TASK 5 - Development History

Date: 29/07/2024

Today, we went through the assignment specification, and started exploring task 1 data. We divided the steps amongs the group members. Both group members had individual files to start doing their work in the separate google collab so that we can see which code would be better. for proof of the work we have put both of our colab link in the following:

Proof of individual google colab

Subbulakshmi :https://colab.research.google.com/drive/1pZqG9bdN1Kd1rrdeK9WnuwjgrCaHhM6d?usp=sharing

Bhavna: https://colab.research.google.com/drive/1XPeflOMF9ghMmRFWDIpulrhFONYSZFZu?usp=sharing

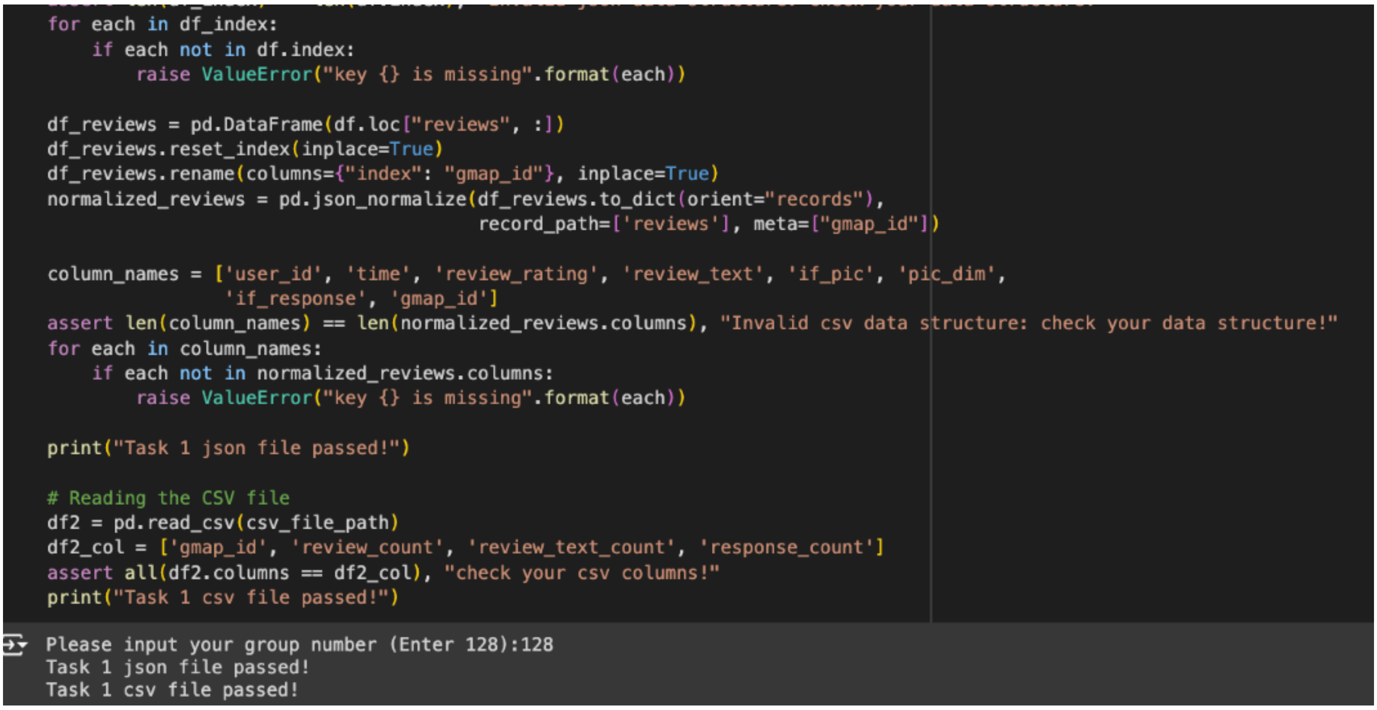
Date: 30/07/2024

Next day we both sat see what we did in task 1 and both of us identified the errors that we encountered to convert into the json and csv. We saw each other task 1 we did not get the proper JSON file and CSV file so we used ChatGPT for debugging issues and errors since our CSV gmap_id and response count came blank so we used ChatGPT to solve the error and then we had our applied class that day so we asked our tutor shauo for the issue and solved the problem. Then we did insert code through the test script JSON and CSV file passed.

Proof

Google Colab : https://colab.research.google.com/drive/1_Rsd6Yo-XqnhLmrFVg2Sltpd2Nqh3nnJ?usp=sharing

Screenshot of the test script



```
    for each in df_index:
        if each not in df.index:
            raise ValueError("key {} is missing".format(each))

    df_reviews = pd.DataFrame(df.loc["reviews", :])
    df_reviews.reset_index(inplace=True)
    df_reviews.rename(columns={"index": "gmap_id"}, inplace=True)
    normalized_reviews = pd.json_normalize(df_reviews.to_dict(orient="records"),
                                           record_path=['reviews'], meta=["gmap_id"])

    column_names = ['user_id', 'time', 'review_rating', 'review_text', 'if_pic', 'pic_dim',
                    'if_response', 'gmap_id']
    assert len(column_names) == len(normalized_reviews.columns), "Invalid csv data structure: check your data structure!"
    for each in column_names:
        if each not in normalized_reviews.columns:
            raise ValueError("key {} is missing".format(each))

    print("Task 1 json file passed!")

    # Reading the CSV file
    df2 = pd.read_csv(csv_file_path)
    df2_col = ['gmap_id', 'review_count', 'review_text_count', 'response_count']
    assert all(df2.columns == df2_col), "check your csv columns!"
    print("Task 1 csv file passed!")
```

