



CAPSTONE

Building Event Extraction and Trending Framework for Twitter

Joyce Lin, General assembly, data science immersive

POWER of Social Media

Social media started with life sharing; through mail, phone, campus internet....

.....And now has gone wild!!






GOAL: REALTIME EVENT DETECTION FRAMEWORK

Question:

What event do we want to know about?

Questions:

How soon can we catch a trending event?



WHY TWITTER

- » **Real-time results.** Because of Twitter's real-time streaming engine, we can quickly get public opinion
- » **Wide reach.** Twitter is used by all age group, many type and size of businesses
- » **Direct feedback.** You hear what people are saying as they say it.



Justin Bieber @justinbieber · Jun 5

I support all sports I'll put ANY jersey from ANY pro team if I'm whack for wearing jerseys they give me out of love then I'm Whack

3.7K 55K 188K



Donald J. Trump @realDonaldTrump · Jun 11

The **#FakeNews** MSM doesn't report the great economic news since Election Day. **#DOW** up 16%. **#NASDAQ** up 19.5%. Drilling & energy sector...

15K 24K 98K



sad i am @samcxeleman · Jun 4
being bored f'ing **SUUUUUCKS**

1




smeargle @fucktationam · 7h
my bladder fcking **suuuuucks**

CHALLENGES –NOISY TEXT

- » **Accumulate quickly.** Twitter is used by all age group, many type and size of businesses
- » **Mostly noise.** And it is often hard to identify which are and aren't!
- » **Unique grammar**
- » **Lexical Variation.** 2mar, tmr, tmoro, tmorrow, tmrw, tomo, tmw, and many.....



MY APPROACH

- » Set up **Data Collection** process and **Data Infrastructure**
 - » Examine different **Natural Language Processing** tools on collected tweets
 - » Create A/B Testing Model on similarity comparison
 - » Use **Time Series Modeling** to catch the trends
 - » Tuning hyperparameters for model improvement
- 

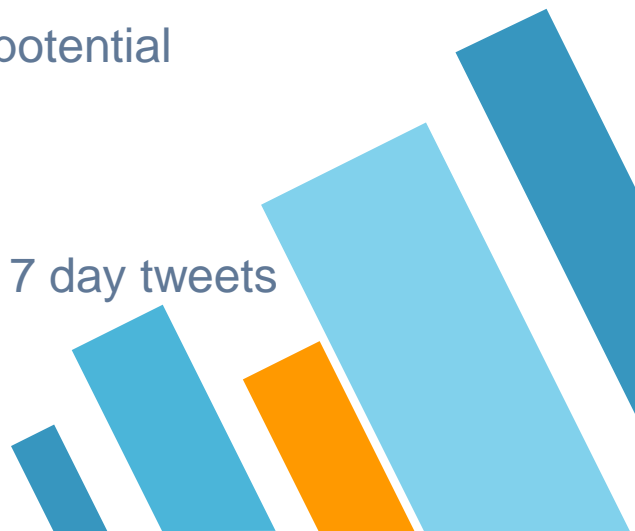


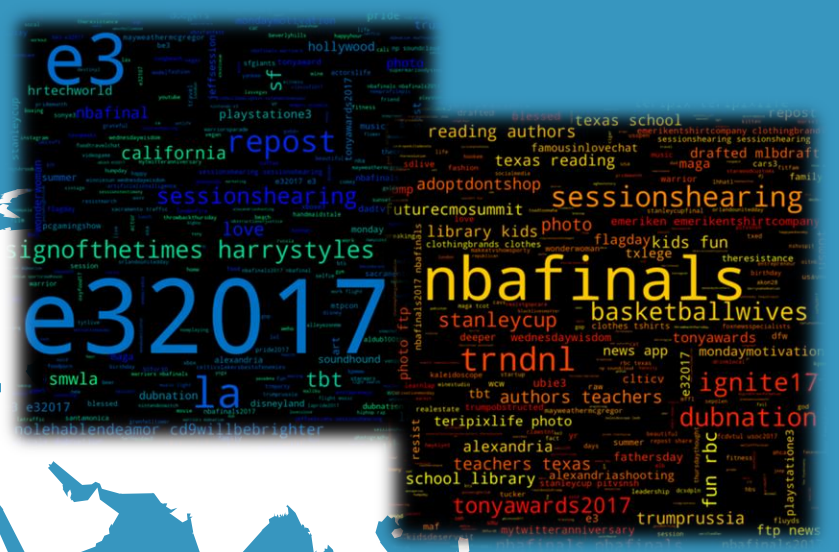
SET UP INFRASTRUCTURE

Created an automated and sustainable **data collection process**

- » Collect and clean tweets real-time
- » Restart the service periodically to clear up potential cache

