



Zip Zap Zapier  
Tom Levy, Medha Aravind, Maggie  
Arellano

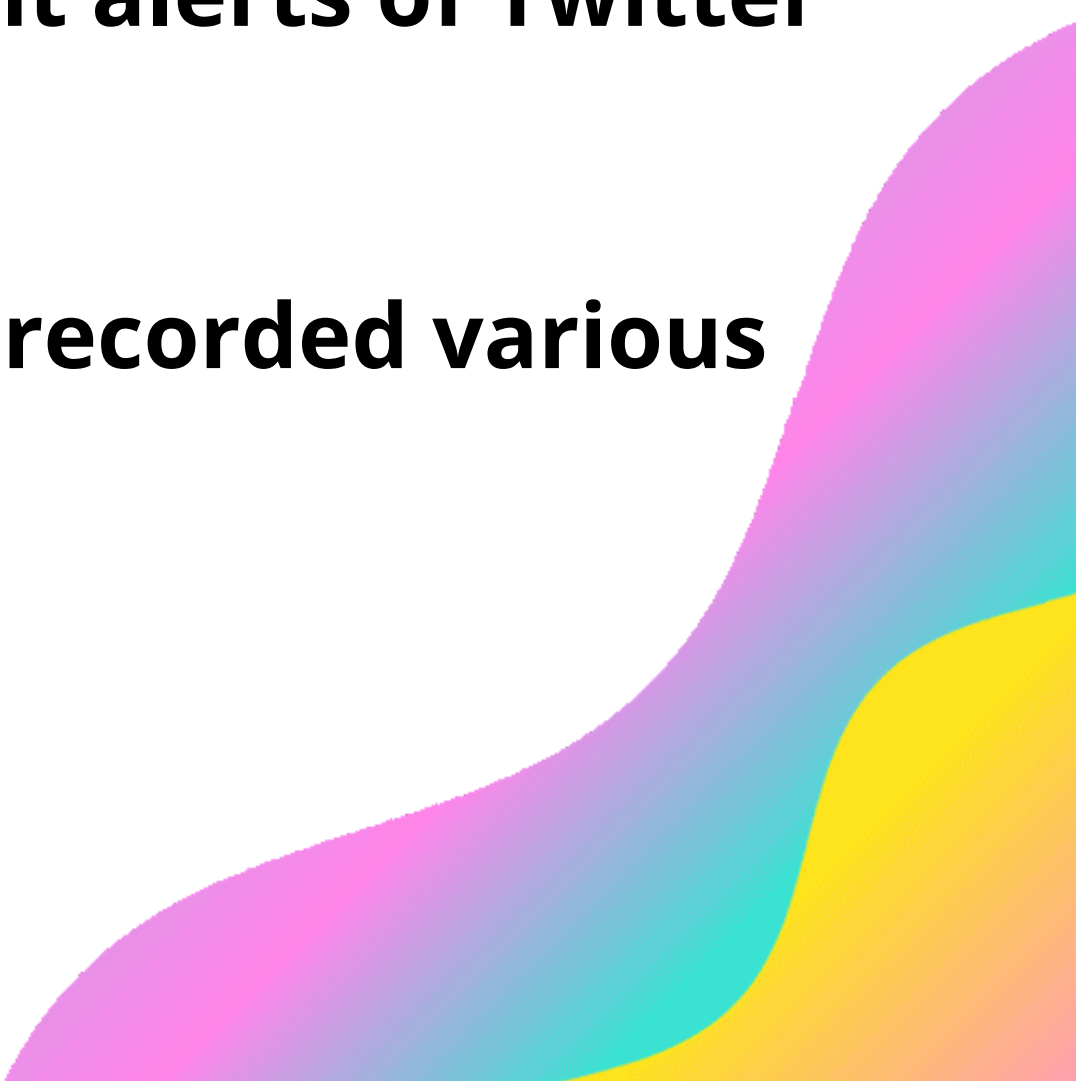
# Motivation & Summary Slide

\*What trends and distributions can we find in Slack message data?

\* Specific questions pertaining to three major aspects of the data we collected, which are text features used in the messages, time frequency of the messages and picture messages that were shared on the platform.

\* Once we were able to analyze the data we determined we had to change the questions we were asking to see the actual trends that were displaying by the DATA

# Data

- **We created a Slack app with the `conversations.history` function and added it to a blank workspace.**
  - **We added a bot called Zapier to our workspace which sent alerts of Twitter messages into multiple channels.**
  - **Through the Slack API, we were able to create a call that recorded various features of each message sent in the workspace.**
- 

# Questions

## Text Feature:

- What can be determined about the various languages of messages in the data?
- What can be determined from the lengths of messages in the data?
- What can be determined from the lengths of the words in the data?
- What can be determined from the types of characters in the data?

# Questions (Con't)

## Data related to the time:

- What can be determined about how many text messages are grabbed per millisecond?
- What format was used to display the timestamps?
- How can we convert the timestamps to general UTC time?

