Find Topics Across Twitter for All Known State Sponsored Hackers

Randy Grant (DSML)

INTRODUCTION

ISSUE

Cyber security professionals have a hard time keeping track of news regarding state sponsored hacker activities due to so many sources of information

SOLUTION

With a list of known state sponsored hacker groups, get all relevant data from Twitter posts with hashtags of the hacker group name.

GOAL

Discover topics for all known state sponsored hackers for which cyber security (blue team) analysts have found evidence of activity.

METHODS

State sponsored actor (hacker groups) cleaned list from original trusted csv made into hashtags

The list was then used to get a custom json from from PhantomBuster that scraped Twitter hashtags

DATA USED

Standard data science Python libraries

PhantomBuster trial subscription

TOOLS USED

NMF and LDA (method 1)

Gensim's TfidfModel and LDA (method 2)

MODELS

DATASET INFO

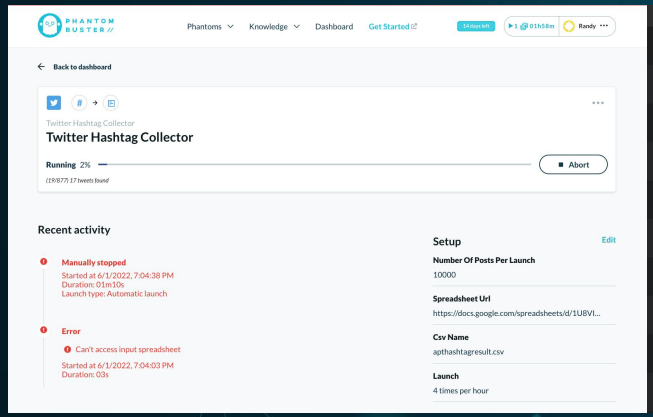
Number of documents (tweets) — **12,411**

271,026 — Aggregated length of corpus

Earliest date — **2010-08-09**

2022-06-02 — Latest date

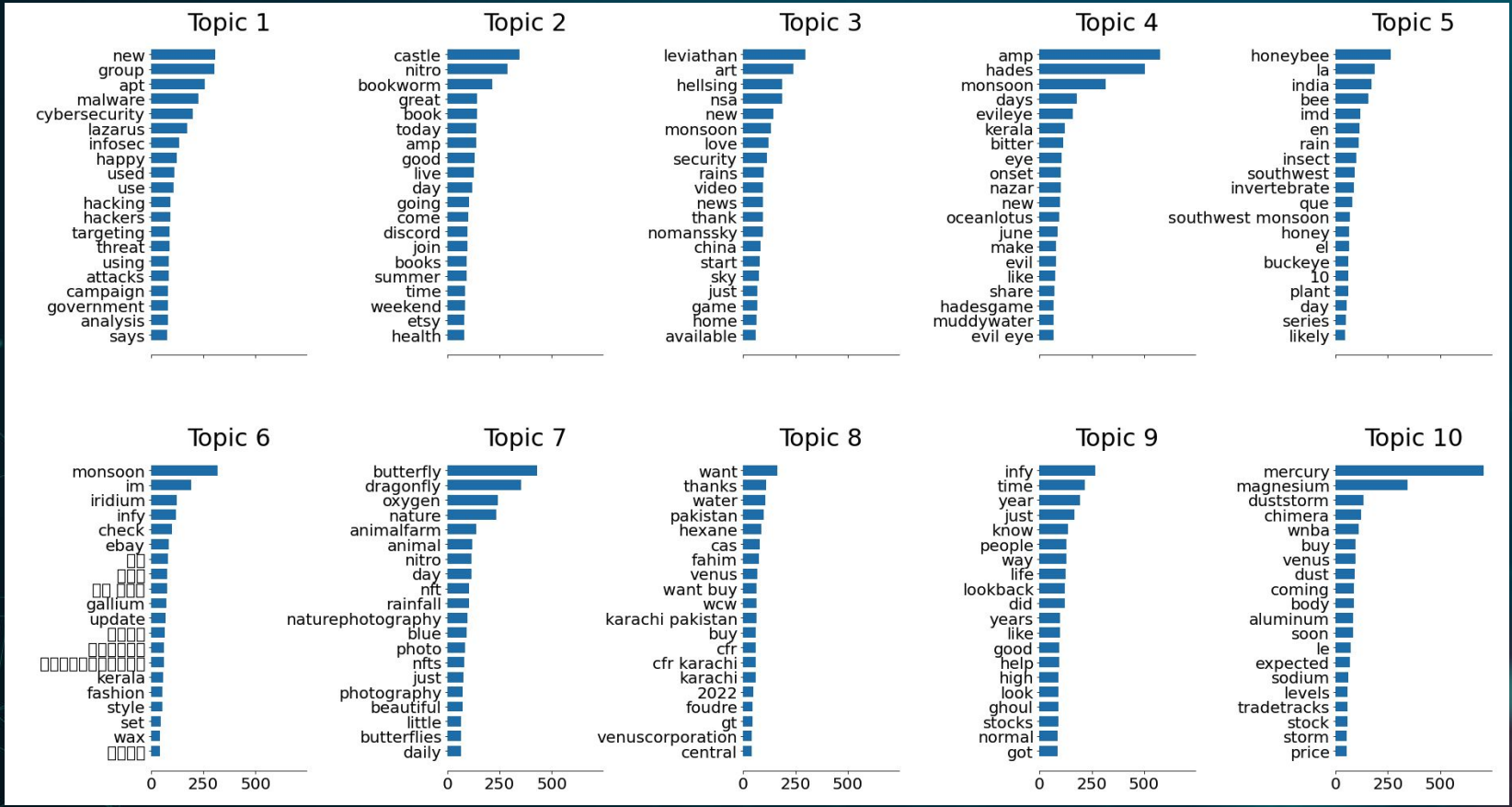
SAMPLE DATA



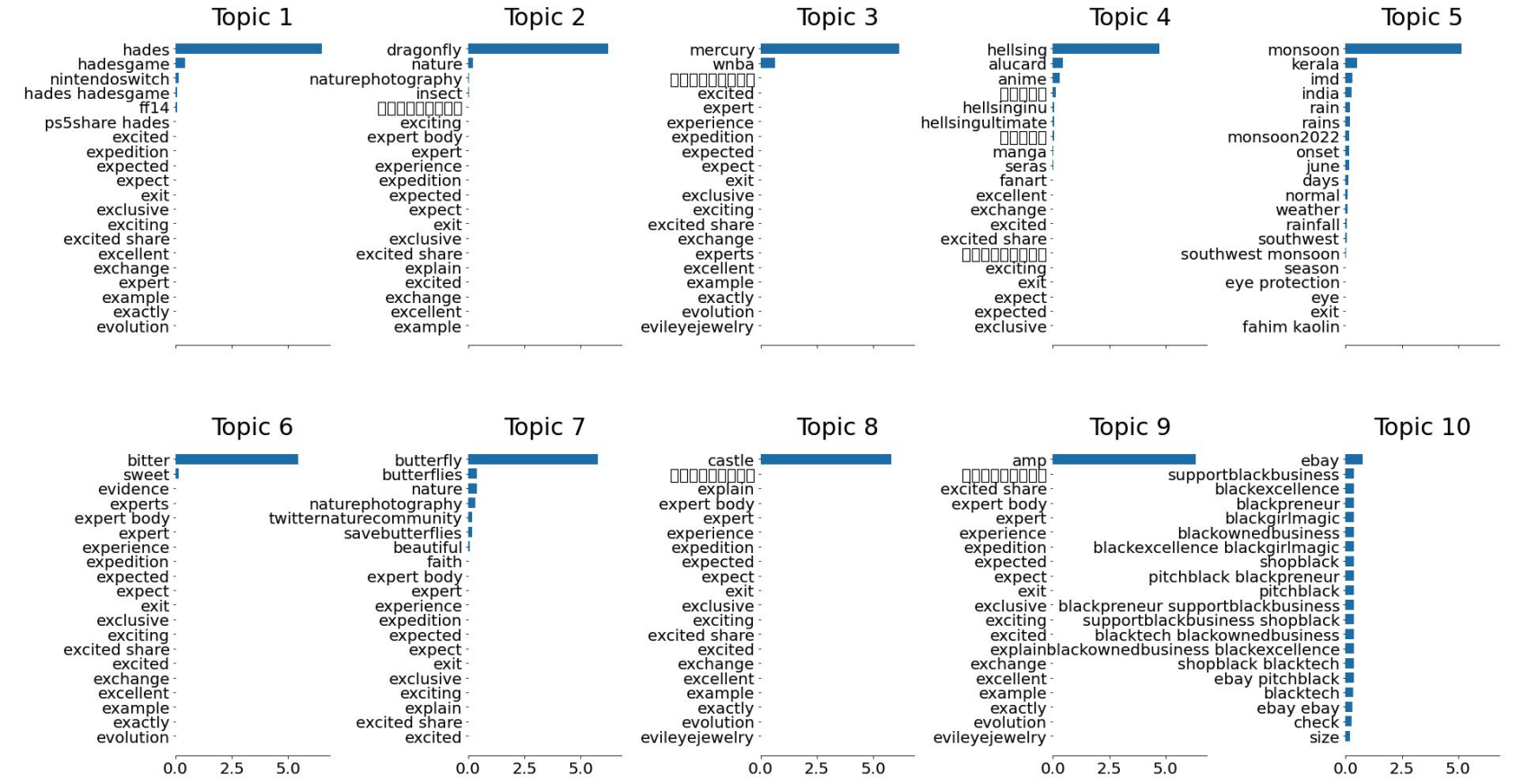
twitterProfile	tweetUrl	timestamp	query
https://twitter.com/SpokePedal	https://twitter.com/SpokePedal/status/1531628869630951430	2022-06-02 02:01:41.260000+00:00	#Oxygen
https://twitter.com/AlohaTemple	https://twitter.com/AlohaTemple/status/1531684233143386113	2022-06-02 01:52:28.585000+00:00	#Lazarus
https://twitter.com/shubhamwalke100	https://twitter.com/shubhamwalke100/status/1531169299984809984	2022-06-02 01:50:26.364000+00:00	#Infy
https://twitter.com/MTUArtsOffice	https://twitter.com/MTUArtsOffice/status/1049324351781265409	2022-06-02 01:45:09.982000+00:00	#Greenbug
https://twitter.com/deborah62538044	https://twitter.com/deborah62538044/status/1530269212945817600	2022-06-02 02:00:28.015000+00:00	#Nitro
https://twitter.com/MKuemmerlen	https://twitter.com/MKuemmerlen/status/1053761593585090562	2022-06-02 01:27:26.830000+00:00	#Blackfly
https://twitter.com/tur_def	https://twitter.com/tur_def/status/1531622737298526208	2022-06-02 02:12:06.548000+00:00	#SideWinder
https://twitter.com/SapphireTech	https://twitter.com/SapphireTech/status/1529100392667152386	2022-06-02 02:00:25.477000+00:00	#Nitro
https://twitter.com/CdiVilgenis	https://twitter.com/CdiVilgenis/status/855384571399024640	2022-06-02 02:07:19.553000+00:00	#RedDjinn
https://twitter.com/sunfloweruluv	https://twitter.com/sunfloweruluv/status/1530940767358836738	2022-06-02 01:27:00.728000+00:00	#BlackTech

	tweetDate	content
9985	Tue May 31 13:29:13 +0000 2022	bicyclesbeyond - #CYCOLOGY 🚲#Cycling is #oxygen for the #soul 🤩Good morning dear #BikeFriends 👫👫👫💚https://t.co/pen0W72I2g
7476	Tue May 31 17:09:13 +0000 2022	#LAZARUS
6867	Mon May 30 07:03:03 +0000 2022	Booking profits early is worst than booking loss I had monthly infy calls😔#infy
5350	Mon Oct 08 15:43:15 +0000 2018	A wrap for today's film screening art(ist FILM, thanks to all who came & showed films for todays short films fest at Rory Gallagher Theatre. Image of #Greenbug by Thomas Spencer & Wilhelmina van der Bent, closing Programme 1 & more coming 2moro for Programme 2 the same venue @ 1. https://t.co/aQtHIVTZaW
9669	Fri May 27 19:26:26 +0000 2022	So thrilled on the updates of #Nitro League. I'm one of the pioneers who are observing this project's direction 🚀
1701	Sat Oct 20 21:35:16 +0000 2018	Nice #surprise! Deployed #temperature loggers as artificial substrates for #blackfly / Simulidae #pupae in the Eulach #river https://t.co/o2ZZfZv4XK
12366	Tue May 31 13:04:51 +0000 2022	🇳🇱🇺🇸The #Netherlands to Acquire AIM-9x Block II Missiles from #USA #sidewinder https://t.co/H0TFX9ofUs
9636	Tue May 24 14:01:57 +0000 2022	Fitted with the latest Angular Velocity Fan Blade Design & constructed in the sleek white Die Casted Aluminum-Magnesium Alloy Frame, experience increased airflow & cooler operation in a strong & scratch resistant special frame with the SAPPHIRE #NITRO+ #AMD #Radeon #RX6950XT PURE https://t.co/KQIHjs73R7
11200	Fri Apr 21 11:35:48 +0000 2017	Prêt pour le #ConcertauCDI ? RV mardi prochain 25 à midi30. Dépaysement assuré, avec #RedDjinn !
1586	Sun May 29 15:54:57 +0000 2022	Check out Neon Snake Skin Print Women Deep V One Piece Swimsuit High Leg Leopard Size XL https://t.co/RpKS5TN7z6 #eBay via @eBay #PitchBlack #Blackpreneur #SupportBlackBusiness #ShopBlack #BlackTech #Blackownedbusiness #Blackexcellence #Blackgirlmagic

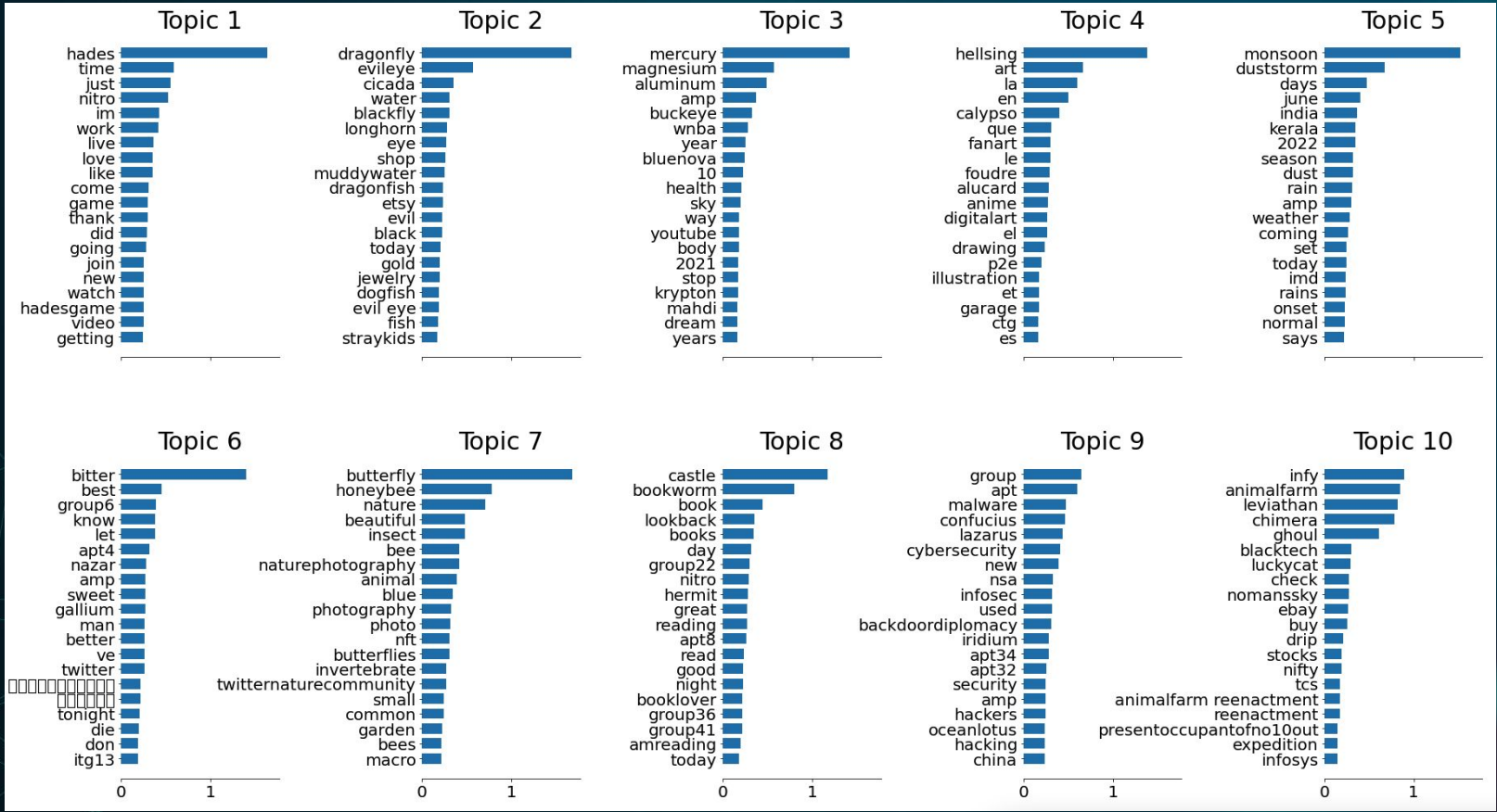
TOPICS IN LDA MODEL



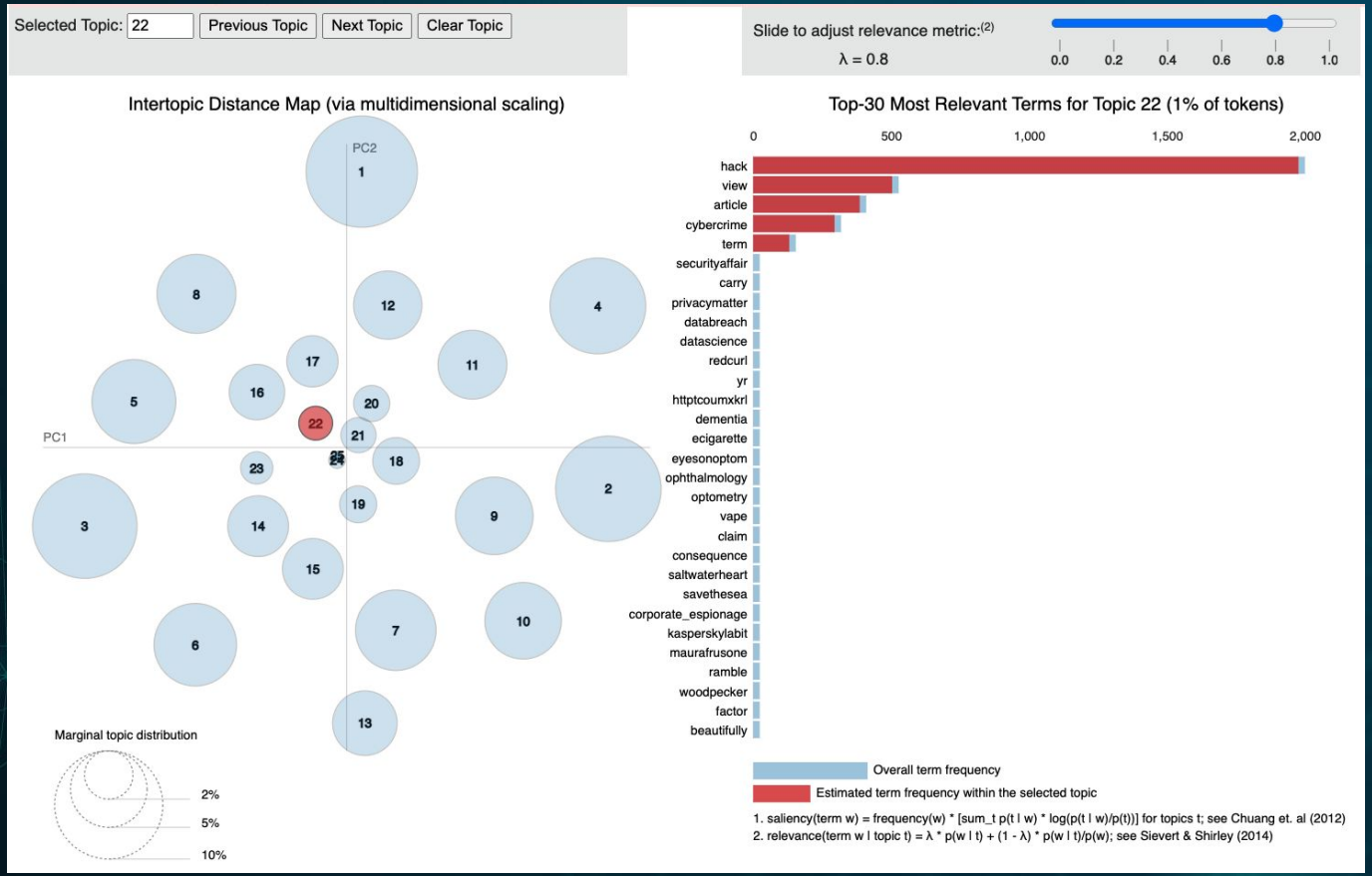
TOPICS IN NMF MODEL (EUCLIDEAN DISTANCE)



TOPICS IN NMF MODEL (RELATIVE ENTROPY)



TOPICS IN LDA MODEL USING GENSIM AND PYLDAVIZ



REFERENCES

https://scikit-learn.org/stable/auto_examples/applications/plot_topics_extraction_with_nmf_lda.html#sphx-glr-auto-examples-applications-plot-topics-extraction-with-nmf-lda-py

https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#1introduction

https://stackoverflow.com/questions/66759852/no-module-named-pyldavis

https://github.com/StrangerealIntel/EternalLiberty/blob/main/EternalLiberty.csv

https://phantombuster.com/

https://stackoverflow.com/questions/46848209/get-the-earliest-date-from-a-column-python-pandas-after-csv-reader

https://stackoverflow.com/a/50084009

https://machinelearningmastery.com/clean-text-machine-learning-python/

https://stackoverflow.com/questions/24386489/adding-words-to-scikit-learns-countvectorizers-stop-list/24386751#24386751

https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a

https://github.com/wjbmattingly/topic_modeling_textbook/blob/main/03_03_lda_model_demo.ipynb

https://www.youtube.com/watch?v=TKjjlp5_r7o

https://www.youtube.com/watch?v=UEn3xHNBXJU&list=PL2VXyKi-KpYttggRATQVmgFcQst3z6OlX&index=11

THANKS!

Questions?

randywgrant@gmail.com