Most Valued Data Science Skills

Guillermo Schneider ● Jonathan Cruz ● Lucas Weyrich ● Richie Rivera

●●●

March 20, 2024

Resources

Reddit

r/dataengineering

u/AutoModerator • 17 days ago

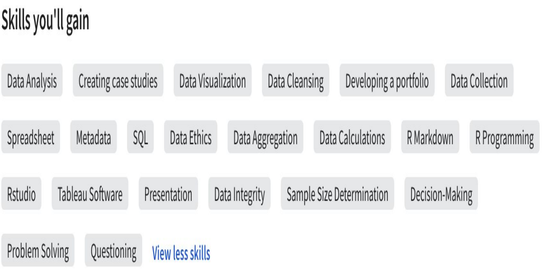
Monthly General Discussion - Mar 2024

Discussion

https://www.reddit.com/r/dataengineering/comments/1b3zb2c/monthly_general_discussion_...

- R Package "RedditExtractoR"
- Scrape comments from the past year
  - r/datascience (1.2m members)
  - r/DataEngineering (169k members)
- Yielded a combined data frame of over 120,000 comments for analysis.
- Project's replicability is somewhat limited due to the dynamic nature of the subreddits

Coursera

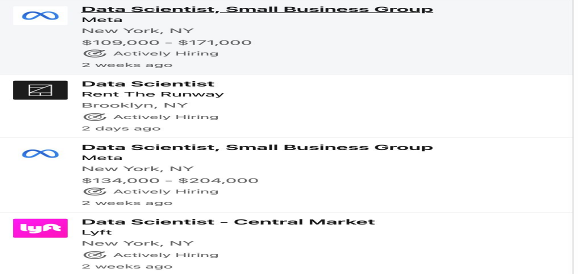


Skills you'll gain

Data Analysis · Creating case studies · Data Visualization · Data Cleansing · Developing a portfolio · Data Collection · Spreadsheet · Metadata · SQL · Data Ethics · Data Aggregation · Data Calculations · R Markdown · R Programming · Rstudio · Tableau Software · Presentation · Data Integrity · Sample Size Determination · Decision-Making · Problem Solving · Questioning

- Python lib BeautifulSoup4 and request
- Extract advertised skills that are to be acquired from most popular data science courses

Linkedin



Data Scientist, Small Business Group
Meta
New York, NY
$109,000 – $171,000
Actively Hiring
2 weeks ago

Data Scientist
Rent The Runway
Brooklyn, NY
Actively Hiring
2 days ago

Data Scientist, Small Business Group
Meta
New York, NY
$134,000 – $204,000
Actively Hiring
2 weeks ago

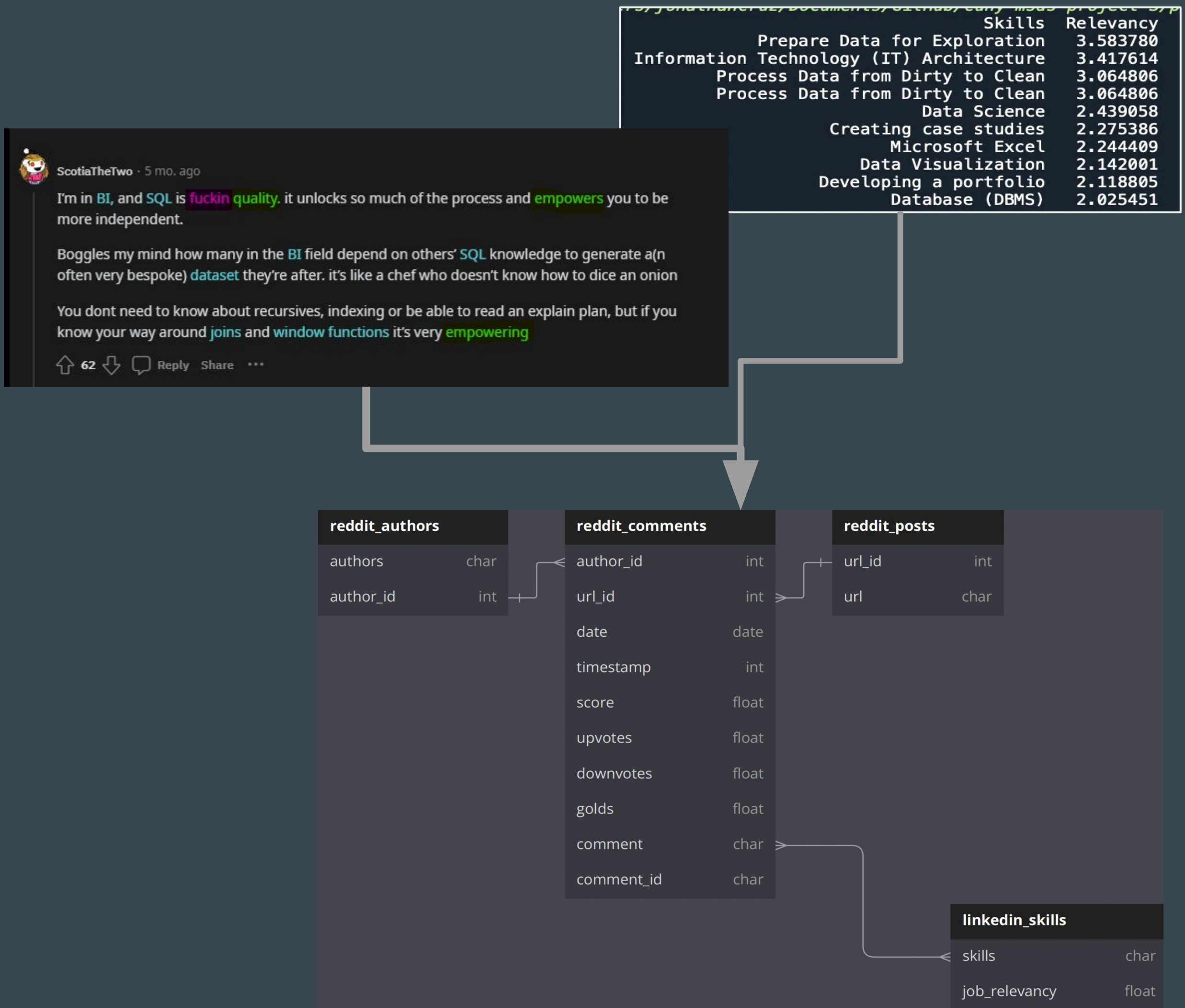
Data Scientist – Central Market
Lyft
New York, NY
Actively Hiring
2 weeks ago