## Exploring the pictures:

- How many post include Emojis?
- Count of "Thumbs up" and "Joy"




# Data Cleanup & Exploration

## General:


- Certain links to pages which redirected us to makeup store webpages, etc. were eliminated.
- After going through all of the data, we found parts that we didn't think were necessary for us to carry out our analysis.


```
messages = []  
  
url_1 = f"https://slack.com/api/conversations.history?token={slack_token}&channel=C01FL12MUN4&pretty=1"  
  
responsel = requests.get(url_1)  
response_jsonl = responsel.json()  
  
for x in response_jsonl["messages"]:  
    messages.append(x)
```



# Data Cleanup & Exploration

## Text:

- We isolated the text from each specific message and created lists pertaining to language, tweet length, word length, frequency of characters
  - Language detecting-function called detect from the langdetect library in Python allowed us to determine the language each message
  - Only alphanumeric characters were analyzed for frequency
- 



# Data Cleanup & Exploration

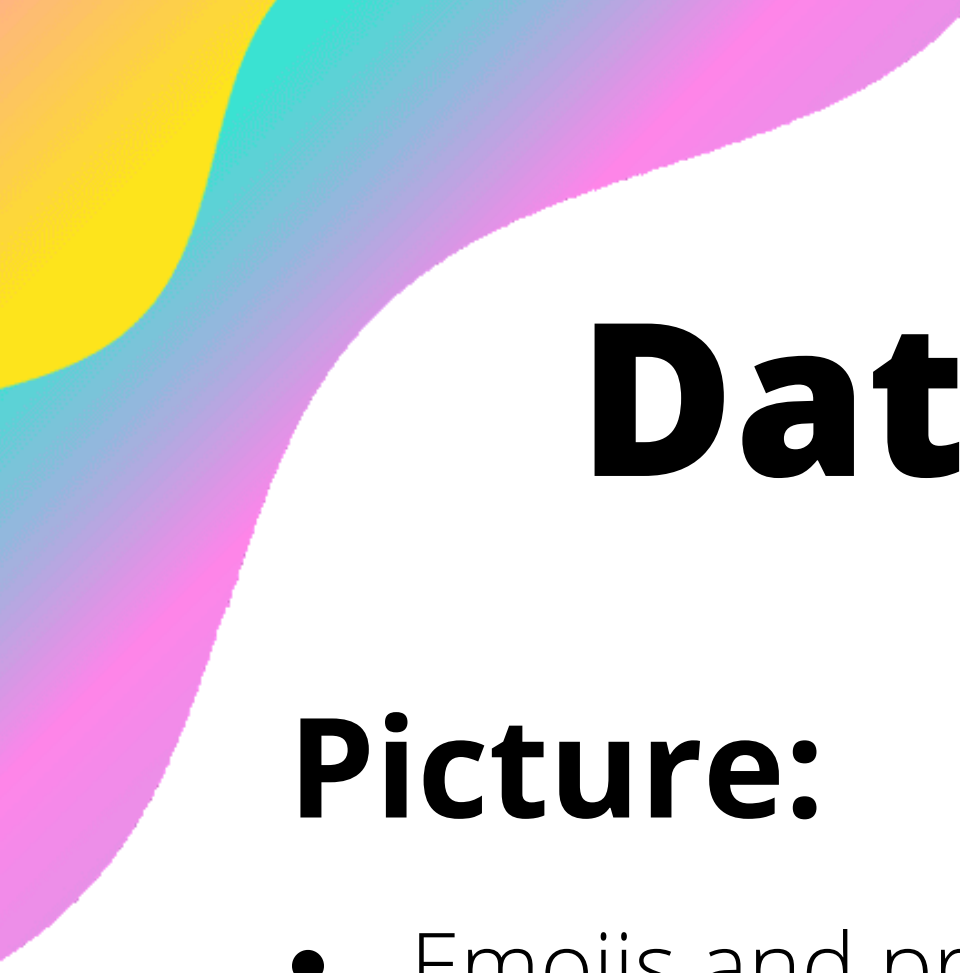
## Time:

- One factor that caught our attention was, the way the timestamps were displayed in the json that we had collected. The timestamps were displayed in Unix time and were not readable
- We figured out how to convert it to normal UTC date and time.

## Picture:

- Emojis and previews appearnt in slack message but did not display string to
- Search through data to identify and isolate emojis
- Isolated the "text" fields of the tweet to find identifiers in jpeg, png forms, none were found





