The News that Shifts the Value of Cryptocurrency

Dongming Jin, dongming.jin@mavs.uta.edu
University of Texas Arlington

BACKGROUND

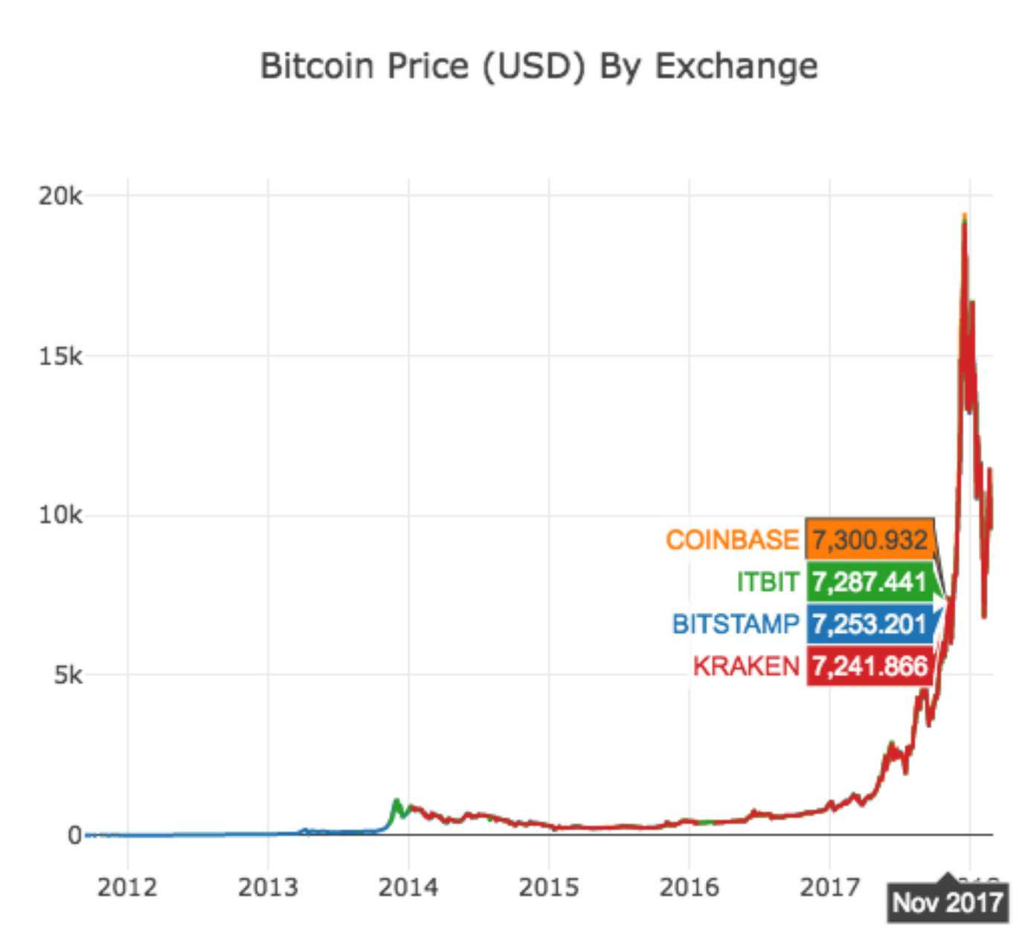
Cryptocurrency is a concept of digital currency in which encryption techniques are used to regular the generation and verify the transfer of funds based on blockchain, a distributed ledger that is inherently resistant to modification of the data.

Ever Since its birth, its nature and value has been highly debated. It is the ideal digital asset in the world of internet, that is decentralized and inherently resistant to modification of the ownership. Even though the nature and value has been highly debated, the combination of security and transparency makes it one of the most important innovation in the era of 'cloud data', where security is the last shield of privacy.

