



—

# BC COMS 2710: Computational Text Analysis

—

## Lecture 14 – Data Collection API's - Reddit



- Readings 04:
  - link posted to course site later today
  - due Sunday
  
- HW 02:
  - Due Wednesday night



- HW03:
  - Scrape CULPA for reviews
  - Identify *words that discriminate categories in textual data*:
    1. <https://wordify.unibocconi.it/index>
    2. Computer PMI
- Tutorial 5.1:
  - Classifying Tweets
- Tutorial 5.2:
  - Amazon Mechanical Turk



- **Project ideation – Friday May 28<sup>st</sup>**
  - <https://www.overleaf.com/read/yzpgxcgsqdvq>
- roughly 250 word overview of what you are interested in

# Final Project – Deliverables



- Project ideation – Friday May 28<sup>st</sup>
  - 5 points
  
- Project proposal – Friday June 4<sup>th</sup>
  - 9 points
  
- Project presentations – Monday June 14<sup>th</sup>
  - 6 points
  
- Project submissions – Friday June 18<sup>th</sup>
  - 15 points
  
- [http://coms2710.barnard.edu/final\\_project](http://coms2710.barnard.edu/final_project)




# Application Programming Interface



Generally: set of protocols that specify how software programs communicate with each other

Programmatically extract and interact with data

Released by companies for data sharing purposes



—

# Reddit: the front page of the internet

—

Slides from Alan Ritter

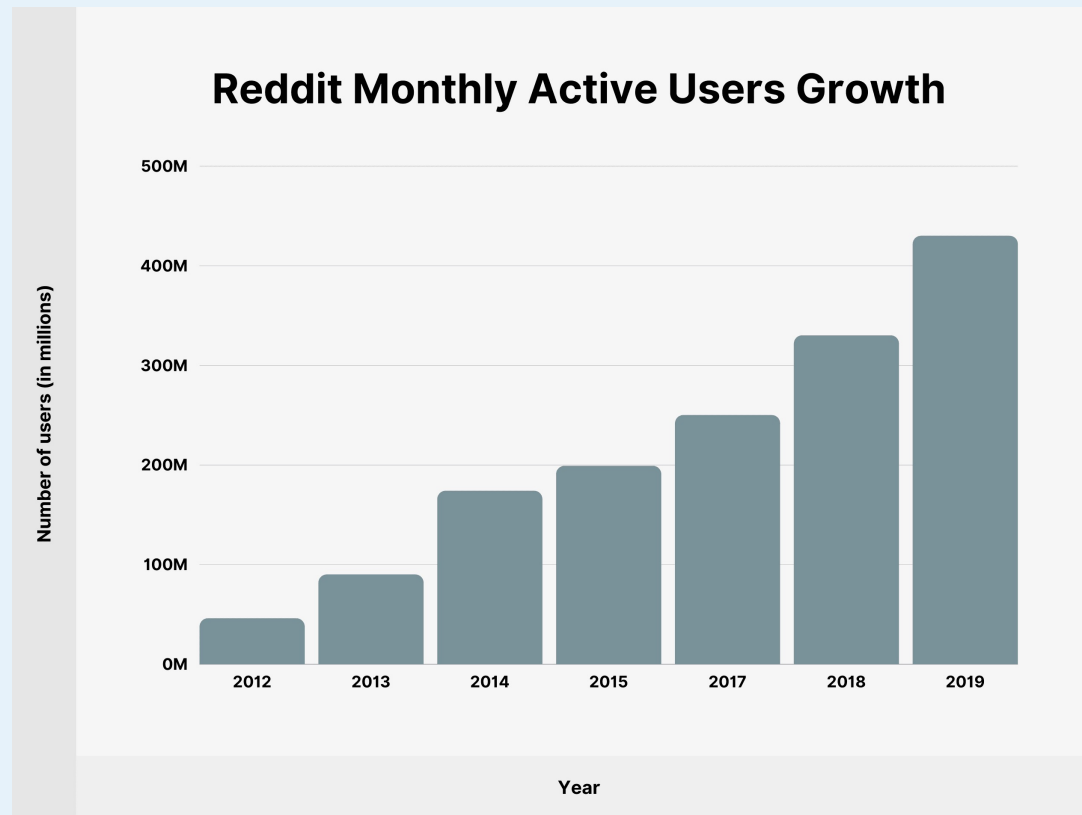


- Co-founded by Steve Huffman, Alexis Ohanian & Aaron Schwartz
- Launched 2005
- Sold to Conde Nast in 2006 for 8 digits (10 – 20M)

