

DS501 - INTRO TO DS

CASE STUDY 1 : REDDIT POST

ANALYZER USING PRAW

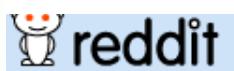
OBJECTIVE

The code is designed to fetch Reddit posts that meet specific criteria and extract information, including post scores, titles, and the top three comments for each post. If a comment is deleted, it replaces it with the next available top comment for the same post.

STEPS:

- Initialize Reddit API Client
- Define subreddit (r/datascience in our case study)
- Create an empty list to store data.
- Fetch submissions from the specified sub reddit on the following criteria:
 1. Posts are only from year 2023
 2. the integer time format of dateconverted into year-month-day
 3. Posts with score > 50 and words in title more than 5 were considered
 4. Minimum 100 posts along with top three comments for each post were collected.
- Create a pandas data frame from the data and convert it into csv

MY TOP 3 POSTS

 reddit [EXPERIENCEDDEVS](#) [comments](#) [other discussions \(1\)](#)

[I investigated the Underground Economy of Glassdoor Reviews](#) ([self.ExperiencedDevs](#))
submitted 4 months ago * by ibsurvivors

1378

Online company reviews are high stakes.
Top reviews on sites like Glassdoor and Google can get thousands of impressions each month and are major drivers of brand perception.
Employers know this. And when I come across multiple 5 star reviews left with no cons, or a Pulitzer worthy essay from a former intern, I become suspicious.
These reviews start to resemble 30 under 30 lists: so artificially constructed that you begin to question their credibility in the first place.
The scrutiny around company reviews is well documented; some companies file lawsuits worth over a million dollars to reveal anonymous reviewers that complain about their jobs.
Whilst it's the flashy lawsuits that make the headlines, there also exists an underground economy of company reviews operating quietly every single day.
In this underground economy, some companies pay over \$150 to freelancers to try and get a negative review removed. If they want "better" results, they go to the plethora of Online Reputation Management services (ORMs) in the United States that can charge retainers worth thousands of dollars.
The supply of positive reviews exists too. My research led me to find companies, including a prominent Y-Combinator backed startup, that solicit fake positive reviews from online freelancers to improve their rating.
Many of these mercenary fake reviewers, often based in South East Asia, make a full time living doing this, netting over \$2,000 per month.
Some of these run such sophisticated operations that they've even created their own pricing tiers (e.g \$35 per original review, \$20 to post an already created review from an email address), a la SaaS offering.
Others operate on a contingency fee agreement model, where they only get paid if they're able to take a negative review down.
The underground economy of company reviews is well and truly alive. And today we're going to find out how it operates.

[+] [roshambo11](#) 103 points 4 months ago
Very thorough analysis OP. I wonder if there's a similar effect with college rankings. I don't know if universities put their thumb on the scale as much, but I think the perception aspect plays into that kind of admissions marketing
[permalink](#) [embed](#) [save](#) [report](#) [reply](#)

[+] [forbiscuit](#) 74 points 4 months ago*
There's an exceptional study by a Math professor in Cornell Columbia who investigated this (and his report later plummeted their rating within U.S. Newsweek from #2 to #18 as of May 2023):
<https://www.math.columbia.edu/~thaddeus/ranking/investigation.html>
[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [reply](#)

[+] [schwarzbaer](#) 24 points 4 months ago
Small correction: Columbia instead of Cornell. You linked to the right page (hosted on [columbia.edu](#)).
[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [reply](#)

[+] [forbiscuit](#) 13 points 4 months ago
Good catch! I think I've been thinking too much about Cornell while writing this :P
[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [reply](#)

[+] [Magrik](#) 4 points 4 months ago
Boner Champ?
[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [reply](#)

[+] [forbiscuit](#) 2 points 4 months ago
My dude, Broccoli Rob was Broccoli Rob!
[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [reply](#)

[+] [bgighijgftuik](#) 157 points 4 months ago*
Incredibly thorough analysis. And that's why I don't trust Glassdoor at all
[permalink](#) [embed](#) [save](#) [report](#) [reply](#)

[+] [ibsurvivors](#) 157 points 4 months ago

Channel	Cost	Timeline	Model	Self reported success rate
Freelancer #1	\$100 per review	3 days	Contingency Agreement Model	100%
Freelancer #2	\$30 per review	7 days	Contingency Agreement Model	100%
Reputation management service #2	\$450 per review	21 business days	Contingency Agreement Model	Unknown
Reputation management service #3	\$1000 per review	Undefined	Contingency Agreement Model	100%
Reputation management service #4 Plan 1	\$550 per review	5-6 weeks	Contingency Agreement Model	50-75%
Reputation management service #4 Plan 2	\$300 Subscription + \$100 per each review removed	Monthly service	Subscription plan	50-75%
Freelancer #3	\$20	Undefined	Pay regardless	Undefined
Freelancer #4	\$500	Undefined	Contingency Agreement Model	Undefined

As you can see, unlike the fake review generation market, the prices vary quite a bit for getting reviews removed. At one end, you have freelancers on gig marketplaces that will attempt to remove a review for less than \$100. And then on the other end, you have ORMs (Online Reputation Management Agencies) that have multiple employees and more comprehensive packages in place. The one constant seems to be that most companies operate on a contingency



DATASCIENCE

comments

Discussion Help me understand how to think about Generative AI on my career (self.datascience)
submitted 3 months ago by brybrydataguy

I have been trying to get my head around Generative AI (GAI) for a few weeks. Specifically how I should future proof myself around it so that I don't find myself in the data science equivalent as a colbot programmer in some bank basement. Here are some of my scatter, and maybe over optimisitic, thoughts so far:

Product Data Science

I'm starting to think that GAI will be a boon for product data scientists that specialize in statistical inference and rigorous product analytics. The number and diversity of products that use GAI are and will continue to explode. These will all be digital products with digital exhausts that have the added complication of variance in product experience around the nature of LLMs and natural language prompt diversity. Product iteration will use more data and have more degrees of freedom for improvements. I suspect exceptional inference/stats will be more important in the future.

Automating away DS Jobs

I am not worried about GAI removing the need for DS even as I see examples of Text -> SQL and summary and visualizations of Data. Personally, my most valuable contributions are not around doing SQL or making visualizations (expert about both), but around the judgement I execute or the insights I have around the outputs. I look forward to using these tools myself and not being asked to do this type of work as one-offs because it's so easy for others in the org to do it themselves.

ML Work

I am a little less clear on how ML works is going to change. On one hand I think this will explode because the number tasks that can have better predictions that generate is going to grow. The economics of costly FP/FN will make sense as the predictions get much better which will create new businesses and business models. On the other hand, with these better and more diverse prediction will come more inference cost, more pipelines, more chains of dependencies, and more product judgment about what to predict and what to do with the predictions. I think this complexity will lead to heterogeneity in outcomes between different companies based on experience/culture. I suspect the cloud providers will be the real winners here as they build tools to help with all the GAI integrations.

save

content policy formatting help

↑ [-] sven_ftw 12 points 3 months ago

↓ Did having really, really good calculators remove the need for engineers who know math?

permalink embed save report reply

↑ [-] brybrydataguy [S] 3 points 3 months ago

↓ Not at all. Excel increased the number of people that did math and had a huge impact on productivity.

GAI is different because excel improved speed and quality. It's not clear to me on what dimensions GAI is faster and more reliable.

permalink embed save parent report reply

↑ [-] sven_ftw 2 points 3 months ago

↓ Getting code pipelines, summarizing code for working docs, etc. All these tasks that had to be done somewhat manually before became trivial to generate mostly right baseline stuff. Huge, huge productivity gains.

permalink embed save parent report reply

↑ [-] brybrydataguy [S] 3 points 3 months ago

↓ 100% agreed summarization will be much faster and allow us more time to do other work. How much free time would that free up for you? Personally I don't see this dimension a huge unlock on most of my work.

I think you posted this comment multiple times.

permalink embed save parent report reply

↑ [-] sven_ftw 1 point 3 months ago

↓ Yeah was weird. My app on phone bugged out. Deleted the others!

permalink embed save parent report reply



what aspects of data science can chat gpt automate



ChatGPT can automate several aspects of data science, including:



1. Data Preparation: Automating tasks such as data cleaning, normalization, and transforming raw data into a format suitable for analysis.
2. Predictive Modeling: Automating the process of building and deploying machine learning models for tasks such as classification, regression, and clustering.
3. Data Visualization: Generating visualizations such as bar charts, line graphs, and heat maps to help make data insights more accessible and understandable.
4. Natural Language Processing (NLP): Automating tasks such as sentiment analysis, named entity recognition, and topic modeling.
5. Data Insights Generation: Automatically generating insights and recommendations based on data analysis and machine learning models.

Note that ChatGPT can assist with these tasks, but a human data scientist is still required to validate the results and make final decisions based on the output.



DATASCIENCE

comments



Discussion What else is left? Should I continue with my masters in DS?

(i.redd.it)

submitted 7 months ago by burralohit01

[↑] [-] mrbrambles 289 points 7 months ago
Chatgpt is good at fluffing it's own resume
[permalink](#) [embed](#) [save](#) [report](#) [reply](#)

[↑] [-] burralohit01 [S] 22 points 7 months ago
Nicee, good one.
[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [reply](#)

[↑] [-] Unlikely_Tie8166 442 points 7 months ago
ChatGPT subscription is 20\$/month. You'll have to work for less to remain competitive in the market. Jokes aside, I can't imagine data science being automated in the foreseeable future. That said, you might have to deal with lesser amount of annoying crap that usually takes 90% of data scientist's time
[permalink](#) [embed](#) [save](#) [report](#) [reply](#)

[↑] [-] renok_archnmy 109 points 7 months ago
It's a Microsoft product so more like \$1500/seat/year to use the output or \$1500/core/year to license for use of feature on servers with a minimum of 32 cores to support 100 headcount active users.
Or you can use azure, where you're charged \$1/compute minute but they don't tell you you gotta pay \$1500/month for VPN endpoint, another \$1500/month for firewall, and if you need any semblance of data privacy and security you gotta go with gov tier account.
Oh, and that \$1/compute minute grows exponentially as all the executives and directors fumble their way through inputting overly vague prompts that don't produce the answers they expect and can't rage out on a robot so they hire us back so we can be the whipping post and expected to translate human language to succinct and precise computer prompts to get the exact data they want for their TPS reports out the computer and somehow this is better than just writing SQL...
[permalink](#) [embed](#) [save](#) [parent](#) [report](#) [reply](#)

“

THANK YOU