NTNU

Norwegian University of Science and Technology

Analyzing the DarkNetMarkets Subreddit for Evolutions of Tools and Trends Using Latent Dirichlet Allocation

DFRWS USA 2018

Kyle Porter

The DarkWeb and Darknet Markets

- The darkweb are websites which can only be accessed through anonymity networks such as Tor.
- Well known for hosting online criminal market places, “Darknet Markets”.
  - Vendors use markets as a platform for selling drugs, weapons, malware, and other illicit items.
- In July 2017, there were significant law enforcement busts that shut down the two most popular markets, Hansa and Alphabay.

This Research

- How can we relatively quickly understand how the bust affected the markets, users, and tools of DNM users?
- Let’s crawl for a years worth of content from a darknet market oriented subreddit (forum), called “darknetmarkets”.
- That’s a of information, let’s try topic modelling on each month of the extracted content.
- The topics produced can be used as a sort of text summarization for a large number of documents.

Topic Modeling?

- Given a corpus of documents and N topics, a Latent Dirichlet Allocation algorithm can generate N topics that the corpus is composed of.
- Trivial topic modelling example:
  - Topic 1: {dog, leash, kibble, walk, cat, …},
    Topic 2: {Trees, nature, walk, park, …}
- These topic-word distributions are one of the latent items learn that we learn.
  - Which we use for our work.

How can we use these topics?

- To quickly see how things changed from pre-bust to post-bust.
- To understand criminal community
- To Identify useful keywords in generated topics (tools, vendors, markets, etc).
  - Hopefully data pops out at us.

Caveats to this story (1)

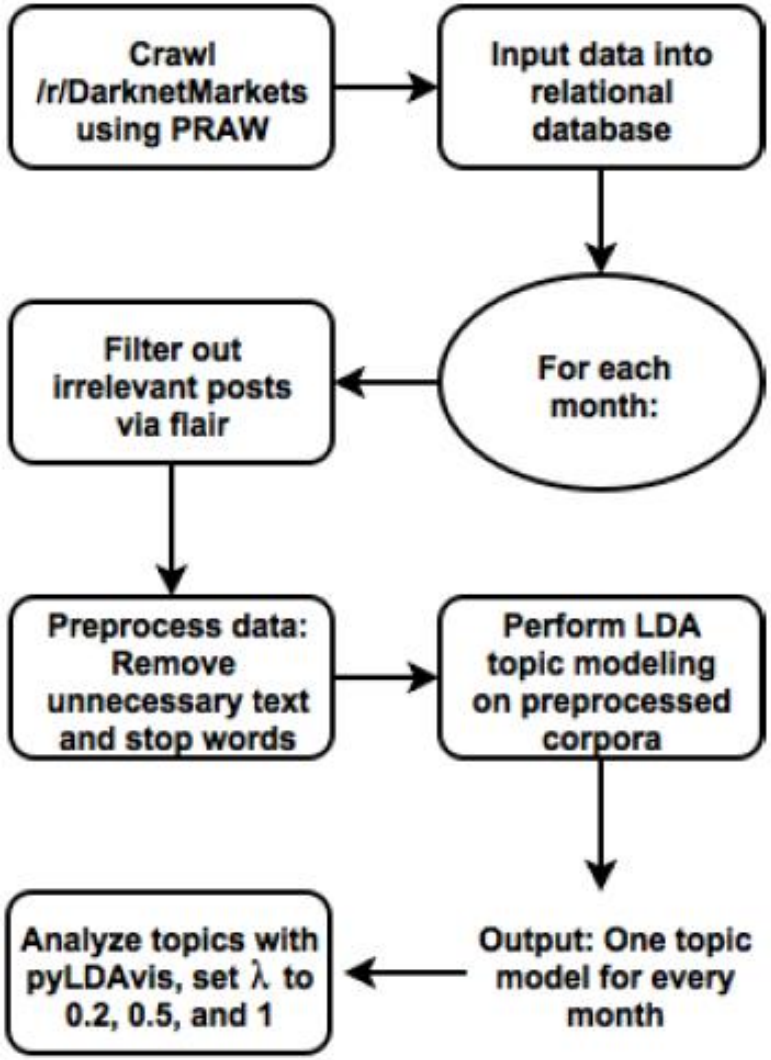
- The DarknetMarkets subreddit was banned by Reddit in March 2018.

https://motherboard.vice.com/en_us/article/ne9v5k/reddit-bans-subreddits-dark-web-drug-markets-and-guns

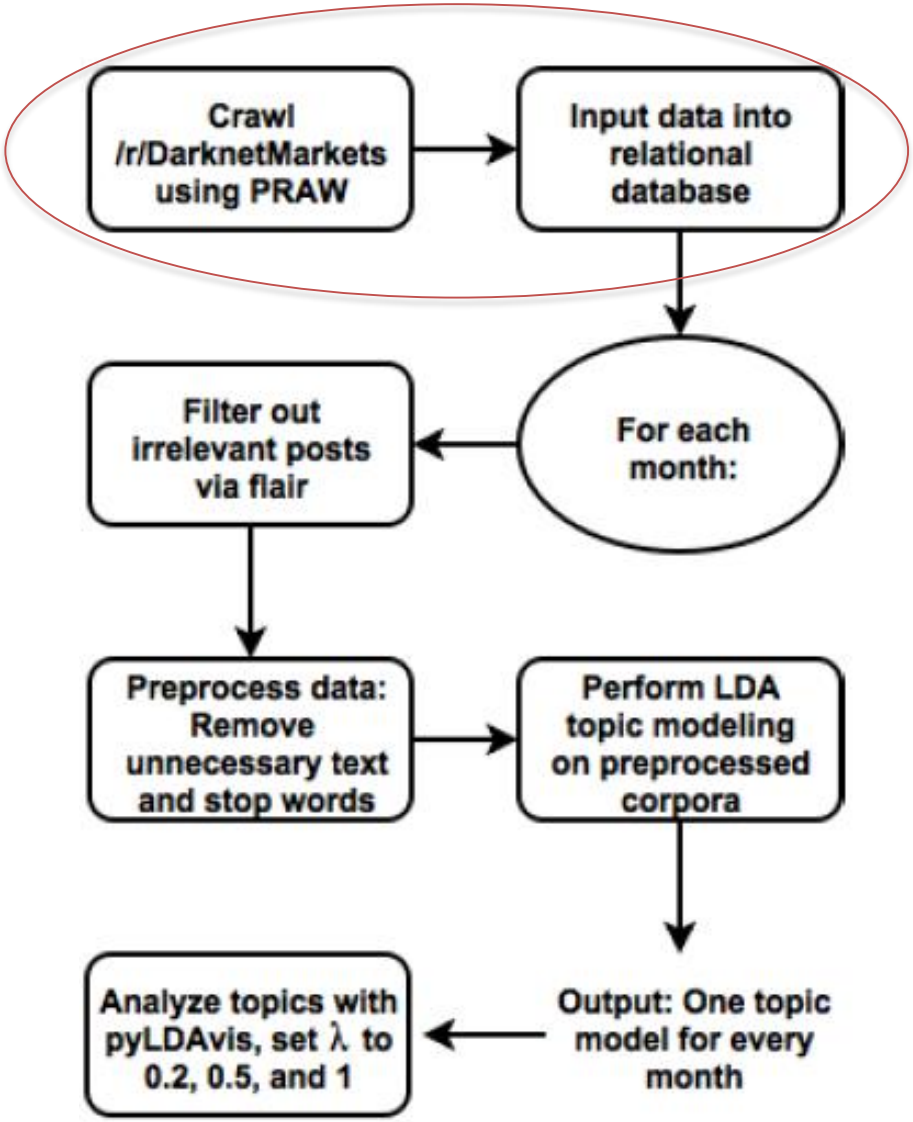
Caveats to this story (2)

- The Reddit Search API no longer allows searching historical posts via timestamps.
- Therefore, PRAW (Python Reddit API Wrapper) cannot get historical data.
  - PushShift API can potentially serve as an alternative?
- Our findings still have potential uses:
  - If you happen to already have Reddit corpora, or build it over time, then the analysis we did here is still possible.