DATA SUMMARY

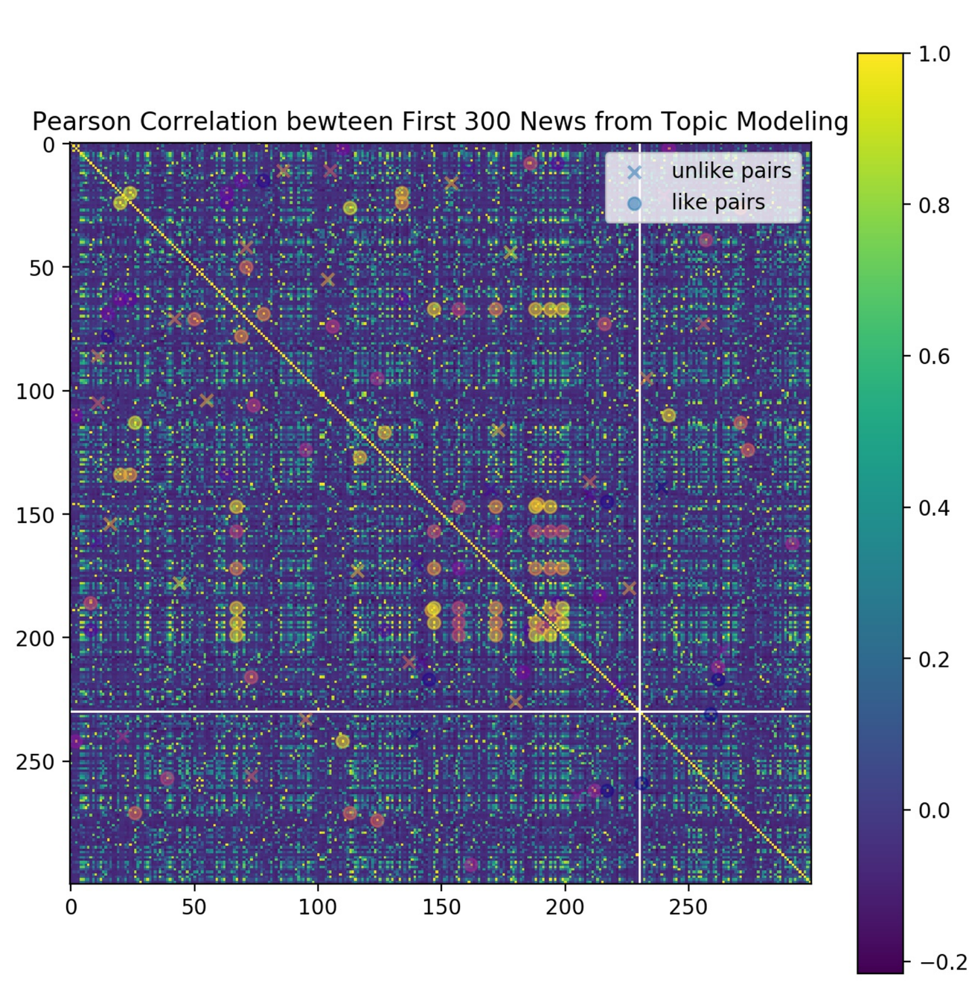
Data

- Bitcoin Price and Volume
  - Kraken, Coinbase, Bitstamp and Itbit; 2012/09 to present
- News Archive
  - 2017 Wall Street Journal, 43,268 valid entries
- Corpus
  - NTLK corpus
  - Wikipedia, 669,539 vocabulary with 35,556,952 documents



The gap in the red curve is due to the maintenance of Kraken exchange.
Due to various reason, the Bitcoin data recorded from four exchanges contains faulty and missing information. Thus, the average Price and Volume are calculated based on the best available and continuous data.
The changes in 2017 are primary used to be consistent with News obtained from WSJ archive.

	BITSTAMP	COINBASE	ITBIT	KRAKEN
Online Days	2339	902	1538	1465



The correlation indicates how much two news articles differ in topics. The white line is caused by special news column named "Pepper… and Salt", which is uncorrelated with topics of all other news. The like pairs are the news in similar topics.

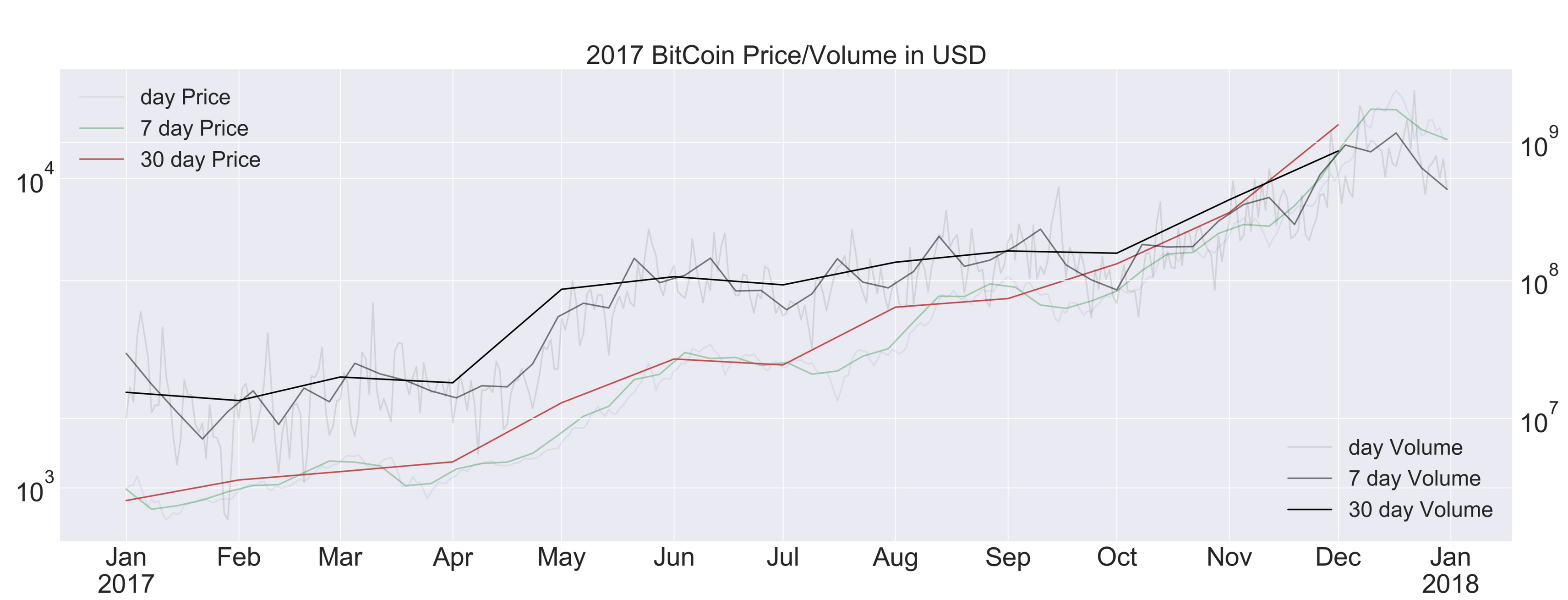
Unlike pair example:

- News id: 23, Riders and Residents Greet Second Avenue Subway's Opening The new Second Avenue Subway was greeted on Monday by Upper East Side residents who welcomed the end of construction and riders who looked forward to easier, less-congested commutes.
- News id: 119, Alcoa to Permanently Close Suriname Alumina Refinery and Bauxite Mines Alcoa will permanently close the Suralco alumina refinery and bauxite mines in Suriname, which were fully curtailed in November 2015 amid a limited bauxite supply and unfavorable market conditions.

Like pair example:

- News id: 8, China Tries to Recalibrate Credit Chinese officials need to tread carefully in the coming months to tighten monetary conditions, curb risky investment practices and keep the yuan from depreciating too far while keeping China's economy healthy.
- News id: 186, Boutique Hotels Bet on China Even as Nation's Glut of Rooms Worsens Flashy hotels targeting the young and hip see opportunity in China even as the country is experiencing an overall glut of hotels that has reduced average room rates.

The examples demonstrate some degree of efficiency in separating news topics by TFIDF and NMF.



METHODS

Method

- Sentiment
  - Textblob: pre-trained on sentimental corpus
- Topic model
  - TF-IDF: term frequency and inverse document frequency
  - NMF: non-negative matrix factorization
- Semantic model
  - Doc2Vec: two layer neural network with vector representation

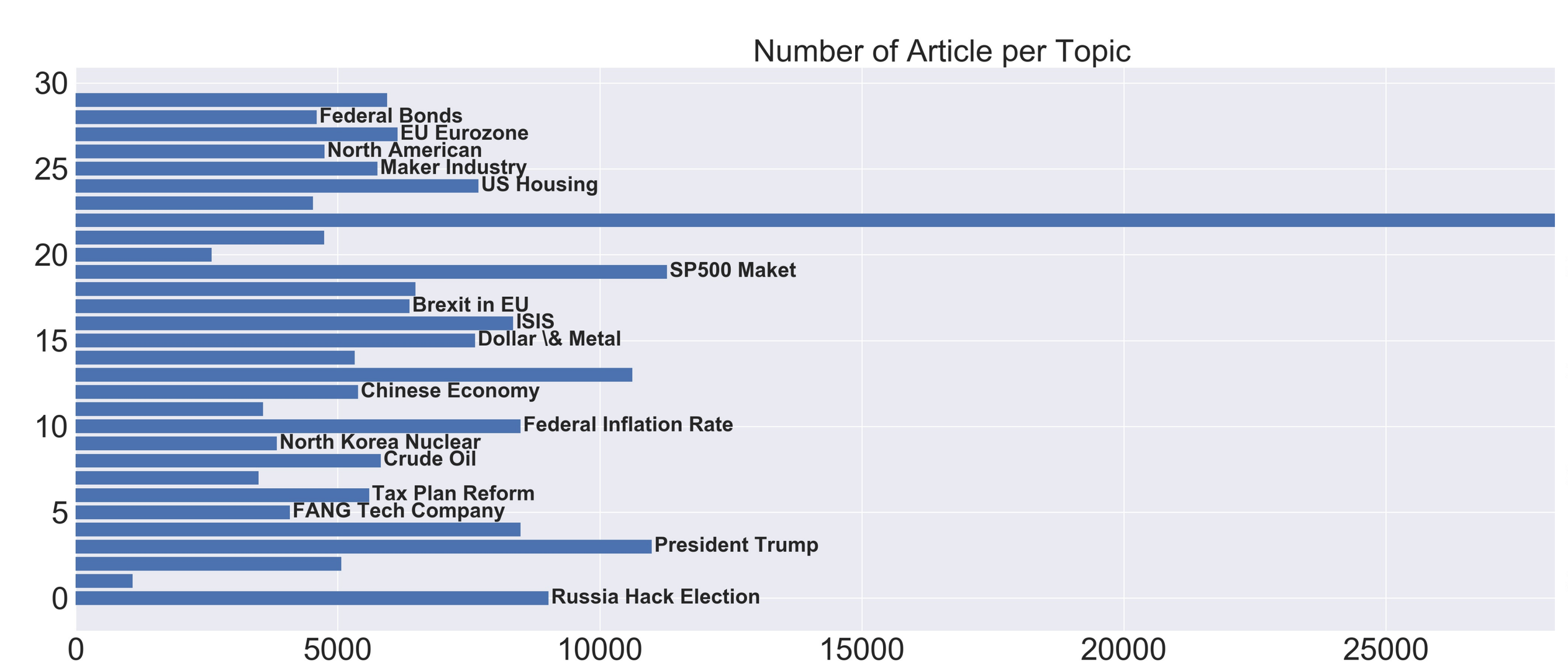
$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

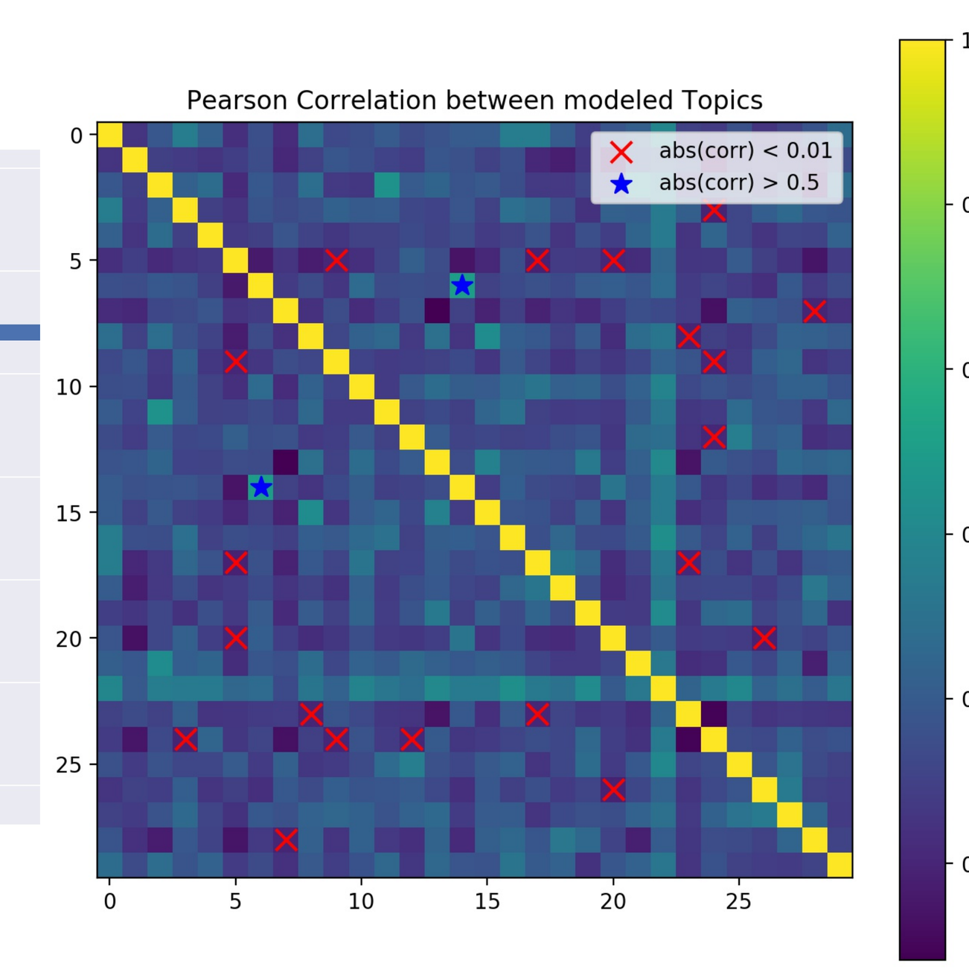
$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