Automate **NLP process**

- » NLP Pipeline
 - » Daily scheduled process to vectorizer pass 7 day tweets to be ready for quick analysis
- 



Real-Time Process



Pre Process
Data Cleaning

Data
Collection
From
twitter

Scheduled Process | On Demand

Pre Process
SPACY

Pre Process
TFIDF

Pre Process
Count
vectorizer

Pre Process
SVD

Redis

NLP


On Demand

Event Detection
Cosine
Similarity

Trending Detection
Time Series Analysis
ARIMA Modeling



TOPIC MODELING

- In this model, we are trying to extract informational tweets (tweets related to a topic or even) from noise.
 - We also try to find similarities between tweets whether on the same topic or not
 - Several Vectorizers were tried in the experiment.
 - ❑ There is no perfect vectorizer from the shelf
 - ❑ We are settle on TFIDF due to it gives us better similarity score
- 

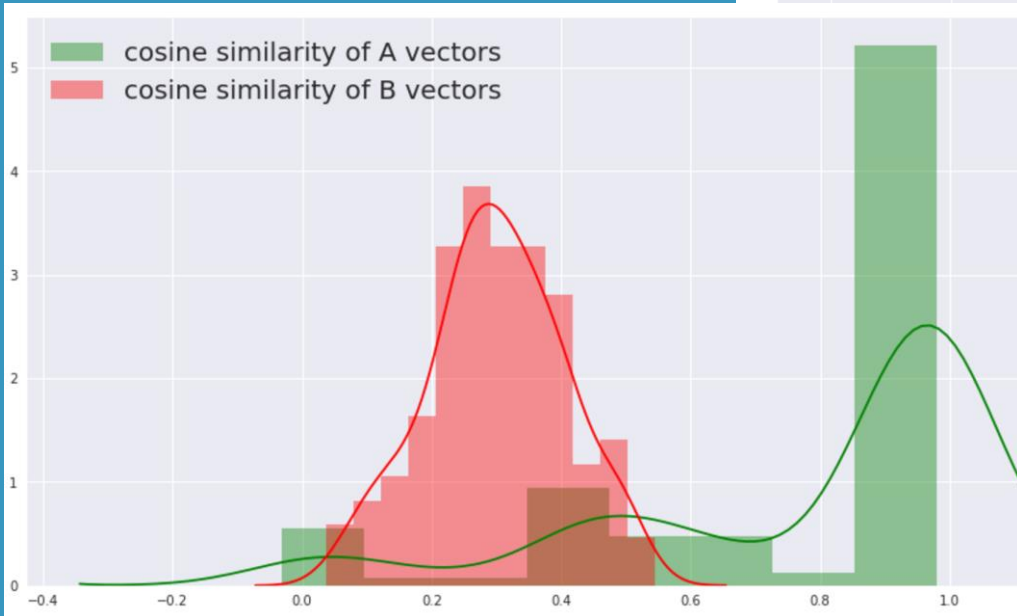
TOPIC MODELING

```
print('Event Tweets (earthquake|quake): ', event_tweet_count('earthquake|quake'))  
A, B = tweets_event_ab_test('earthquake|quake')
```

Event Tweets (earthquake|quake): 1012

Cosine Similarity

[A1_mean A2_mean] :	0.999		mean of [A1_mean A2] :	0.78		STD of [A1_mean A2] :	0.294
[B1_mean B2_mean] :	0.969		mean of [B1_mean B2] :	0.3		STD of [B1_mean B2] :	0.105
[A1_mean B2_mean] :	0.291		mean of [A1_mean B2] :	0.102		STD of [A1_mean B2] :	0.154