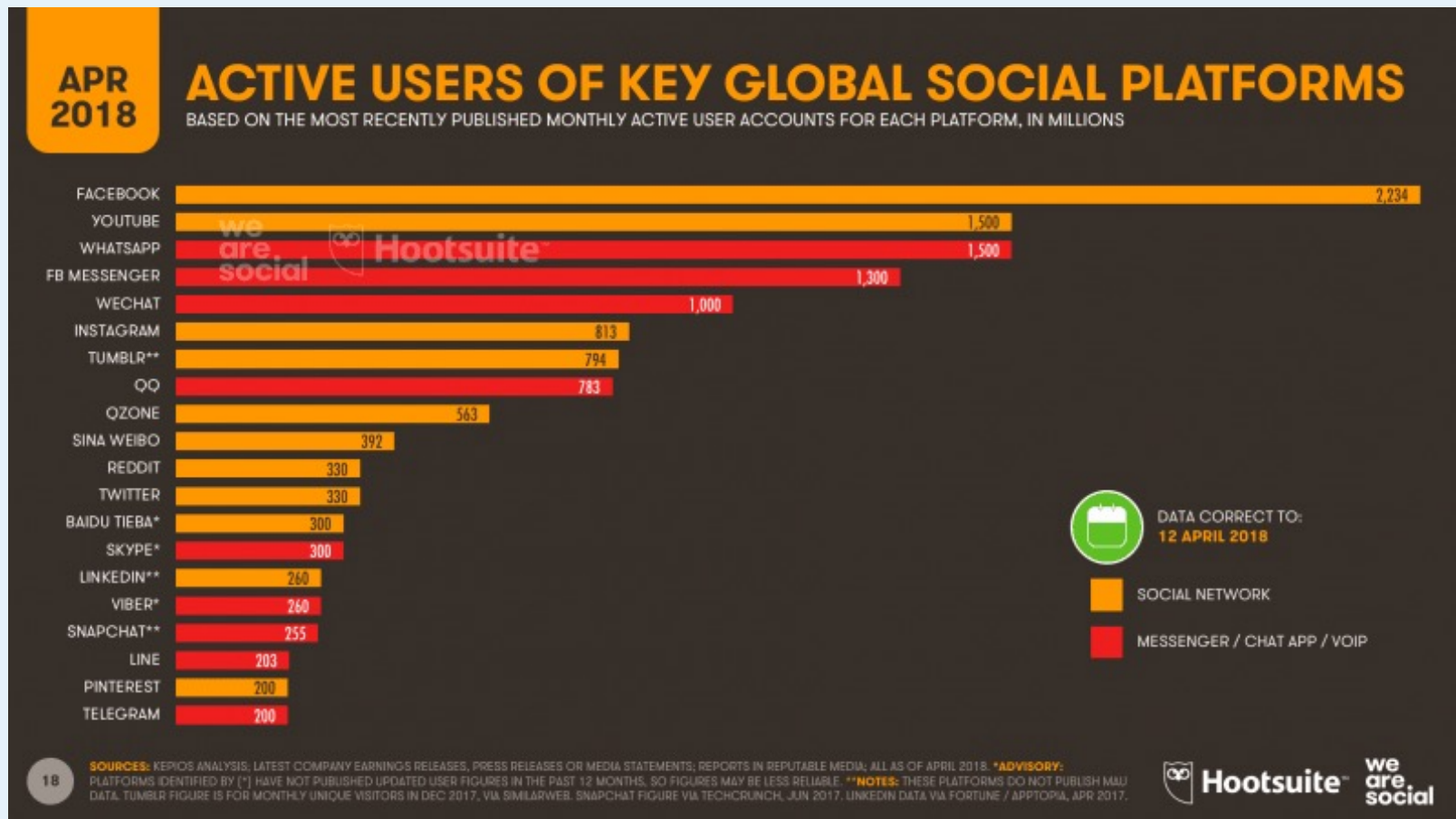


Twitter thread: <https://twitter.com/alexisohanian/status/1322927843421179904>

- 430 million monthly active users  
52 million daily active users



## 7<sup>th</sup> most popular social media site in the US

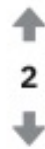




- 2020 Stats:
  - 303.4 Million posts
  - 2 Billion comments
- > 100,000 active communities
- > 2.6 Million subreddit
- More 2020 stats:  
<https://redditblog.com/2020/12/08/reddits-2020-year-in-review/>