# Data Cleanup & Exploration

## Picture:

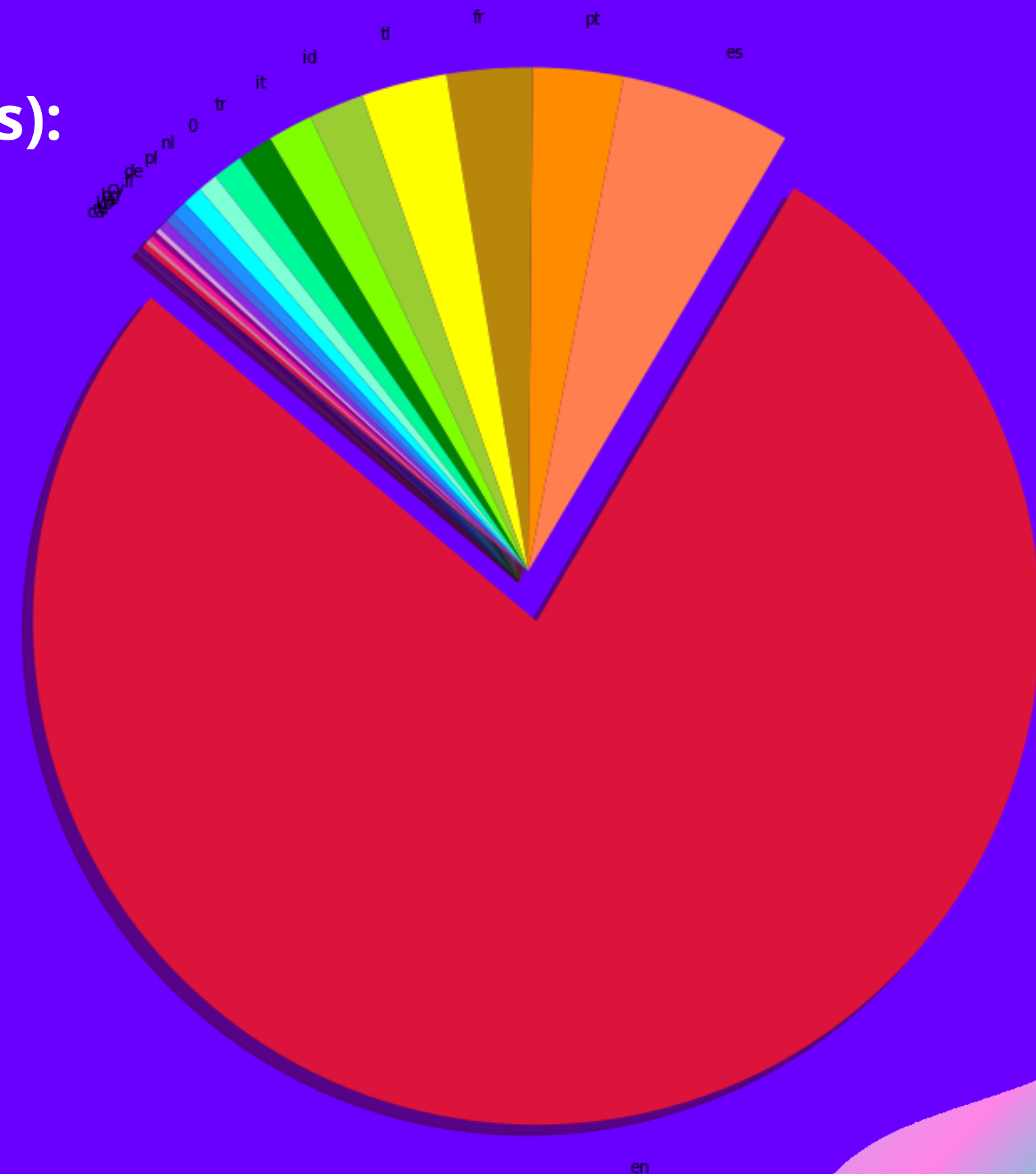
- Emojis and previews apparent in slack message
- Isolated the "text" fields of the tweet to find identifiers in jpeg, png forms, none were found
- Search through data to identify and isolate emojis, focused on one emoji "joy"

# Data Analysis - Text

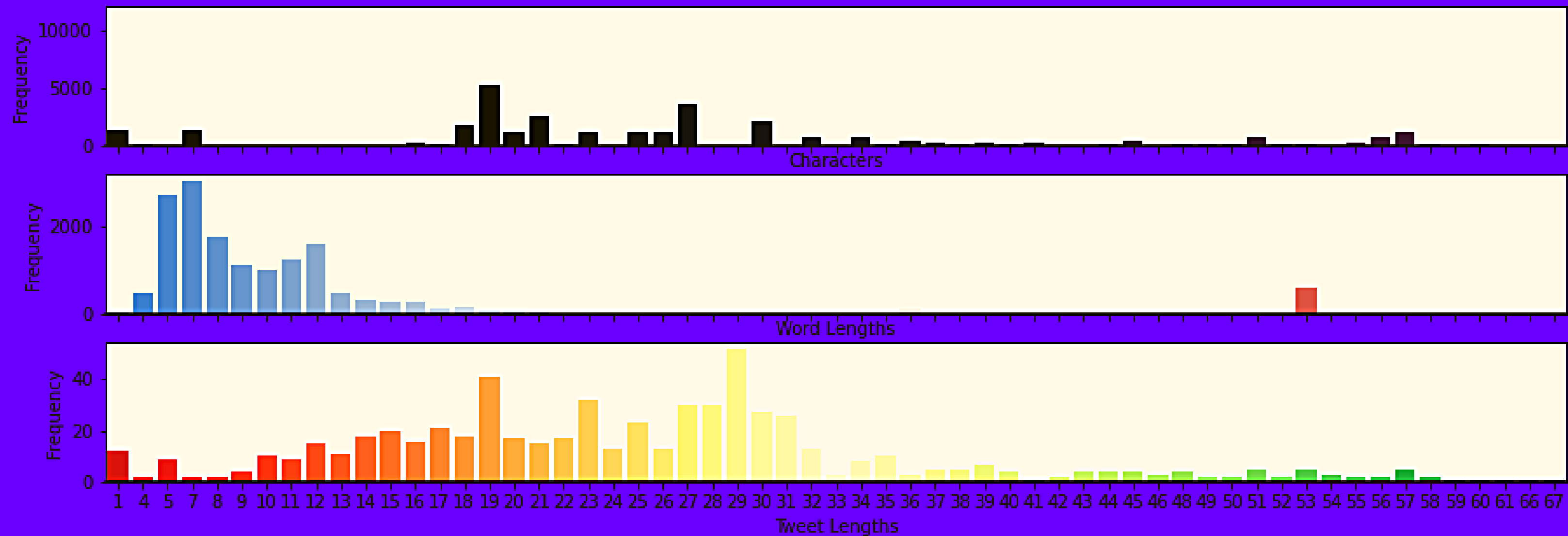
## Top 5 Languages (from 620 detectable messages):

1. English (77.6%)
2. Spanish (5.5%)
3. Portuguese (2.9%)
4. French (2.7%)
5. Tagalog (2.7%)

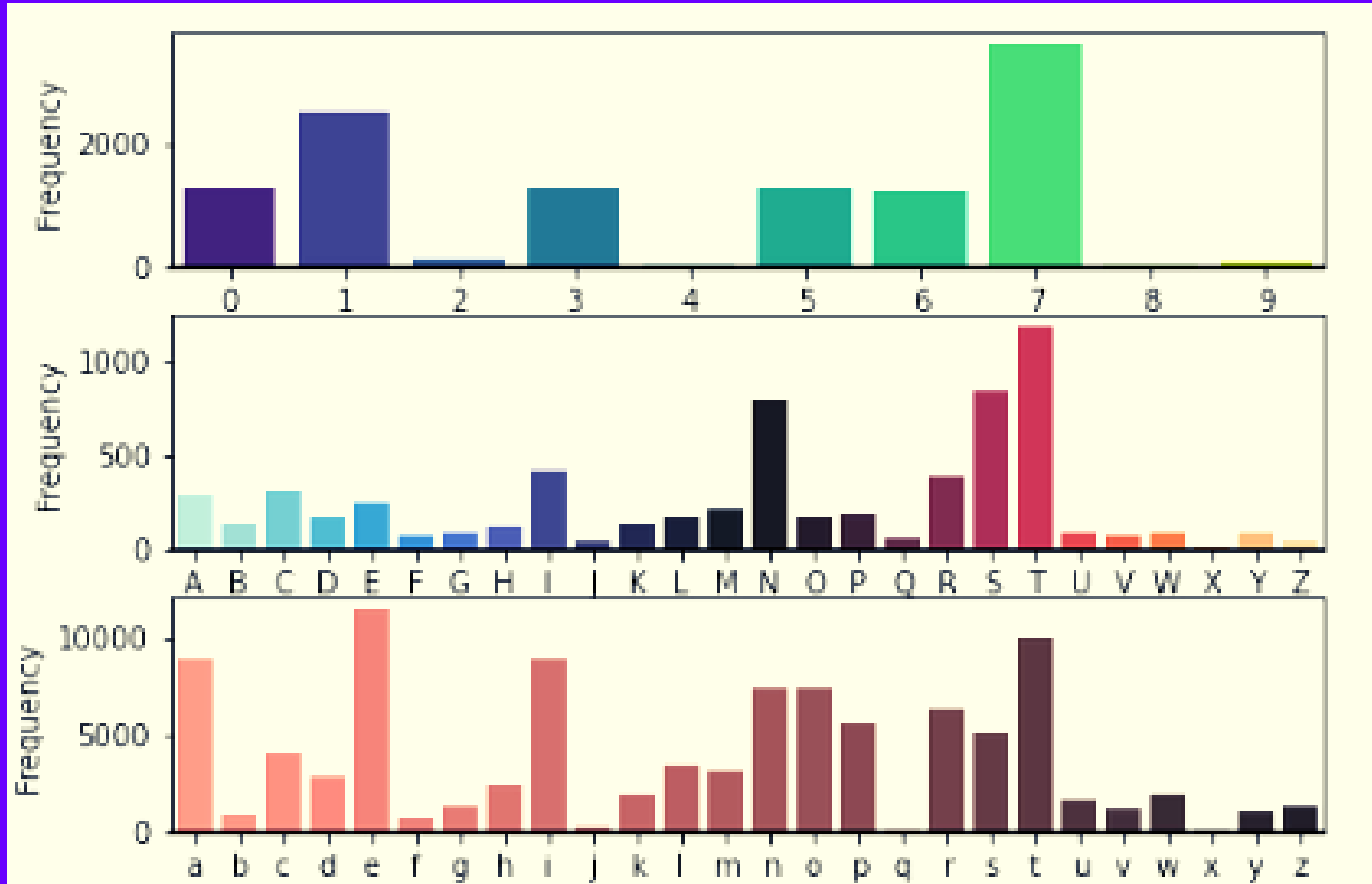
```
languages = []  
  
for y in text_files:  
    try:  
        b = detect(y)  
        languages.append(b)  
    except:  
        languages.append("0")  
    pass
```



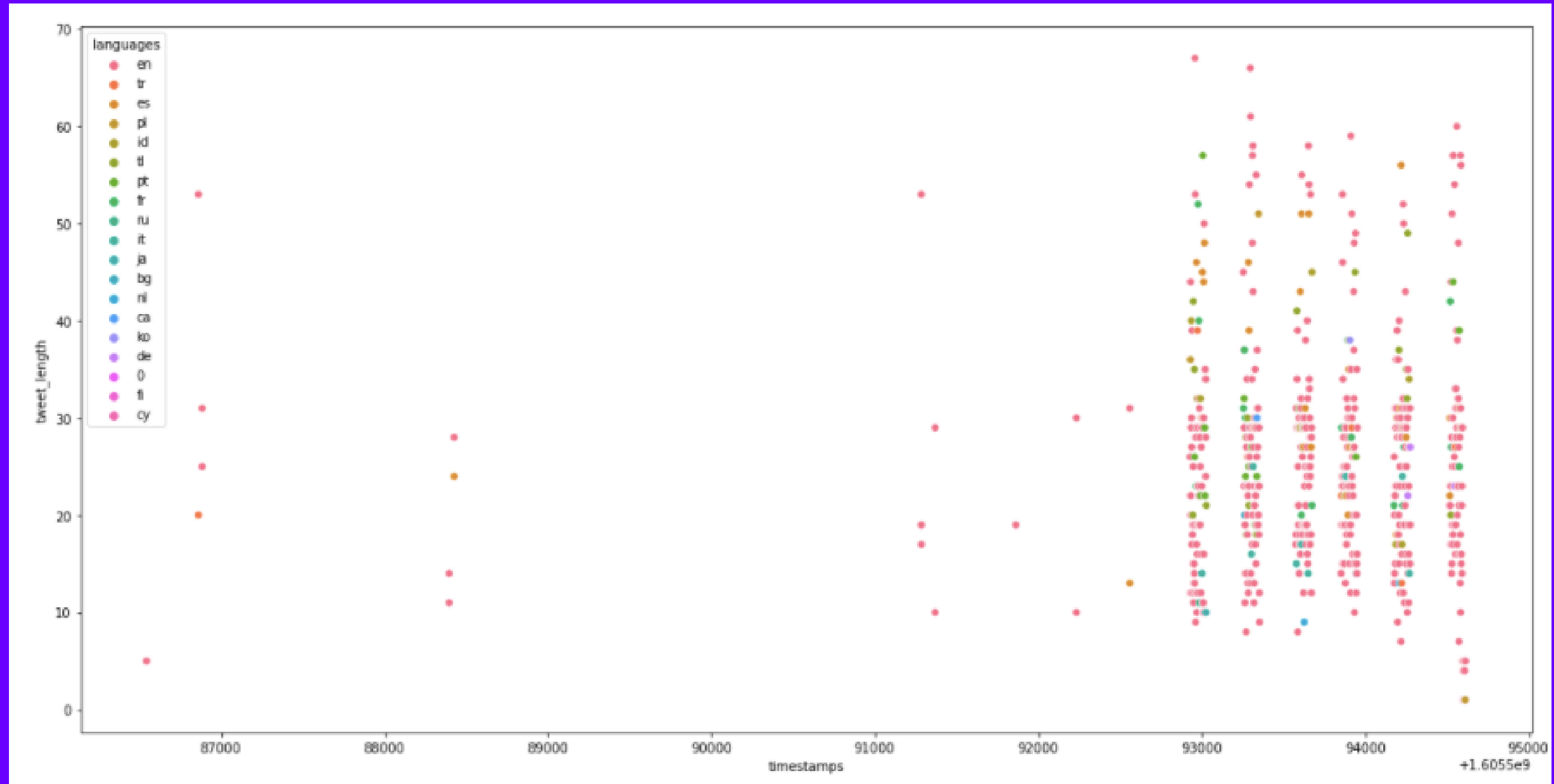
# Data Analysis - Text (con.)



# Data Analysis - Text (con.)

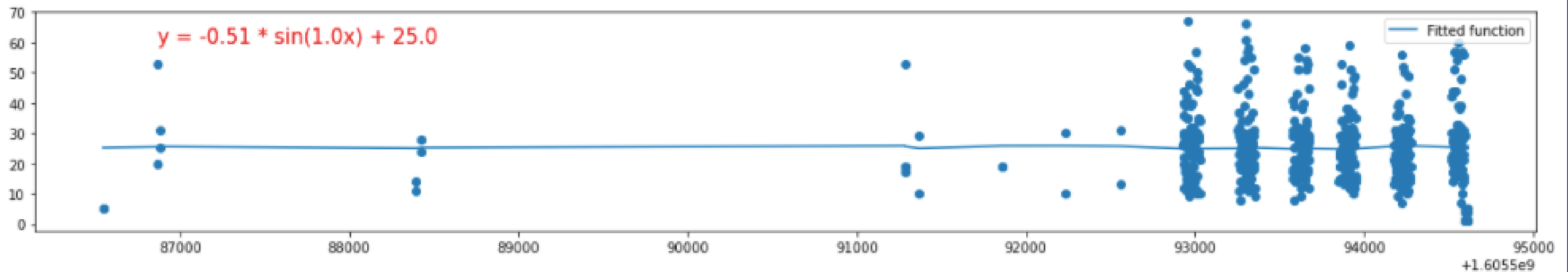
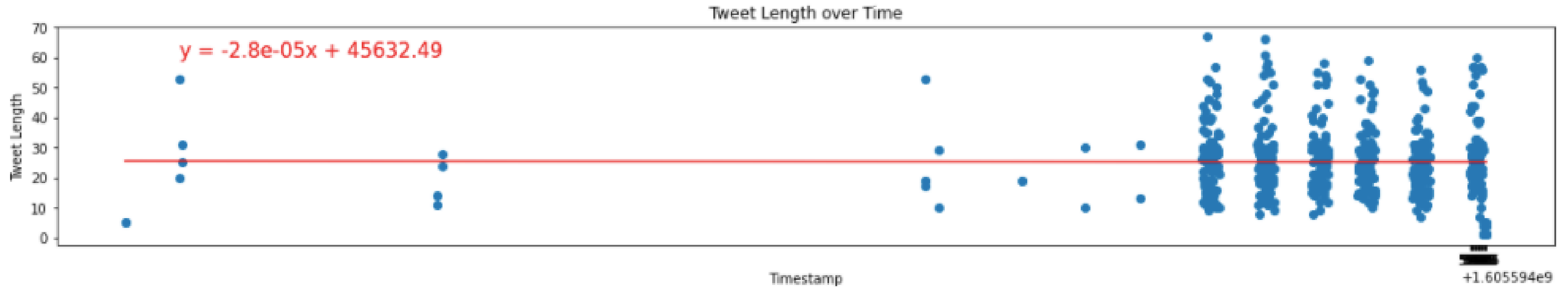


# Data Analysis - Text (con.)



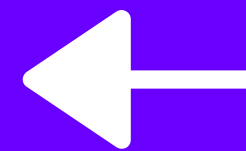


# Data Analysis - Text (con.)



# Data Analysis - Time

```
timestamps = []  
  
for x in response_json["messages"]:  
    timestamps.append(float(x["ts"]))  
    timestamps.sort()  
  
timestamps
```



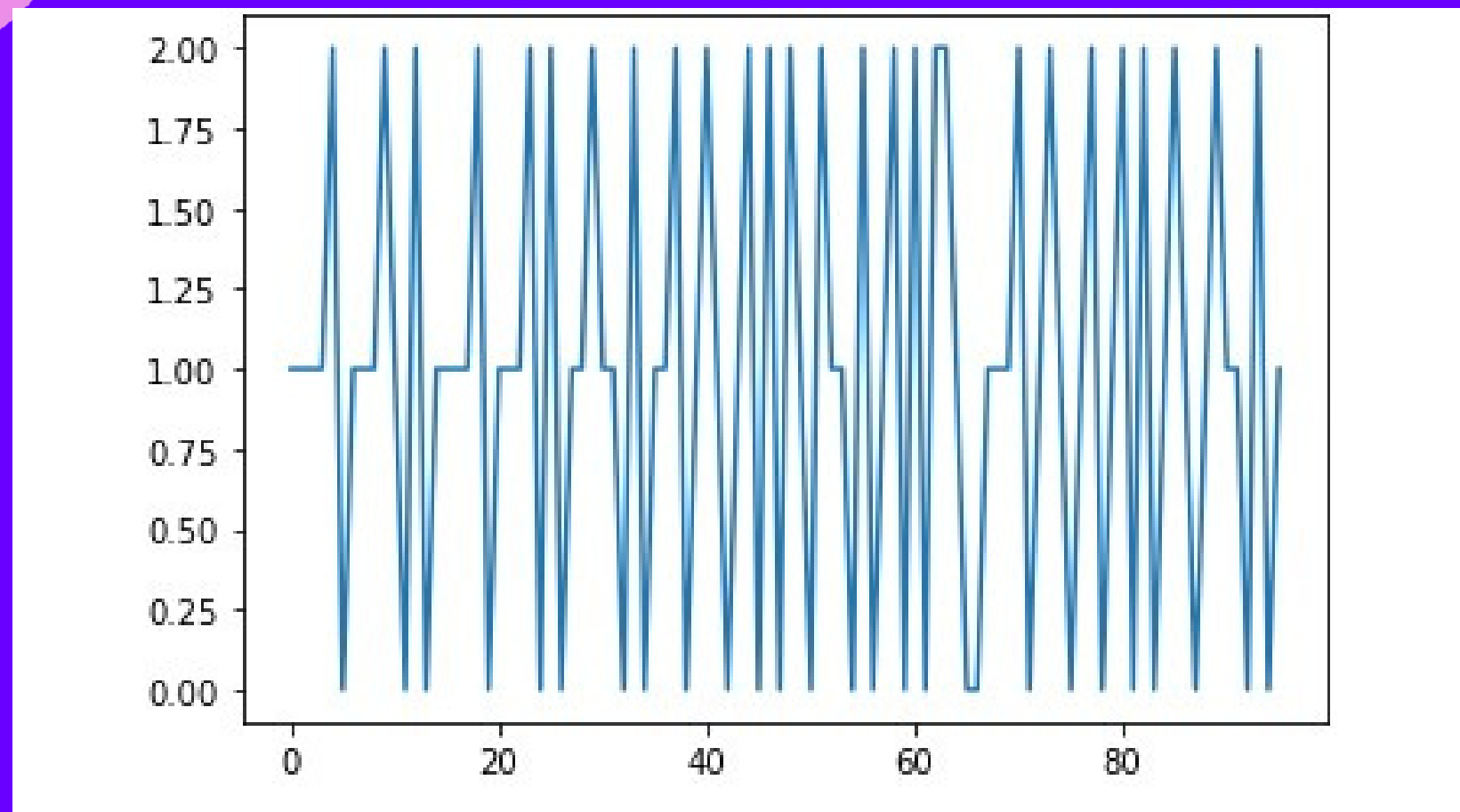
This piece of code pulled out the list of all the timestamps that were in the JSON file and we observed that the time was displayed in a format that was not comprehensive.

# Data Analysis - Time

```
for ts in timestamps:  
    local_timezone = tzlocal.get_localzone()  
    local_time = datetime.fromtimestamp(ts, local_timezone)  
    utc_time = datetime.utcfromtimestamp(ts)  
    print(utc_time.strftime("%Y-%m-%d %H:%M:%S.%f+00:00 (UTC)"))
```

- We then used the datetime module to convert the time from unix time format to normal UTC date time format and display it.

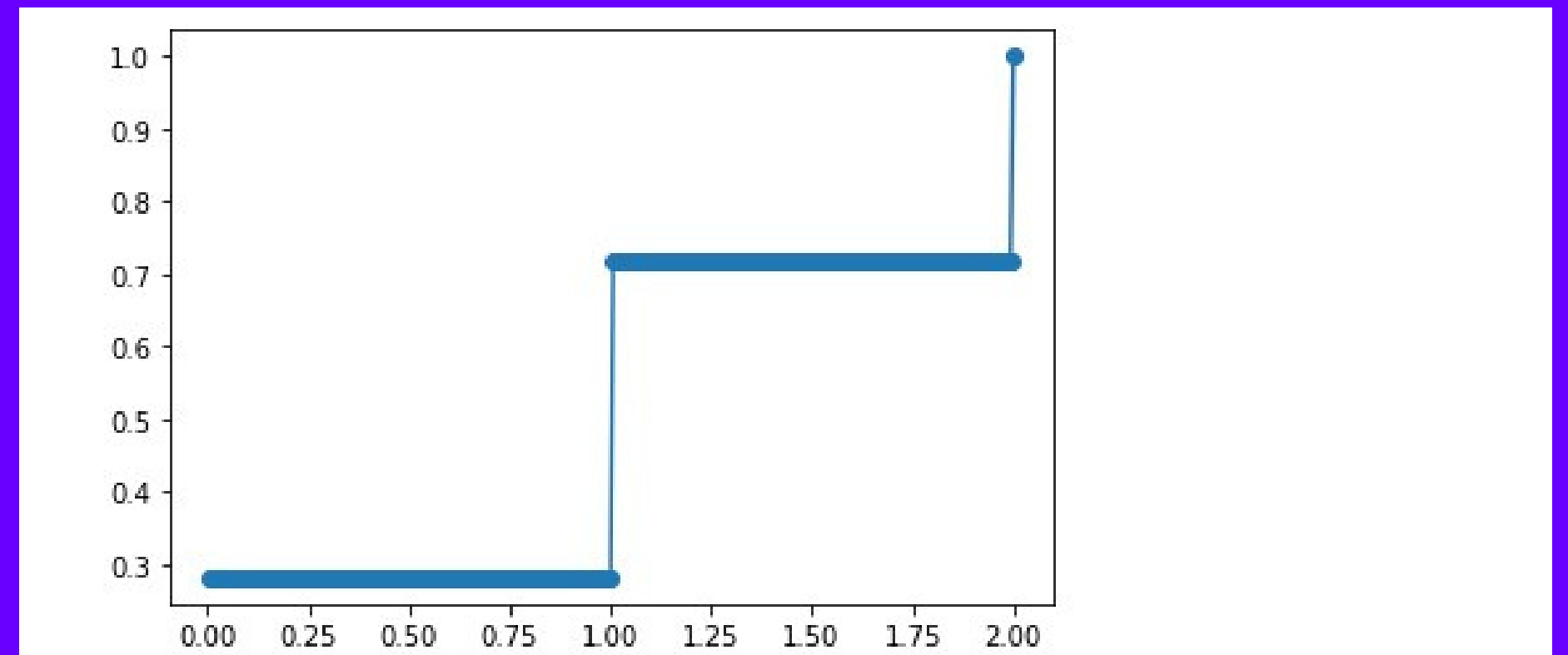
# Data Analysis - Time



- After displaying all the timestamps, we went on to try to figure out how many messages were sent in a specific time frame.
- For this we had to first find out the sequential differences of the timestamps from the previous timestamps and plot the values.

# Data Analysis - Time

- We then plotted a cumulative distributive function that showed us the concentration of messages that were exchanged per minute.







# Data Analysis - Pictures

New mention in Twitter: RT @YoungQwan: Doug Williams got a copyright claim on this video and it's scrubbed from the internet 😂😂😂 <https://t.co/ba5x3mGHEX>

Sent via [zapier.com/app/editor/105137767#slack](https://zapier.com/app/editor/105137767#slack)

```
'New mention in Twitter: RT @YoungQwan: Doug Williams got a copyright claim on this video and it's scrubbed from the internet
:joy::joy::joy: <https://t.co/ba5x3mGHEX>\n\n_Sent via <https://zapier.com/app/editor/105137767#slack|zapier.com/app/editor/10
5137767#slack>_',
```





# Data Analysis - Pictures

```
'New mention in Twitter: maganda lang pakinggan pag tinanong ka kung ano internet niyo tas sagot mo "converge" pero kapag tinanong ka na kung maganda ba connection mapapamura ka na lang eh :+1:\n\n_Sent via <https://zapier.com/app/editor/105137767#slack|zapier.com/app/editor/105137767#slack>_',
```

New mention in Twitter: maganda lang pakinggan pag tinanong ka kung ano internet niyo tas sagot mo "converge" pero kapag tinanong ka na kung maganda ba connection mapapamura ka na lang eh 👍

Sent via [zapier.com/app/editor/105137767#slack](https://zapier.com/app/editor/105137767#slack)

maganda lang ... ▼

1	maganda lang pakinggan pag tinanong ka kung ano internet niyo tas sagot mo "converge" pero kapag tinanong ka na kung maganda ba connection mapapamura ka na lang eh 👍
---	---



# Discussion and Conclusions:

- Tweet length F-statistic ( $df = 619$ ) = 3244.13, p-value=0.0
- Word length F-statistic ( $df=15640$ ) = 477184.03, p-value $\approx$ 0.0
- Two-sample T-test between lower/upper case letter frequencies
  - ( $df=50$ ): T-statistic = -5.39, p-value =  $1.29 \times 10^{-05}$
- Timestamp v tweet length:
  - linear r-value: -0.002, sine r-value:  $-7.95 \times 10^{-13}$
- Tweet/word length have large differences between expected/observed values
- There is little to no correlation between time and tweet length
- Distributions between lower/uppercase letters are significantly different, but not as extremely significantly different as prev. tests

# Discussion

- From plotting the sequential difference of the timestamps, we see that messages were being sent out with a time difference of just around 1 millisecond per message.
- From plotting the Cumulative distribution function of the timestamps, we see that approximately around 300 messages have been exchanged in just 2 minutes.
- The Cumulative distribution function gives us the spread of the time difference between tweets that were based on one topic.
- Together, the plotted sequential differences and the CDF, help us estimate the median time difference and also helps us estimate the range of values in the interquartile area of 25th percentile to 75th percentile.
- We also understood what Unix time is. Simply put, it represents the number of seconds that have passed since midnight on 1st January, 1970.

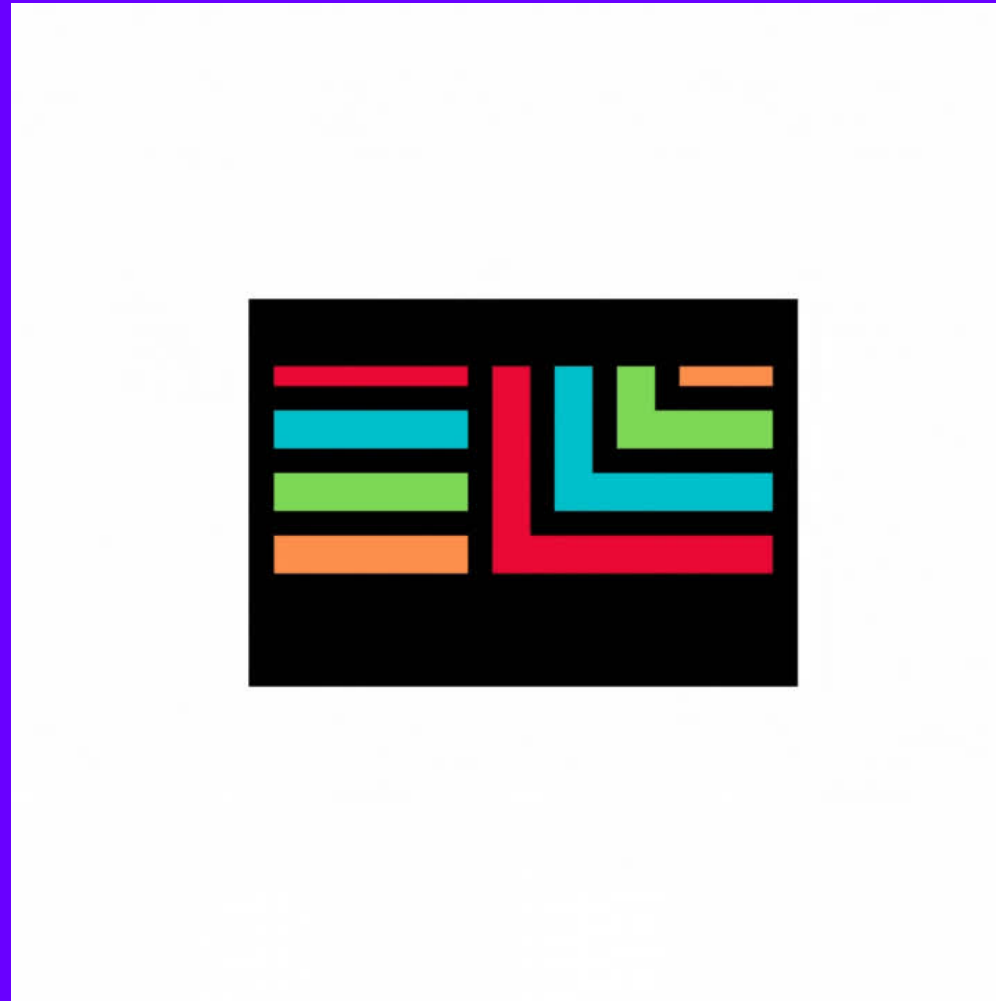
# Discussion

- We were expecting to find jpeg/png files to be able to display or count when from the slacks "preview"
- Emojis are used to draw attention, highlight importance



# Post Mortem

- Limited amount of usable data due to free Zapier task replay limit
- Lang detect function not consistent
- Small sample size of non-Latin alphanumeric characters
- Could not enable geo-tagging task for finding out how many distinct languages were used in the messages specifically sent out from the USA.
- Time stamp from Tweeter Post and analyzed a time trend for a specific channel.
- Limited amount of data also led to limitations in the time-distribution analysis since all messages that we could collect messages that were exchanged for only about 5 minutes.
- Lack of Organic Data
  - Analyze true usage of emojis for communication purposes
  - Catch trends for specific users (Since we only had one user)
- Greater knowledge of identifying picture strings and formats



Questions??