Using
COSINE SIMILARITY

EVENT DETECTION

COSINE SIMILARITY USING TFIDF/SVD

```
Top_scored_tweets_in_B('paris|climate', n = 15000)
```

score

0.972668	i'm proud to be a ca resident, ca leads the us and some of the world in pollution standards. we won't give up!
0.972082	duh, what all of the news people have failed to recognize - a coup by trump and russia. if it doesn't...
0.971221	roberts seems to have lost all faith in romo. if this wasn't the spot for him, i don't know what is. only dodger not to pitch in the series.
0.967501	i'm sure a lot of people don't ask for fame. they seek it as regret it after a period of time. rappers especially, it's in their lyrics.
0.966892	this is making me feel less irate. end of the day however trump has 2 go. he's a menace and threatening not on the...
0.966647	i don't even mind working all day. in order to get ahead in life you have to pay your dues! it's all apart of the process. trust in it.
0.966576	actually, i'd take it outside the u.s. pretty sure he's in the top five for worst leaders of any country. ever. jfc.
0.966342	being a man or a woman has nothing to do with #success or #power. both could reach the top of the world if they use their minds - s.ali
0.966090	i do not get it. this show is so fucking important. money is not the issue, they have more of it than god.
0.965310	many people left in this world, sad ! there is a saying: why is it always the good ones go first?
0.965040	#mondaymotivation many of us lived through the bush years, even though we did not vote for him, without obstructing the gov't. why now?
0.964966	it's up to individual states now, people have to lobby their reps. california is not going along with this!...
0.964887	i just wish ppl would see whats in front of their eyes as real, believe it is real and work as hard as they can to fix it and make it better
0.964403	a15: kinda reminds me of a man called 45. all that power and doesn't use it the right way. #mochagirlsread