—  
**Subreddit**  
—



Posted by u/[deleted] 5 years ago 

## what is a subreddit?

Resolved

please make it simple and easy to understand either im retarded or stupid because other sites cant explain it to me thanks for answering

 4 Comments  Share  Save  Hide  Report

75% Upvoted



NumberTwoFan · 5y

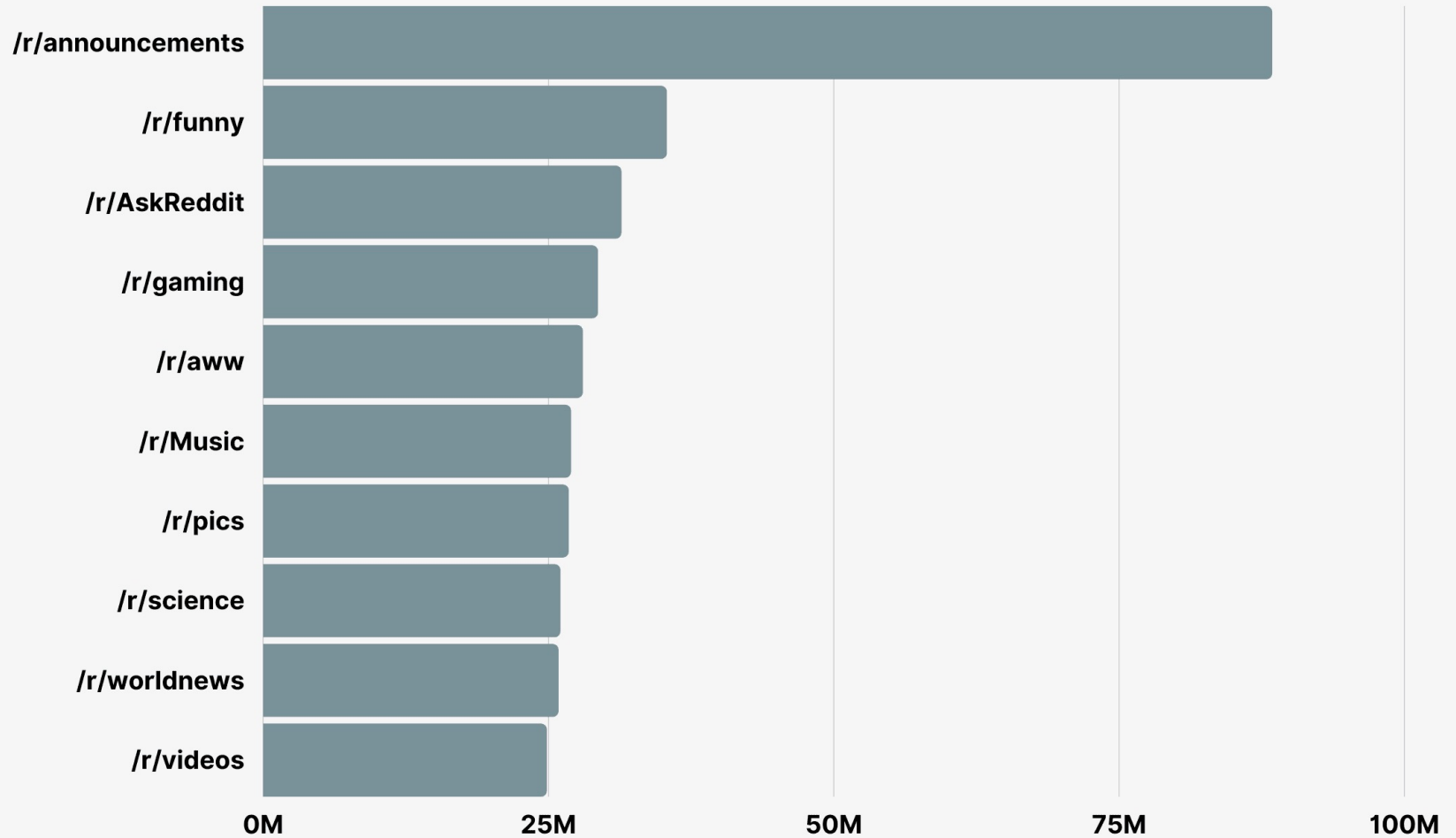
It is a web forum of a particular topic where you can post links or create a self post and discuss. You can subscribe if you like that topic or you can unsubscribe if you don't.

