BC COMS 2710: Computational Text Analysis

Lecture 14 – Data Collection API’s - Reddit
Announcements – Assignments

- Readings 04:
  - link posted to course site later today
  - due Sunday

- HW 02:
  - Due Wednesday night
Remaining Assignments

- **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
Final Project – Deliverables

- Project ideation – Friday May 28th
  - [https://www.overleaf.com/read/yzpgxcgsqdvp](https://www.overleaf.com/read/yzpgxcgsqdvp)

- roughly 250 word overview of what you are interested in
Final Project – Deliverables

- Project ideation – Friday May 28th
  - 5 points

- Project proposal – Friday June 4th
  - 9 points

- Project presentations – Monday June 14th
  - 6 points

- Project submissions – Friday June 18th
  - 15 points

- 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
Reddit History

- 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

Copyright © 2016 Barnard College
430 million monthly active users
52 million daily active users

Source: https://backlinko.com/reddit-users
7th 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:
what is a subreddit?

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

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.
Community organized around posts

Subreddits for basically everything
  • > 100,000 active communities
  • > 2.6 Million subreddit
What are the most popular subreddits?

- /r/announcements
- /r/funny
- /r/AskReddit
- /r/gaming
- /r/aww
- /r/Music
- /r/pics
- /r/science
- /r/worldnews
- /r/videos

Live list: http://redditlist.com/sfw
Finding subreddit statistics

- https://subredditstats.com/
- https://frontpagemetrics.com/r/

Question: Why might these stats be helpful
Research based on Reddit data
Research Venues

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

The Web Conference (WWW) - [https://www2021.thewebconf.org/](https://www2021.thewebconf.org/)

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

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

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

1. What are the self-reported demographics of posters on r/AskDocs?
2. What health topics are asked 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?

Weekly Discussion/General Questions Thread - May 24, 2021

Posted by u/AutoModerator 1 day 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
How do they discover demographics?
- Regular Expressions

How do they discover topics?
- LDA

How do they measure empathy?
- Modification of LIWC

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

https://www.reddit.com/r/AcademicReddit/
Reddit API and Pushshift
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/
Submission:
  • A post on a subreddit
Posts on a subreddit

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!

---

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?
Submissions & Comments

- **Submission:**
  - A post on a subreddit

- **Comments:**
  - Replies to a submission
Threaded Comments

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 shits.

I also feel that at this point, there's virtually nothing that can impede the Game 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 Game 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 moving, 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 Game. I've been holding a boatload of the other since Feb, and a modest amount of Game since Jan, but all of my money recently has gone into Game. Thanks again for the breakdown. This is not financial advice, no one should listen to anything I say, I am a total amateur.

---

32

Reply
Give Award
Share
Report
Save

Kuwait

5h

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.

---

5

Reply
Give Award
Share
Report
Save

BlackPancakeKid

1h

True that. But for some reasons, I feel super confident that most Game 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.

---

2

Reply
Give Award
Share
Report
Save

Nemojose

1h

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 stocks after being unable to cope with their regrets.

---

1

Reply
Give Award
Share
Report
Save

Ph Nobel

1h

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 MDAX and need to take measures to mitigate my anxiety.

AMC is cheap (I only deal in whole Game 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
### Search parameters

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

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
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