Text Analysis for Reddit in Big Data

Group 12 - Jahyeon Jee, Pi-Te Chen, Audrey Jung, Seunggyun Shin, Hyun Ji Lee

BADM 558
Big Data Infrastructure

Agenda

- Objectives
- Pipeline
- Performance Metrics
- Data Explanation
- Pre-processing and Cleaning
- Analysis
- Conclusions

Objectives

reddit

- The project will utilize AWS-based workflow for data collection, preprocessing, and analysis using visualization.
- The project represents a convergence of technology and social science, bridging the gap between Big Data analytics and online community engagement.
- The objective is to develop a robust Big Data Pipeline to extract actionable insights from the subreddit's content.

Objectives

The aim is to contribute meaningfully to the advancement of knowledge in employment dynamics and offer practical solutions to challenges faced by job seekers.

r/jobsearchhacks

This will pass too...

Your mental state after 12 months of unemployment and 500+ rejections

Job Search Hacks

Forget traditional job searching - improve your odds with good tips, tricks and tactics that help you stand out.

113K Members

47 Online

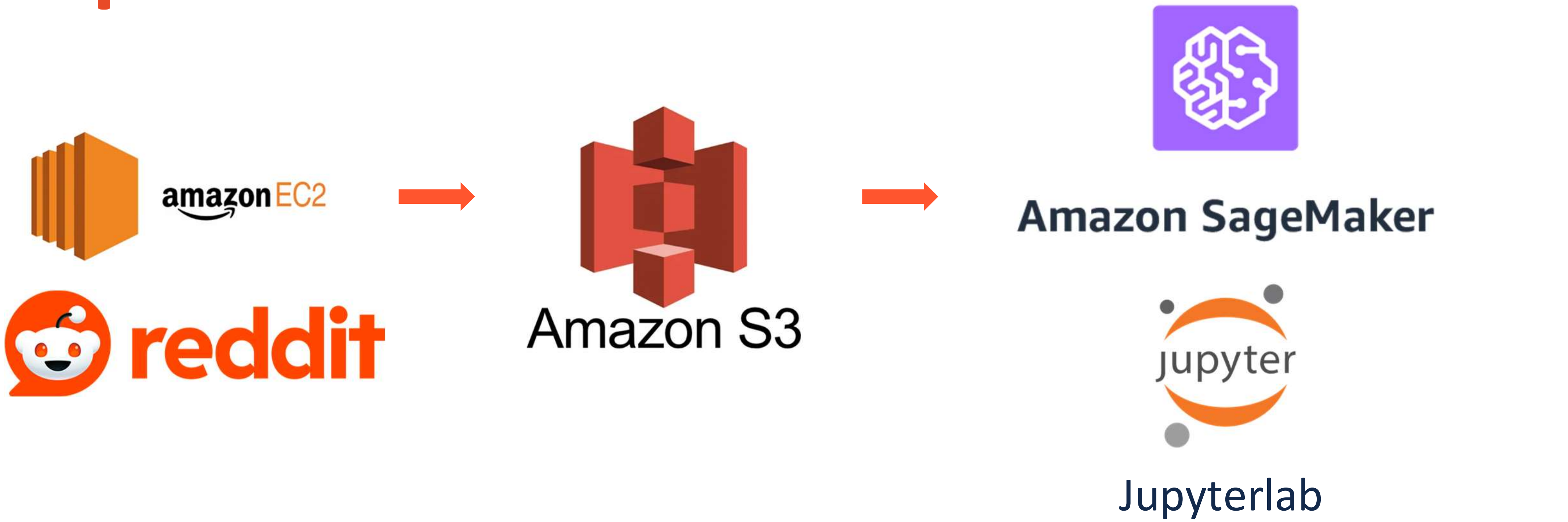
Top 2% Rank by size

RULES

1 No links to job listings are allowed

2 Link Posting Rules

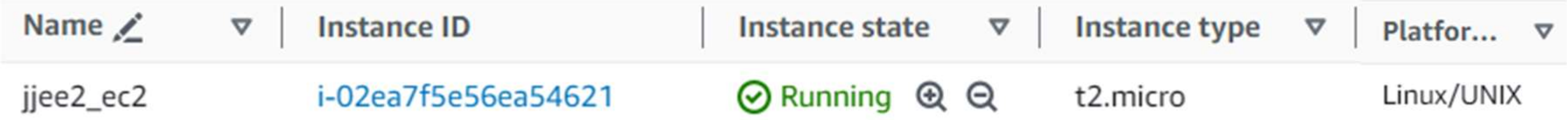
Pipeline



Data Extracting | **Data Storing** | **Preprocessing, Cleaning, Analysis**

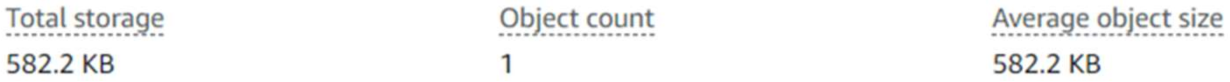
Performance Metrics - Resources & Cost

EC2 (2 weeks usage)