 4  Give Award  Share  Report  Save



- Community organized around posts
- Subreddits for basically everything
  - > 100,000 active communities
  - > 2.6 Million subreddit

# What are the most popular subreddits?



Live list: <http://redditlist.com/sfw>

Number of subscribers (in millions)





- <https://subredditstats.com/>
- <https://frontpagemetrics.com/r/>
- Question: Why might these stats be helpful



—

# Research based on Reddit data

—



International Conference on Web & Social Media (IWCSM) - <https://www.icwsm.org/>

The Web Conference (WWW) - <https://www2021.thewebconf.org/>

<https://cscw.acm.org/>  
Conference on Computer-Supported Cooperative Work and Social Computing -



**International Conference on Web & Social Media (IWCSM) - <https://www.icwsm.org/>**

The Web Conference (WWW) - <https://www2021.thewebconf.org/>

[Conference on Computer-Supported Cooperative Work and Social Computing - https://cscw.acm.org/](https://cscw.acm.org/)

# Examining Peer-to-Peer and Patient-Provider Interactions on a Social Media Community Facilitating Ask the Doctor Services

Alicia L. Nobles,<sup>1</sup> Eric C. Leas,<sup>2</sup> Mark Dredze,<sup>3</sup> John W. Ayers<sup>1</sup>

<sup>1</sup>Department of Medicine, University of California San Diego

<sup>2</sup>Department of Family Medicine and Public Health, University of California San Diego

<sup>3</sup>Department of Computer Science, Johns Hopkins University

{alnobles, ecl eas}@health.ucsd.edu, mdredze@cs.jhu.edu, ayers.john.w@gmail.com

1. What are the self-reported demographics of posters on r/AskDocs?
2. What health topics are ask about on r/AskDocs, how does this vary across demographics?
3. Does receipt of a response vary across demographics?
4. Does the empathy of responses vary across demographics or health topics?



# ASK A DOCTOR



## Medical Questions

r/AskDocs

Join



Create Post



Hot



New



Top



2



 PINNED BY MODERATORS

Posted by u/AutoModerator 1 day ago

### Weekly Discussion/General Questions Thread - May 24, 2021

 38 Comments  Award  Share  Save ...



33



Posted by u/googoohaha **Layperson/not verified as healthcare professional** 3 hours ago

### I vomit everytime I sneeze

29F, 5'7, 200 lbs, 115 mg methadone for seven years now, had high blood pressure since 13 but stopped taking medication for it 10 years ago and have had normal blood


### About Community



Having a medical issue? Ask a doctor or medical professional on Reddit! All flaired medical professionals on this subreddit are verified by the mods.

307k  
Members

1.3k  
Online

 Created Jul 10, 2013

Create Post

COMMUNITY OPTIONS



# Examining Peer-to-Peer and Patient-Provider Interactions on a Social Media Community Facilitating Ask the Doctor Services



How do they discover demographics?

- Regular Expressions

How do they discover topics?

- LDA

How do they measure empathy?

- Modification of LIWC

Category	Examples
Personal pronouns	<i>I, his, their</i>
Impersonal pronouns	<i>it, that, anything</i>
Articles	<i>a, an, the</i>
Conjunctions	<i>and, but, because</i>
Prepositions	<i>in, under, about</i>
Auxiliary verbs	<i>shall, be, was</i>
High-frequency adverbs	<i>very, rather, just</i>
Negations	<i>no, not, never</i>
Quantifiers	<i>much, few, lots</i>

Table 2: Functional word categories for LSM as defined by LIWC 2015 (Pennebaker et al. 2015)



## About Community



A subreddit of curated academic articles, pre-prints, books and conference papers on Reddit. With grey literature allowed on a case-by-case basis if relevant. Covering works which study Reddit or use the site as a proxy to investigate other social phenomena.


