UC San Diego

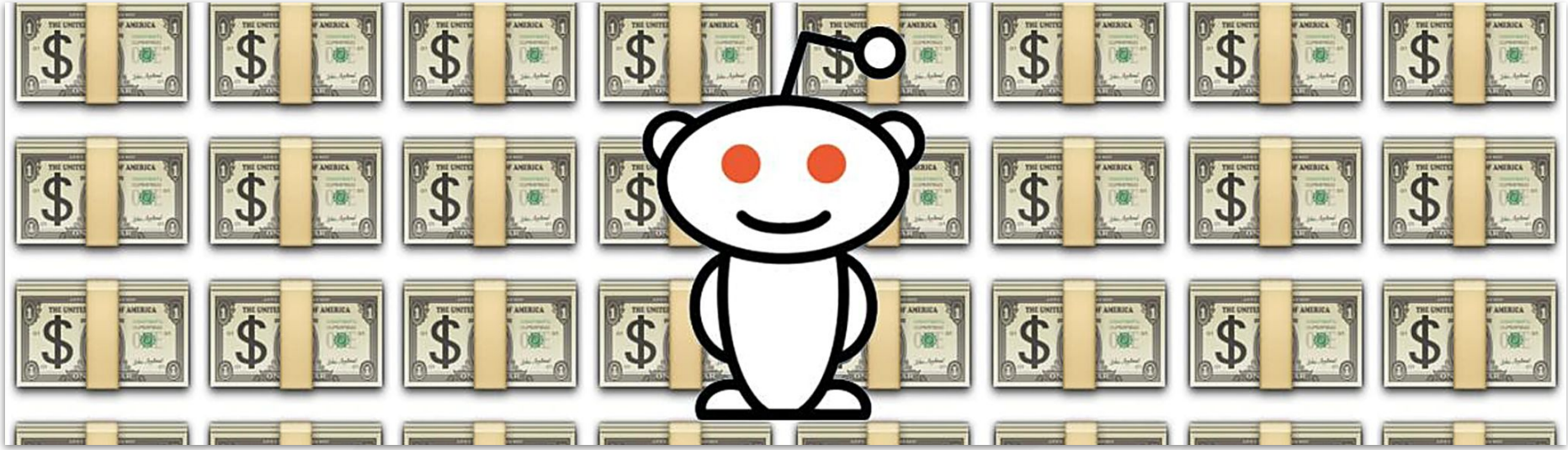
Social Media Talks Stocks
2021-12-10

Amit Sachan, Mustafa Burny, Sam Courtney

Overview

- Every day, millions of people discuss stock prices and financial markets on social media.
- Recently, social media has started to significantly impact stock prices.
- We wanted to investigate the relationship between financial markets and social media activity.
- We wanted to investigate the engagement period.

Problem



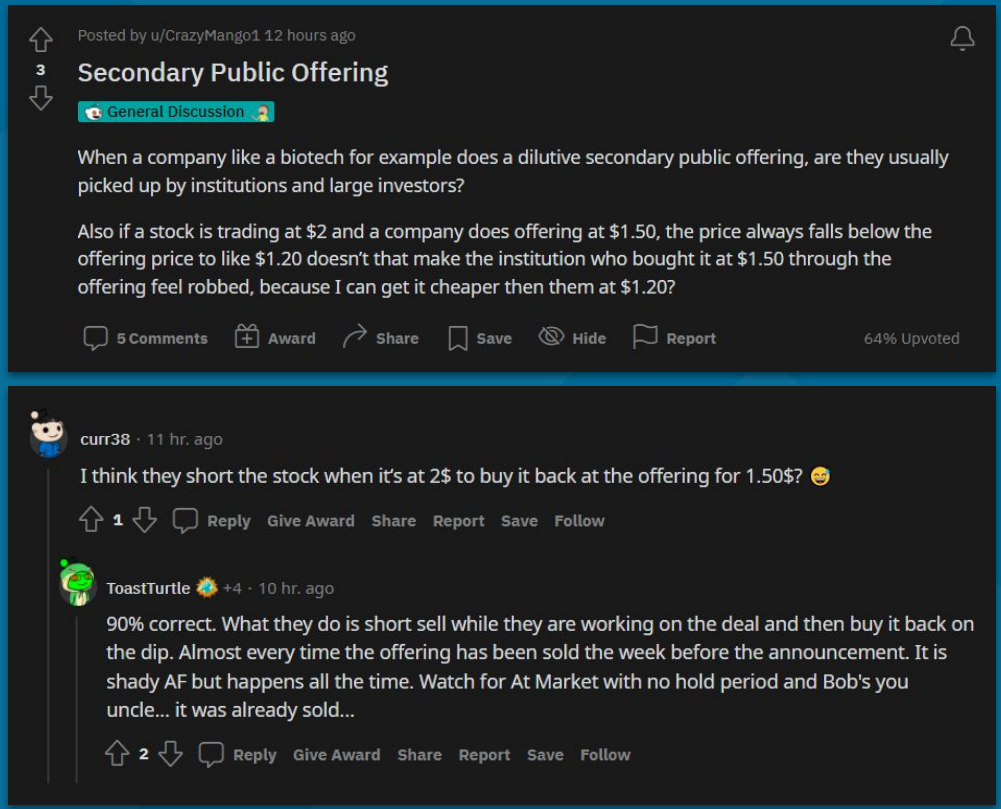
1. Can we extract what companies and stocks are mentioned in Reddit discussions?
2. Can we correlate the Reddit discussions to stock price movements?
3. Can we understand the topics being mentioned in Reddit discussions?
4. Can we understand relation between posts and comments lifetime?

UC San Diego

DATASET OVERVIEW

Reddit

- Anonymous users can create Reddit submissions in sub-reddits associated with specific topics.
- Each Reddit submission has an author, a title, and (optional) body.
- Each Reddit submission can have a collection of comments and sub-comments.
- Reddit submissions and Reddit comments have a score determined by how many Reddit users “liked” or “disliked” the content. It’s called “karma”



Extracting Data from Reddit - Reddit API & PRAW

- Reddit provides a REST API to retrieve submissions, comments, and user information
- Requires creating application/bot credentials via Reddit web interface
- praw (Python Reddit API Wrapper) is a Python package that facilitates Reddit API access
- API is rate-limited to about 1 request per second

https://www.reddit.com/dev/api/
https://praw.readthedocs.io/

```
# Initialize Reddit instance
reddit = praw.Reddit(
    client_id = REDDIT_CLIENT_ID,
    client_secret = REDDIT_CLIENT_SECRET,
    user_agent = REDDIT_USER_AGENT
)
```

```
# Get submission details
submission = reddit.submission(id='<SUBMISSION ID>')
print(f'  Title = {submission.title}')
print(f' Author = {submission.author.name}')
print(f'Created = {submission.created}')
print(f'  Score = {submission.score}')

# Get comment details
comment = reddit.comment(id='<COMMENT ID>')
print(f'Created = {comment.created}')
print(f' Author = {comment.author.name}')
print(f'   Body = {comment.body}')
print(f'  Score = {comment.score}')
```

Extract Data from Reddit - PushShift API & PSAW

- PushShift is a third-party archive of Reddit data
- Provides useful features that are not available via official Reddit API:
  - Find top-scoring submissions or comments in a specific date range
  - Search multiple subreddits in a single API call
  - Filter submissions and comments by a simple wild-card search (e.g. `$AAPL|AAPL|"apple inc"`)

https://pushshift.io/api-parameters/
https://psaw.readthedocs.io

```python
# Initialize Pushshift API instance
api = PushshiftAPI()

# Search submissions that include the search query `q`
submissions = api.search_submissions(
    after = start_epoch,
    before = end_epoch,
    q = '$AAPL|AAPL|"apple inc"',
    subreddit = 'wallstreetbets,wallstreetbetsnew,Investing',
    sort_type = 'num_comments',
    sort = 'desc',
    filter = [
        'id', 'url', 'subreddit', 'created', 'author',
        'num_comments', 'num_crossposts', 'title', 'selftext'
    ],
    limit = 10
)

for submission in submissions:
    pass  # Do something with each submission
```

Extracting Data from Reddit - Kaggle

- Ref: https://www.kaggle.com/pavellexyr/wallstreetbets-posts-and-comments-for-august-2021
- Contains 2 csv files:
  - **wsb-aug-2021-posts** ['type', 'id', 'subreddit.id', 'subreddit.name', 'subreddit.nsfw','created_utc', 'permalink', 'domain', 'url', 'selftext', 'title','score']

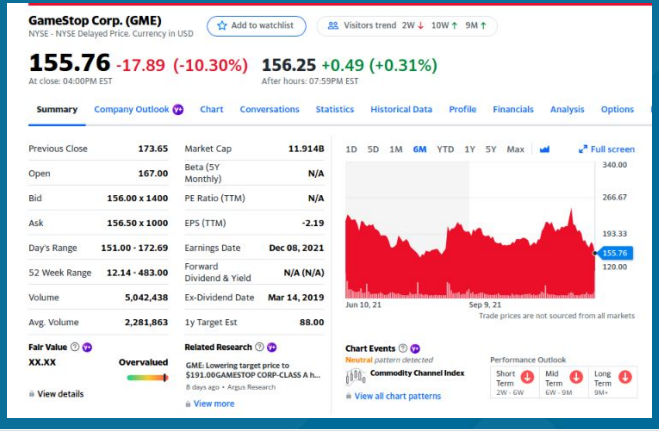
    shape - (25,751 rows, 12 columns)
  - wsb-aug-2021-comments ['type', 'id', 'subreddit.id', 'subreddit.name', 'subreddit.nsfw','created_utc', 'permalink', 'body', 'sentiment', 'score']

    shape - (1,001,160 rows, 10 columns)

Yahoo! Finance

- yfinance is a python package that provides information about publicly traded companies
- This is used to replace an API that was decommissioned by Yahoo!
- Provides up to 151 fields of information - including opening, closing, and daily delta
- Used to match and validate legitimate tickers extracted via RAKE

https://pypi.org/project/yfinance/

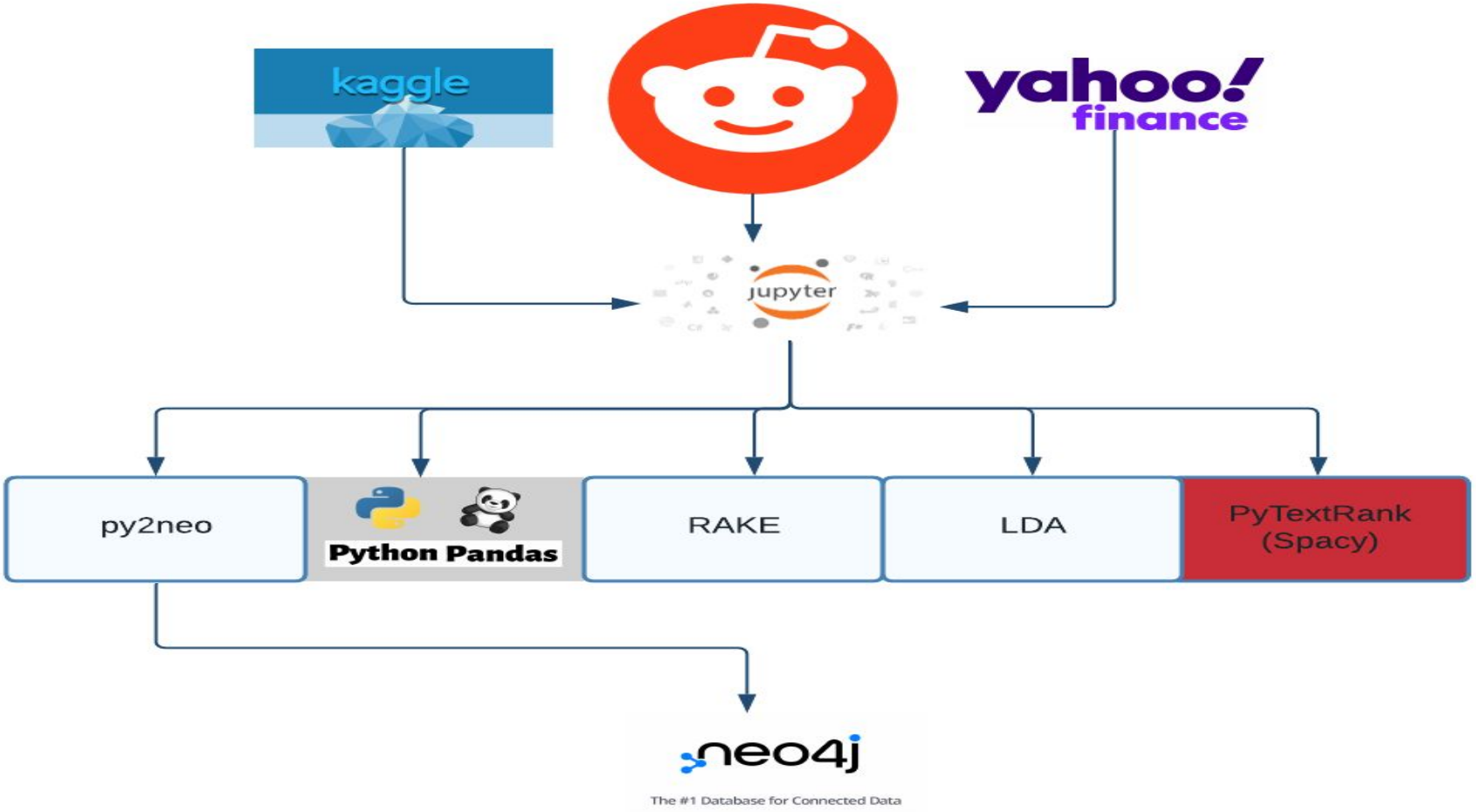



```
[5]: yf.Ticker('GME').info

[5]: {'zip': '76051',
 'sector': 'Consumer Cyclical',
 'fullTimeEmployees': 12000,
 'longBusinessSummary': 'GameStop Corp., a specialty retailer, provides games and entertainment products through its e-commerce properties and various stores in the United States, Canada, Australia, and Europe. The company sells new and pre-owned video game platforms; accessories, such as controllers, gaming headsets, virtual reality products, and memory cards; new and pre-owned video game software; and in-game digital currency, digital downloadable content, and full-game downloads, as well as network points cards, and prepaid digital and subscription cards. It also sells collectibles comprising licensed merchandise primarily related to the video game, television, and movie industries, as well as pop culture themes. The company operates its stores and e-commerce sites under the GameStop, EB Games, and Micromania brands; and collectibles stores under the Zing Pop Culture and ThinkGeek brand, as well as offers Game Informer, a print and digital video game publication featuring reviews of new title releases, game tips, and news regarding the video game industry. As of January 30, 2021, it operated 4,816 stores. The company was formerly known as GSC Holdings Corp. GameStop Corp. was founded in 1996 and is headquartered in Grapevine, Texas.',
 'city': 'Grapevine',
 'phone': '817 424 2000',
 'state': 'TX',
 'country': 'United States',
 'companyOfficers': [],
 'website': 'https://www.gamestop.com',
 'maxAge': 1,
 'address1': '625 Westport Parkway',
 'industry': 'Specialty Retail',
 'ebitdaMargins': -0.01587,
 'profitMargins': -0.026080001,
 'grossMargins': 0.24123,
 'operatingCashflow': -159200000,
 'revenueGrowth': 0.291,
 'operatingMargins': -0.028250001,
 'ebitda': -93300000,
 'targetLowPrice': 24,
 'recommendationKey': 'underperform',
```

UC San Diego

THE ANALYSIS

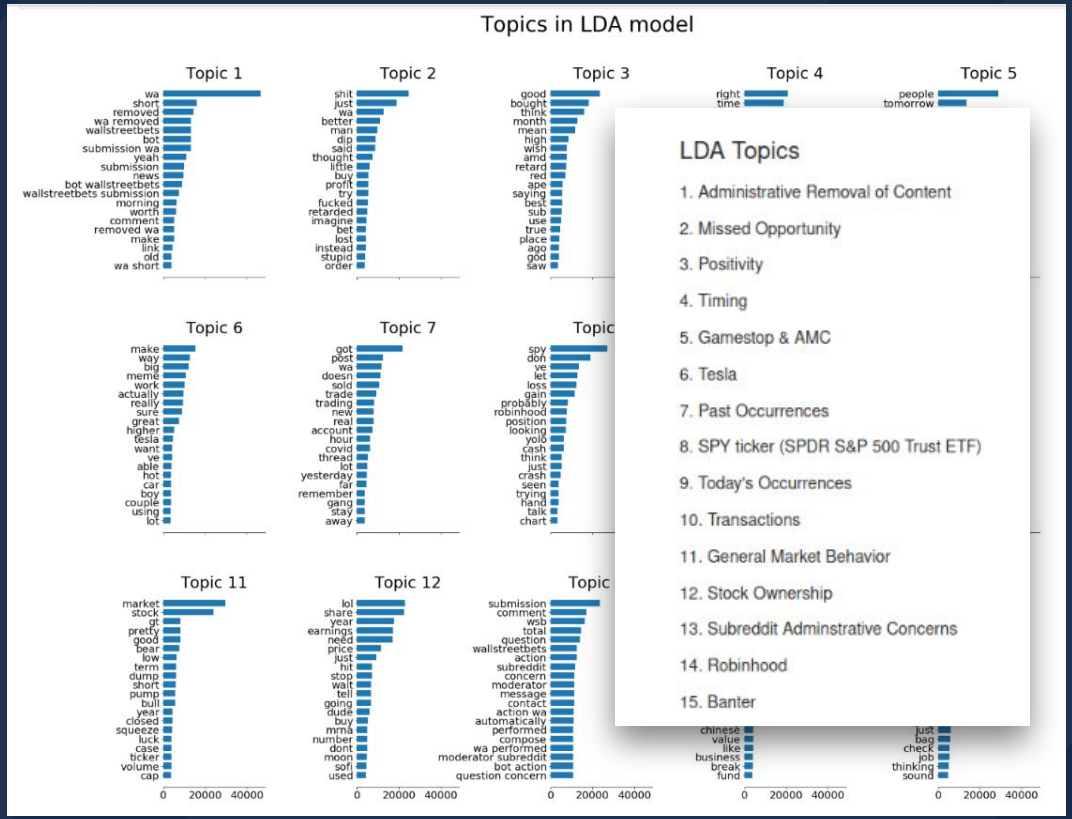

kaggle
yahoo! finance
jupyter
py2neo
Python Pandas
RAKE
LDA
PyTextRank (Spacy)
neo4j
The #1 Database for Connected Data

Keywords extraction

- To keep the extraction output manageable, we extracted all keywords from posts and comments and than ran a regex to only look for keywords with $ prefix.
- 1456 company keywords extracted through RAKE.
- 294 stock symbols remained after validating against yahoo finance.

```
['sdc', 'leg', 'adm', 'crm', 'grwg', 'd', 'tt', 'coty', 'cpop',
 'open', 'gdrx', 'abnb', 'amd', 'cvna', 'rgbp', 'pcg', 'fubo',
 'jblu', 'dq', 'nio', 'mj', 'u', 'rrgb', 'alb', 'wmt', 'crsr',
 'low', 'zim', 'psfe', 'dfen', 'roku', 'tup', 'panw', 'gme', 'dnut',
 'clov', 'cvs', 'paya', 'wix', 'dash', 'zm', 'spce', 'sndl', 'bark',
 'aso', 'unh', 'bud', 'ms', 'coin', 'base', 'mrna', 'lmt', 'ecvt',
 'spot', 'nkla', 'tal', 'lc', 'nvda', 'vxx', 'tr', 'corn', 'sony',
 'csco', 's', 'astr', 'dkng', 'uvxy', 'save', 'mnst', 'x', 'eat',
 'dis', 'pypl', 'snps', 'sq', 'igt', 'crnt', 'carr', 'coke', 'rblx',
 'me', 'clf', 'hood', 'azn', 'grom', 'play', 'luv', 'bili', 'nflx',
 'fvrr', 'maxr', 'ejh', 'geni', 'mdlz', 'fsm', 'amc', 'khc', 'ater',
 'b', 'baba', 'cat', 'ge', 'jmia', 'jack', 'su', 'vzio', 'amzn',
 'xl', 'ed', 'azz', 'uwmc', 'webr', 'spy', 'tsla', 'qs', 'zbra',
 'afrm', 'pbya', 'pton', 'stla', 'manu', 'clne', 'root', 'gaymf',
 'gld', 'mmnd', 'asml', 'mu', 'et', 'atvi', 'gm', 'hyln', 'ftch',
 'pltr', 'viac', 'jd', 'car', 'tdoc', 'alto', 'you', 'hog', 'cvsi',
 'arw', 'm', 'ibkr', 'axsm', 'y', 'vym', 'hut', 'fas', 'spxu',
 'jepi', 'simp', 'gt', 'cmp', 'pins', 'wish', 'ha', 'crsp', 'goog',
 'ship', 'sklz', 'mo', 'xlf', 'bngo', 'mrvl', 'rgr', 'aal', 'twlo',
 'lcid', 'cron', 'ba', 'mrk', 'dow', 'dude', 'job', 'rtx', 'fxlv',
 'cook', 'flws', 'posh', 'nvax', 'ftnt', 'snap', 'se', 'msft',
 'cyxt', 'tch', 'tlt', 'tgt', 'ndaq', 'rklb', 'hexo', 'sgmd',
 'edit', 'crwd', 'sens', 'wfc', 'muds', 'bntx', 'sbux', 'shop',
 'etsy', 'sli', 'ally', 'chwy', 'land', 'hd', 'ebet', 'yinn',
 'dooo', 'bb', 'abt', 'arkk', 'asts', 'dal', 'dole', 'sesn', 'dats',
 'goev', 'four', 'wkhs', 'gis', 'ride', 'fnma', 'body', 'dnmr',
 'bbig', 'riot', 'net', 'dia', 'mara', 'jnj', 'v', 'rkt', 'didi',
 'upst', 'sava', 'woof', 'avpt', 'sofi', 'wen', 'ptra', 'tm',
 'tlry', 'tell', 'wing', 'aapl', 'ida', 'zy', 'slqt', 'tcehy',
 'rdnt', 'f', 'ntnx', 'xpo', 'celh', 'wwe', 't', 'apps', 'bac',
 'nakd', 'ayx', 'nok', 'ely', 'indi', 'gold', 'pdd', 'amat', 'view',
 'antm', 'zg', 'doyu', 'ap', 'googl', 'el', 'ccl', 'aci', 'bidu',
 'ebay', 'pfe', 'j', 'negg', 'evgo', 'fb', 'vrns', 'tmf', 'mvst',
 'gotu', 'pg', 'path', 'penn', 'znga', 'mcfe'], dtype=object)
```

Analysis Details - LDA - Latent Dirichlet Allocation



- LDA is a dimensionality reduction algorithm
- It is used to generate synthetic topics from a corpus of documents and a dictionary of terms
- The end result is a set of topics that can be used to classify documents based on how well they matches the topics in the trained set

Analysis Details - LDA (cont.)

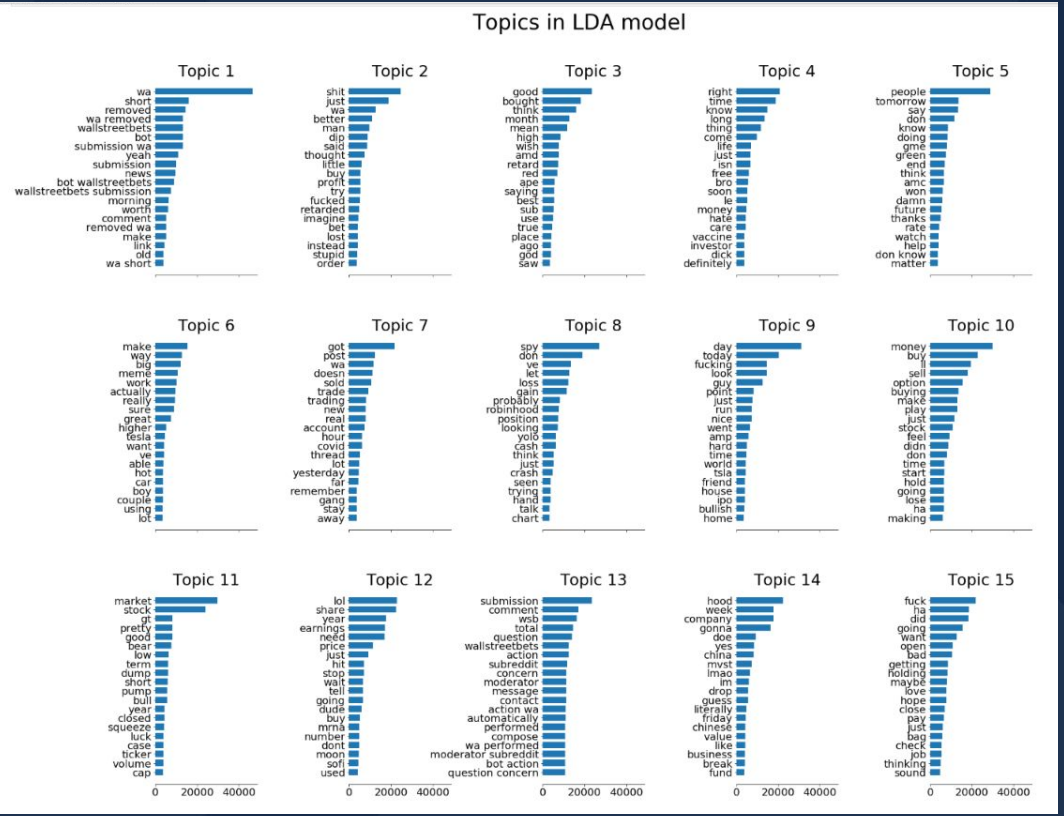
['amp x200b position geni dd genius retard position not yolo know right geni tldr cathie ark fund nfl season incoming earnings nfl exclusive partner huge moat dkng possible deal espn called repos39 guy called negg more doe geni do act intermediary providing data sport league sportbooks help sportsbook draftkings provide live prop bet will person strike bat end are technology company utilizes official sport data lives play play statistic analytics intrinsic value sportsbook partner medium organization sport league etc bet cathie wood arkw geni buy cathie ha been doing overtime buying shit geni aug share bought buy arkw aug share bought buy arkw aug share bought buy arkw aug share bought buy arkw aug share bought buy arkw aug share bought buy arkw nfl season start september stock dkng geni etc are know run event pair wa upcoming earnings report geni could see crazy run start nfl season earnings report quarter earnings geni report earnings september pre market geni ha huge moat exclusive deal nfl growth is fantastic ebitda geni ha only scratched surface majority revenue is sport betting moment currently geni get cut take rate sportsbooks are looking increase even were increase would give major profit increase will lead faster industry growth dkng ha possible deal espn worth geni dkng move close together according great dd wlsrtjp dkng espn may have massive partnership coming would allow investor see volatile move upwards share price similar amzn afrm source deal summary', 'started year ago when son wa year old started supplement picture book cartoon beginner option book martin shkreli live video how pick pharmaceutical stock course year feel ha retained absolutely nothing even have spent waking minute trying make understand have done almost everything have thought including having jerome powell speech play is sleeping only having warren buffet tv try make understand market felt got breaking point when couldn differentiate straddle strangle even went different strategy almost month straight finally convinced wife were doing right thing when said will soon be wendy worker begging wife husband weekly allowance will never amount be anything couldn fathom raising kid wa not able able make profit trading option time wa said anyone want year old child is shitty market strategy check ernal sunny orphanage omaha nebraska maybe luck will be better mine']

Text Pre-Processing

- Removed missing and deleted posts
- Removed html code
- Lower case
- Remove numbers
- Limit parts-of-speech
- Tokenize
- Remove single letter words
- Lemmatization
- Remove words occurring in >95% of documents
- Remove words that only occur in one document
- Remove stop words
- Create bi-grams
- Vectorize

Analysis Details - LDA - Latent Dirichlet Allocation

- Processed vector fed into Latent Dirichlet Allocation model
- 15 component LDA model
- 1000 features used
- Trained on 814k posts
- Trained model used to classify posts
- Packages used - natural language toolkit (nltk) , sci kit learn (sklearn)



Analysis Details - LDA - Latent Dirichlet Allocation

LDA Topics

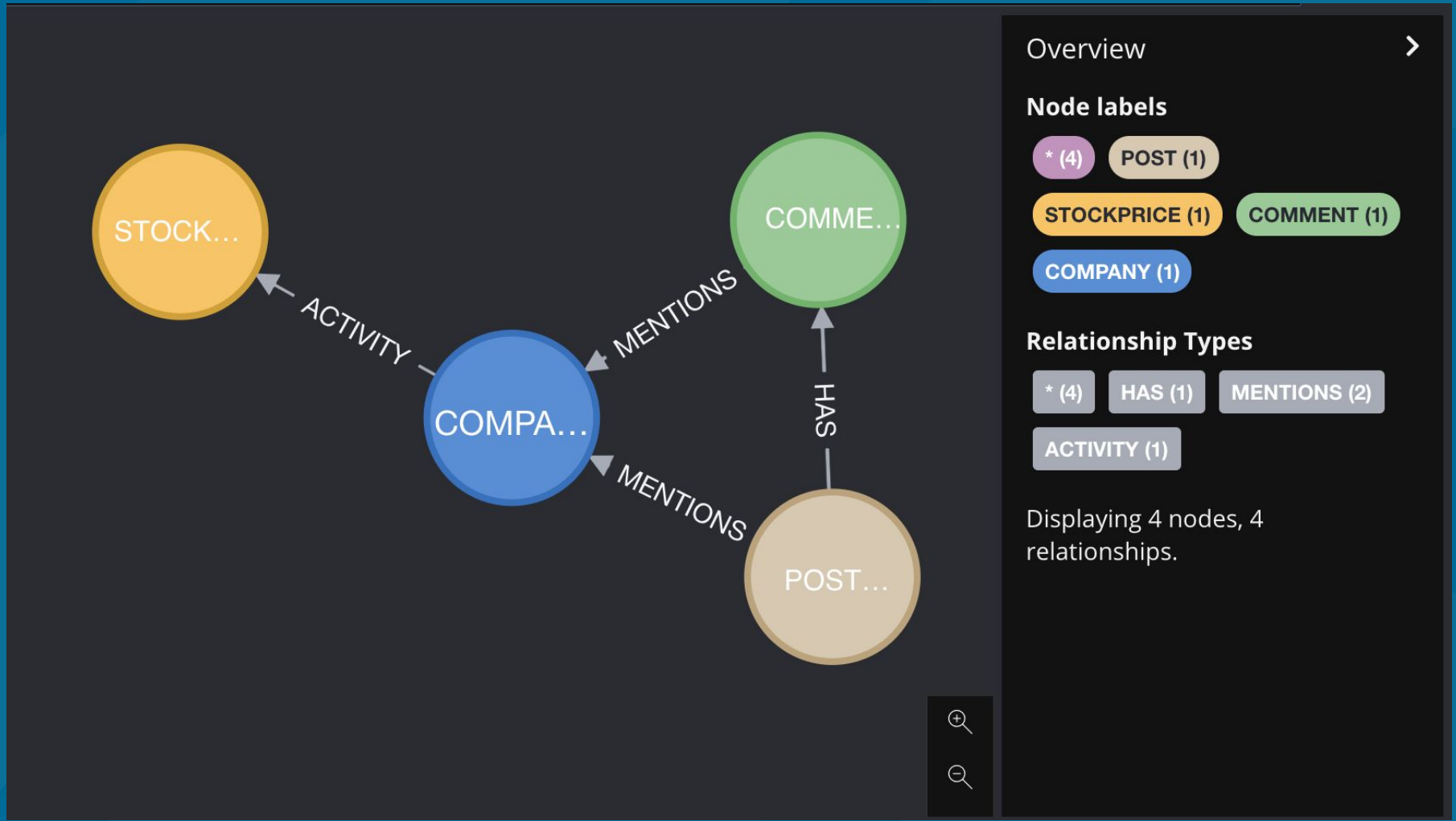
1. Administrative Removal of Content
2. Missed Opportunity
3. Positivity
4. Timing
5. Gamestop & AMC
6. Tesla
7. Past Occurrences
8. SPY ticker (SPDR S&P 500 Trust ETF)
9. Today's Occurrences
10. Transactions
11. General Market Behavior
12. Stock Ownership
13. Subreddit Adminstrative Concerns
14. Robinhood
15. Banter

1	14.960851
2	6.792915
3	5.931684
4	5.483008
5	6.613543
6	3.787083
7	5.312113
8	5.854775
9	6.882847
10	8.630618
11	5.394674
12	7.046002
13	2.887152
14	7.880081
15	6.542654

"c.topic"	"num"
"Administrative Removal of Content"	19579
"Robinhood"	10621
"Transactions"	9568
"Missed Opportunity"	9202
"Todays Occurrences"	8749
"Gamestop & AMC"	8495
"Mixed Banter"	8089
"Stock Ownership"	7933
"Positivity"	7220
"Timing"	7153
"SPY ticker"	7070

UC San Diego

KNOWLEDGE GRAPH


STOCK…
ACTIVITY
COMPA…
MENTIONS
COMME…
HAS
MENTIONS
POST…
Overview
Node labels
* (4)
POST (1)
STOCKPRICE (1)
COMMENT (1)
COMPANY (1)
Relationship Types
* (4)
HAS (1)
MENTIONS (2)
ACTIVITY (1)
Displaying 4 nodes, 4 relationships.

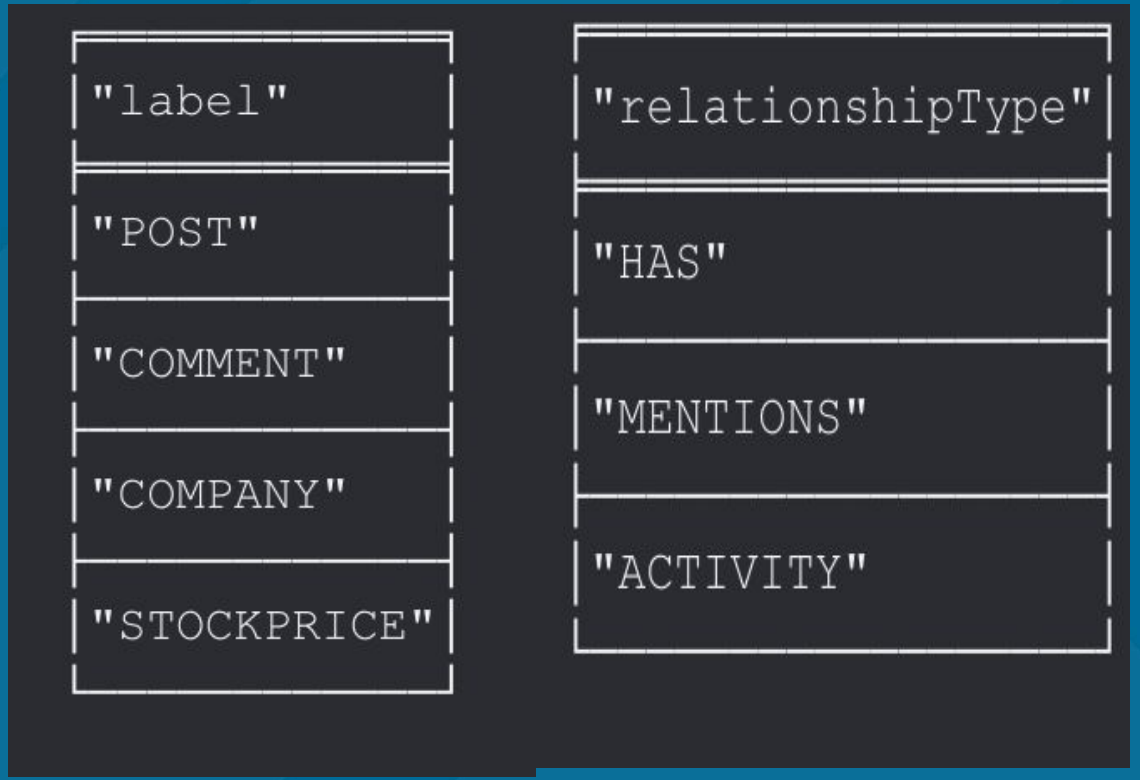
Knowledge Graph

"labels(n)"	"SampleSize"	"Avg_PropertyCount"	"Min_PropertyCount"	"Max_PropertyCount"	"Avg_RelationshipCount"	"Min_RelationshipCount"	"Max_RelationshipCount"
["COMMENT"]	12347	6.0	6	6	1.0091520207337765	1	4
["COMPANY"]	33	1.0	1	1	23.57575757575757	22	53
["STOCKPRICE"]	596	5.0	5	5	1.0	1	1
["POST"]	1	4.0	4	4	12606.0	12606	12606

- On average, each post has ~12k comments
- On average, each comment mentions 1 company.
- On average, each company has 23 relationships owing to number of open market days.
- No null property values across nodes.

Graph Stats

- Total Number of nodes - 128936
- Total Number of relationships - 129910



UC San Diego

THE RESULTS

Most popular post

	p.created	p.title	p.score
1	"24"	"Daily Discussion Thread for August 24, 2021"	"502"

Most popular comment

	p.c_id	p.body	p.score
1	"halme9b"	"Hit half mil net worth today, thank you JPOW and WSB."	"98"

Most commented post

	p.p_id	p.title	num
1	"p3sv76"	"Weekend Discussion Thread for the Weekend of August 13, 2021"	19259

Top 10 mentioned companies

	Name	num
1	"hood"	199
2	"spy"	78
3	"wish"	69
4	"gme"	56
5	"pfe"	41
6	"nvda"	34
7	"root"	32
8	"baba"	31
9	"clov"	23

Lifetime of posts

	p.title	daysdiff	num
1	"Daily Discussion Thread for August 19, 2021"	0	13272
2	"Daily Discussion Thread for August 24, 2021"	0	12550
3	"Daily Discussion Thread for August 04, 2021"	0	12277
4	"What Are Your Moves Tomorrow, August 19, 2021"	0	10341
5	"Weekend Discussion Thread for the Weekend of August 13, 2021"	1	9127
6	"Weekend Discussion Thread for the Weekend of August 06, 2021"	1	8506
7	"Weekend Discussion Thread for the Weekend of August 20, 2021"	1	8000
8	"Weekend Discussion Thread for the Weekend of August 27, 2021"	1	7199
9	"Weekend Discussion Thread for the Weekend of August 06, 2021"	0	6802

Lifetime of posts

	po.title	numComments	Company	CreatedPost	StockDate	val
1	"Weekend Discussion Thread for the Weekend of August 27, 2021"	5	"bbig"	"27"	"27"	50.14164967962419
2	"Weekend Discussion Thread for the Weekend of August 27, 2021"	1	"ater"	"27"	"27"	47.14423931778589
3	"Daily Discussion Thread for August 04, 2021"	164	"hood"	"04"	"04"	29.27456088771682
4	"Daily Discussion Thread for August 24, 2021"	44	"gme"	"24"	"24"	26.459800214369984
5	"Daily Discussion Thread for August 24, 2021"	11	"amc"	"24"	"24"	19.01048657137274
6	"Daily Discussion Thread for August 24, 2021"	1	"bbig"	"24"	"24"	17.60000228881836
7	"Daily Discussion Thread for August 24, 2021"	3	"gotu"	"24"	"24"	16.24999379118308
8	"Daily Discussion Thread for August 24, 2021"	2	"tal"	"24"	"24"	14.98973380344713
9	"Daily Discussion Thread for August 19, 2021"	3	"dnut"	"19"	"19"	14.692306518554691

UC San Diego

DEMO

Takeaways

- Social activity goes up when there is a big change in stock price.
- Whether the volume of social activity influences stock price is inconclusive.
- Engagement with posts is only active for t + 1 days. Moderators need to frequently create new posts to get traffic.
- Blue chip companies are frequently mentioned alongside meme stocks.

Lessons Learned & Future Work

LESSONS LEARNED

- Power of keywords extraction.
- Power of nltk for text preprocessing.
- py2neo over neo4j driver.
- Graph tuning.
- Node and relationship creation in batch.
- Remove bot activity.
- There are different “degrees” to unstructured data!
- Everything about GRAPH!!

POTENTIAL FUTURE WORK

- Switching to a dataset (e.g. from /r/Investing_Advice) that contains more grammatically well-formed unstructured data may allow us to refine our topic modeling and keyword extraction efforts.
- Focusing down on “Daily Discussion Threads” may help to better understand time-specific trends in topics and sentiment.

UC San Diego

QUESTIONS?

Amit Sachan
Mustafa Burny
Sam Courtney