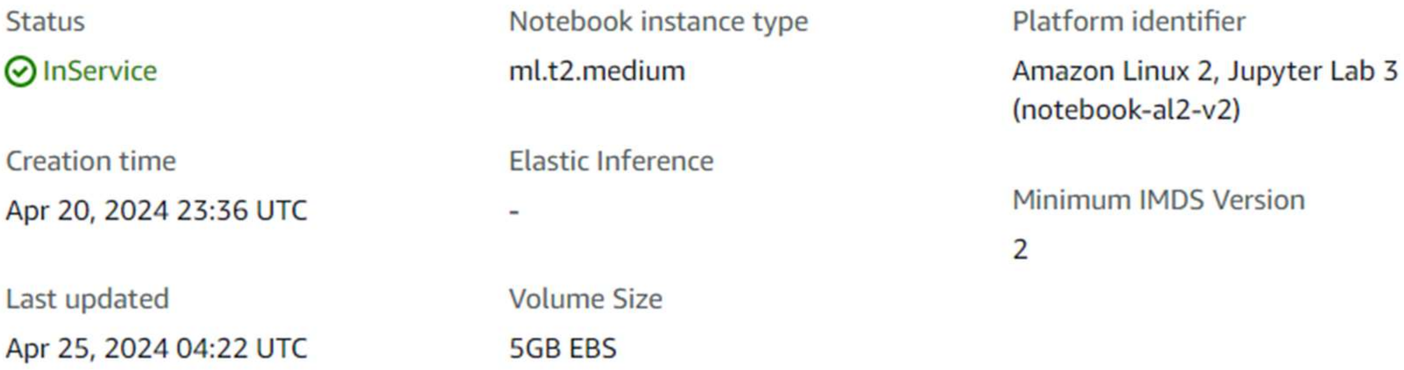
Name	Instance ID	Instance state	Instance type	Platfor...
jjee2_ec2	i-02ea7f5e56ea54621	✓ Running	t2.micro	Linux/UNIX

S3 (2 weeks usage)



Total storage	Object count	Average object size
582.2 KB	1	582.2 KB

SageMaker (2 weeks usage)



Status

✓ InService

Creation time

Apr 20, 2024 23:36 UTC

Last updated

Apr 25, 2024 04:22 UTC

Notebook instance type

ml.t2.medium

Elastic Inference

-

Volume Size

5GB EBS

Platform identifier

Amazon Linux 2, Jupyter Lab 3 (notebook-al2-v2)

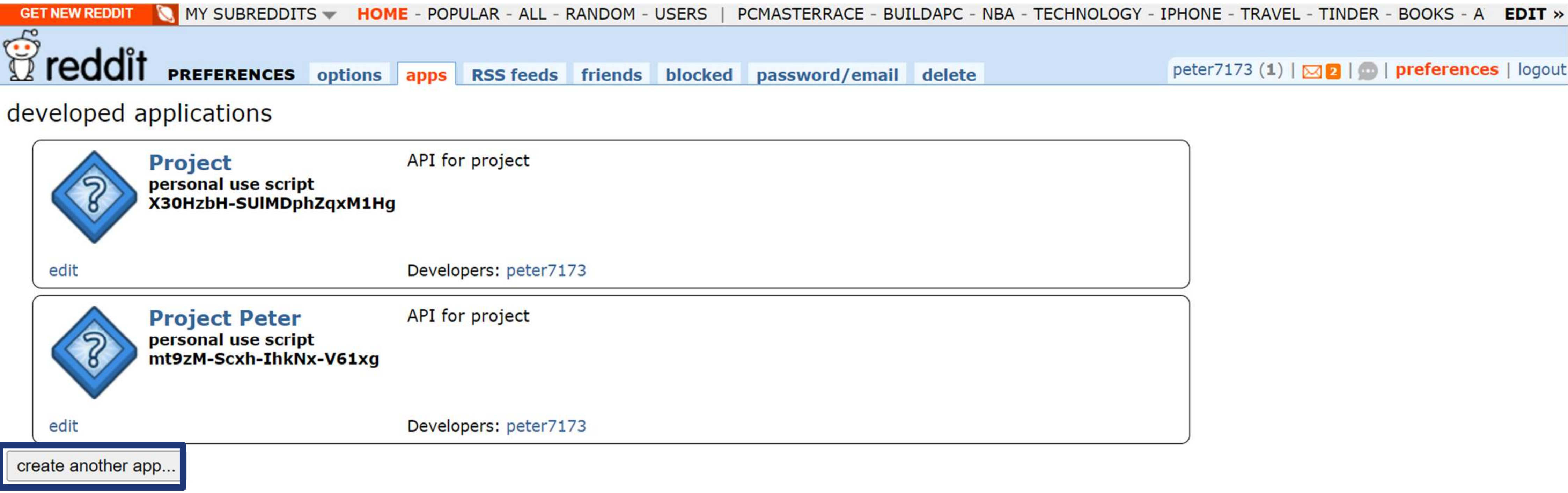
Minimum IMDS Version

2

Cost

- AWS: **Free**
- Reddit API: **Free**

Data Explanation



GET NEW REDDIT
MY SUBREDDITS
HOME - POPULAR - ALL - RANDOM - USERS | PCMASTERRACE - BUILDAPC - NBA - TECHNOLOGY - IPHONE - TRAVEL - TINDER - BOOKS - A
EDIT »
reddit
PREFERENCES
options
apps
RSS feeds
friends
blocked
password/email
delete
peter7173 (1) | 2 | | preferences | logout

