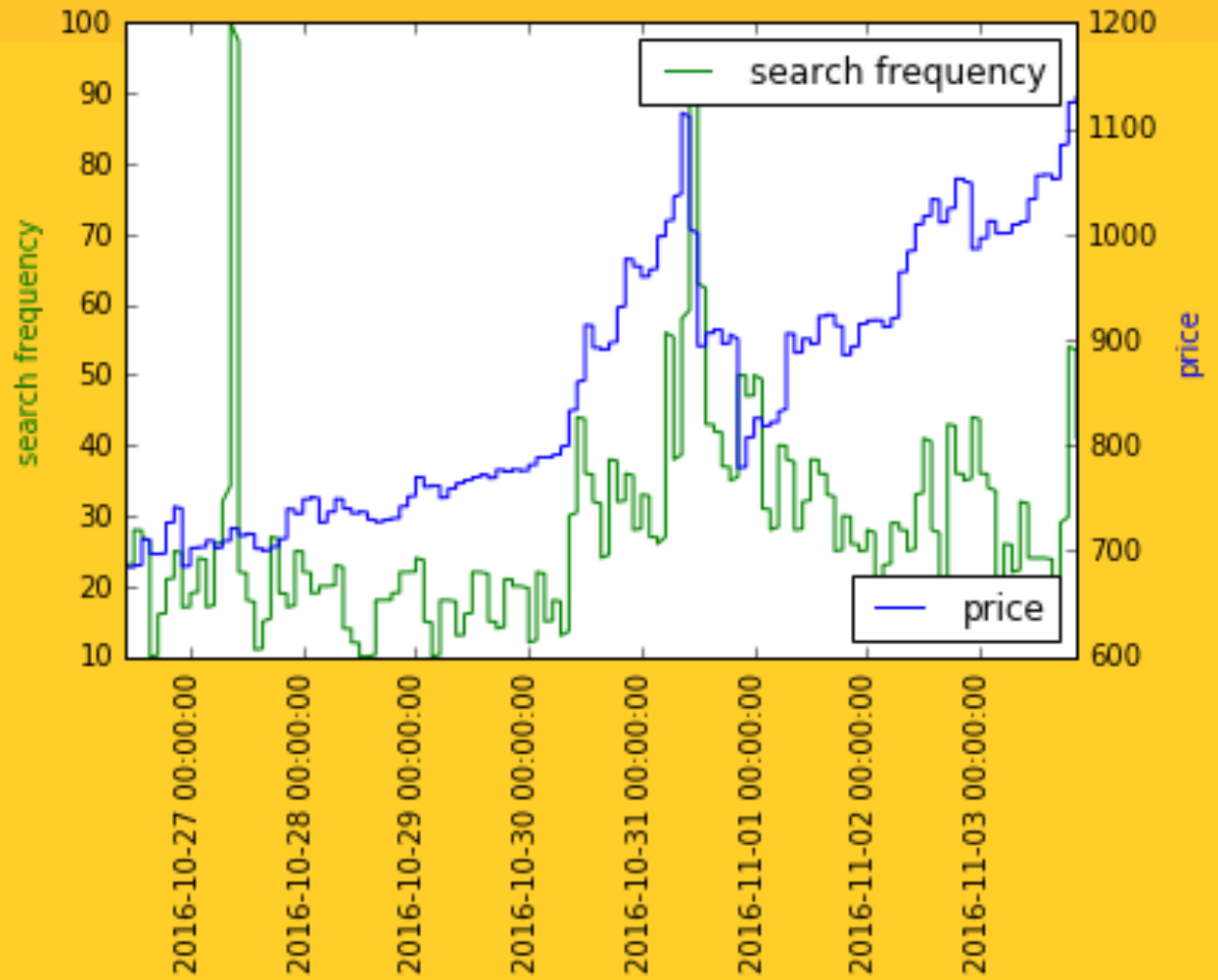
EDA

which country has the highest search frequency for bitcoin in 2017?



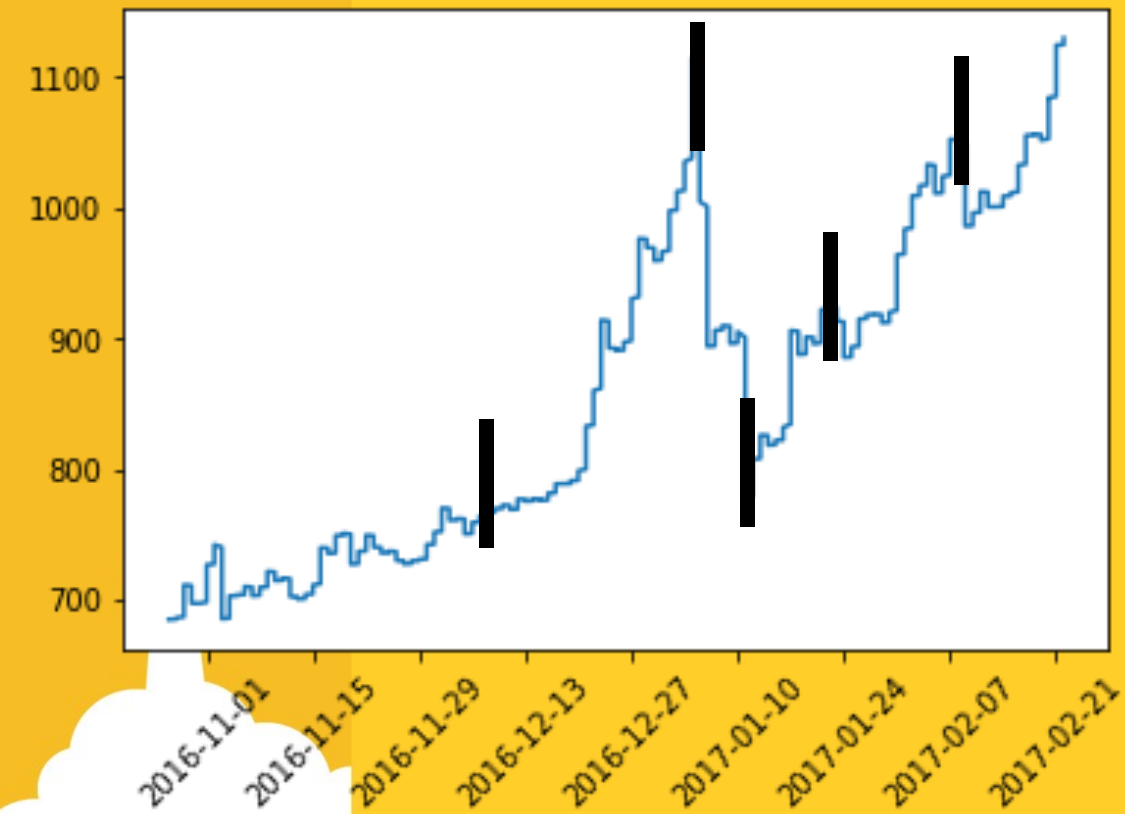
EDA

correlation between search frequency and bitcoin price



Nlp-topic modeling

divide price into different cycles



#cycle1: 2016/10/27 - 2016/12/02 - fluctuation
#cycle2: 2016/12/03 - 2017/01/04 - up
#cycle3: 2017/01/05 - 2017/01/11 - down
#cycle4: 2017/01/12 - 2017/01/20 - up
#cycle5: 2017/01/21 - 2017/02/08 - up
#cycle6: 2017/02/09 - 2017/02/22 - up

Nlp-topic modeling(NMF)

fluctuation : 1 {0: ['buy', 'sell', 'india', 'instantly', 'potential', 'try', 'currency']}