Experimental Outline

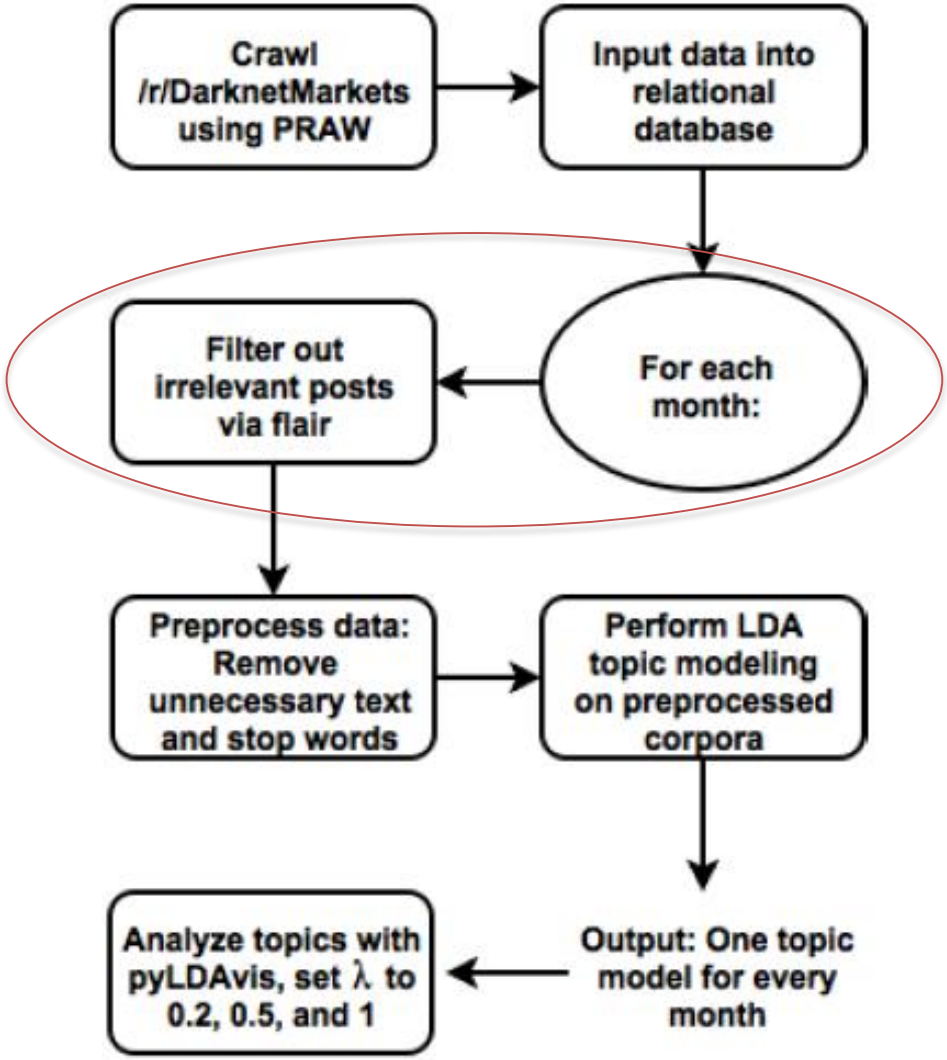


Experimental Outline

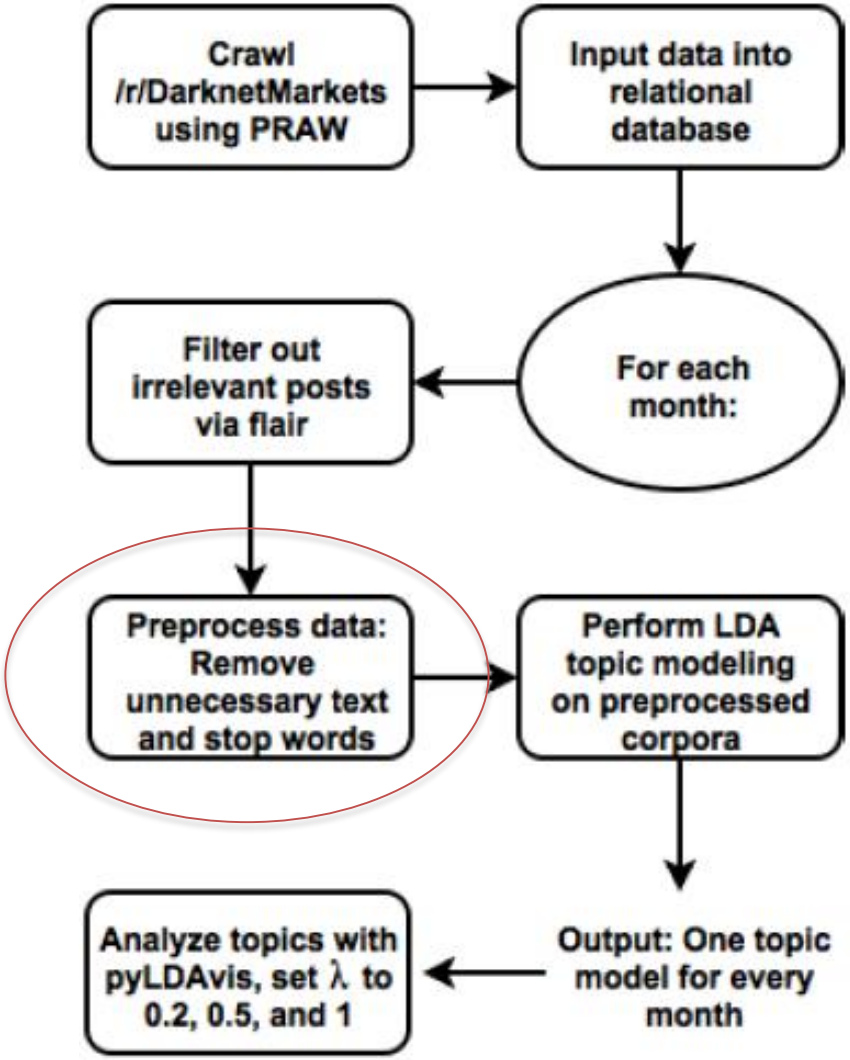


*https://praw.readthedocs.io/en/latest

Experimental Outline

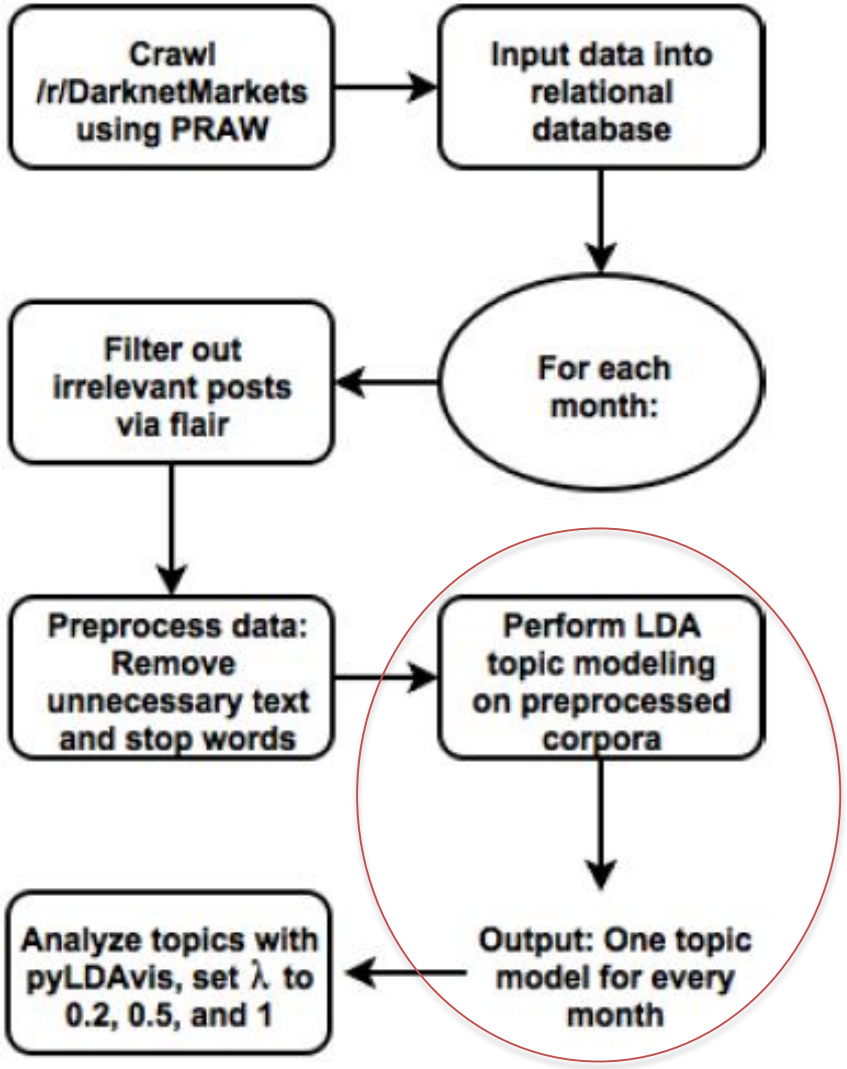


Experimental Outline



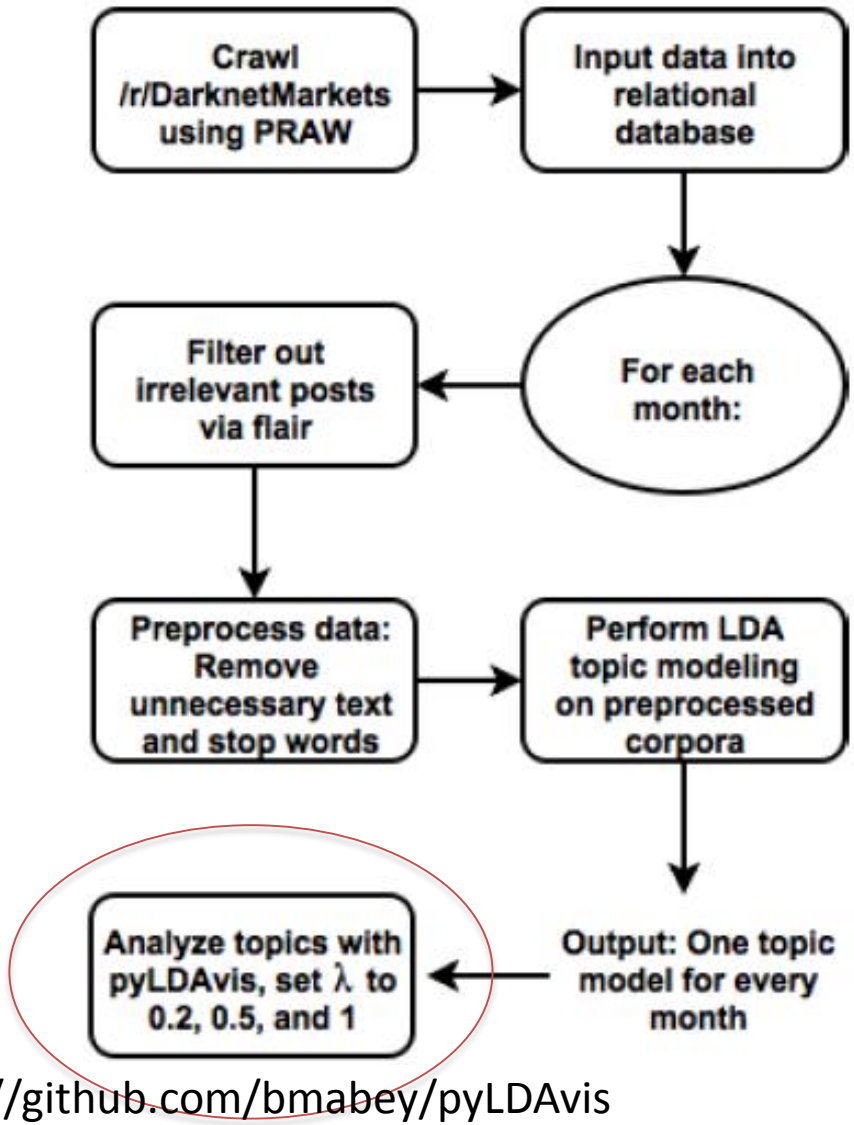
*Standard stop-words from Python NLTK libary (http://www.nltk.org/)

Experimental Outline



*LDA using gensim https://radimrehurek.com/gensim/

Experimental Outline



*pyLDAvis: https://github.com/bmabey/pyLDAvis

Relevancy Metric (1)

- After generating the topic model of a corpus, we can adjust the weight λ to influence word ranking per topic according to relevance.

$$\text{rel}(\text{term w} \mid \text{topic t}) = \lambda * p(w \mid t) + (1 - \lambda) * p(w \mid t)/p(w)$$

- λ = 1 is standard ranking (conditional probability of word given a topic). As λ approaches 0, words with high overall probability are ranked lower.
- We set lambda to λ = 1, 0.5, 0.2 to explore topics.

Results

- In general, the topics did not change significantly from month to month.
  - Largest topics were usually discussions about vendors/markets.
  - Cryptocurrency usually was its own topic
  - Security/anonymity was not always a topic
  - If a news story was large enough, it usually ended up as a topic
- The significant changes were the topic-word distributions.