developed applications

Project
personal use script
X30HzbH-SUlMDphZqxM1Hg

API for project

edit

Developers: peter7173

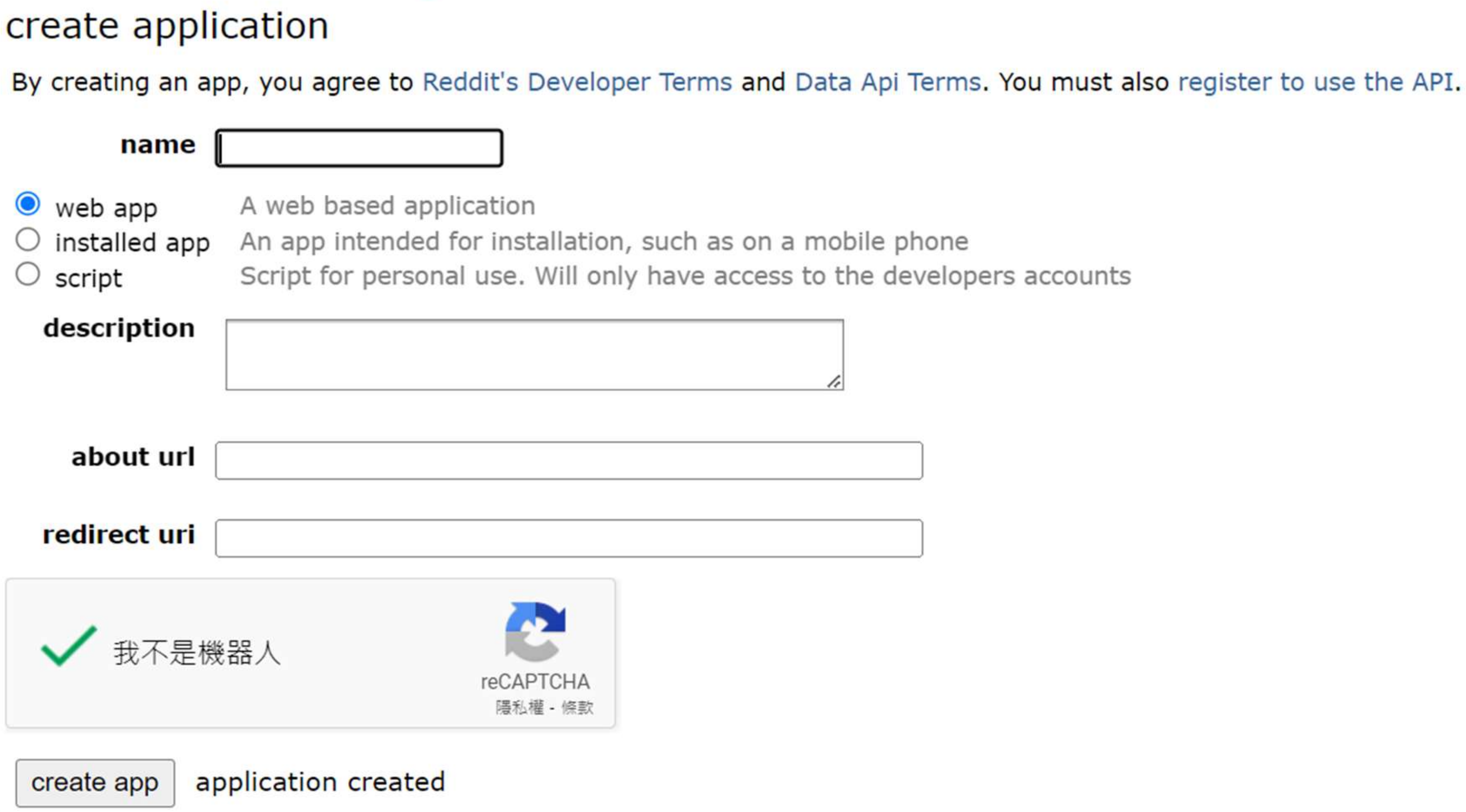
Project Peter
personal use script
mt9zM-Scxh-IhkNx-V61xg

API for project

edit

Developers: peter7173

Data Explanation



create application

By creating an app, you agree to Reddit's Developer Terms and Data Api Terms. You must also register to use the API.

name

● web app — A web based application

○ installed app — An app intended for installation, such as on a mobile phone

○ script — Script for personal use. Will only have access to the developers accounts

description

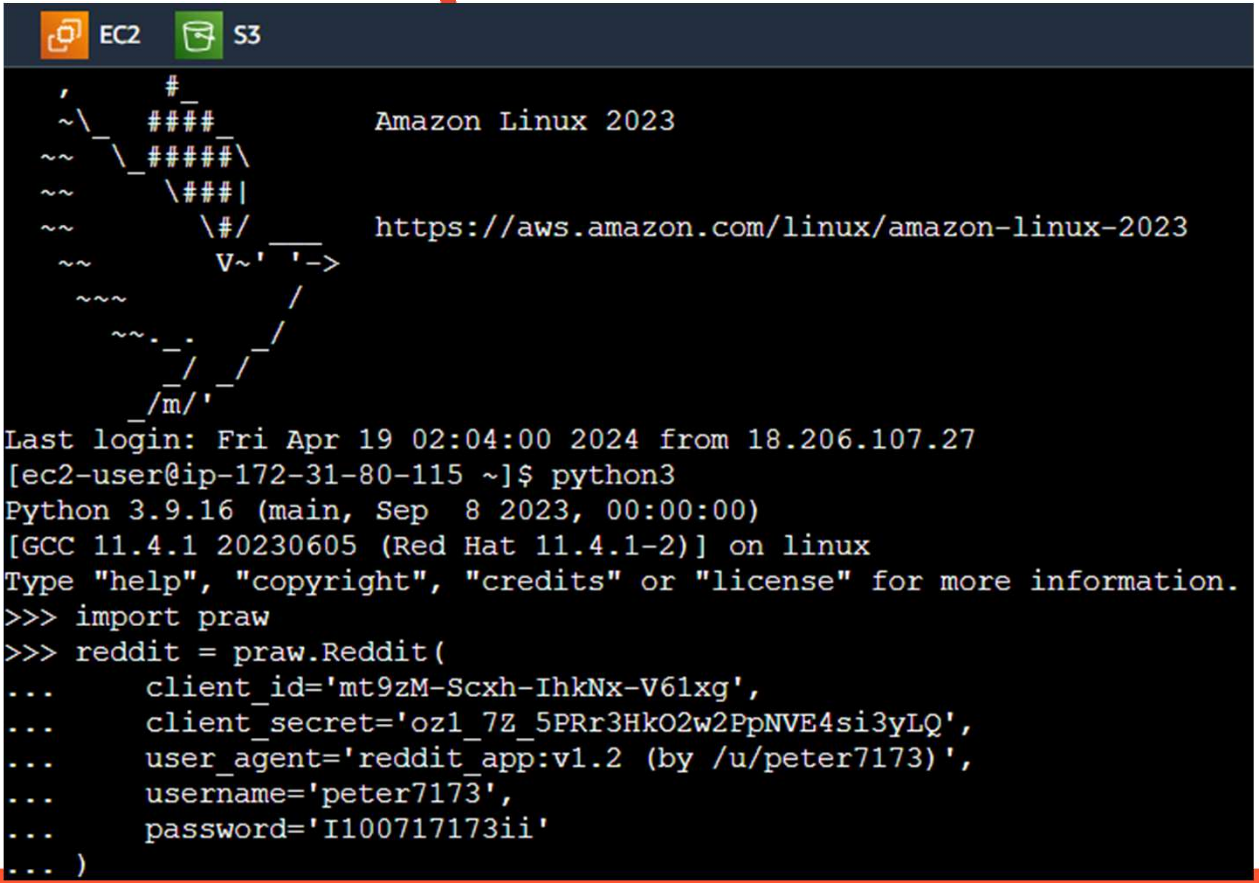
about url

redirect uri

✓ 我不是機器人

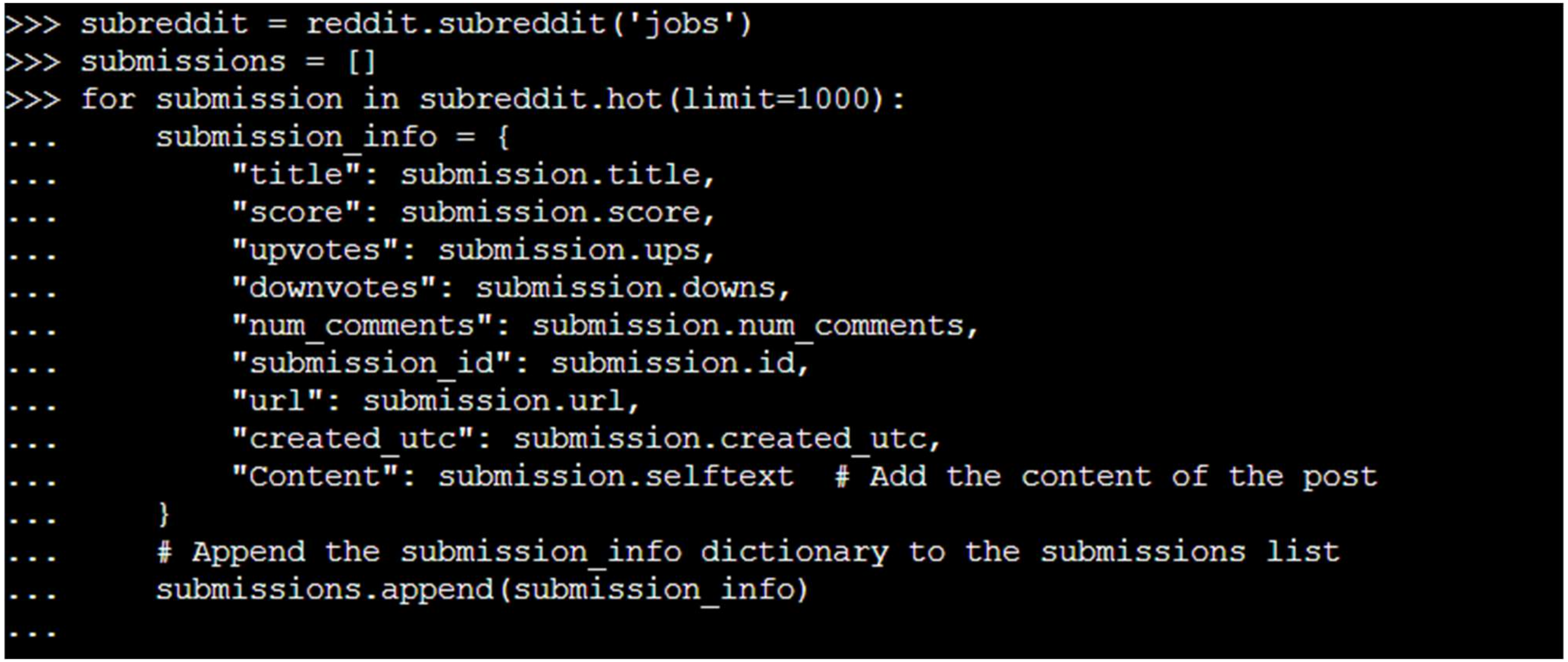
application created

Data Explanation



```
   ,     #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'
Last login: Fri Apr 19 02:04:00 2024 from 18.206.107.27
[ec2-user@ip-172-31-80-115 ~]$ python3
Python 3.9.16 (main, Sep  8 2023, 00:00:00)
[GCC 11.4.1 20230605 (Red Hat 11.4.1-2)] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import praw
>>> reddit = praw.Reddit(
...     client_id='mt9zM-Scxh-IhkNx-V61xg',
...     client_secret='oz1_7Z_5PRr3HkO2w2PpNVE4si3yLQ',
...     user_agent='reddit_app:v1.2 (by /u/peter7173)',
...     username='peter7173',
...     password='I100717173ii'
... )
```

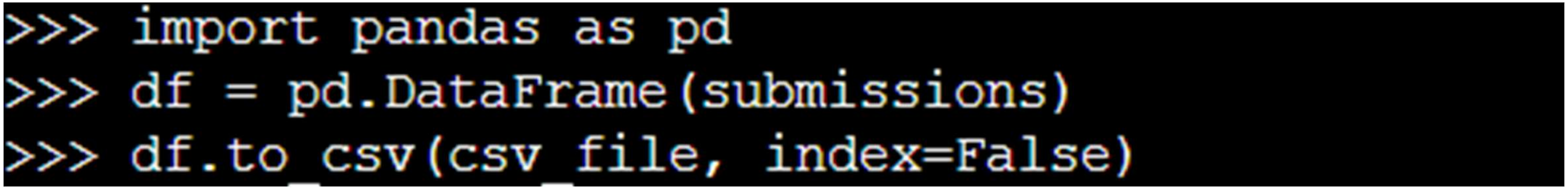
Data Explanation



```
>>> subreddit = reddit.subreddit('jobs')
>>> submissions = []
>>> for submission in subreddit.hot(limit=1000):
...     submission_info = {
...         "title": submission.title,
...         "score": submission.score,
...         "upvotes": submission.ups,
...         "downvotes": submission.downs,
...         "num_comments": submission.num_comments,
...         "submission_id": submission.id,
...         "url": submission.url,
...         "created_utc": submission.created_utc,
...         "Content": submission.selftext  # Add the content of the post
...     }
...     # Append the submission_info dictionary to the submissions list
...     submissions.append(submission_info)
...
```

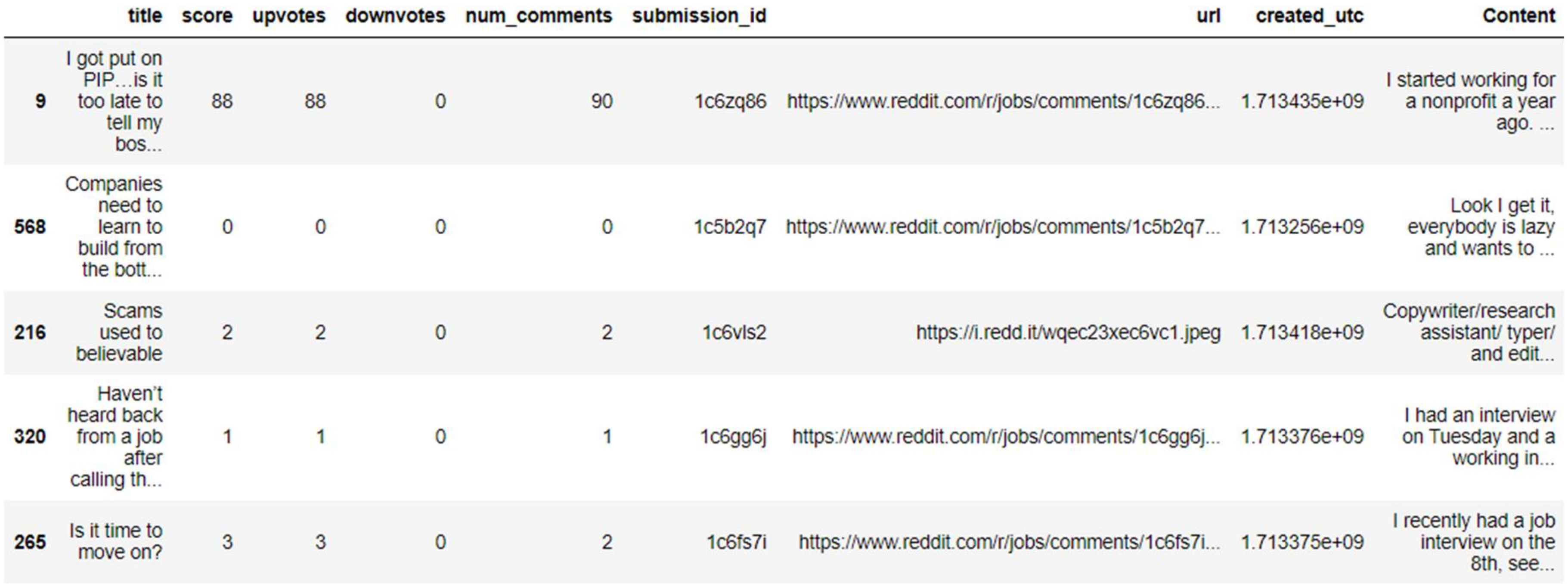
Data Explanation

Transform the data to CSV file



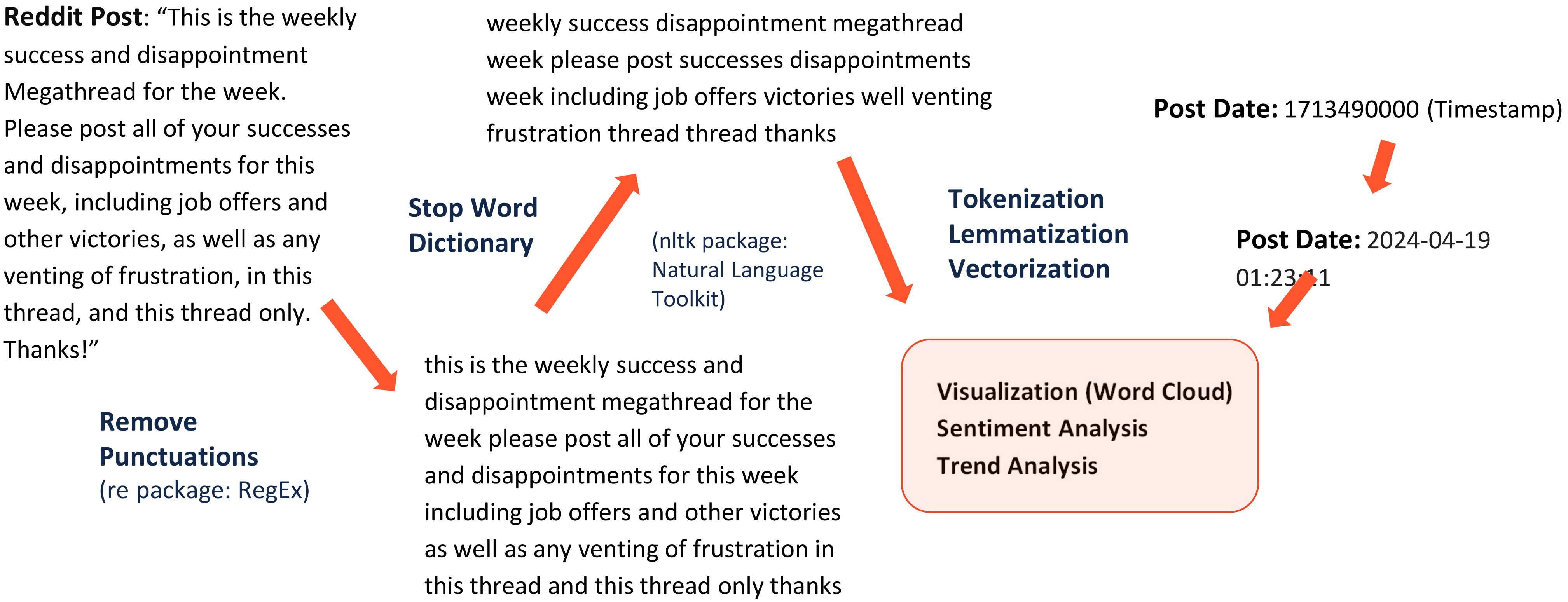
```
>>> import pandas as pd
>>> df = pd.DataFrame(submissions)
>>> df.to_csv(csv_file, index=False)
```

Data Explanation



	title	score	upvotes	downvotes	num_comments	submission_id	url	created_utc	Content
9	I got put on PIP…is it too late to tell my bos...	88	88	0	90	1c6zq86	https://www.reddit.com/r/jobs/comments/1c6zq86...	1.713435e+09	I started working for a nonprofit a year ago. ...
568	Companies need to learn to build from the bott...	0	0	0	0	1c5b2q7	https://www.reddit.com/r/jobs/comments/1c5b2q7...	1.713256e+09	Look I get it, everybody is lazy and wants to ...
216	Scams used to believable	2	2	0	2	1c6vls2	https://i.redd.it/wqec23xec6vc1.jpeg	1.713418e+09	Copywriter/research assistant/ typer/ and edit...
320	Haven't heard back from a job after calling th...	1	1	0	1	1c6gg6j	https://www.reddit.com/r/jobs/comments/1c6gg6j...	1.713376e+09	I had an interview on Tuesday and a working in...
265	Is it time to move on?	3	3	0	2	1c6fs7i	https://www.reddit.com/r/jobs/comments/1c6fs7i...	1.713375e+09	I recently had a job interview on the 8th, see...

Pre-processing & Cleaning



Reddit Post: “This is the weekly success and disappointment Megathread for the week. Please post all of your successes and disappointments for this week, including job offers and other victories, as well as any venting of frustration, in this thread, and this thread only. Thanks!”

Remove Punctuations
(re package: RegEx)

this is the weekly success and disappointment megathread for the week please post all of your successes and disappointments for this week including job offers and other victories as well as any venting of frustration in this thread and this thread only thanks

Stop Word Dictionary

(nltk package: Natural Language Toolkit)

weekly success disappointment megathread week please post successes disappointments week including job offers victories well venting frustration thread thread thanks

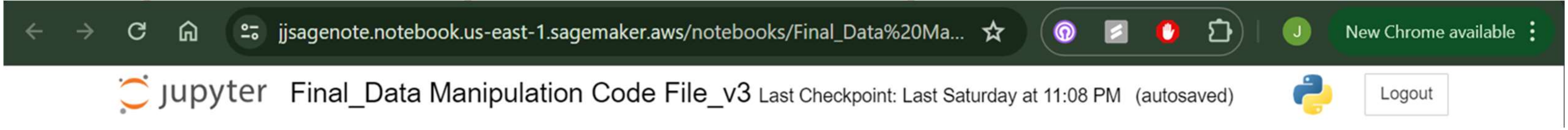
Tokenization
Lemmatization
Vectorization

Post Date: 1713490000 (Timestamp)

Post Date: 2024-04-19 01:23:11

Visualization (Word Cloud)
Sentiment Analysis
Trend Analysis

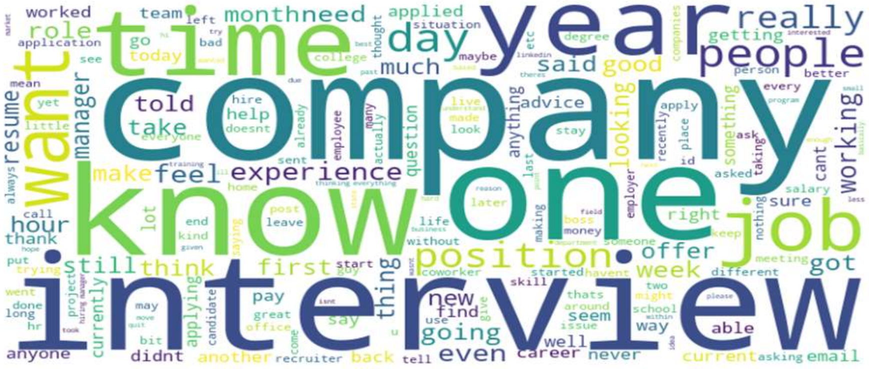
Exploratory Data Analysis



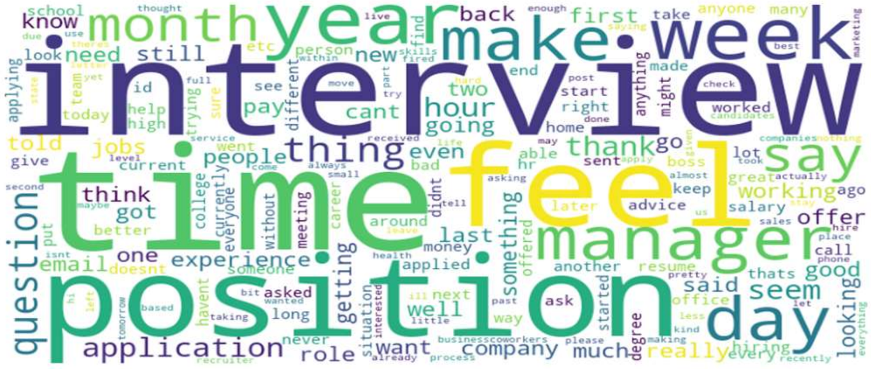
jjsagenote.notebook.us-east-1.sagemaker.aws/notebooks/Final_Data%20Ma...

Final_Data Manipulation Code File_v3 Last Checkpoint: Last Saturday at 11:08 PM (autosaved)

Word Cloud

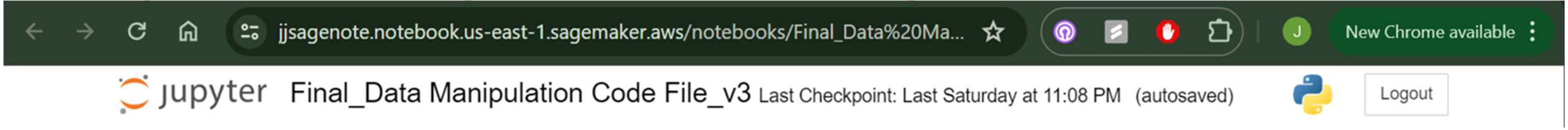


General words:
company, interview, know



Refined words:
Interview, time, position

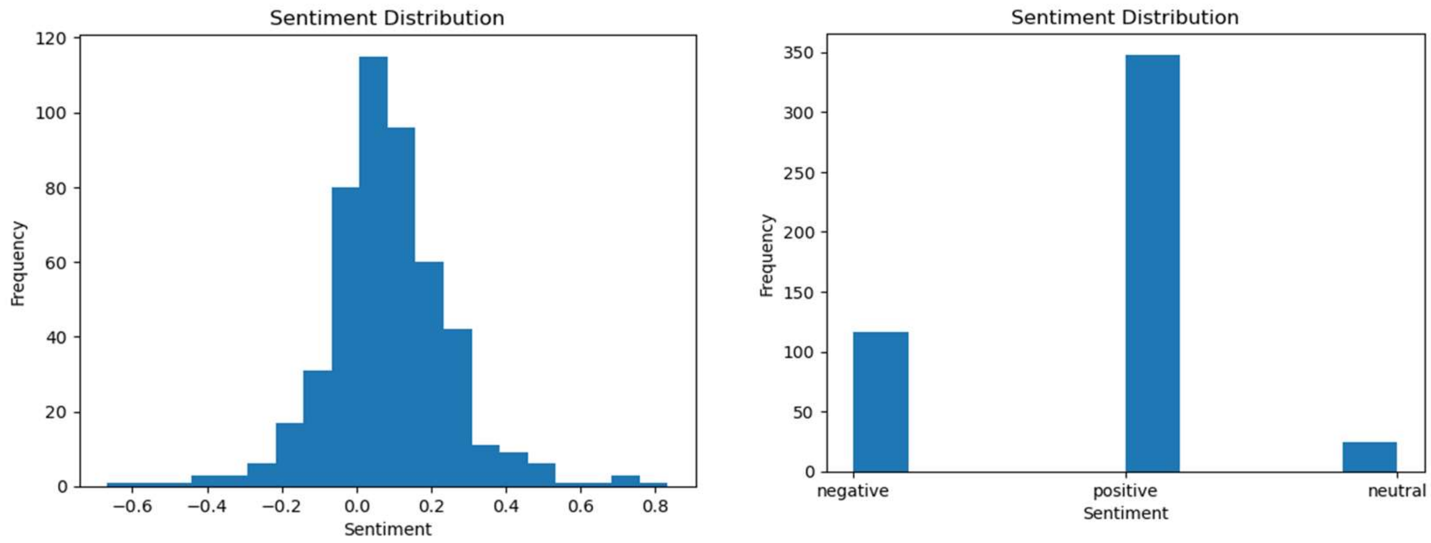
Exploratory Data Analysis



jjsagenote.notebook.us-east-1.sagemaker.aws/notebooks/Final_Data%20Ma...

Final_Data Manipulation Code File_v3 Last Checkpoint: Last Saturday at 11:08 PM (autosaved)

Sentiment Analysis: Overview



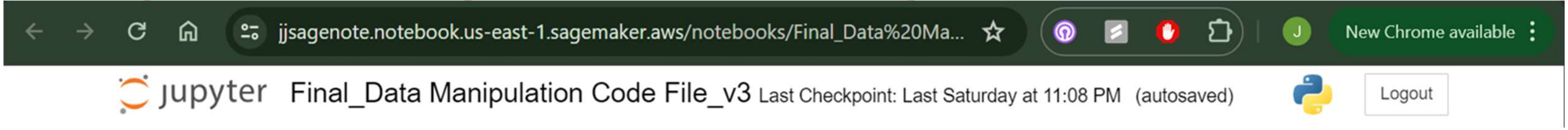
Method

- Barplot with sentiment score/type as x and frequency as y

Analysis

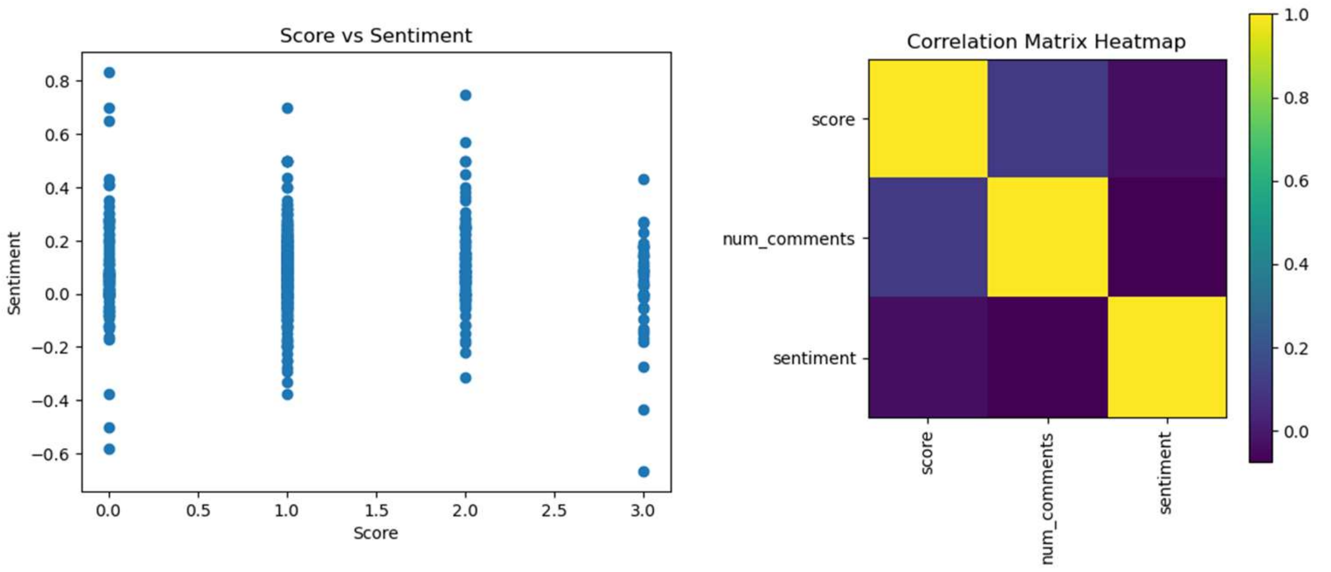
- Almost normal distribution
- More positive reactions

Exploratory Data Analysis


jjsagenote.notebook.us-east-1.sagemaker.aws/notebooks/Final_Data%20Ma...

Final_Data Manipulation Code File_v3 Last Checkpoint: Last Saturday at 11:08 PM (autosaved)

Bivariate Analysis: Score vs. Sentiment



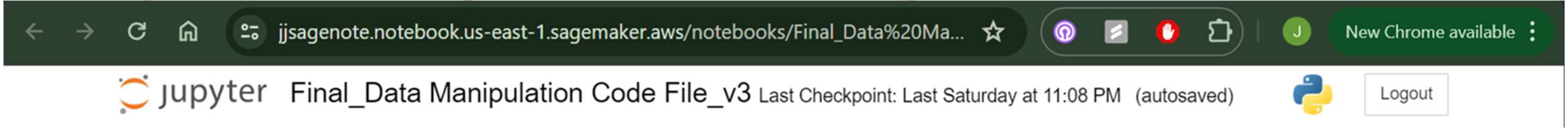
Method

- Boxplot and Heatmap

Analysis

- Weak pattern of worsening the sentiment by the score
- Weak correlation between score and number of comments

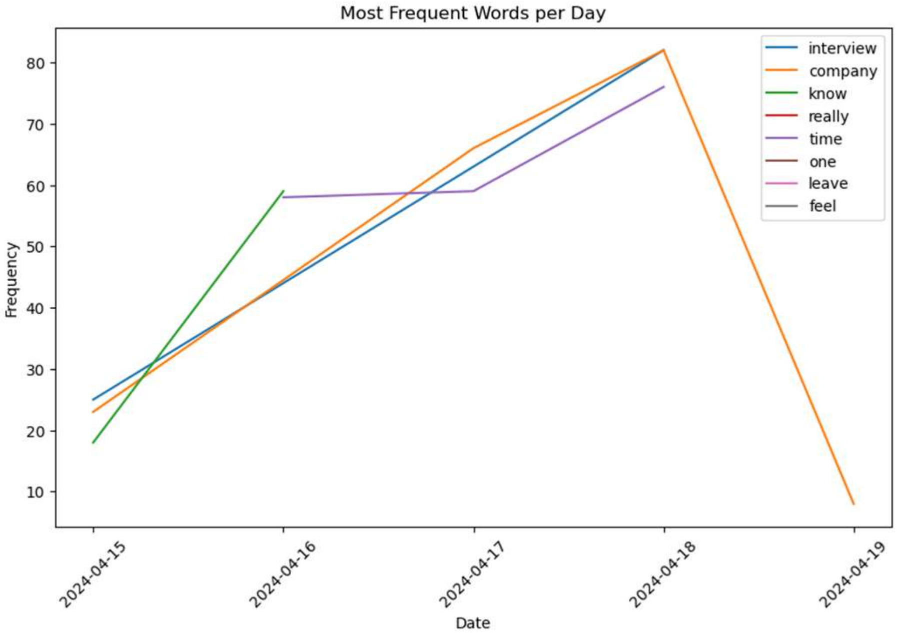
Exploratory Data Analysis



jjsagenote.notebook.us-east-1.sagemaker.aws/notebooks/Final_Data%20Ma...
J
New Chrome available
jupyter
Logout

Final_Data Manipulation Code File_v3 Last Checkpoint: Last Saturday at 11:08 PM (autosaved)

Trend Analysis: Frequent Words by day



Method

- Plotting the word frequency data with the date as x, frequency as y with legend of words

Analysis

- Overlapping words for several days
- Similar trend with the total number of posts

Conclusions

Summary

By harnessing a range of tools and services on the AWS platform, including the Reddit API, we delved into the intricacies of big data infrastructure and established efficient pipelines. Our analysis provided valuable insights into the prevailing trends in web-generated content related to job searches, underscoring its immediate relevance and utility for our objectives.

Innovation

Advanced Data Collection Methods: API usage is efficient and real-time

Natural Language Processing (NLP): Word frequency, Sentiment analysis, and the Trend

Scalability and Flexibility: Through AWS platform, data usage can be easily extended

Future Potential

- More detailed and long term analysis
- Diverse opinions

Thank You

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN