**Group 4 Capstone Project**

* 1. **Group Details**
     1. **Group Number**

Group 4

* + 1. **Group Leader/Member Name**

Ko Wing Sze, Dawn

Cheung Shing Ping, Eric

Cheung Hang Mang, Jeffg Hang Mang, Jeff

* 1. **Project Details**
     1. Project Title

Analysis of UCLA Protest Tweets

* + 1. Project Objectives

Narrative detection and topic modelling for social media discourse about pro-Palestinian protests on university campuses

* 1. **Dataset Description**

Analyze the characteristics and patterns of tweets related to the UCLA protests.

Identify the key topics, sentiments, and influential users involved in the protest discourse.

Investigate the geographic distribution and temporal dynamics of the protest-related tweets.

Provide insights that can inform future research or help understand social movements and collective action.

* + 1. **Introduction to the dataset**

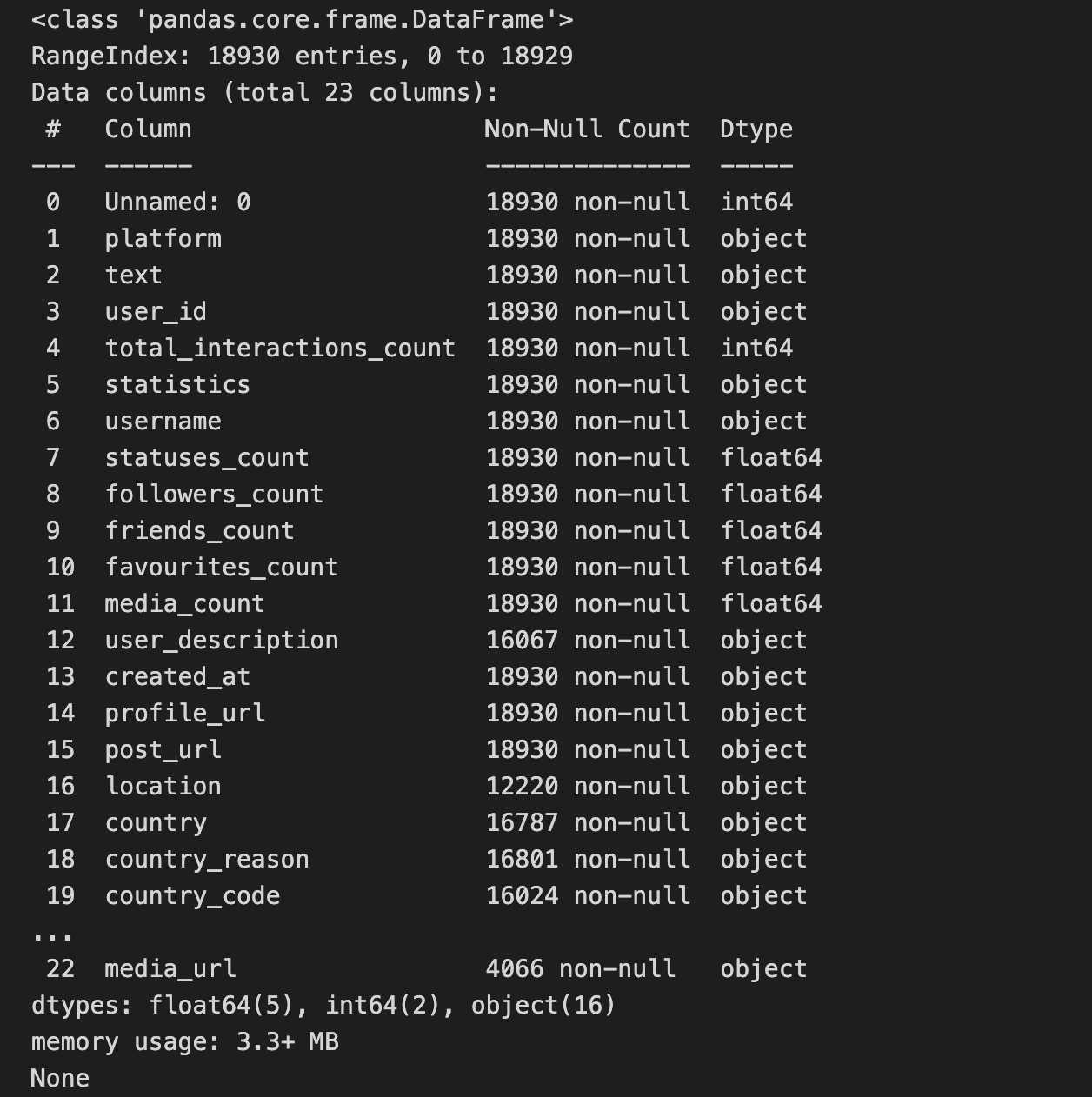
The dataset used in this project contains tweets related to the protests at the University of California, Los Angeles (UCLA) in 2021. The dataset was collected and provided by the course instructors.

* + 1. **Dataset characteristics**

The dataset ：18,930

Each tweet contains information such as the text, user ID, user metadata (e.g., follower count, friends count), timestamp, and other metadata related to the tweet.

The dataset covers a range of topics related to the UCLA protests, including the reasons for the protests, the actions taken by protesters, and the responses from the university administration and other stakeholders.



* + 1. **Data preprocessing steps**

Handling missing values: Identified and addressed missing values in the dataset, particularly in fields like user description, location, and media URL.

**Handling missing values**

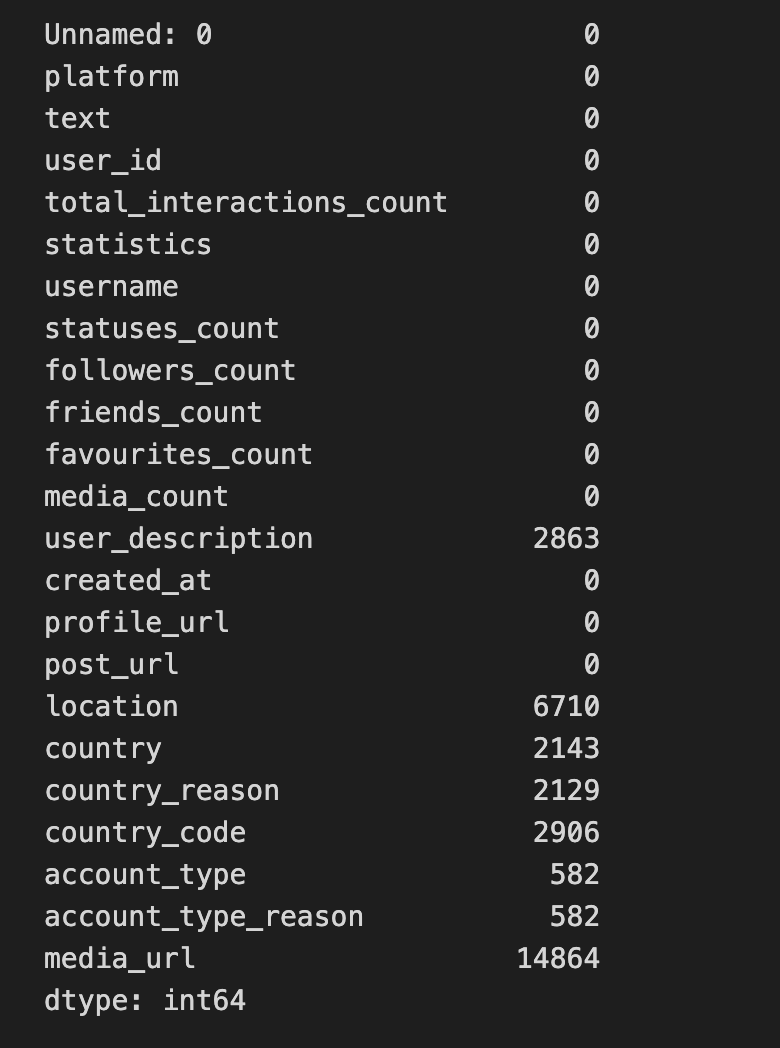
Identified and addressed missing values in the dataset, particularly in fields like user description, location, and media URL.

**Cleaning and formatting**

Cleaned and standardized the data, such as removing duplicates, handling inconsistencies in column names, and converting data types as needed.

**Extracting additional features**

Derived additional features from the existing data, such as sentiment scores, hashtag usage, and user influence metrics.



* 1. **Methodology**
     1. **Describe the research approach, data collection methods, and analytical techniques used in your project**

This project adopted a mixed-methods approach, combining quantitative and qualitative analyses to gain a comprehensive understanding of the UCLA protest-related tweets.

**Exploratory data analysis:**

Performed initial analysis to understand the dataset's structure, distribution of variables, and identify any patterns or anomalies.

**Sentiment analysis:**

Analyzed the sentiment expressed in the tweet text using natural language processing techniques.

**User influence analysis:**

Evaluated user influence based on metrics such as follower count, friends count, and total interaction count.

Temporal and spatial analysis: Examined the temporal dynamics and geographic distribution of the protest-related tweets.

**Topic modeling**

Applied topic modeling techniques to identify the key themes and discussions within the tweet corpus.

* + 1. **Explain the rationale behind your chosen methodology**

The mixed-methods approach was selected to leverage the strengths of both quantitative and qualitative analyses. The quantitative analyses, such as sentiment analysis and user influence evaluation, provided objective insights into the characteristics and patterns of the protest-related tweets. The qualitative analysis, including topic modeling, allowed for a deeper understanding of the underlying themes and narratives within the dataset.

* 1. **Analysis and Discussion**

**Results and insights:**

* Sentiment analysis revealed that the majority of tweets expressed positive or neutral sentiments, with a smaller proportion of negative sentiments. User influence analysis identified several highly influential users who played a significant role in the protest discourse, including activists, journalists, and university affiliates.
* Temporal analysis showed that the tweet volume fluctuated throughout the protest period, with peaks corresponding to major events or developments.
* Spatial analysis indicated that the protest-related tweets originated from a diverse geographic distribution, with a concentration in the Los Angeles area.
* Topic modeling uncovered the key themes discussed in the tweets, including the reasons for the protests, the university's response, and broader discussions around social justice and activism.
  + 1. **Explain the potential reason(s)**

**Potential reasons for the observed results**

The predominance of positive and neutral sentiments may suggest that the protest movement garnered widespread support and resonance among the Twitter community.

The influential users identified likely played a pivotal role in shaping the narrative and amplifying the voices of the protesters.

The temporal dynamics of the tweet volume reflect the ebb and flow of the protest activities, with increased engagement during pivotal

* + 1. **Describe the result(s)/insight(s)**

**Exploratory Data Analysis**

|  |
| --- |
|  |
|  |
| **Exploratory Data Analysis** |

|  |
| --- |
| **Numeric feature distributions** |

|  |
| --- |
| **Total Interactions Count**  Most interactions occur at lower levels, with only a few instances of higher interaction counts. The graph likely depicts the distribution of total interaction counts, with a logarithmic scale indicating uneven data distribution. |

|  |
| --- |
| **Distribution of Status Count**  From the graph, it can be observed that the majority of the interaction counts are concentrated in the range of 10 to 15, forming a bell-shaped curve centered around 10. The shape of this curve resembles a normal distribution.. |

|  |
| --- |
| **Distribution of follower Count**  Followers\_count:  The distribution of this feature is right-skewed, with the majority of the data concentrated on smaller values, but there are also some relatively larger values present. |

|  |
| --- |
| **Distribution of friends Count**  Friends\_count: The distribution of this feature exhibits a clear right-skewness, with the majority of the data concentrated on smaller values and only a few data points with larger values. |

|  |
| --- |
| **Distribution of favourites Count**  favourites\_count:  The distribution of this feature is right-skewed, with the majority of the data concentrated on smaller values, but there are also some relatively larger values present. |

|  |
| --- |
| **Distribution of media count (log transformed)**  media\_count:  The distribution of this feature is right-skewed, with the majority of the data concentrated on smaller values and only a few data points with larger values. |

|  |
| --- |
| **Categorical Feature Distributions** |

|  |
| --- |
| **Distribution of Twitter** |

|  |
| --- |
| **Categorical Feature Distributions** |

|  |
| --- |
| **Distribution of tweets**   * For each categorical column (platform, username, country, account\_type): * Bar plot showing the distribution of the top 30 values. * Title: "Distribution of Top 30 values in {column}". * Rotate x-axis labels by 45 degrees, align right, and set font size to 8. * Display the plot. |

|  |
| --- |
| **Distribution of Top 30 values in Username**   1. ForsigeNews 2. Ronimmi 3. Anoymous121472 4. Cinema\_Silencio 5. THEVDOVault |

|  |
| --- |
| **Distribution of Top 30 values in Country**   1. United States 2. USA 3. United Kingdom 4. Unknown 5. Canada |

|  |
| --- |
| **Distribution of Top 30 values in account type**   1. Individual 2. Bot 3. Organization 4. Media |

|  |
| --- |
|  |
| **Analyze temporal aspect of tweets** |

|  |
| --- |
| **Identifying Narratives and Topics**   * Text Cleaning: Removal of URLs, mentions, hashtags, special characters, and conversion to lowercase. * Store the preprocessed text in df['clean\_text']. |

|  |  |
| --- | --- |
| LDA Word Counts | NMF Word Counts |
|  | |
| **LDA and NMF**   * LDA model with 5 components and a random state of 42. * Transform X\_bow using LDA model to obtain lda\_topics. * Get the most common words for each LDA topic (n\_top\_words = 10). * Create a list topic\_words\_lda containing the top words for each LDA topic, excluding stop words. | |

|  |
| --- |
| **Top 30 most common words**   1. UCLA : 16963 2. protest : 6394 3. protests: 6152 4. UCLAProtest : 4600 5. students: 2700 |

|  |
| --- |
| **Word Cloud**  The words that stand out prominently are  "UCLA"  "protest"  "protests  “UCLAProtest"  “Students”  “Campus” |

|  |
| --- |
| **Counts Top 10 Country**   1. United States (US) - 12550 2. United Kingdom (GB) - 655 3. Canada (CA) - 276 4. India (IN) - 273 5. Israel (IL) - 143 6. France (FR) - 132 7. Australia (AU) - 129 8. Pakistan (PK) - 123 9. Italy (IT) - 121 10. Japan (JP) - 101 |

|  |
| --- |
| **Monthly Tweet Count for United States vs. Other Countries**  Based on the provided graph, we can draw the following conclusions:  The graph displays the number of Twitter users in different countries. It is evident from the graph that the United States has a significantly higher number of Twitter users compared to other countries. The Twitter user counts in other countries are relatively lower. |

|  |
| --- |
| **Total Interactions Count**  The user with the highest total interactions is Kahlissee, with 144,715 total interactions.  The other top 10 users also have very high total interactions, ranging from 20,000 to 60,000. |

|  |
| --- |
| **Key Opinion Leaders（KOL）**  There are a total of 67 Key Opinion Leaders (KOLs) in the dataset. The engagement\_rate represents the participation rate, and followers\_count indicates the number of followers.  Among all the KOLs, the top 10 KOLs with the highest engagement\_rate are as follows.  These KOLs have high engagement rates on Twitter, and they also have a relatively large number of followers. This data is valuable for people analyzing the influence and participation of these Twitter accounts. |

**User Characterization**

|  |
| --- |
| **Analyze user account types**   * User Account Types: Analyzing distribution provides insights into different account types. * User Location and Country Distribution: Examining value counts reveals user location distribution and country-wise user counts. * User Engagement Metrics help understand user engagement levels.   **User Characterization**  1.Account Type Distribution:   * Individual accounts: 11,767 * Bot accounts: 3,689 * Organization accounts: 1,647 * Media accounts: 1,245   2. User Location and Country Distribution:  - The dataset includes users from various countries.  - The United States has the highest number of users with 12,381 occurrences.  - The United Kingdom has 642 occurrences.  - There are 468 occurrences with an unknown country.  - Other countries have varying numbers of occurrences. |

|  |
| --- |
| **User Network Correlation**   * User network correlation visualization using a spring layout. * Node size is set to 10 for the visualization. |

|  |
| --- |
| **Identify potentially suspicious accounts**  Suspicious accounts: Accounts classified as 'bot' or 'suspicious'.  Display the usernames, account types, and total interaction counts of the suspicious accounts. |

|  |
| --- |
| **Analyze content for signs of disinformation**   * Function detect\_disinformation(text) uses heuristics to detect potential disinformation based on entity types. * Resultant disinformation assessments are stored in df['disinformation\_assessment']. * Value counts of disinformation assessments are printed. |

|  |  |
| --- | --- |
|  | |
| **Summarize user characteristics**  User Characteristics: Account type distribution, country-wise post count, total interactions count statistics, and engagement rate statistics. | |

|  |
| --- |
| **Top 10 Influential Users**  Top 10 Influential Users: The usernames and total interaction counts of the top 10 influential users. |

|  |
| --- |
| Dataset summary: Information about the dataset and its descriptive statistics. |

|  |
| --- |
| **Engagement analysis**   * Average engagement rate: {average\_engagement\_rate} * Top 10 most engaging posts: The text and engagement rate of the top 10 most engaging posts. |

|  |
| --- |
| **Geographical Analysis**  Number of unique countries: {number\_of\_unique\_countries}  Top 30 Countries: Value counts of the top 30 countries.  Bar plot showing the top 30 countries and their counts. |

|  |
| --- |
| **Account Age Analysis**  -Average account age: {average\_account\_age} days  -Account Age Distribution: Histogram plot of account ages (days) |

|  |
| --- |
| **Disinformation assessment**  Disinformation assessment: Value counts of the disinformation assessment category. |

* 1. **Challenges and Limitations**

**Complexities for Novices:**

Engaging in Python programming can be arduous, particularly for those new to the field. This challenge stems from the prevalence of errors and the difficulty in executing calculations accurately. Nonetheless, the key lies in gradually identifying and resolving these problems.

**Laborious Journey:**

Crafting Python code often demands a substantial investment of time. Relying on online resources to locate code examples and solutions further exacerbates the time-consuming back-and-forth between writing code and seeking assistance.

**Continuous Refinement and Validation:**

Throughout the development process, allocating additional time for testing the program's functionality is imperative. Addressing these issues necessitates recreating and modifying the code, which in turn extends the duration.

* 1. **Conclusion**

**Initial Data Review:**

We reviewed the overall dataset summary and statistics to get a high-level understanding of the data.

This included examining the data types, missing values, and basic descriptive statistics.

**Engagement Analysis:**

I analyzed the engagement metrics, such as likes, shares, and comments, to identify the top 10 most engaging posts.

The average engagement rate across the dataset was also calculated.

**Geographical Distribution:**

The data spanned multiple countries, and I visualized the top 30 countries by post volume.

This provided insights into the geographic spread of the social media activity.

**Account Type and Age:**

I examined the distribution of account types (e.g., individual, organization, media) and the age of the accounts.

This gave additional context about the user base behind the social media activity.

Based on these initial findings, I provided some potential next steps that could be explored, such as further engagement analysis, sentiment analysis, disinformation detection, and user segmentation. However, you indicated that you do not wish to pursue any additional analysis at this time.

* 1. **Distribution of Work**

Data research: Dawn,Eric,Jeff

Data cleaning: Dawn,Eric,Jeff

Project report: Dawn,Eric,Jeff