Topic #0	Topic #1	Topic #2	Topic #3	Topic #4	Topic #5	Topic #6	Topic #7	Topic #8	Topic #9	Topic #10	Topic #11	Topic #12	Topic #13	Topic #14
9022	1086	5066	10989	8486	4087	5605	3489	5819	3836	8491	3572	5389	10620	5321
russia	corrections	gerard	trump	new	news	tax	review	oil	korea	fed	photos	china	quarter	health
election	amplifications	baker	president	york	media	plan	outlook	prices	north	federal	day	chinese	earnings	care
russian	edition	editor	donald	jersey	marketing	overhaul	reform	opec	south	reserve	editors	beijing	profit	gop
presidential	2017	scoops	administration	cuomo	advertising	gop	trump	production	korean	rates	wall	xi	sales	act
senate	jan	guided	obama	gov	cmo	reform	obama	crude	missile	rate	journal	world	revenue	affordable
putin	oct	stories	immigration	art	roundup	republicans	reviews	energy	nuclear	inflation	selected	jinping	watch	senate
probe	aug	tour	order	andrew	happenings	senate	obamacare	output	pyongyang	yellen	street	economy	growth	republicans
campaign	sept	point	policy	times	today	code	gop	gas	sanctions	raise	chosen	global	results	insurance
2016	nov	personal	said	office	morning	taxes	good	saudi	kim	economy	sunday	yuan	scheduled	republican
french	july	10	elect	book	industry	corporate	political	week	military	policy	thursday	foreign	market	repeal

Topic #15	Topic #16	Topic #17	Topic #18	Topic #19	Topic #20	Topic #21	Topic #22	Topic #23	Topic #24	Topic #25	Topic #26	Topic #27	Topic #28	Topic #29
7618	8347	6368	6487	11284	2591	4742	31490	4528	7688	5761	4747	6143	4599	5941
dollar	state	brexit	billion	investors	house	best	company	court	million	street	trade	bank	government	city
gold	islamic	european	debt	stocks	white	books	chief	supreme	home	heard	free	central	bonds	york
prices	tillerson	eu	deal	market	republicans	week	business	judge	estate	wall	mexico	ecb	bond	mayor
data	syria	union	treasury	stock	republican	web	executive	appeals	real	big	canada	banks	yields	police
rose	forces	minister	buy	markets	security	selling	said	gorsuch	property	banks	nafta	policy	treasurys	blasio
higher	iraq	prime	sell	shares	adviser	ended	ceo	case	pay	apple	american	european	prices	nyc
fell	military	theresa	auction	funds	director	data	world	ban	homes	growth	deficit	rate	yield	council
copper	said	britain	comprising	dow	staff	bookscan	financial	ruling	lists	good	agreement	inflation	strengthened	officials
month	iraqi	europe	fund	rally	gop	npd	companies	travel	settle	investors	surplus	monetary	debt	schools
weaker	attack	talks	sold	bond	said	children	year	justice	asks	main	exports	eurozone	note	man





Topic Modeling:

1. The popular Brown corpus from NLTK is used to exclude the stop words, in additional to the common stop words in English.
2. The rest of the words are built into a vocabulary and all of the 2017 WSJ news are then factorized based on the vocabulary.
3. TF-IDF is adopted to generate count statistics. Top 20 weighted key words are saved to summarize each topics.
4. We use NMF to decompose news vectors into 30 dimensions and convert the key words back. See the top 10 of each topic in table.

A few example on topic #9 North Korea Nuclear:

- 20525 South Korea: Stance on U.S. Missile-Defense System Hasn't Shifted A top South Korean national-security official sought to tamp down concerns about the government's commitment to a controversial U....
- 18901 U.S. Plans Missile Defense Test Amid Concern Over North Korea The Pentagon is planning to conduct a test next week of a system designed to shoot down intercontinental missiles, U.S. defense offici...
- 14274 North Korean Nuclear Threat Fuels Outsider's Election Hopes in South Hong Joon-pyo has found hope in escalating tensions with Pyongyang over its nuclear program that have changed the dynamics of t...

The news are grouped by the topics and then aggregated to each day of the year. The average polarity is taken to indicate the sentiment of the day. A few examples are listed.