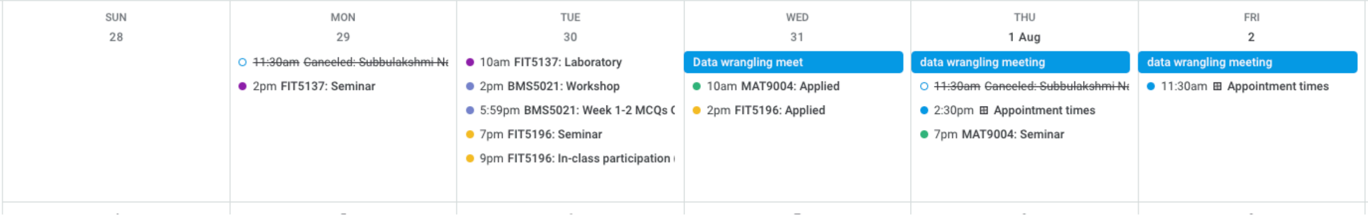
```
Please input your group number (Enter 128):128
Task 1 json file passed!
Task 1 csv file passed!
```

Date: 31/07/2024 - 2/08/2024

Both members were working on task 2 as before we did our work in our colab notebooks and we made sure that we communicate during these days if we encountered any problems. We made sure to document all these records through our own colab so that we can show it in our documentation history. So during these days both members have communicated through google meet to get the update about each other work.

Proof:

Meeting Link



SUN 28	MON 29	TUE 30	WED 31	THU 1 Aug	FRI 2
	○ ~~11:30am Canceled: Subbulakshmi N~~ ● 2pm FIT5137: Seminar	● 10am FIT5137: Laboratory ● 2pm BMS5021: Workshop ● 5:59pm BMS5021: Week 1-2 MCQs (● 7pm FIT5196: Seminar ● 9pm FIT5196: In-class participation	Data wrangling meet ● 10am MAT9004: Applied ● 2pm FIT5196: Applied	data wrangling meeting ○ ~~11:30am Canceled: Subbulakshmi N~~ ● 2:30pm ⊞ Appointment times ● 7pm MAT9004: Seminar	data wrangling meeting ● 11:30am ⊞ Appointment times

Date: 3/08/2024 - 8/08/2024

During these we were working on the task 2, we had encountered some problems in task 2 since the count vec files came empty when made sure our task 1 was wrong so we went back to the task 1 and rectify the problem that we had and rectify it using some applied class exercise as well as attending consultation with Kiara and Shawna to get the error problem solved. Apart from that we did encounter some errors due to the csv files inserting links which didnt allows to complete on time. We realized that we had only 2 weeks to go so we made sure that we speed up we have to redo our test 1 again since task 2 was coming errors

Proof of individual google colab

Subbulakshmi :https://colab.research.google.com/drive/1pZqG9bdN1Kd1rrdeK9WnuwjgrCaHhM6d?usp=sharing

Bhavna: https://colab.research.google.com/drive/1XPeflOMF9ghMmRFWDIpulrhFONYSZFZu?usp=sharing

Test files Google Colab : https://colab.research.google.com/drive/1_Rsd6Yo-XqnhLmrFVg2Sltpd2Nqh3nnJ?usp=sharing

Date: 9/08/2024 - 15/08/2024

After a while we figured the mistakes and errors that we encountered and we got the error right it was json file which we got as the structure was not following as per specification as we didnt insert the earliest date as well as latest date in the json which was good finding so we made sure we include these dates and do it. So we used the sample output as a reference to make the JSON files as well as csv output to make it right. Then we both started to work on the task 2 and the code finally worked and count vec text and vocab text came properly as per the sample out and then we tested the code as per the testing script but then the structure was not right.

We made sure that do the coding as per the sample output and then it worked so our task 2 got over and we were converting all the code into task_128_final version colab

Proof:

Task_128_final_version: https://colab.research.google.com/drive/1_mAFSv8zm8ZZMSFn2eZ5MN7ZMDOicSgh?usp=sharing

Date: 16/08/2024 - 20/08/2024

After that we did the task 3 where we decided what all we can include in this task 3 we had some confusion in the template so we went to the consulatation to resolve the problem as well as the consultation was helpful for us to get to know what we can include in our task 3. From that we splited the task 3 we divided among each other like subbulakshmi doing the csv and json eda analysis and then bhavna did the auxiliary meta data.

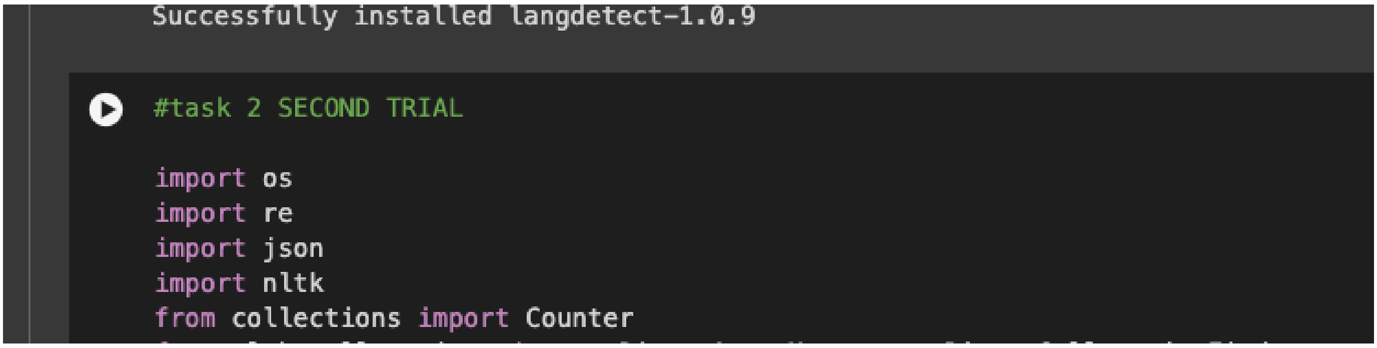
Proof:

Task_128_final_version:
https://colab.research.google.com/drive/1_mAFSv8zm8ZZMSFn2eZ5MN7ZMDOicSgh?usp=sharing

Date: 21/08/2024 - 26/08/2024

Team members were rechecking the outputs of Task 1 to see if it aligned with the output formats and requirements.Then we split and begun working on task-2

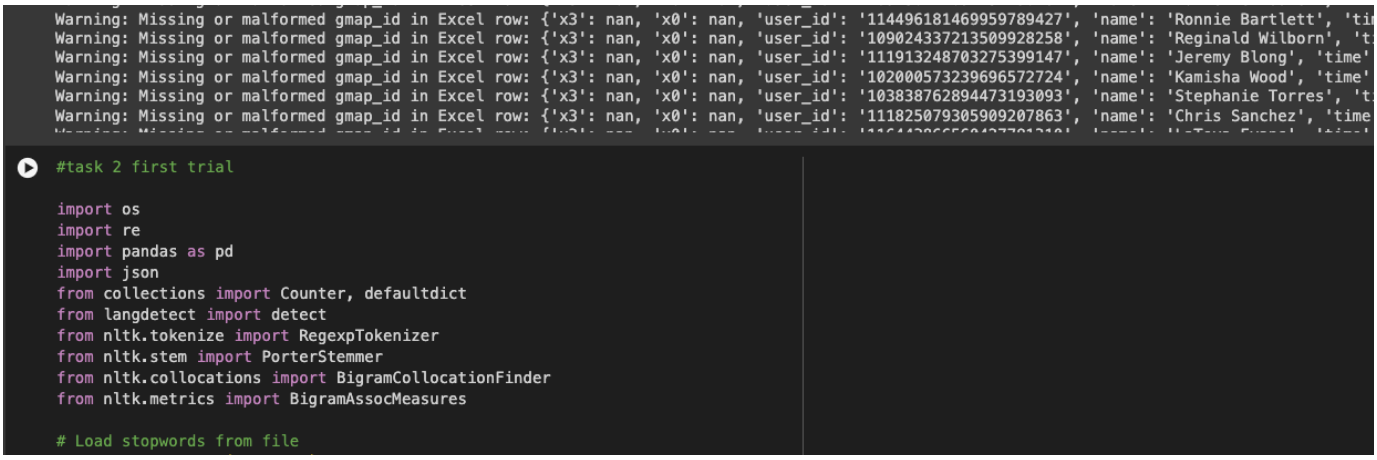
- Ran independent codes to look for best version
- Debugged task-2 and ran multiple trials
- Assessed the errors and worked together to correct them.
- Finalise the last version for the task-2
- Confirm the outputs aligned with the requirements.
- Run the test scripts to pass the outputs.



```
Successfully installed langdetect-1.0.9

#task 2 SECOND TRIAL

import os
import re
import json
import nltk
from collections import Counter
```



```
Warning: Missing or malformed gmap_id in Excel row: {'x3': nan, 'x0': nan, 'user_id': '114496181469959789427', 'name': 'Ronnie Bartlett', 'ti
Warning: Missing or malformed gmap_id in Excel row: {'x3': nan, 'x0': nan, 'user_id': '109024337213509928258', 'name': 'Reginald Wilborn', 't
Warning: Missing or malformed gmap_id in Excel row: {'x3': nan, 'x0': nan, 'user_id': '111913248703275399147', 'name': 'Jeremy Blong', 'time'
Warning: Missing or malformed gmap_id in Excel row: {'x3': nan, 'x0': nan, 'user_id': '102000573239696572724', 'name': 'Kamisha Wood', 'time'
Warning: Missing or malformed gmap_id in Excel row: {'x3': nan, 'x0': nan, 'user_id': '103838762894473193093', 'name': 'Stephanie Torres', 't
Warning: Missing or malformed gmap_id in Excel row: {'x3': nan, 'x0': nan, 'user_id': '111825079305909207863', 'name': 'Chris Sanchez', 'time
```

```
#task 2 first trial

import os
import re
import pandas as pd
import json
from collections import Counter, defaultdict
from langdetect import detect
from nltk.tokenize import RegexpTokenizer
from nltk.stem import PorterStemmer
from nltk.collocations import BigramCollocationFinder
from nltk.metrics import BigramAssocMeasures

# Load stopwords from file
```

Date: 27/08/2024 - 30/08/2024

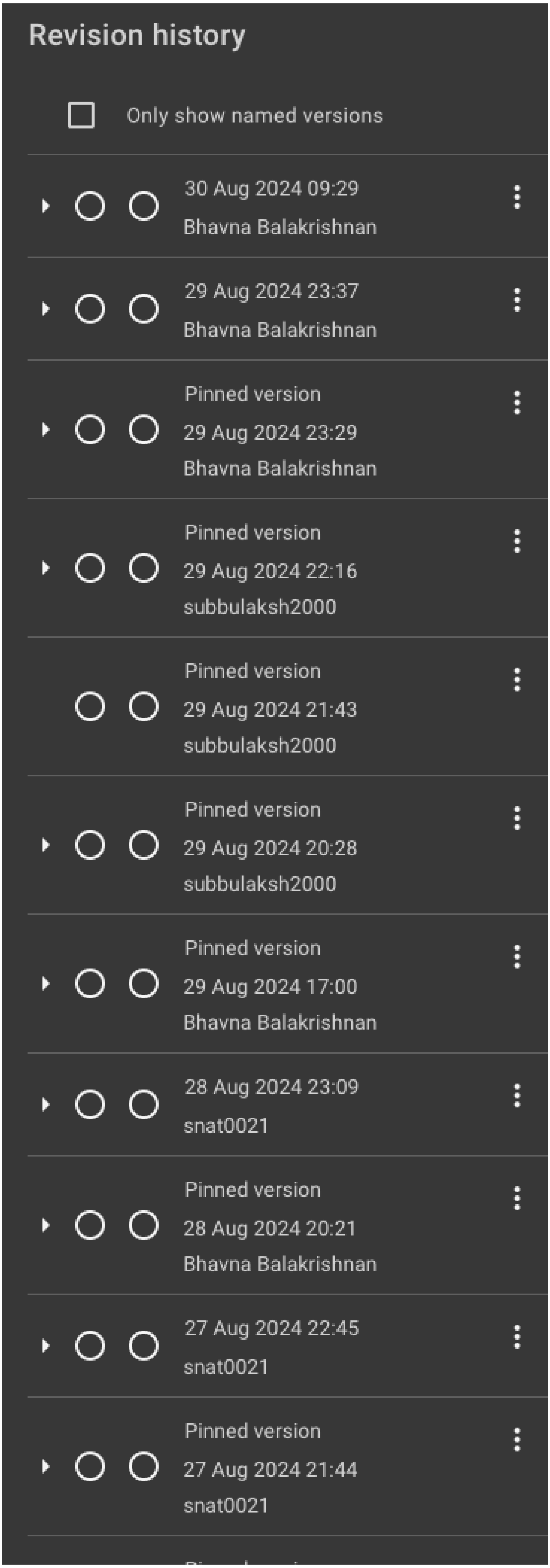
Finally tem members discussed and split the following:

- Consildate the final versions of the codes.
- Do EDA
  - split the portions each member would do.
- Shoot the presentation video
- Video editing
- Formatting and index
- Finalise all the codes in templates per specifications

Version History

Below we can loook at all th colab revision histories. Please note that team members have also worked on separate sheets independently and then worked to combine them later on.

Below version history for link : https://colab.research.google.com/drive/1_mAFSv8zm8ZZMSFn2eZ5MN7ZMDOicSgh?usp=sharing



Revision history

☐ Only show named versions

○ ○ 30 Aug 2024 09:29
Bhavna Balakrishnan

○ ○ 29 Aug 2024 23:37
Bhavna Balakrishnan

○ ○ Pinned version
29 Aug 2024 23:29
Bhavna Balakrishnan

○ ○ Pinned version
29 Aug 2024 22:16
subbulaksh2000

○ ○ Pinned version
29 Aug 2024 21:43
subbulaksh2000

○ ○ Pinned version
29 Aug 2024 20:28
subbulaksh2000

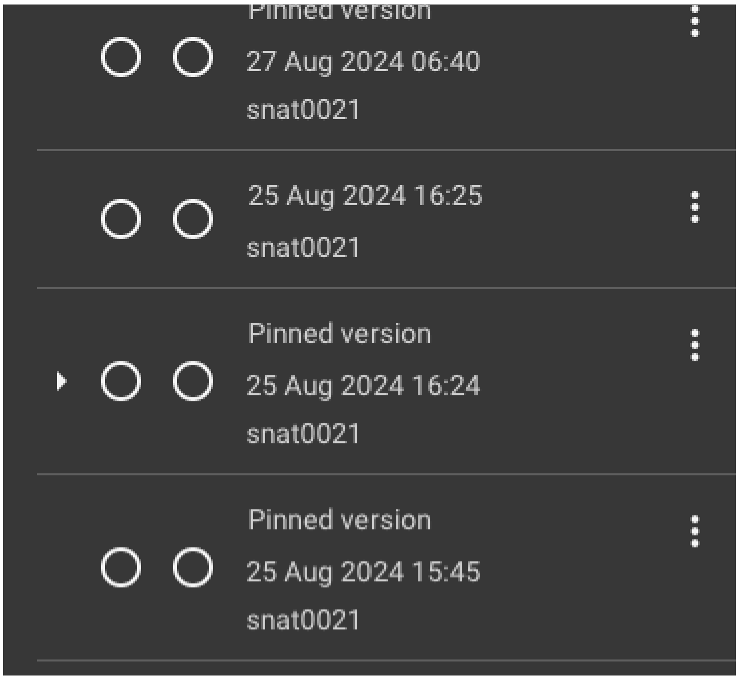
○ ○ Pinned version
29 Aug 2024 17:00
Bhavna Balakrishnan

○ ○ 28 Aug 2024 23:09
snat0021

○ ○ Pinned version
28 Aug 2024 20:21
Bhavna Balakrishnan

○ ○ 27 Aug 2024 22:45
snat0021

○ ○ Pinned version
27 Aug 2024 21:44
snat0021



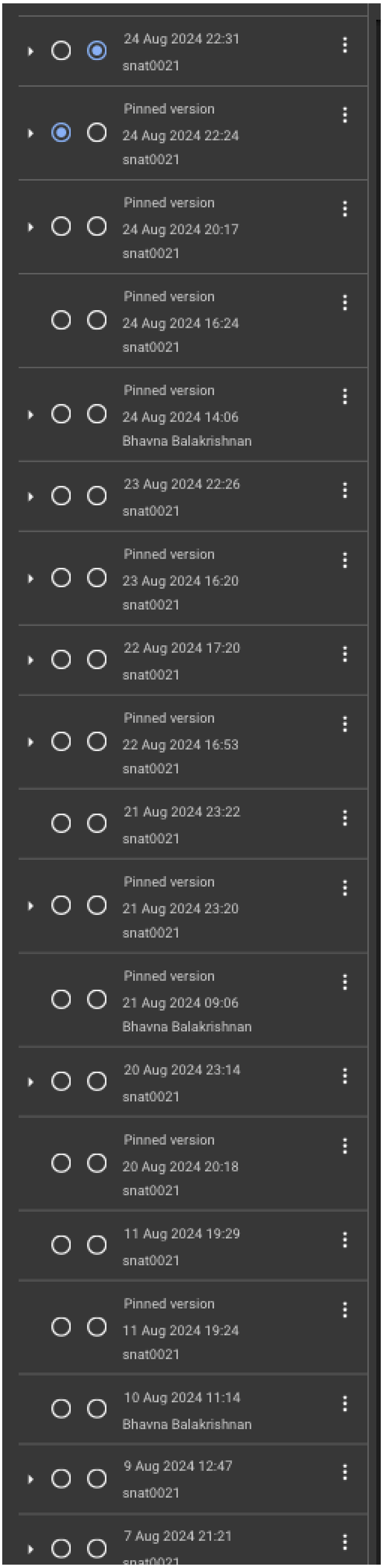
Pinned version
27 Aug 2024 06:40
snat0021

25 Aug 2024 16:25
snat0021

Pinned version
25 Aug 2024 16:24
snat0021

Pinned version
25 Aug 2024 15:45
snat0021

Now for colab with link:
https://colab.research.google.com/drive/1j8_4xTxcU2_XvwV9trs3fS5ASoFDccdy?usp=sharing



○ ● 24 Aug 2024 22:31
snat0021

● ○ Pinned version
24 Aug 2024 22:24
snat0021

○ ○ Pinned version
24 Aug 2024 20:17
snat0021

○ ○ Pinned version
24 Aug 2024 16:24
snat0021

○ ○ Pinned version
24 Aug 2024 14:06
Bhavna Balakrishnan

○ ○ 23 Aug 2024 22:26
snat0021

○ ○ Pinned version
23 Aug 2024 16:20
snat0021

○ ○ 22 Aug 2024 17:20
snat0021

○ ○ Pinned version
22 Aug 2024 16:53
snat0021

○ ○ 21 Aug 2024 23:22
snat0021

○ ○ Pinned version
21 Aug 2024 23:20
snat0021

○ ○ Pinned version
21 Aug 2024 09:06
Bhavna Balakrishnan

○ ○ 20 Aug 2024 23:14
snat0021

○ ○ Pinned version
20 Aug 2024 20:18
snat0021

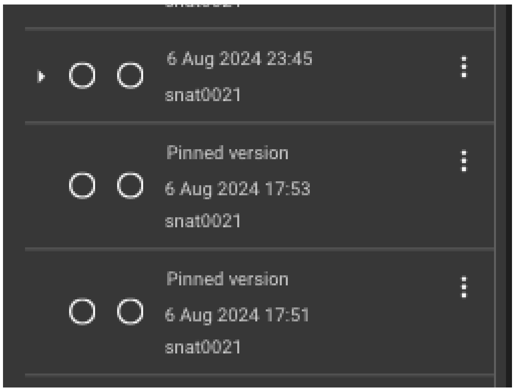
○ ○ 11 Aug 2024 19:29
snat0021

○ ○ Pinned version
11 Aug 2024 19:24
snat0021

○ ○ 10 Aug 2024 11:14
Bhavna Balakrishnan

○ ○ 9 Aug 2024 12:47
snat0021

○ ○ 7 Aug 2024 21:21
snat0021



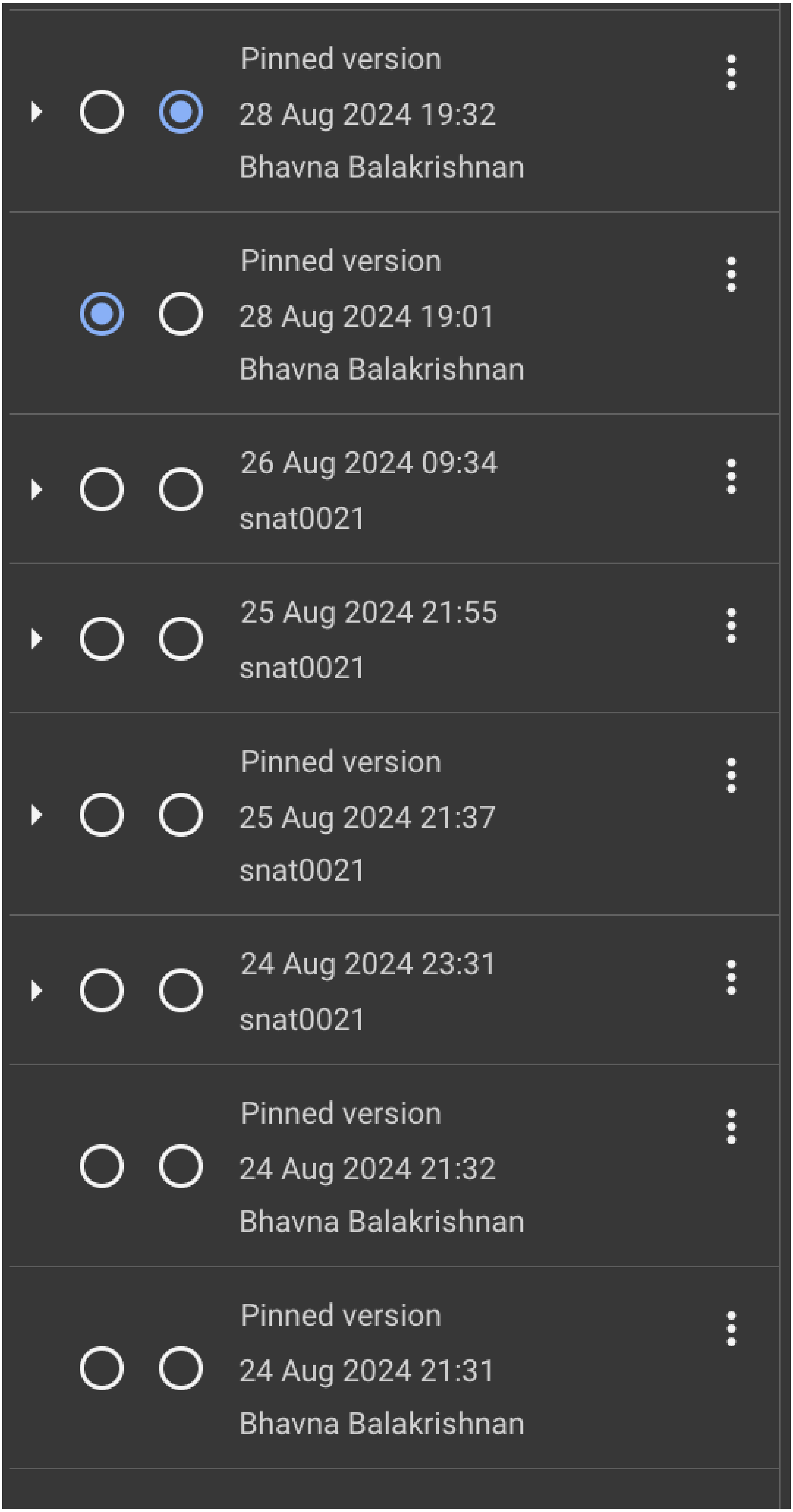
snat0021

▸ ○ ○ 6 Aug 2024 23:45
snat0021

○ ○ Pinned version
6 Aug 2024 17:53
snat0021

○ ○ Pinned version
6 Aug 2024 17:51
snat0021

colab link: https://colab.research.google.com/drive/1XPeflOMF9ghMmRFWDIpulrhFONYSZFZu?



▸ ○ ● Pinned version
28 Aug 2024 19:32
Bhavna Balakrishnan

● ○ Pinned version
28 Aug 2024 19:01
Bhavna Balakrishnan

▸ ○ ○ 26 Aug 2024 09:34
snat0021

▸ ○ ○ 25 Aug 2024 21:55
snat0021

▸ ○ ○ Pinned version
25 Aug 2024 21:37
snat0021

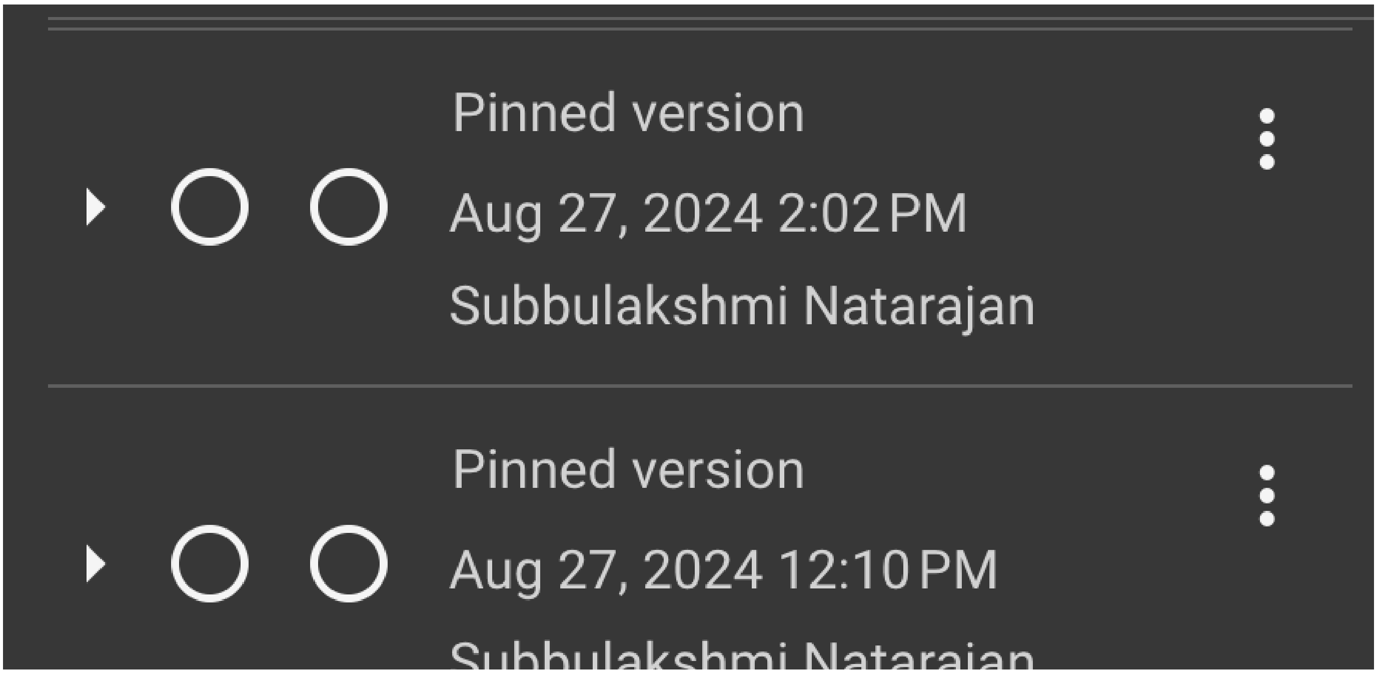
▸ ○ ○ 24 Aug 2024 23:31
snat0021

○ ○ Pinned version
24 Aug 2024 21:32
Bhavna Balakrishnan

○ ○ Pinned version
24 Aug 2024 21:31
Bhavna Balakrishnan

usp=sharing

Google Colab Link : https://colab.research.google.com/drive/1pZqG9bdN1Kd1rrdeK9WnuwjgrCaHhM6d?usp=sharing



Pinned version

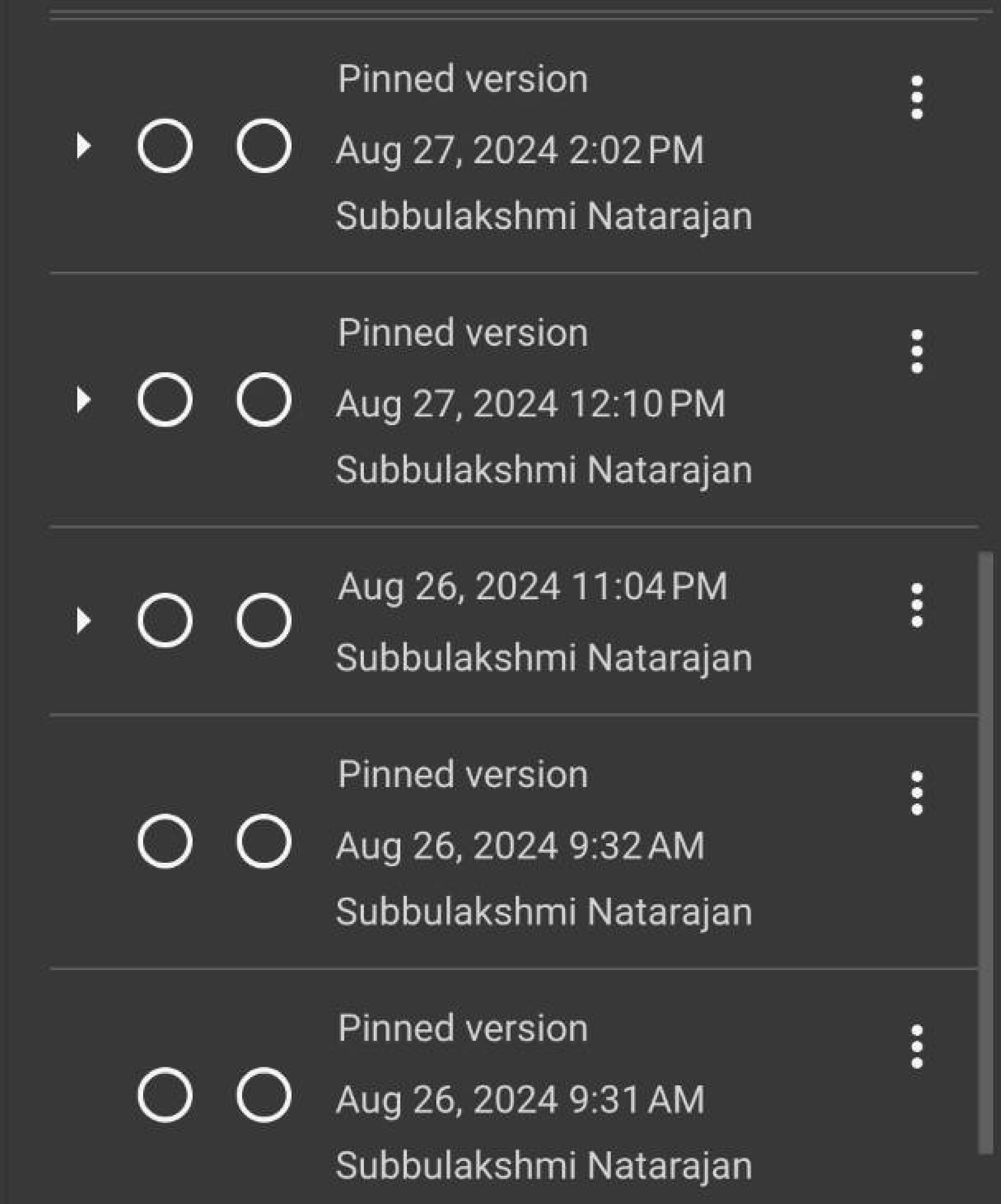
Aug 27, 2024 2:02 PM

Subbulakshmi Natarajan

Pinned version

Aug 27, 2024 12:10 PM

Subbulakshmi Natarajan



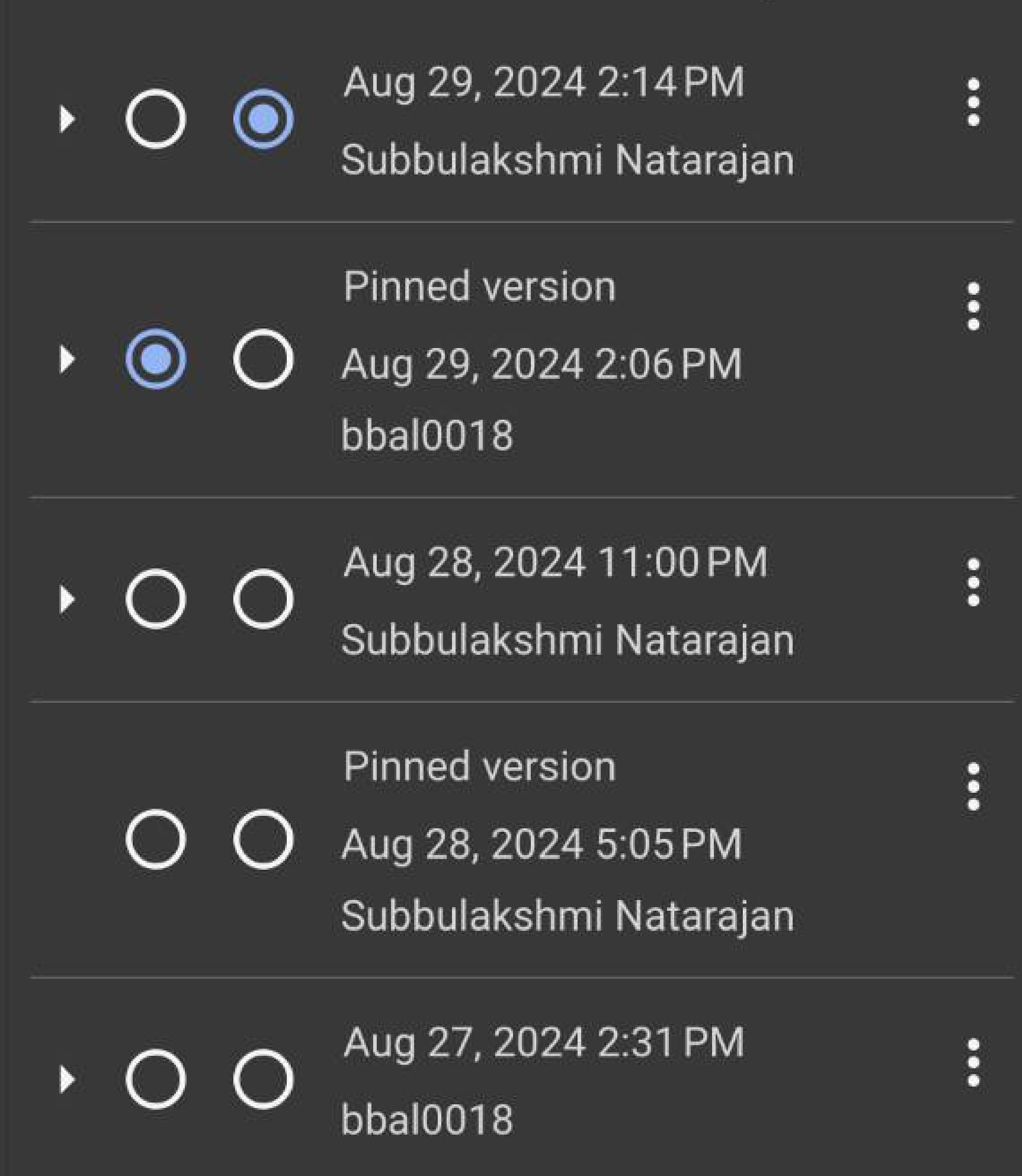
○ ○ Pinned version
Aug 27, 2024 2:02 PM
Subbulakshmi Natarajan

○ ○ Pinned version
Aug 27, 2024 12:10 PM
Subbulakshmi Natarajan

○ ○ Aug 26, 2024 11:04 PM
Subbulakshmi Natarajan

○ ○ Pinned version
Aug 26, 2024 9:32 AM
Subbulakshmi Natarajan

○ ○ Pinned version
Aug 26, 2024 9:31 AM
Subbulakshmi Natarajan



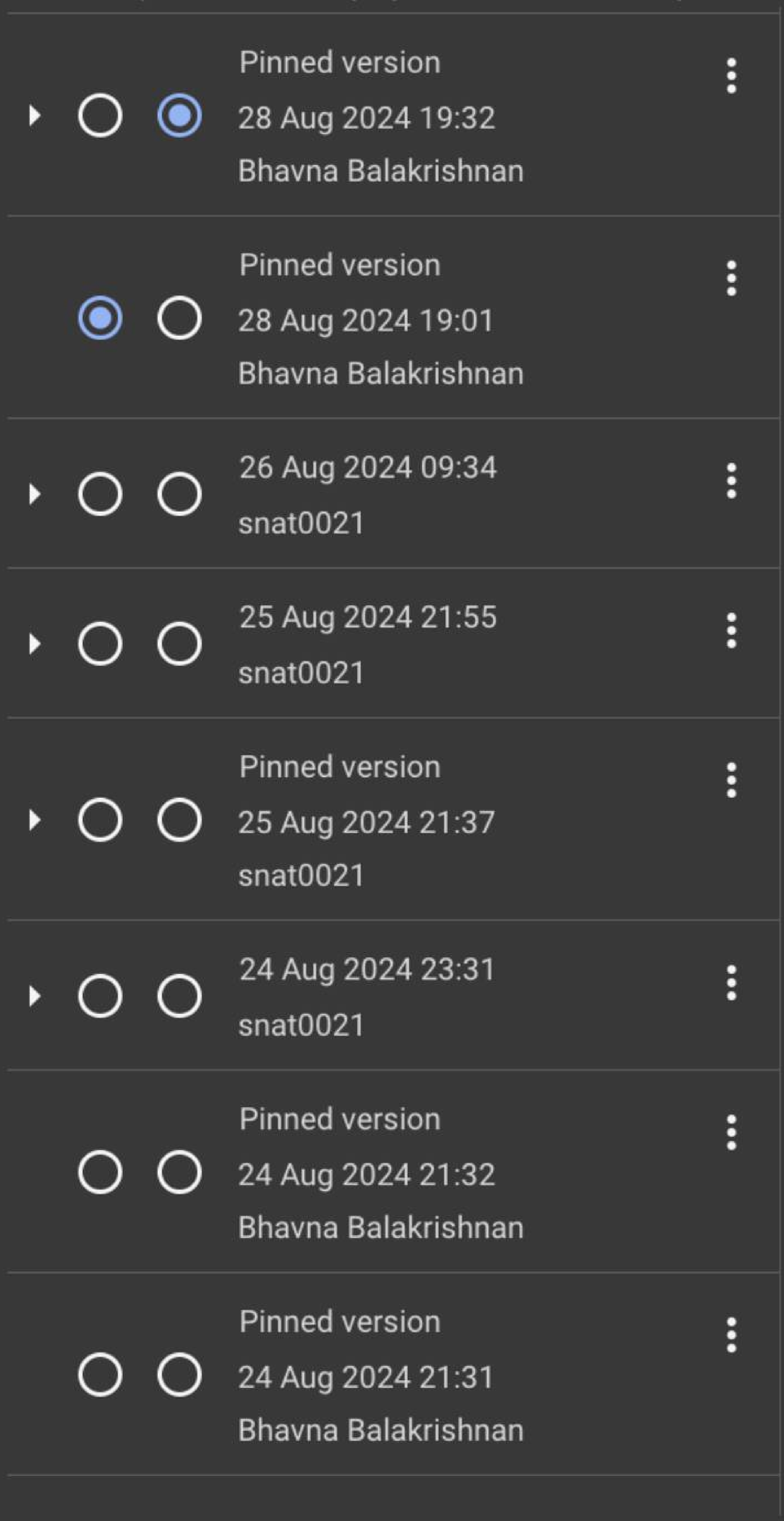
○ ● Aug 29, 2024 2:14 PM
Subbulakshmi Natarajan

● ○ Pinned version
Aug 29, 2024 2:06 PM
bbal0018

○ ○ Aug 28, 2024 11:00 PM
Subbulakshmi Natarajan

○ ○ Pinned version
Aug 28, 2024 5:05 PM
Subbulakshmi Natarajan

○ ○ Aug 27, 2024 2:31 PM
bbal0018



▸ ○ ◉ Pinned version
28 Aug 2024 19:32
Bhavna Balakrishnan ⋮

◉ ○ Pinned version
28 Aug 2024 19:01
Bhavna Balakrishnan ⋮

▸ ○ ○ 26 Aug 2024 09:34
snat0021 ⋮

▸ ○ ○ 25 Aug 2024 21:55
snat0021 ⋮

▸ ○ ○ Pinned version
25 Aug 2024 21:37
snat0021 ⋮

▸ ○ ○ 24 Aug 2024 23:31
snat0021 ⋮

○ ○ Pinned version
24 Aug 2024 21:32
Bhavna Balakrishnan ⋮

○ ○ Pinned version
24 Aug 2024 21:31
Bhavna Balakrishnan ⋮