up: 2 {0: ['buy', 'high', 'currency', 'time', 'billion', 'worth', 'gold'}

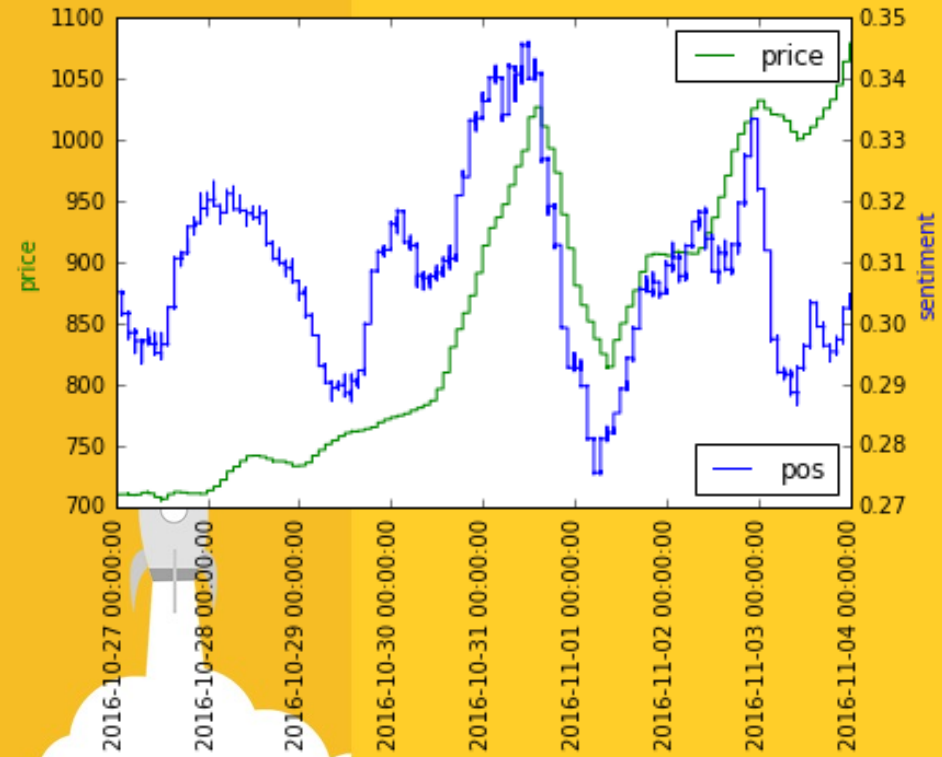
down : 3 {0: ['china', 'bank', 'central', 'currency', 'exchange', ‘high','warning']}

up: 4 {0: ['china', 'trading', 'bank', ‘central’, 'buy', 'currency', ‘india']}

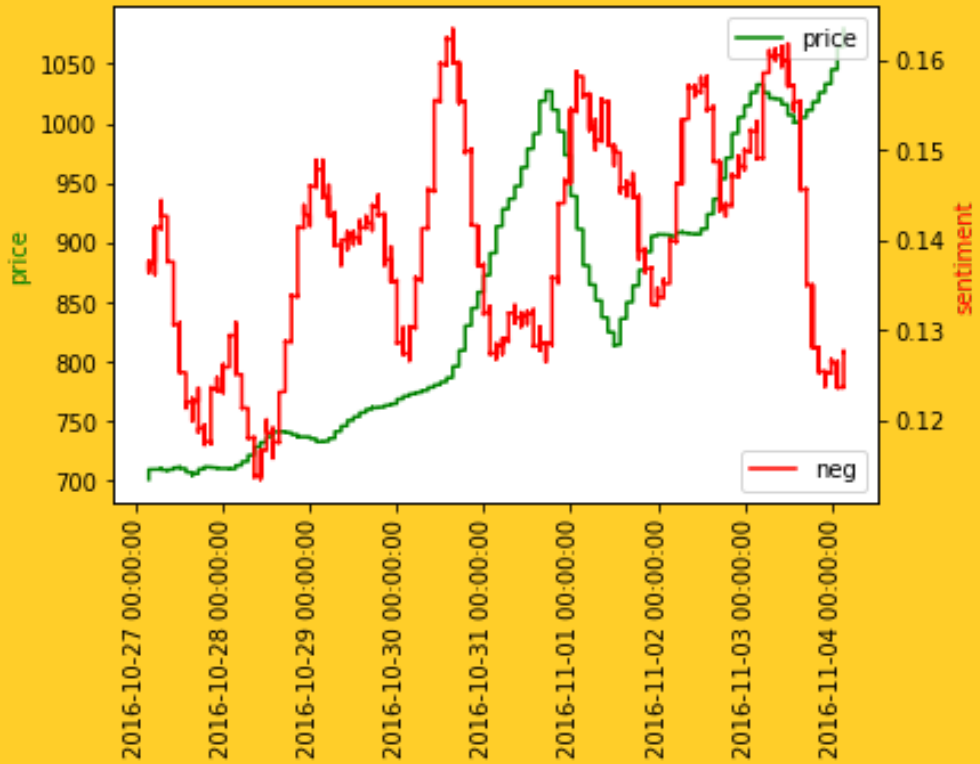
up: 5 {0: ['buy', 'trading', 'make', 'china', 'great', 'gold', 'mining']}

up: 6 {0: ['bank', 'china', 'central', 'buy', 'first', 'trading', 'exchange']}

positive sentiment vs price



negative sentiment vs price



NLP -Sentiment Analysis

some negative topics about bitcoin

3: ['via', 'stolen', ‘block’,’millions','hard','phone', ‘scam’,'size','unconfirmed','buy'],
bitcoin theft

4: [‘pay','ransom', ‘buy’,’invest','mobile','data','us',‘money','check','cash'],
kidnapping

5: [‘news','price','china','currency','demand','bank','warning','ban','stop']}
Warnings from Chinese bank

Investigation on China? China! China.(Word2Vec)

devaluation

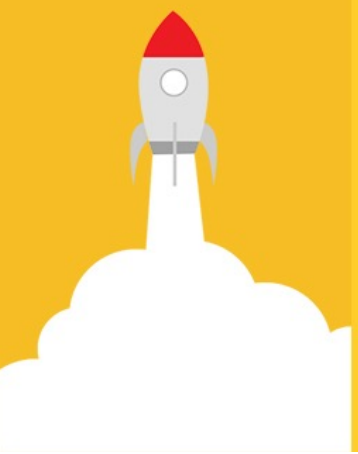
india

worsen

China

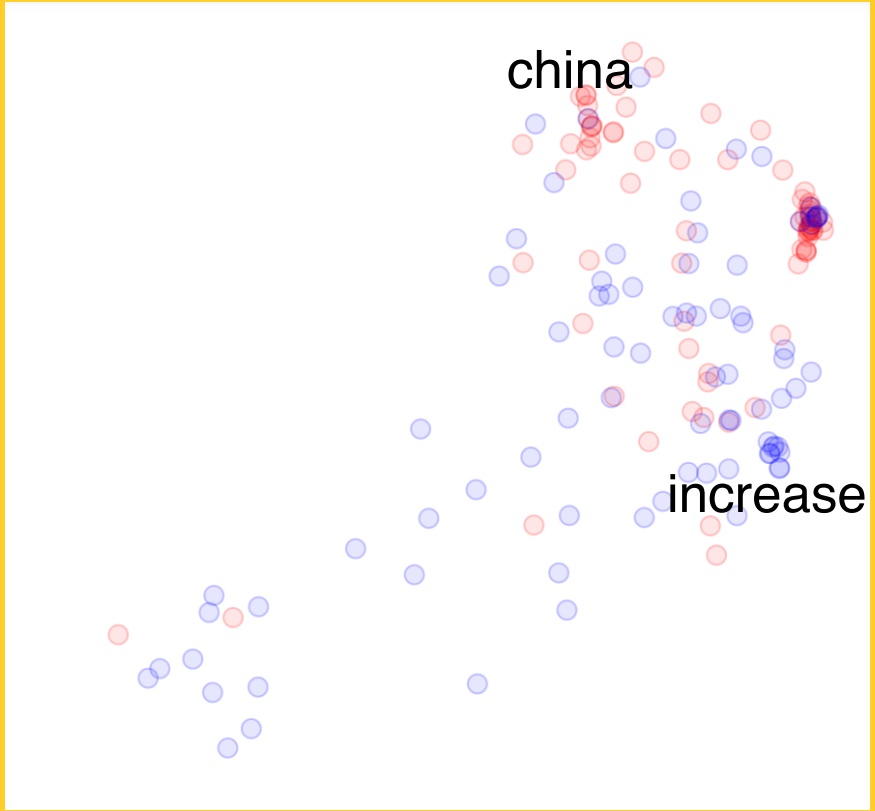
pressure

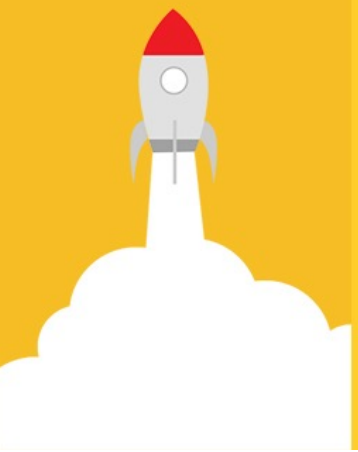
halt



TSNE

words clustered around ‘China’, ‘increase’





Prediction –base model Arima

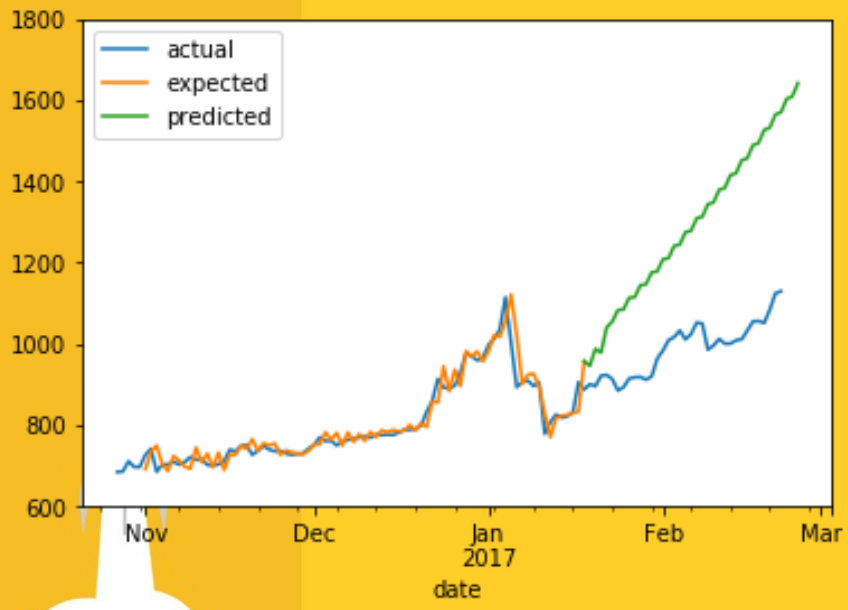
features: lag-3 time series, 6 topics, search trends, average sentiment score for the day

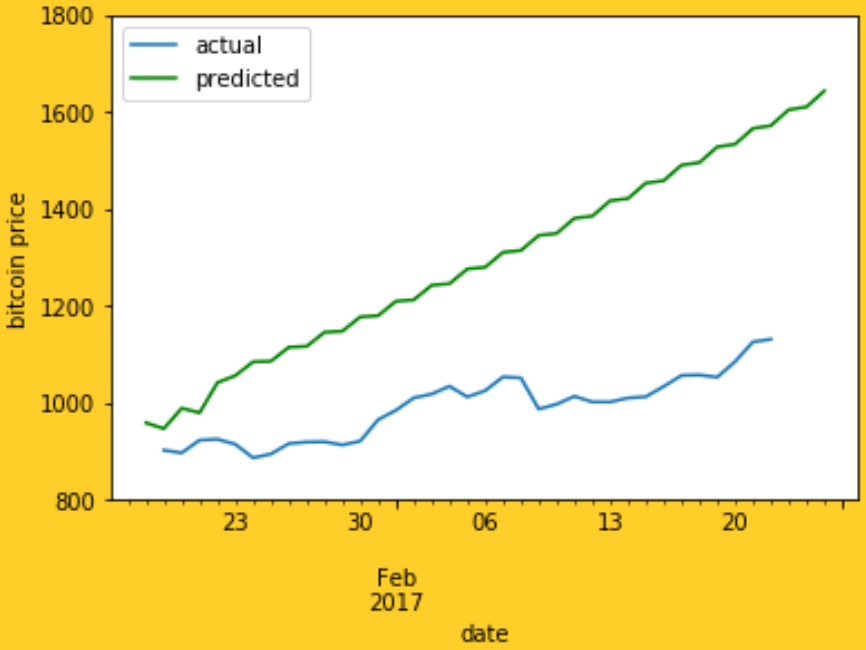
ARMA Model Results

Dep. Variable:	Close	No. Observations:	83
Model:	ARMA(3, 8)	Log Likelihood	-406.605
Method:	css-mle	S.D. of innovations	28.948
Date:	Thu, 01 Mar 2018	AIC	839.210
Time:	17:50:13	BIC	870.655
Sample:	10-27-2016 - 01-17-2017	HQIC	851.843

Prediction - Arima

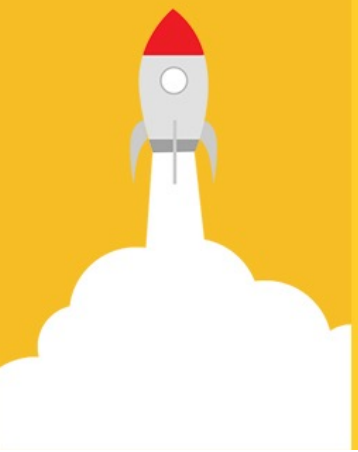
Test MSE: 104.142

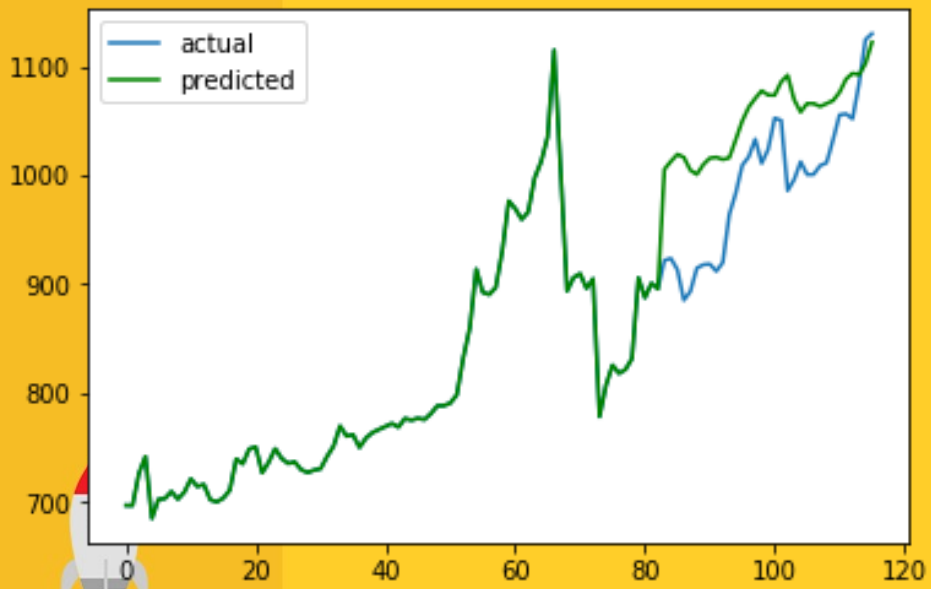
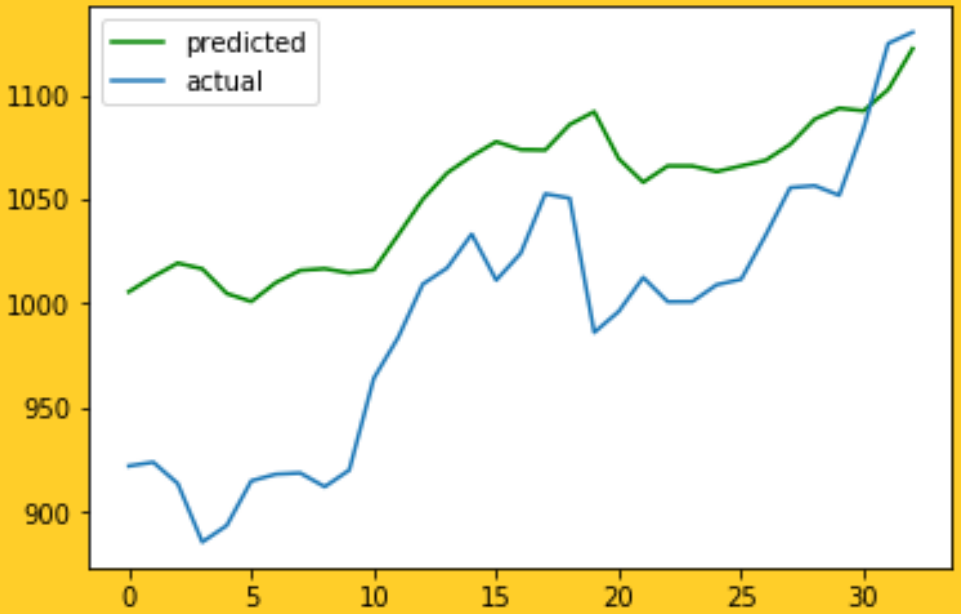




Prediction -LSTM

1. create look back(t1, t2, t3)
2. scale data into (-1,1)
3. reshape X, Y
4. create layers(Dense Layer)
5. pass on data, fit, predict
6. inverse transform output to original scale



Prediction -LSTM

Test MSE: 47.262