- Python lib BeautifulSoup4 and request
- Extract "Skills Required" section from over 50 of the latest job postings on LinkedIn

Data Loading

- Two files were created to scrape Reddit comments and LinkedIn job postings

- Data was tidied and then normalized into a SQLite database
    - reddit_comments
    - reddit_authors
    - reddit_posts
    - linkedin_skills



```
                                   Skills  Relevancy
               Prepare Data for Exploration   3.583780
Information Technology (IT) Architecture   3.417614
             Process Data from Dirty to Clean   3.064806
             Process Data from Dirty to Clean   3.064806
                                 Data Science   2.439058
                         Creating case studies   2.275386
                              Microsoft Excel   2.244409
                          Data Visualization   2.142001
                        Developing a portfolio   2.118805
                              Database (DBMS)   2.025451
```

ScotiaTheTwo · 5 mo. ago

I'm in BI, and SQL is fuckin quality. it unlocks so much of the process and empowers you to be more independent.

Boggles my mind how many in the BI field depend on others' SQL knowledge to generate a(n often very bespoke) dataset they're after. it's like a chef who doesn't know how to dice an onion

You dont need to know about recursives, indexing or be able to read an explain plan, but if you know your way around joins and window functions it's very empowering

62 Reply Share ···

reddit_authors	
authors	char
author_id	int

reddit_comments	
author_id	int
url_id	int
date	date
timestamp	int
score	float
upvotes	float
downvotes	float
golds	float
comment	char
comment_id	char

reddit_posts	
url_id	int
url	char

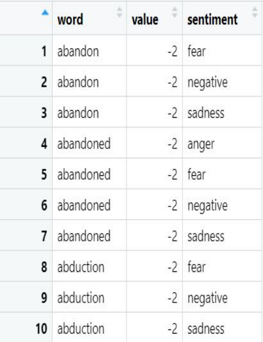
linkedin_skills	
skills	char
job_relevancy	float

Sentiment Analysis Prep

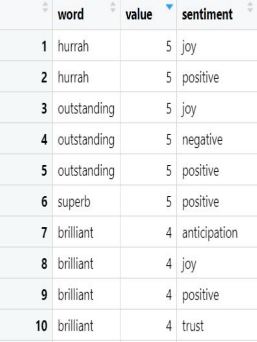
Step 1 - Sentiments

AFINN by Finn Årup Nielsen, **bing** by Bing Liu and collaborators, **nrc** by Saif Mohammad and Peter Turney.

These dictionaries categorize **sentiments** (positive/negative) and **emotions** (joy/anger/disgust/etc) on a scale from -5 to 5, ranking positivity.



	word	value	sentiment
1	abandon	-2	fear
2	abandon	-2	negative
3	abandon	-2	sadness
4	abandoned	-2	anger
5	abandoned	-2	fear
6	abandoned	-2	negative
7	abandoned	-2	sadness
8	abduction	-2	fear
9	abduction	-2	negative
10	abduction	-2	sadness

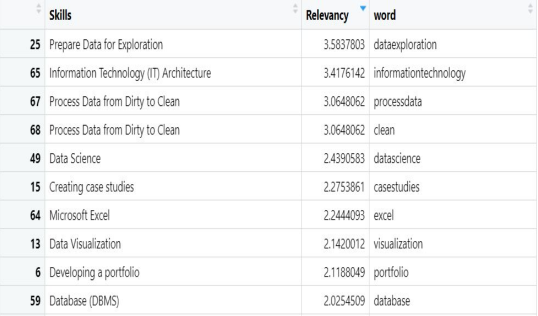


	word	value	sentiment
1	hurrah	5	joy
2	hurrah	5	positive
3	outstanding	5	joy
4	outstanding	5	negative
5	outstanding	5	positive
6	superb	5	positive
7	brilliant	4	anticipation
8	brilliant	4	joy
9	brilliant	4	positive
10	brilliant	4	trust

Step 2 - Two-word Skills

Two-word skills were merged into single words (e.g., **"Data Science"** became **"datascience"**)

We applied the same transformation to all the Reddit comments collected by Lucas using gsub.

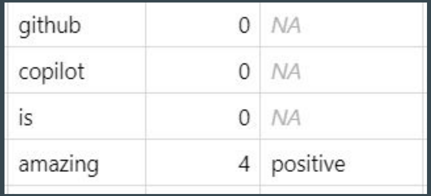


	Skills	Relevancy	word
25	Prepare Data for Exploration	3.5837803	dataexploration
65	Information Technology (IT) Architecture	3.4176142	informationtechnology
67	Process Data from Dirty to Clean	3.0648062	processdata
68	Process Data from Dirty to Clean	3.0648062	clean
49	Data Science	2.4390583	datascience
15	Creating case studies	2.2753861	casestudies
64	Microsoft Excel	2.2444093	excel
13	Data Visualization	2.1420012	visualization
6	Developing a portfolio	2.1188049	portfolio
59	Database (DBMS)	2.0254509	database

Step 3 - Separate Words

Comments were **unnested** into individual words (grouped by comment ID).

Using the sentiment dictionaries, each word was assigned a **score** and a **sentiment**.



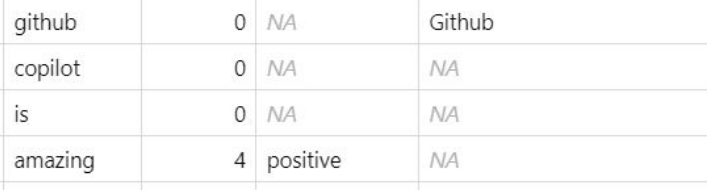
github	0	NA
copilot	0	NA
is	0	NA
amazing	4	positive

Sentiment Analysis Results

Step 4 - Filtering for Skill words

We filtered for words that matched our Skills.

Using those comment IDs, we then filtered the full comment scrape to just include comment IDs that were in that list.

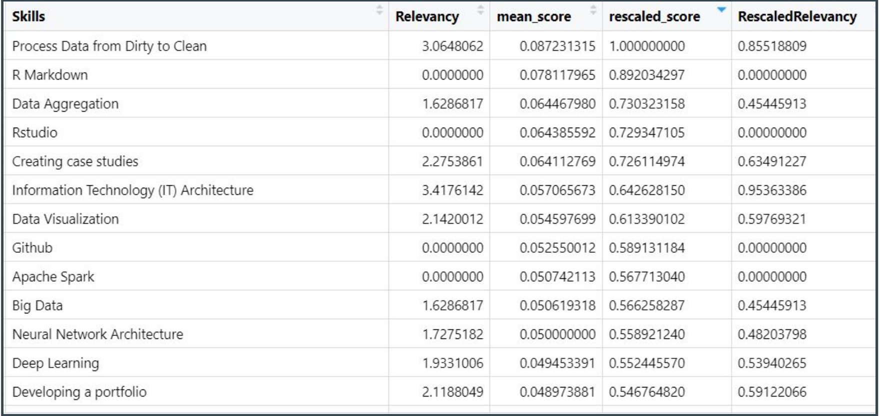


github	0	NA	Github
copilot	0	NA	NA
is	0	NA	NA
amazing	4	positive	NA

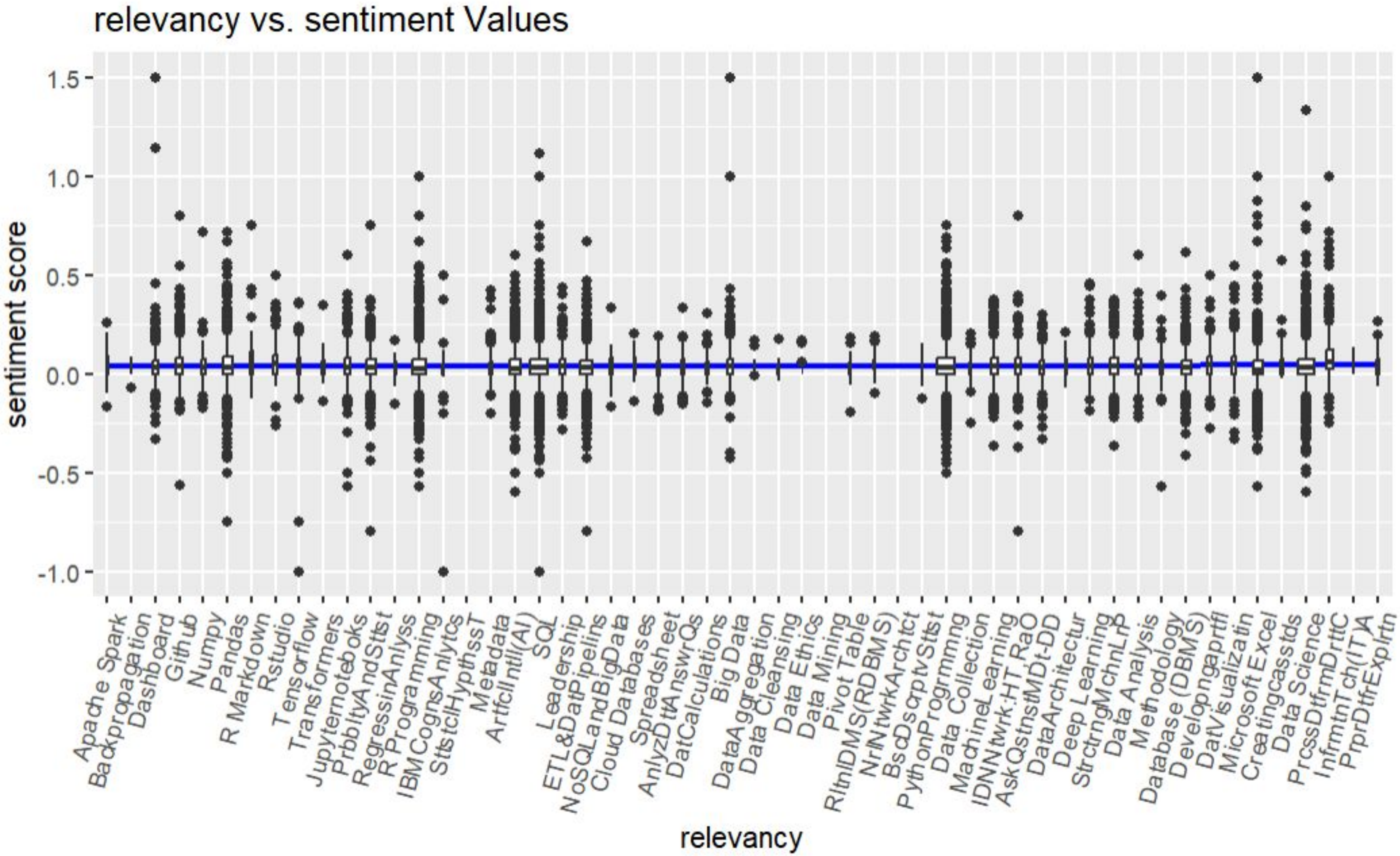
Step 5 - Average Skill Sentiment Score

We grouped words first by comment ID for **Average Comment Sentiment Score**, and then by Skills for **Average Skills Sentiment Score.**

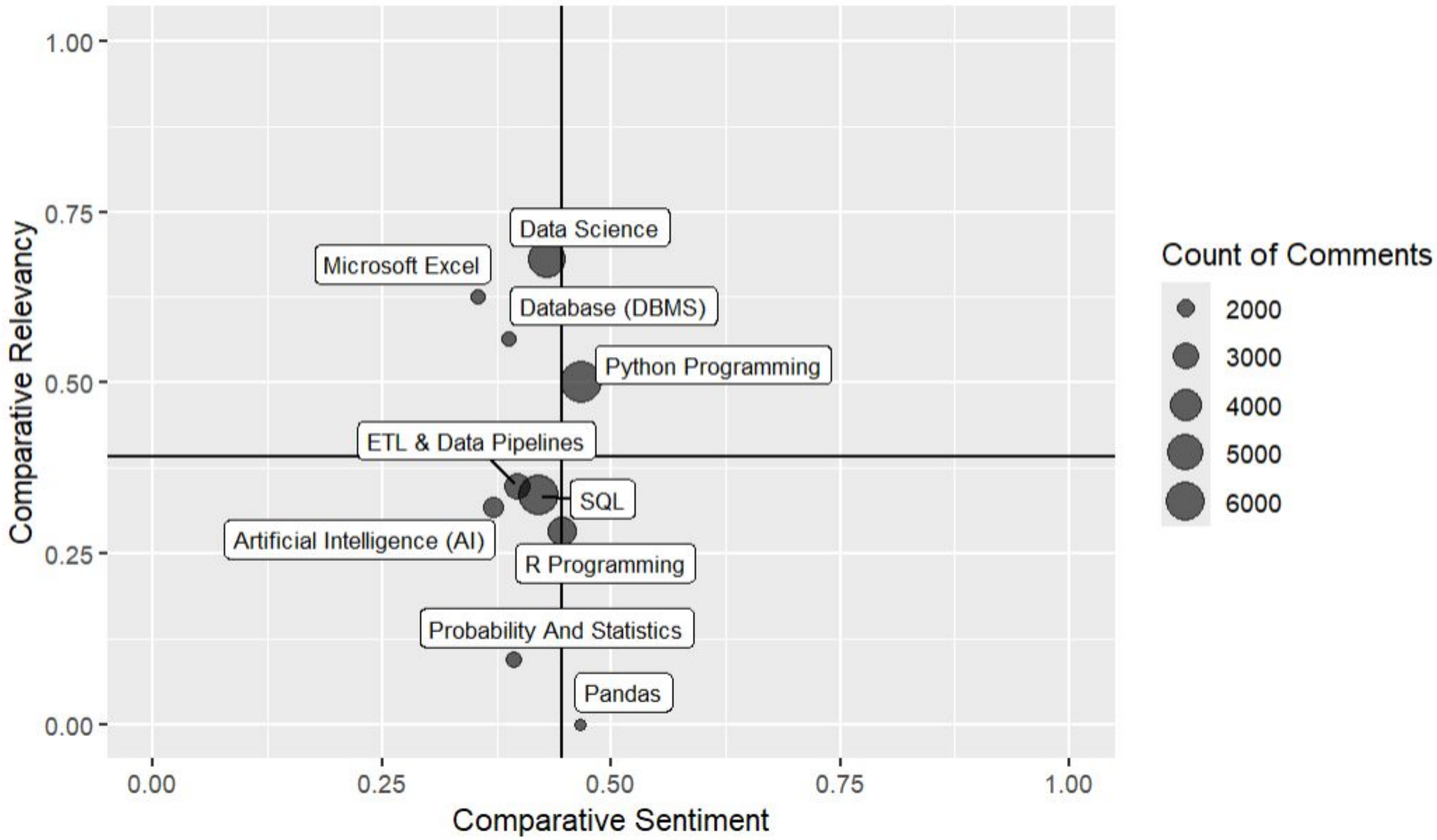
'Sum of sentiment scores in a comment' / 'Total words in a comment' =Average Comment Sentiment Score



Skills	Relevancy	mean_score	rescaled_score	RescaledRelevancy
Process Data from Dirty to Clean	3.0648062	0.087231315	1.000000000	0.85518809
R Markdown	0.0000000	0.078117965	0.892034297	0.00000000
Data Aggregation	1.6286817	0.064467980	0.730323158	0.45445913
Rstudio	0.0000000	0.064385592	0.729347105	0.00000000
Creating case studies	2.2753861	0.064112769	0.726114974	0.63491227
Information Technology (IT) Architecture	3.4176142	0.057065673	0.642628150	0.95363386
Data Visualization	2.1420012	0.054597699	0.613390102	0.59769321
Github	0.0000000	0.052550012	0.589131184	0.00000000
Apache Spark	0.0000000	0.050742113	0.567713040	0.00000000
Big Data	1.6286817	0.050619318	0.566258287	0.45445913
Neural Network Architecture	1.7275182	0.050000000	0.558921240	0.48203798
Deep Learning	1.9331006	0.049453391	0.552445570	0.53940265
Developing a portfolio	2.1188049	0.048973881	0.546764820	0.59122066


relevancy vs. sentiment Values
sentiment score
1.5
1.0
0.5
0.0
-0.5
-1.0
Apache Spark
Backpropagation
Dashboard
Github
Numpy
Pandas
R Markdown
Rstudio
Tensorflow
Transformers
Jupyternoteboks
PrbbltyAndStst
RegressinAnlyss
R Programming
IBMCognsAnlytcs
SttstclHypthssT
Metadata
ArtfclIntll(AI)
SQL
Leadership
ETL&DatPipelins
NoSQLandBigData
Cloud Databases
Spreadsheet
AnlyzDttAnswrQs
DatCalculations
Big Data
DataAggregation
Data Cleansing
Data Ethics
Data Mining
Pivot Table
RltnlDMS(RDBMS)
NrlNtwrkArchtct
BscDscrptvStst
PythonProgrmmng
Data Collection
MachineLearning
IDNNtwrk:HT RaO
AskQstnstMDt-DD
DataArchitectur
Deep Learning
StrctrngMchnLrP
Data Analysis
Methodology
Database (DBMS)
Developngaprtfl
DatVisualizatin
Microsoft Excel
Creatingcasstds
Data Science
PrcssDtfrmDrttC
InfrmtnTch(IT)A
PrprDtfrExplrtn
relevancy

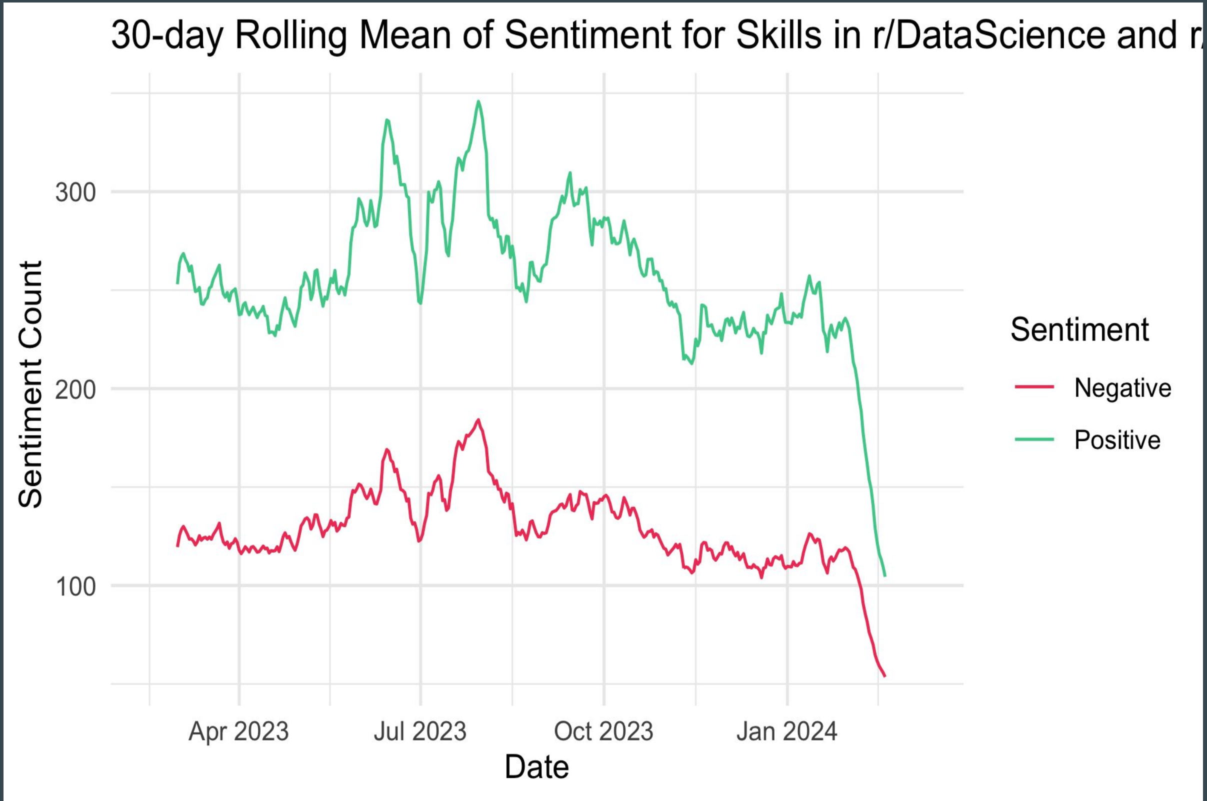
Relevancy vs Sentiment in Most Commented Skills



Results

Findings

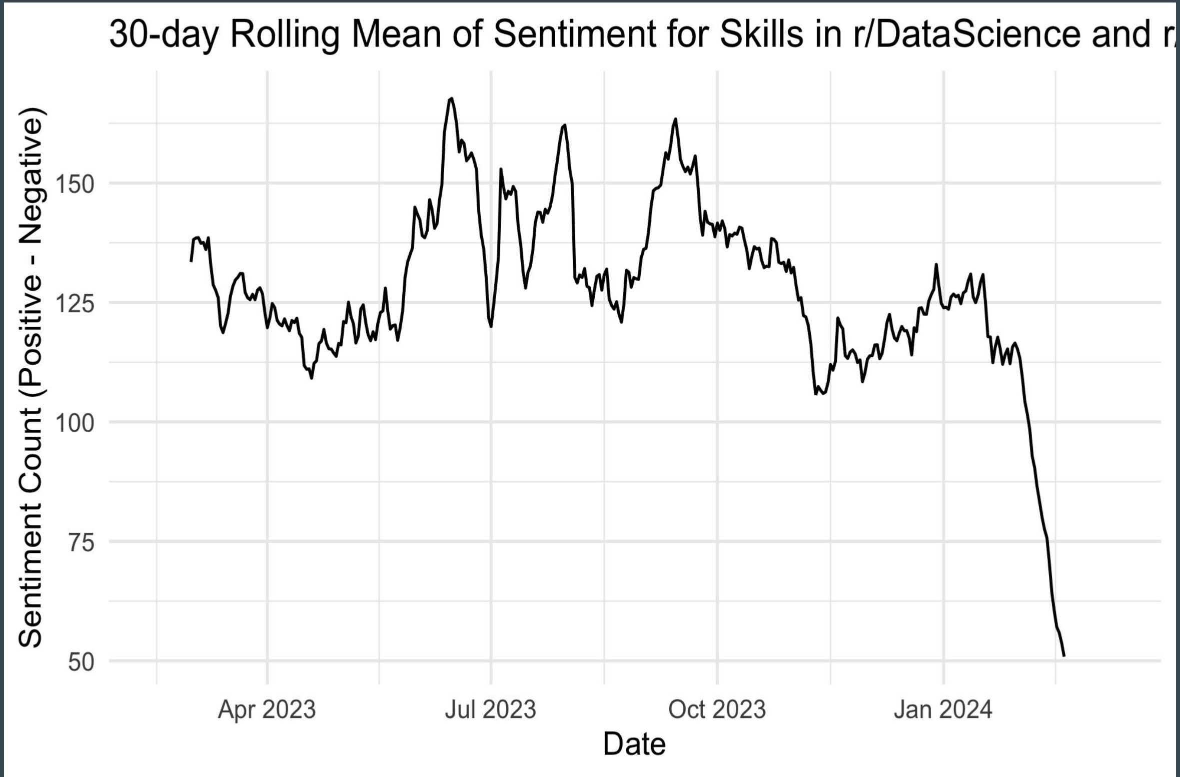
- On average there are more positive sentiments than negative sentiments in past year
- Drop off due to decreasing total comment amounts for recent comments (> 1 month)



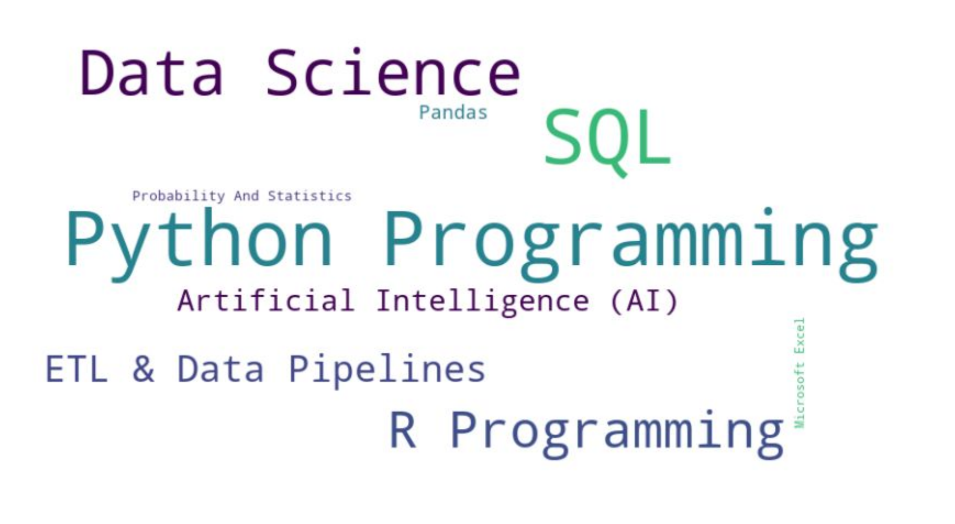
Results

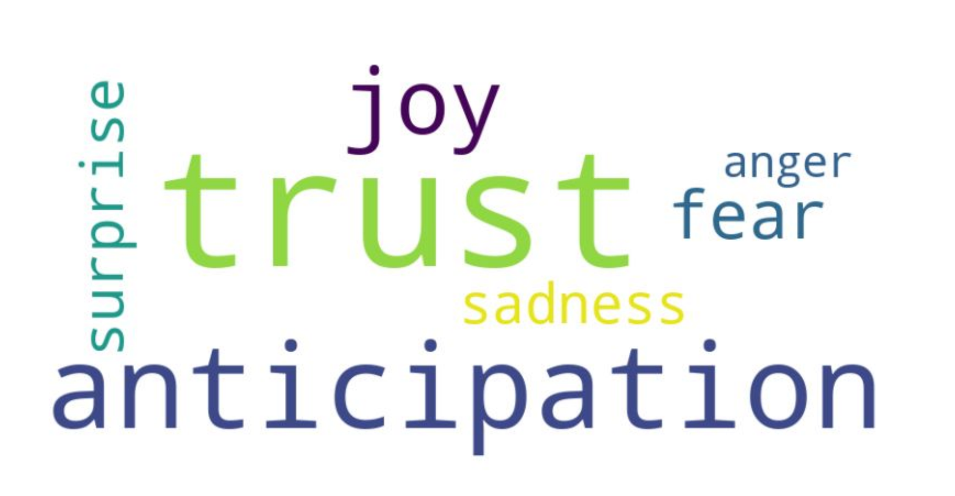
Findings

- Net sentiments usually ranges from +110 to +170

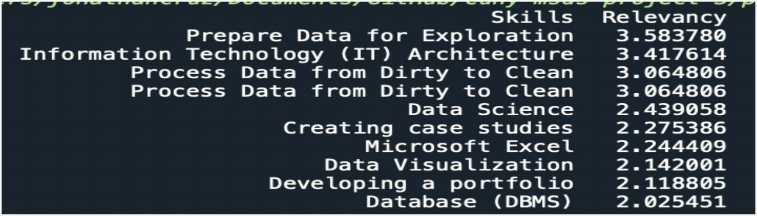


Conclusion

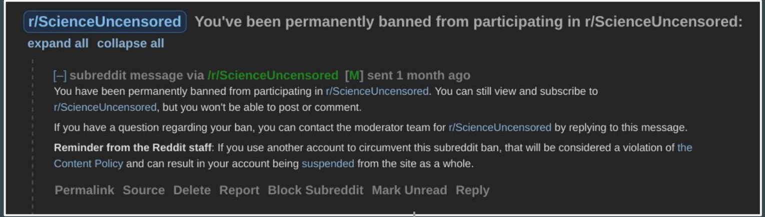

Data Science
Pandas
SQL
Probability And Statistics
Python Programming
Artificial Intelligence (AI)
ETL & Data Pipelines
Microsoft Excel
R Programming


joy
surprise
trust
anger
fear
sadness
anticipation

Data Loading



```
rs/jonathancruz/Documents/Github/cuny-msds-project-3/p
                                   Skills  Relevancy
              Prepare Data for Exploration  3.583780
 Information Technology (IT) Architecture  3.417614
           Process Data from Dirty to Clean  3.064806
           Process Data from Dirty to Clean  3.064806
                               Data Science  2.439058
                       Creating case studies  2.275386
                            Microsoft Excel  2.244409
                         Data Visualization  2.142001
                      Developing a portfolio  2.118805
                           Database (DBMS)  2.025451
```



r/ScienceUncensored You've been permanently banned from participating in r/ScienceUncensored:

[–] subreddit message via /r/ScienceUncensored [M] sent 1 month ago

You have been permanently banned from participating in r/ScienceUncensored. You can still view and subscribe to r/ScienceUncensored, but you won't be able to post or comment.

If you have a question regarding your ban, you can contact the moderator team for r/ScienceUncensored by replying to this message.

Reminder from the Reddit staff: If you use another account to circumvent this subreddit ban, that will be considered a violation of the Content Policy and can result in your account being suspended from the site as a whole.

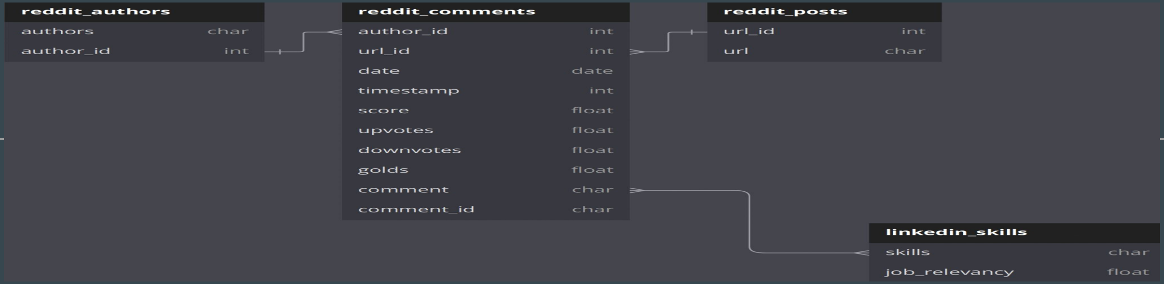
Permalink Source Delete Report Block Subreddit Mark Unread Reply

- Skills script executed and results saved in tab delimited text file
- Reddit_scrape.r file executed and results saved in tab delimited text file
- Dates and numeric columns converted to proper data types
- Reddit dataset encoded to create reference tables for:

A . reddit_comments B. reddit_authors C . reddit_posts

- Four datasets written to tables in same sqlite database:

A . reddit_comments: comments information and content ● **B** . reddit_authors: author_id and author's name ● **C**. reddit_posts: url_id and url of the post ● **D** . linkedin_skills: skills outlined in job posting and relevancy score



reddit_authors	
authors	char
author_id	int

reddit_comments	
author_id	int
url_id	int
date	date
timestamp	int
score	float
upvotes	float
downvotes	float
golds	float
comment	char
comment_id	char

reddit_posts	
url_id	int
url	char

linkedin_skills	
skills	char
job_relevancy	float

Analysis Methods Section 3

Step 1

- Computed 30-day rolling average time-series for sentiments
    - Show absolute numbers and differences

Data Sources & Process

- Reddit comments from the r/datascience and r/dataengineering subreddits
- What linked in jobs did we scrape?
- How did we orchestrate our scripts together and create the database file we used?
- I think Richie wrote this in the Google doc within the section where we went over how we got our data.