Results (General)

- General state of the DNM (from the view of Reddit users) went from relatively casual to concerned, uncertain, and more security-minded after the July 2017 busts.
- In particular, we saw in increase in the use of law enforcement terms

Results (May 2017)

1	General vendor discussion	$\lambda = 1$: vendor, order, time, ordered, pack, vendors, never, days, lol, said, name, deleted, dispute, package, drugs, day, anything, back, orders, feedback $\lambda = 0.5$: order, vendor, dispute, ordered, pack, packs, feedback, vendors, bars, orders, name, box, mail, lol, ordering, never, said, xanax, package, house

Results (July 2017)

1	Markets and vendors	$\lambda = 1$: market, vendor, vendors, **hansa**, **dream**, time, **le**, **pgp**, order, key, **ab**, markets, deleted, never, new, back, address, used, 2, message $\lambda = 0.2$: vendors, **dream**, key, **pgp**, **compromised**, **le**, vendor, **hansa**, market, encrypt, order, buyers, auto, **2fa**, address, markets, **escrow**, **ab**, orders, decrypt

Results (August 2017)

1	Markets and vendors	$\lambda = 1$: market, vendor, **dream**, vendors, time, order, markets, **le**, never, account, 2, **pgp**, site, back, post, **hansa**, deleted, address, money, anyone $\lambda = 0.2$: market, **dream**, **multisig**, markets, password, site, **hansa**, vendors, **sig**, support, **scam**, **compromised**, **phishing**, **2fa**, **multi**, **sourcery**, **le**, passwords, admin, **escrow**

Results (October 2017)

1	Markets and vendors	$\lambda = 1$: market, vendor, **dream**, vendors, time, markets, never, money, **tr**, back, order, **scam**, **le**, **exit**, new, **aero**, site, reddit, 2, something $\lambda = 0.5$: market, **dream**, vendors, markets, money, time, **scam**, **exit**, **tr**, never, back, **le**, site, **aero**, order, **escrow**, everyone, coins, **btc**

Tools: Cryptocurrency (July 2017)

- Popular cryptocurrencies are Bitcoin and Monero.
- Identified popular mixing services and cryptocurrency exchanges.

7	Cryptocurr.	$\lambda = 0.2$: **bitcoin**, aktif, **bcc**, blockchain, **fork**, **coinbase**, bch, **btc**, **monero**, exchanges, **electrum**, **bitmixer**, **localbitcoins**, wallet, **shapeshift**, **tumbling**, **helix**, aug, **tumble**, uahf

Tools: Anonymity (March 2017)

- Common operating system is Tails (all software configured to connect to internet through Tor).
- Common use of VPN, and PGP.

5	Anonymity, drugs, and general conversation	$\lambda = 1$: **tor**, **tails**, time, deleted, 2, cocaine, **vpn**, first, 1, new, used, every, 3, 5, 0, around, water, **cia**, work, man $\lambda = 0.2$: **tails**, **vpn**, **isp**, **usb**, **browser**, **miners**, **tor**, **wpa**, **windows**, **fingerprinting**, **qubes**, **vpns**, cia, **reaver**, **xiopan**, **linux**, persistence, **veracrypt**, filter, democrats

Tools: General

- Tools did not seem to evolve.
- The only trend in tool use we could see is that they become more popular in discussion when real world events (busts/exit-scams/bit-coin price hikes) happen.

Benefits of analyzing topics

- Useful for developing hypotheses of content within the subreddit, that can later be confirmed by searching for it.
- Most useful: the topics put many terms into context.
  - There are many words for markets, users, tools, and services we would not have recognized if not contextualized by the topics.
  - Perhaps can be used as keywords for further investigation

Limitations of this Research

- The generated topics are only made practical when paired with the original data source.
  - Easy to misinterpret.
- Choice of subreddit is important.
- Applying topic modelling on large datasets can take hours.
  - Applying it to our corpora took a matter of minutes due to its size.
- Even though analyzing topics can hasten the search processes, the analysis still takes a generous amount of time.

Conclusion

- Our analysis showed a shift in tone (of Reddit users) from more casual to being more uncertain, concerned, and security-minded after the busts.
- Tools didn’t seem to evolve.
  - The trend is that their discussions occur in reaction to real world events.
- This information may be useful to law enforcement to understand how real world events have effects on online criminal communities or find keywords in a relatively quick manner.

Questions?

- kyle.porter@ntnu.no