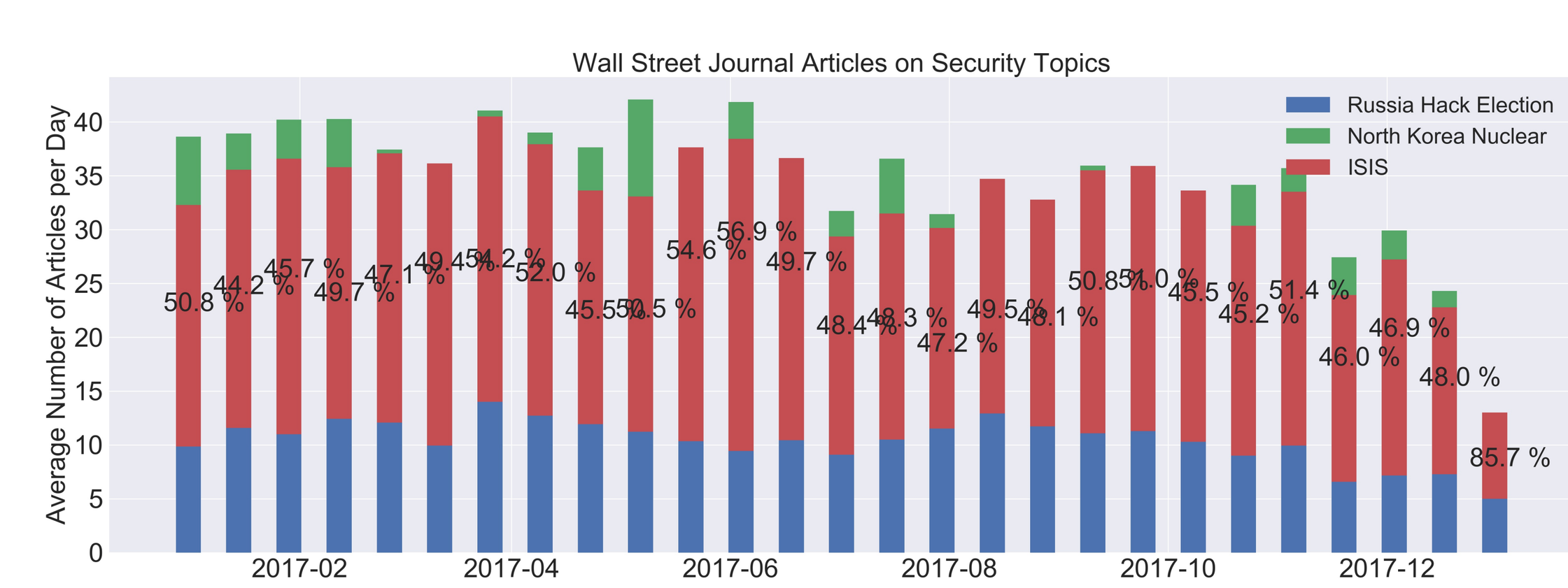
News with the lowest 1% polarity are:

- 2036 Tomlin Calls Brown's Postgame Video 'Selfish' The Steelers coach expresses regret over Facebook post that caught Tomlin disparaging Patriots
- 34279 It's Time to End Ineffective And Cruel Animal Testing Testing drugs on animals is notoriously ineffective.
- 9694 Process of Probate Is Used Often for Planned Estates Probate gets a bad name based on the relatively few estates that go through ugly and sometimes years-long will contests in probate courts.

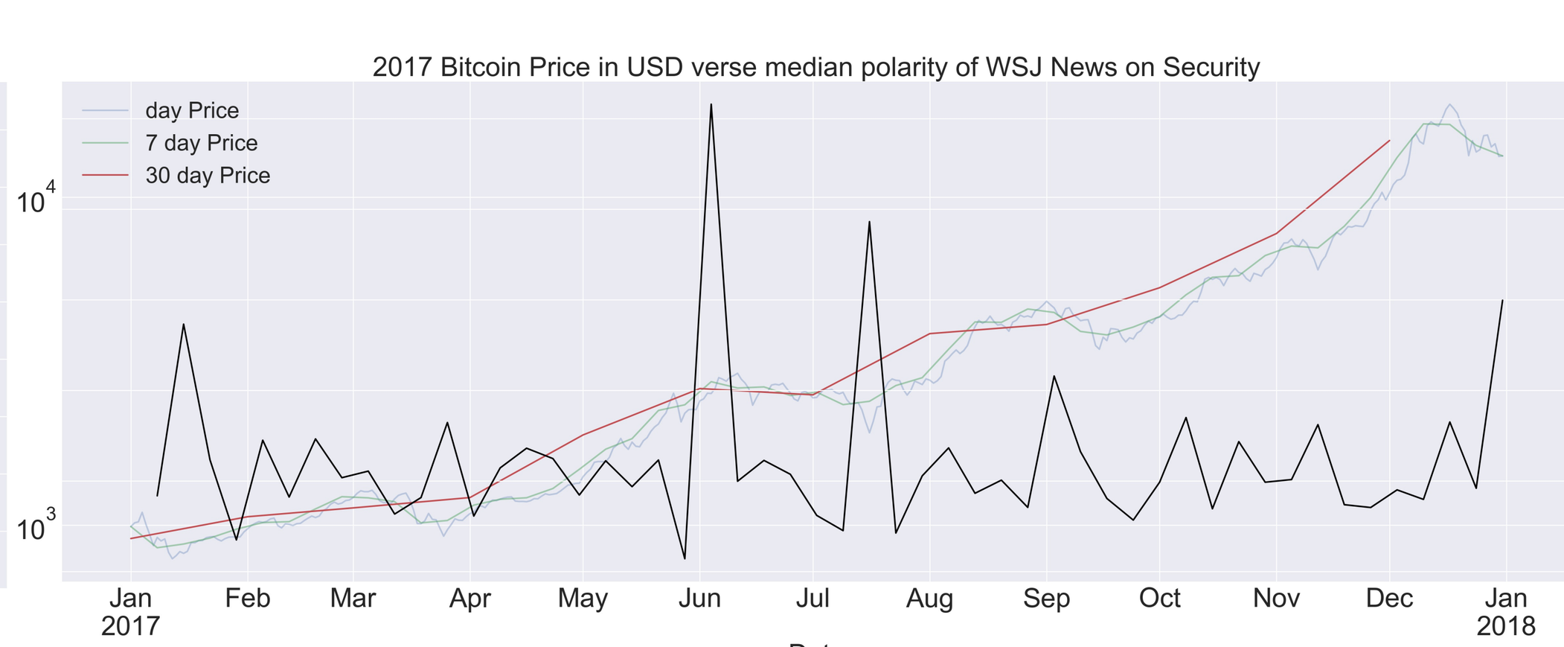
News with highest 1% polarity are:

- 25536 The Voice of Kenya Sam Sacks on Ngugi wa Thiong'o, a perennial favorite for the Nobel Prize and one of Africa's greatest writers.
- 15364 Verizon Invests in Self-Driving Car Startup Renovo Verizon has invested in Renovo Auto, an autonomous-vehicle tech startup best known for turning a vintage DeLorean into a self-driving car.

As a result, higher correlation means topics are having similar polarities throughout 2017, which means the topics are more alike. The markers highlight such alike and unlike topics as reference to determine topic categories.



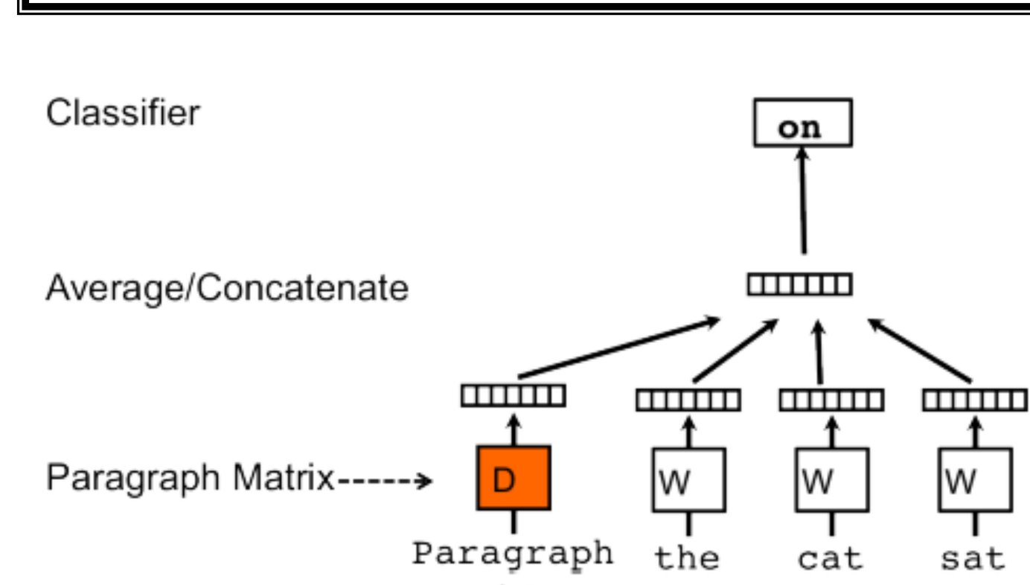
We notice the average daily amounts of news articles across all topics are declining over the year. There are three topics in the security catalog, covering the updates about hack activities exposed during US president election, global security concern about North Korea nuclear weapon development and terrorism of ISIS. The marked numbers show the relative percentage of security related news across all the daily articles. Each news article contains multiple topic tags.



We suspect negative news will have opposite effect on the price change of Bitcoin. To inspect the hypotheses, the median polarity carried out by sentiment analysis is computed and averaged over a period of 7 days to smooth out the violence. It is hard to directly conclude on any correlation by eye. There is also a concerning underling fact from the graph. The average polarity is biased, which indicates the WSJ News unconsciously prefer positive signals to its reader.

Doc2Vec

- Representation
  - Vector dimension: 300 float
- Architectures
  - Distributed bag-of-words: window size 15
- Softmax pulling
  - Negative sampling: 5 noise words will be drawn
- Learning algorithm
  - Gradient descent



The stripe patterns of unlike pairs are caused by special news columns, which is uncorrelated with all other news. The like pairs are the news with similar semantics.

Most frequent in unlike pairs:

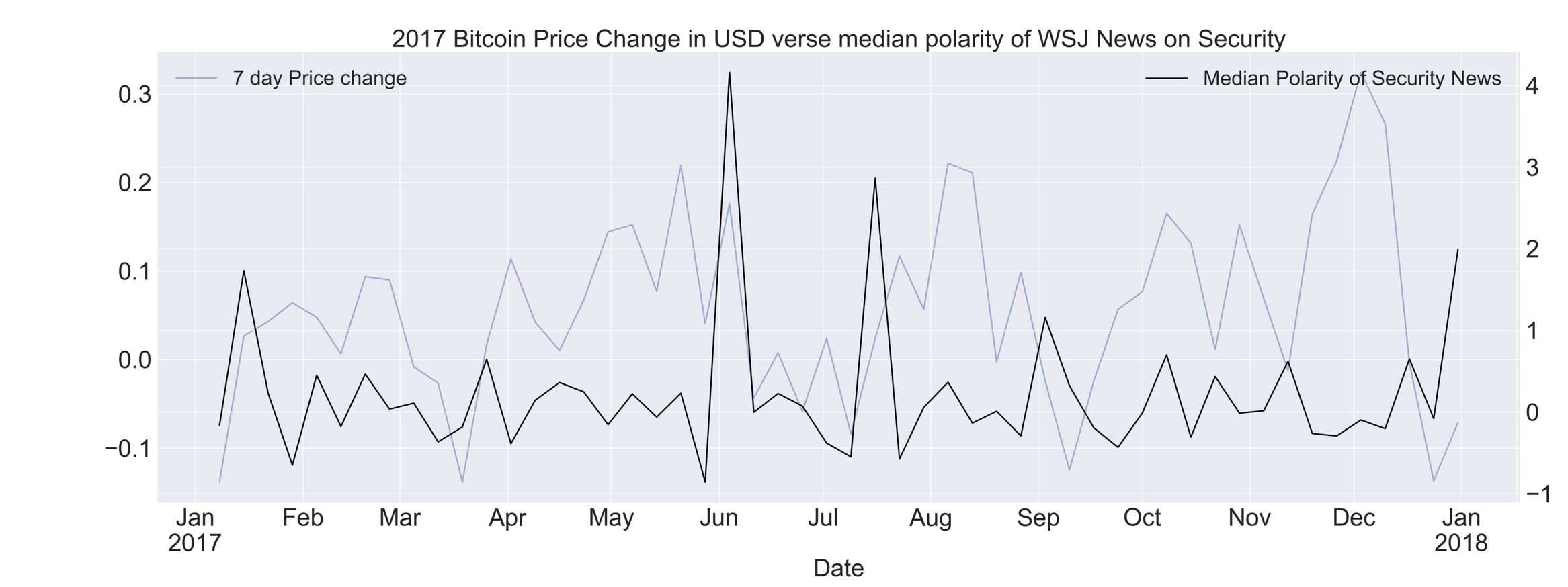
- News id: 230, 299 times, Pepper…and Salt
- News id: 52, 297 times, What's News: World-Wide
- News id: 94, 266 times, Business Watch
- News id: 0, 4 times, Europe's Fate in 2017: Caught…

Like pair example:

- News id: 86, Brazil Prison Biot Leaves Scores Dead A flight between rival drug gangs sparked a prison riot that left about 60 inmates dead in the Brazilian city of Manaus, according to local authorities.
- News id: 137, Brazil Hunts for Scores of Escaped Inmates More than 130 escaped prisoners in Brazil were on the run Tuesday after breaking out during a bloody prison riot that left 60 inmates dead and sparked criticism of chronic overcrowding in the country's jails.

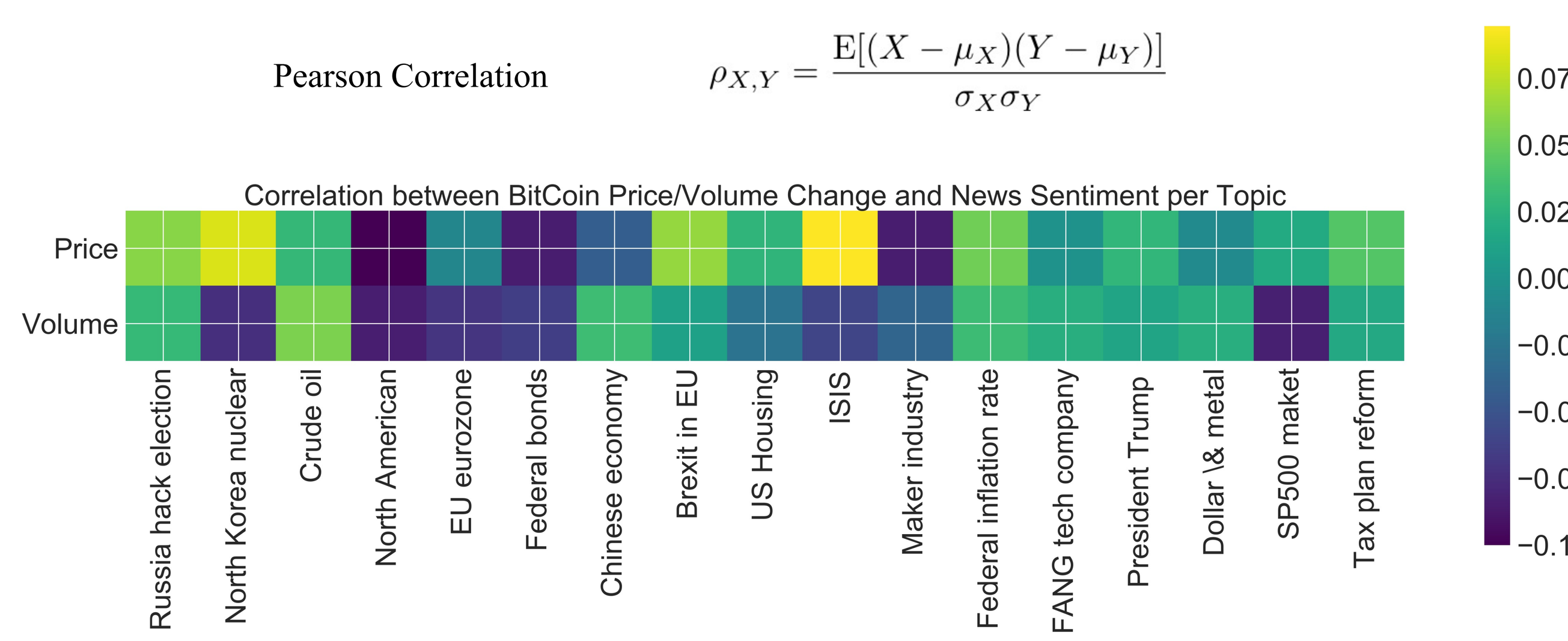
In comparison, Doc2Vec captured two more special news columns and the examples demonstrate semantic understanding from the model.

RESULTs



As we can tell from the graph, it is subjective to draw any conclusion about correlation between the 7 day price change and the 7 day average of the daily median polarity for news in security topics.
To quantitatively describe the correlation between Bitcoin and WSJ news in 2017, the total sentiment of a news topic is defined to be the median value of the normalized polarity to minimize the bias we find in WSJ news. Then we used Pearson's equation to calculate the correlation efficient between the Price/Volume change and the daily news sentiment in each topic.

Pearson Correlation
$$\rho_{X,Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$



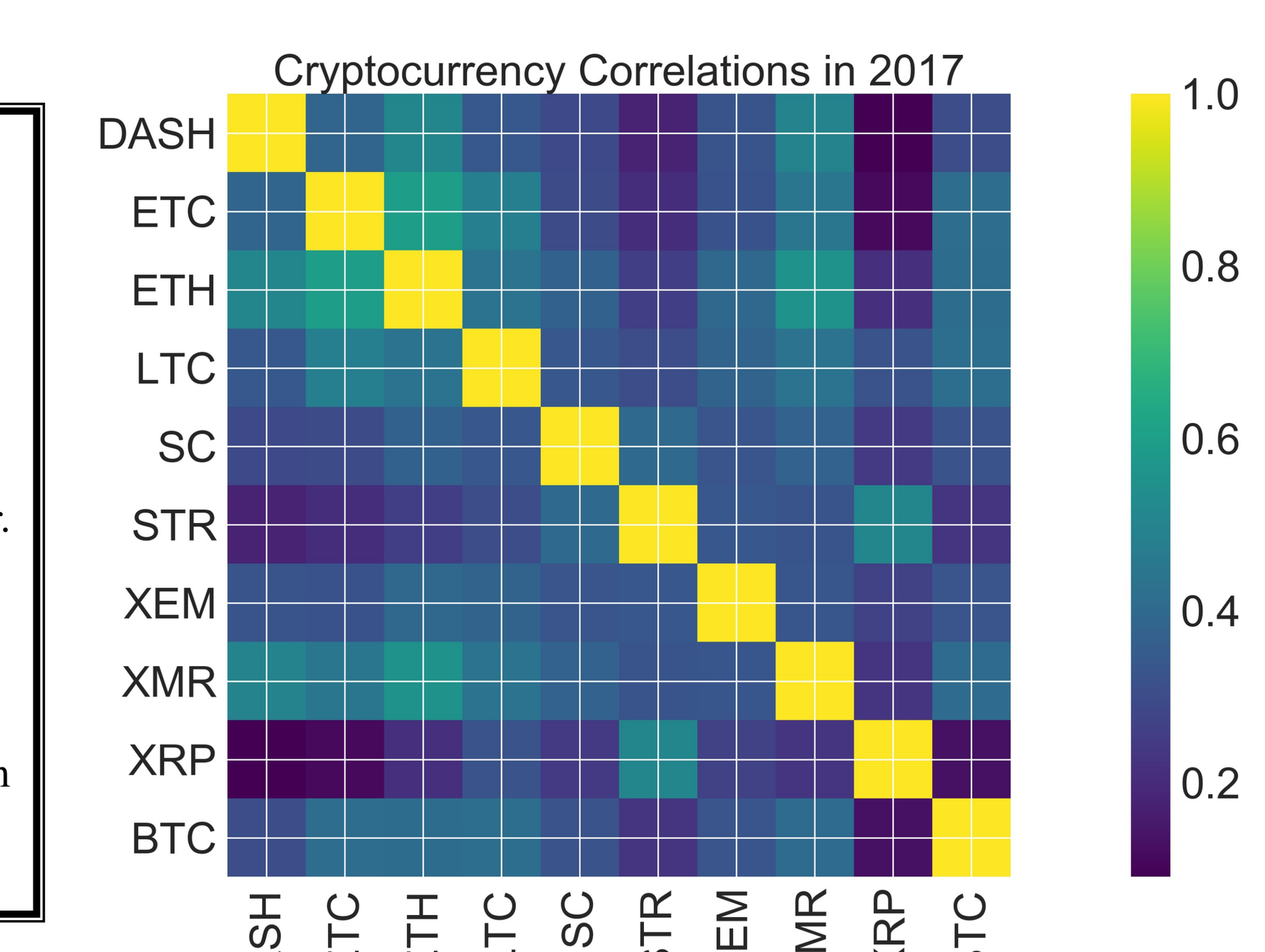
The strongest correlation is found between the Bitcoin price change and the sentiment of ISIS related news, which indicates that fear of terrorism might be the potential driving force of Bitcoin price increase.
The strongest negative correlation between the Bitcoin price increase and free trade policy in North American may also be well explained as the Bitcoin community is less favored in central government and regulations.
<u>But all of which depends on the less efficient topic model and the scale of the color voided the foundation to make any reliable conclusion.</u>

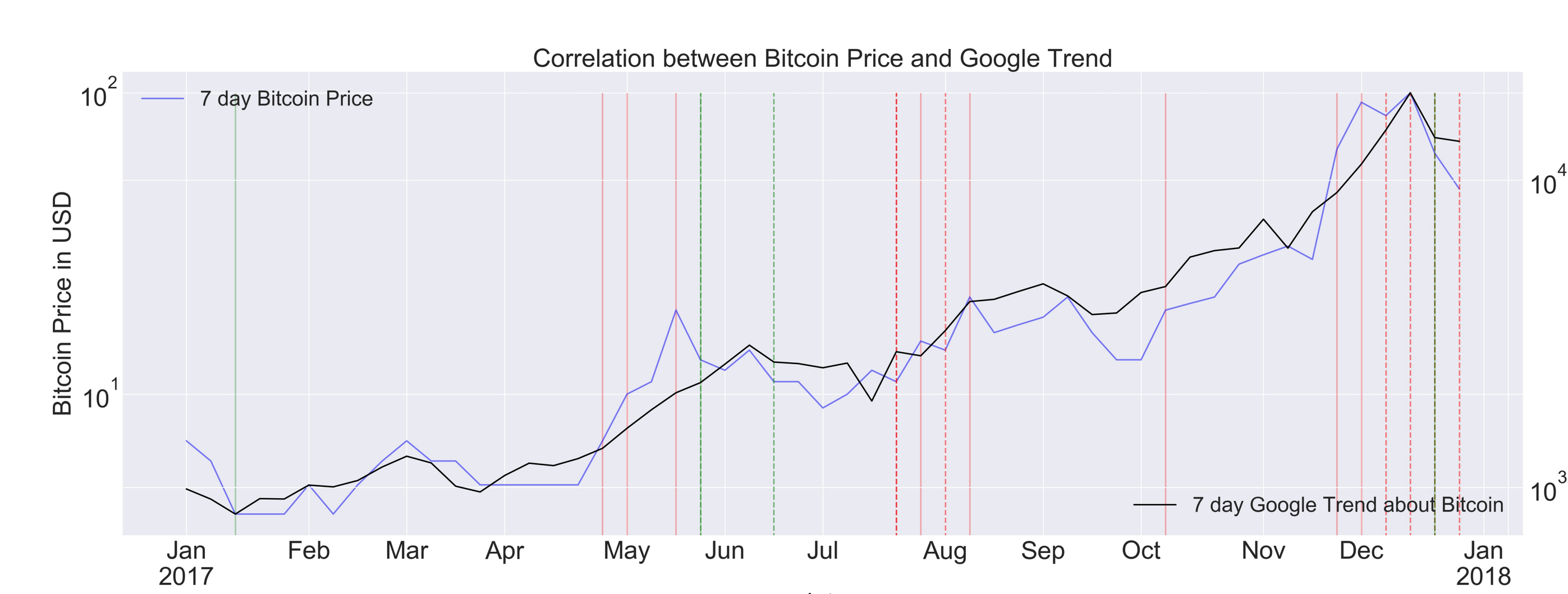
NEW HOPE

Conclusions

As we demonstrated,

1. Doc2Vec can model news semantic more precise and accurate than TFIDF.
2. WSJ News are biased on sentiment and topics. Google Trend and Twitter will be better.
3. It is too early and vague to connect Bitcoin with unlawful crime and security exposure.
4. Correlation between different cryptocurrency exists and worth exploring.





Ramage, Daniel, Susan T. Dumais, and Daniel J. Liebling. "Characterizing microblogs with topic models." ICWSM 10, no. 1 (2010): 16.
Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." In International Conference on Machine Learning, pp. 1188-1196. 2014.