**116**

Members

**2**

Online





# — Reddit API and Pushshift —

# The Pushshift Reddit Dataset

**Jason Baumgartner**<sup>1,\*</sup>, **Savvas Zannettou**<sup>2,Ⓜ</sup>, **Brian Keegan**<sup>3</sup>, **Megan Squire**<sup>4</sup>, **Jeremy Blackburn**<sup>5,Ⓜ</sup>

<sup>1</sup>Pushshift.io, <sup>2</sup>Max Plank Institute, <sup>3</sup> University of Colorado Boulder, <sup>4</sup>Elon University, <sup>5</sup>Binghamton University

\*Network Contagion Research Institute, ⓂiDRAMA Lab

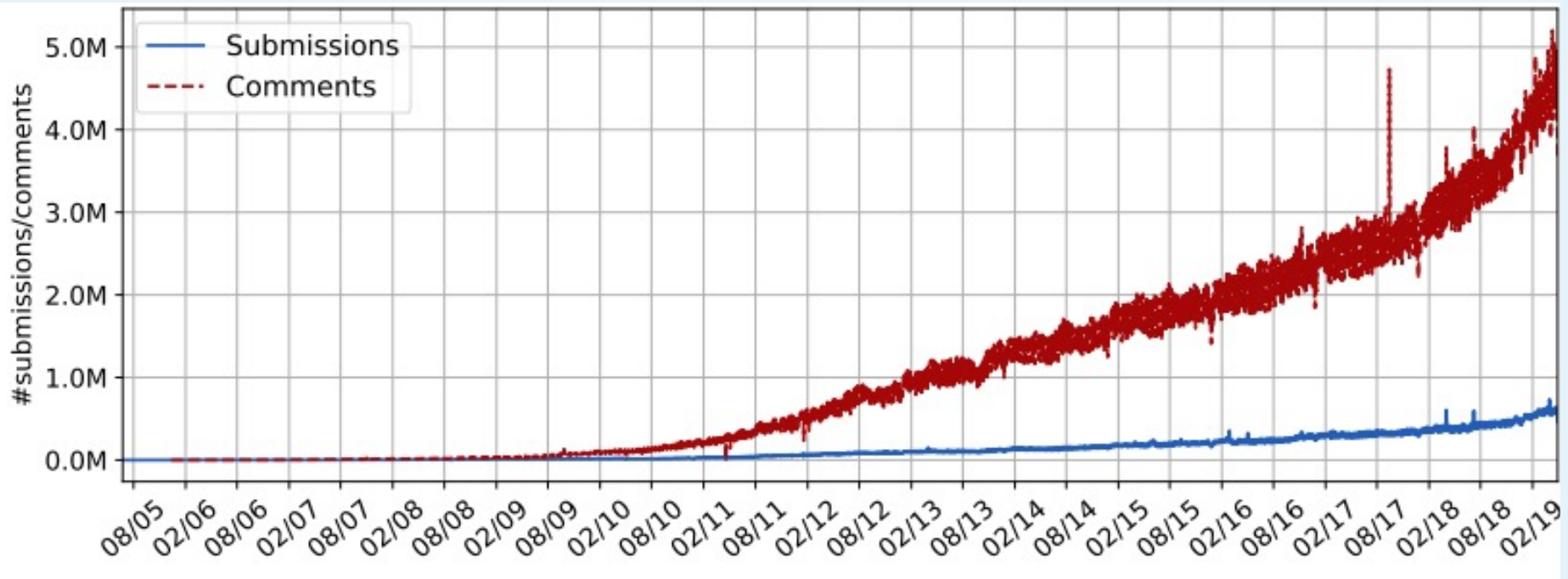
jason@pushshift.io, szannett@mpi-inf.mpg.de, brian.keegan@colorado.edu, msquire@elon.edu, blackburn@cs.binghamton.edu

*Social media data has become crucial to the advancement of scientific understanding ...*

*collecting large-scale social media data involves a high degree of engineering skill set and computational resources ...*

*research is often times gated by data engineering problems that must be overcome before analysis can proceed*

# Pushshift data



[Reddit data on Pushshift:](https://files.pushshift.io/reddit/)  
<https://files.pushshift.io/reddit/>



- Submission:
  - A post on a subreddit

# Posts on a subreddit



r/learnpython

Search

0 Comments Award Share Save

Posted by u/thenovastar17 6 hours ago

**19** **Newbie looking to learn**

Hello!