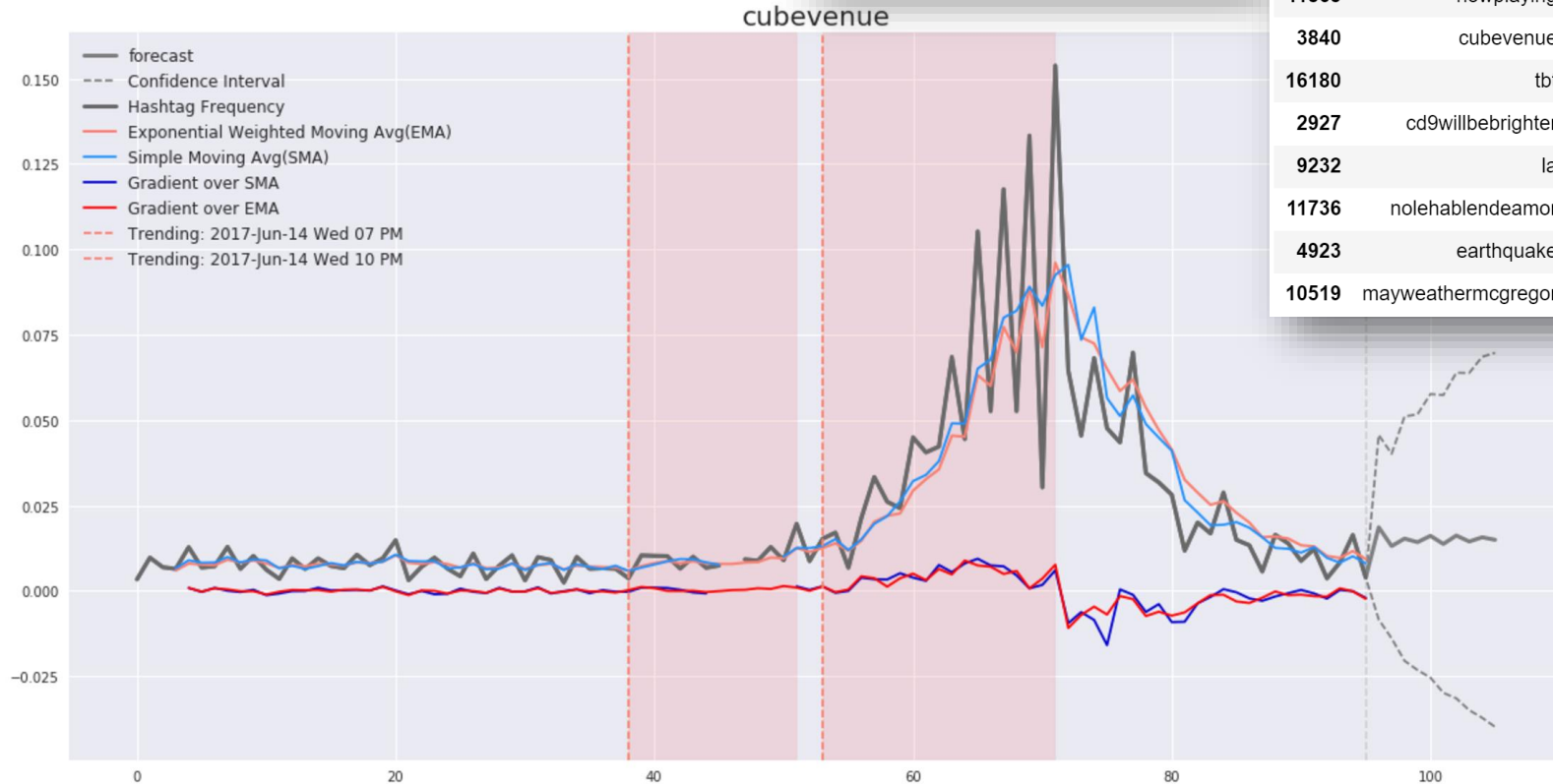
HASHTAG TIME SERIES MODELING

- Trending analysis is highly time dependent.
- We could look at past hashtag (positive or negative) trending pattern to determine event impact for both strength and duration
- We also hope to predict near term topic trend for running commercial or managing social media activities

start time: 2017-06-14 19:02:51
end time: 2017-06-15 23:02:41
total hours: 23
time lag: 0:15:00
time gap: 0:15:00
time windows: 96

hashtag_freq_df.head(10)

	hashtag	occurrences	frequency
4873	e32017	782	0.018867
4870	e3	567	0.013680
11865	nowplaying	255	0.006152
3840	cubevenue	240	0.005790
16180	tbt	231	0.005573
2927	cd9willbebrighter	230	0.005549
9232	la	205	0.004946
11736	nolehablendeamor	200	0.004825
4923	earthquake	177	0.004270
10519	mayweathermcgregor	160	0.003860



e32017

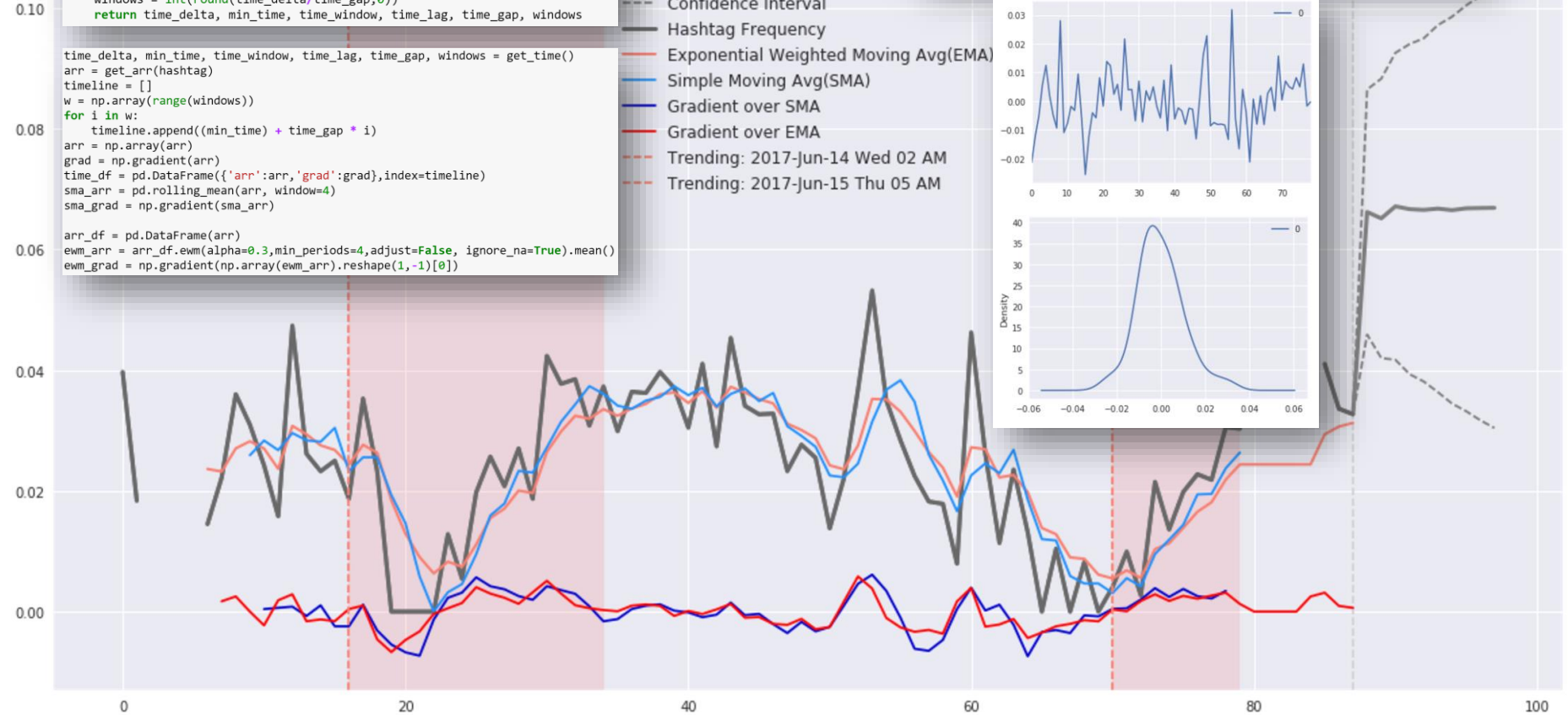
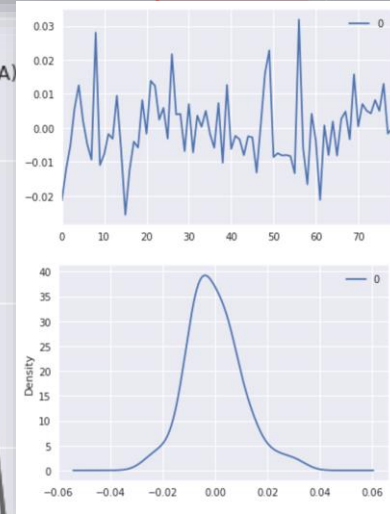
```
def get_time():
    time_delta = (max(df['created_datetime']) - min(df['created_datetime']))
    min_time = min(df['created_datetime'])
    time_window = time_delta.components.days*24 + time_delta.components.hours
    time_lag = timedelta(hours = .5)
    time_gap = timedelta(hours = .5)
    windows = int(round(time_delta/time_gap,0))
    return time_delta, min_time, time_window, time_lag, time_gap, windows
```

```
time_delta, min_time, time_window, time_lag, time_gap, windows = get_time()
arr = get_arr(hashtag)
timeline = []
w = np.array(range(windows))
for i in w:
    timeline.append((min_time) + time_gap * i)
arr = np.array(arr)
grad = np.gradient(arr)
time_df = pd.DataFrame({'arr':arr,'grad':grad},index=timeline)
sma_arr = pd.rolling_mean(arr, window=4)
sma_grad = np.gradient(sma_arr)

arr_df = pd.DataFrame(arr)
ewm_arr = arr_df.ewm(alpha=0.3,min_periods=4,adjust=False, ignore_na=True).mean()
ewm_grad = np.gradient(np.array(ewm_arr).reshape(1,-1)[0])
```

```
from statsmodels.tsa.arima_model import ARIMA
def fit_arima(arr):
    arima_df = pd.DataFrame({'perf':arr})
    series = arima_df.dropna()
    model = ARIMA(np.array(series.values), order=(4,1,0))
    model_fit = model.fit()
    return model_fit
```

- forecast
- - - Confidence Interval
- Hashtag Frequency
- Exponential Weighted Moving Avg(EMA)
- Simple Moving Avg(SMA)
- Gradient over SMA
- Gradient over EMA
- - - Trending: 2017-Jun-14 Wed 02 AM
- - - Trending: 2017-Jun-15 Thu 05 AM



FUTURE WORK

Topic Modeling

- » Explore other vectorizers such as “tweet2vec”

Time Series Analysis

- » Implement custom Weighted Moving Average
- » Find optimal learning constant (α) for EMV

Geospatial Modeling

- » Expand geographical coverage to more cities and states



THANK YOU

You can find me @

GitHub: **joyce-lin**

Linkedin: **joyce-ch-lin**

Gmail: joyce.ch.lin@gmail.com

GA Profile:

