1. **Data Extraction:**

* Extracting tweets using Twitter API - v2 with academic research credentials
* Created developer account
* Created a project ‘twitter-api-research-project’ and connected it with an app.
* Generated keys and tokens
* Passed the bearer token to the environment variable
* Defined the parameters in the header and url
* Search\_url: Defined the endpoint for access. Currently set to full-archive search
* Defined the endpoint query parameters
* Export to csv

1. **Data Wrangling**
   1. **Cleaning:**
      1. Import to pandas dataframe
      2. Handling duplicate tweets:

a. Retweeted/ Quoted/ Replied\_to (referenced\_tweets\_type = NULL)

* + 1. Removal of \n, \t, &gt;
    2. Hyperlinks - New column with the hyperlink, remove the hyperlink from the tweet
    3. Hashtag removal (WIP)
    4. Emoticon conversion (WIP) - Update the tweet replacing the emoji with the relevant words + new column with the relevant words
    5. Normalizing text
    6. Handling contractions
    7. Removing unicode characters
    8. Spell checker
  1. **Transformation**
     1. Case conversion - lower
     2. Lemmatization
     3. Stop word removal

1. **Data Analysis**
   1. N-Gram
   2. Word Cloud
   3. Sentiment Analysis (Polarity & Subjectivity)
   4. Top positive & top negative tweets
   5. Comparison across all medical conditions
   6. Emoji analysis