I have been wanting to learn Python for a while but have struggled to find a good website to commit to. I am a beginner in Python but have experience with other coding languages (Java, HTML, Ruby, etc.). What would you all recommend would be a good place to learn the basics + create projects? I have tried DataCamp, but it wouldn't let me continue with my course until I got the full access.

Thanks again!

16 Comments Award Share Save ...

Posted by u/buckale5 17 hours ago

**172** **How much do you worry about optimization?**

I write code almost exclusively to help myself with work tasks (large file manipulation, organizing files, spreadsheet analysis, etc...). I'm a very rank coder who mostly specializes in copy and pasting. I recently wrote some code to help with a process that I previously did in Excel. The Excel version took 6 minutes and the Python/Pandas version takes less than 2. However, I'm sure there are some major holes in my logic and I'm sure hope optimizations could be made to get it quicker. But I only need it twice a day and it's already significantly faster than my last version.

Do you worry about getting code as fast as possible or is working pretty well good enough for you?

If you do optimize, how do you recommend finding where your bottlenecks are?



66 Comments Award Share Save ...



- Submission:
  - A post on a subreddit
  
- Comments:
  - Replies to a submission

# Threaded Comments






 BlackRussianJedi · 5h  
Voted 

While I don't entirely agree, thank you so much for this very well-reasoned breakdown. I don't entirely disagree either. I hold both, but did from the beginning. Perhaps I misunderstand the shitadel and friends situation, but it seems like having a lot of people hold both puts a lot more strain on the shfs.

I also feel that at this point, there's virtually nothing that can impede the GME squeeze from going to infinity. Not paperhands, not the other stock, nothing. If we trust the DD, the paperhands (if there are still any left) won't impact the price even if they sell. We own the float many times over, if I understand correctly.


I agree with your other points. Thank you for the breakdown in value. I think the other stock feels more accessible for people, it certainly did for me, but you're absolutely right... looking at the numbers, A single GME share likely goes higher than 100 of the other one. But I think early on, people like me weren't confident about either stock actually mooning, because it looked like the squeeze was killed in January, and there really wasn't much meaningful DD back then compared to now. If I was afraid of losing money at first, and wasn't sure about a squeeze for either, the lower barrier of entry would be much more attractive. Now that there is plenty of DD, in my opinion, the logical play is GME. I've been holding a boatload of the other since Feb, and a modest amount of GME since Jan, but all of my money recently has gone into GME. Thanks again for the breakdown. This is not financial advice, no one should listen to anything I say, I am a total amateur.



↑ 12 ↓  Reply Give Award Share Report Save

 Kuzuuryu1 · 5h  
Voted 


Good points, the fact that they need to fight on multiple fronts is a good one, however we do not know how much money it actually costs them to suppress the prices, for all we know it might actually be really cheap especially the non-existent interest rates. So while it may be a good point we do not know if it actually hurts them that much to fight 2 sides.


The other point being that we hold the float, yes the definitely do, however what TRULY matters is, does diamond hands hold the float, you can hold 10000% of the float, if everyone sells at 1000, then the squeeze is over. All that matters is for minimum 100% of the float to be held by people who refuse to sell until millions.

↑ 9 ↓  Reply Give Award Share Report Save


 BlackRussianJedi · 4h  
Voted 


True that. But for some reason, I feel super confident that most GME holders are diamond handed maniacal apes. I personally know that I will hold a decent amount of my shares indefinitely. Hell, selling 1 or maybe 2 on the way down will be more than enough for me. I will hold onto the rest out of spite. I think many others feel this way too, but to your point, I guess there's really no way to know. But I believe the majority will hold. Cheers buddy

↑ 4 ↓  Reply Give Award Share Report Save

 Kuzuuryu1 · 3h  
Voted 

You're very optimistic, that's good, however as you've probably seen, I'm more on the pessimistic side, right now people are holding with a diamond grip on their shares because they have nothing to lose and haven't gained anything yet, however once the price rises, and people suddenly have something to lose, many will paperhand at the smallest loss because they never had any control over their emotions and won't ever have any. Then the sudden wave of regret when they sell and the prices rises right back up, trust me many will invest in \$rope after being unable to cope with their regrets.

↑ 3 ↓  Reply Give Award Share Report Save

 Mundane-Answer · 5h  
Voted 

I've generally been pretty lukewarm on AMC, but have actually bought in a few recently as an anti-panic measure. Because I will be a complete mess during MOASS and need to take measures to mitigate my anxiety.

AMC is cheap (I only deal in whole GME and I'm not a wealthy ape) and with the level of control and manipulation the hedgefunds have here, I would be very surprised if it didn't moon before



## Submissions

Parameter	Description	Default	Accepted Values
ids	Get specific submissions via their ids	N/A	Comma-delimited base36 ids
q	Search term. Will search ALL possible fields	N/A	String / Quoted String for phrases
q:not	Exclude search term. Will exclude these terms	N/A	String / Quoted String for phrases
title	Searches the title field only	N/A	String / Quoted String for phrases
title:not	Exclude search term from title. Will exclude these terms	N/A	String / Quoted String for phrases
selftext	Searches the selftext field only	N/A	String / Quoted String for phrases
selftext:not	Exclude search term from selftext. Will exclude these terms	N/A	String / Quoted String for phrases
size	Number of results to return	25	Integer <= 500
fields	One return specific fields (comma delimited)	All Fields	String or comma-delimited string (Multiple values allowed)
sort	Sort results in a specific order	"desc"	"asc", "desc"
sort_type	Sort by a specific attribute	"created_utc"	"score", "num_comments", "created_utc"
aggs	Return aggregation summary	N/A	["author", "link_id", "created_utc", "subreddit"]
author	Restrict to a specific author	N/A	String or comma-delimited string (Multiple values allowed)
subreddit	Restrict to a specific subreddit	N/A	String or comma-delimited string (Multiple values allowed)
after	Return results after this date	N/A	Epoch value or Integer + "s,m,h,d" (i.e. 30d for 30 days)
before	Return results before this date	N/A	Epoch value or Integer + "s,m,h,d" (i.e. 30d for 30 days)
score	Restrict results based on score	N/A	Integer or > x or < x (i.e. score=>100 or score=<25)
num_comments	Restrict results based on number of comments	N/A	Integer or > x or < x (i.e. num_comments=>100)
over_18	Restrict to nsfw or sfw content	both allowed	"true" or "false"
is_video	Restrict to video content	both allowed	"true" or "false"
locked	Return locked or unlocked threads only	both allowed	"true" or "false"
stickied	Return stickied or unstickied content only	both allowed	"true" or "false"
spoiler	Exclude or include spoilers only	both allowed	"true" or "false"
contest_mode	Exclude or include contest mode submissions	both allowed	"true" or "false"
frequency	Used with the aggs parameter when set to created_utc	N/A	"second", "minute", "hour", "day"
metadata	display metadata about the query	false	["true", "false"]

## Comments

Parameter	Description	Default	Accepted Values
q	Search term.	N/A	String / Quoted String for phrases
ids	Get specific comments via their ids	N/A	Comma-delimited base36 ids
size	Number of results to return	25	Integer <= 500
fields	One return specific fields (comma delimited)	All Fields Returned	string or comma-delimited string
sort	Sort results in a specific order	"desc"	"asc", "desc"
sort_type	Sort by a specific attribute	"created_utc"	"score", "num_comments", "created_utc"
aggs	Return aggregation summary	N/A	["author", "link_id", "created_utc", "subreddit"]
author	Restrict to a specific author	N/A	String
subreddit	Restrict to a specific subreddit	N/A	String
after	Return results after this date	N/A	Epoch value or Integer + "s,m,h,d" (i.e. 30d for 30 days)
before	Return results before this date	N/A	Epoch value or Integer + "s,m,h,d" (i.e. 30d for 30 days)
frequency	Used with the aggs parameter when set to created_utc	N/A	"second", "minute", "hour", "day"
metadata	display metadata about the query	false	"true", "false"





## Submissions:

- ids: Get specific submissions via their ids
- q: search term
- title: search the title field
- Selftext: searches text
- subreddit
- Author
- After/before
- Num\_comments

## Comments:

- ids
- q
- Author
- Subreddit
- size

[Documentation can be found here:](https://github.com/pushshift/api)  
<https://github.com/pushshift/api>



# What's in a Submission



- author
- title
- Selftext
- Created\_at
- Score
- num\_comments
- Subreddit
- url



# Whats in a comment



- author
- title
- Selftext body
- Created\_at
- Score
- num\_comments
- Subreddit
- url