Practical 1.1: Handling big data

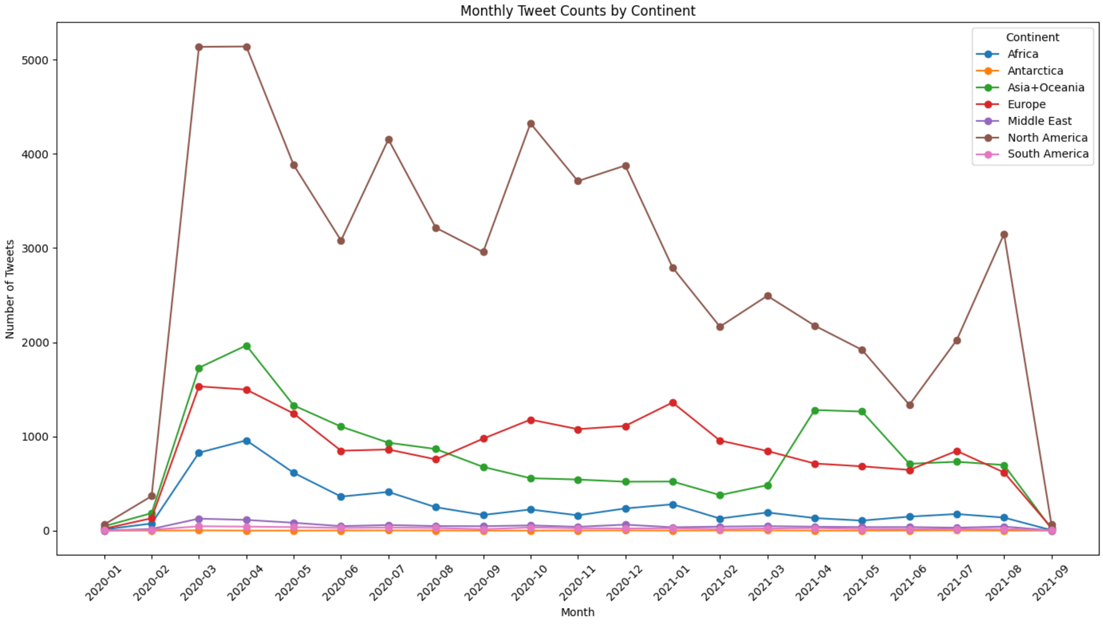
Group8 : Fulin ZHANG, Shiyao GU, Ruicong WANG

Section 1: Tweet Volume by Continent from January 28th, 2020 to September 1st, 2021

In the first part of our analysis, we aimed to visualize the tweet volume across different continents from January 28th, 2020, to September 1st, 2021, using a line chart. The approach involved leveraging the **country/region** column in the original dataset to identify the continent associated with each tweet. To achieve this, I sourced a JSON file from the internet that maps country names to their respective continents. This file served as the continent mapping reference.

We performed a group-by operation on the data to aggregate the tweet volumes per continent for each month. The resulting line chart revealed several key insights:

1. **Significant Growth in February 2020:** There was a notable spike in tweet volumes in February 2020 across North America, Asia-Pacific, Europe, and Africa. North America's increase was particularly pronounced, exceeding the combined tweet volumes of the other continents.
2. **Sustained Dominance of North America:** North America consistently maintained a significantly higher tweet volume compared to other continents throughout the analyzed period.
3. **Stable Low Volumes in South America, Middle East, and Antarctica:** These regions showed minimal growth and consistently low tweet volumes, likely due to a smaller user base and lower engagement on Twitter.

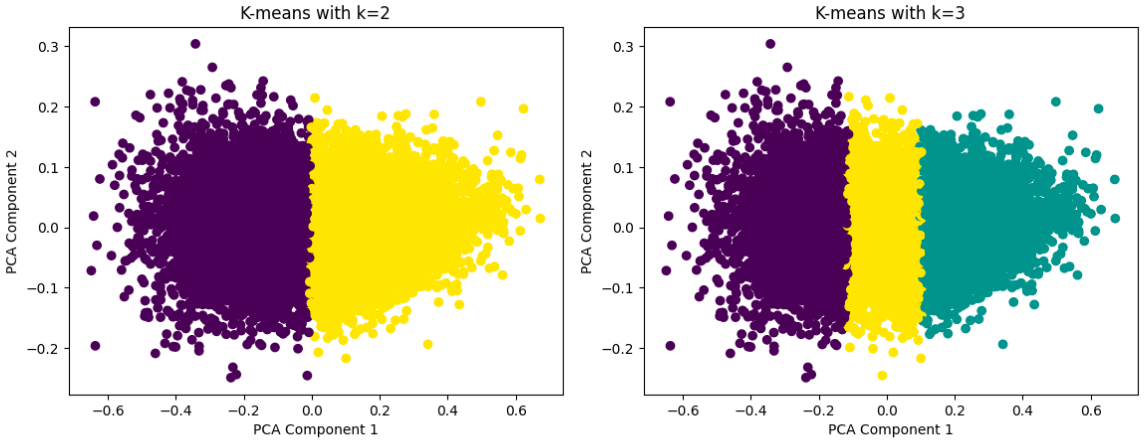


The observed peaks in tweet volumes could reflect significant COVID-19-related events and milestones that sparked widespread public discussion, such as worsening or improving health conditions, government interventions, or other impactful developments.

These preliminary findings suggest potential areas for deeper analysis, focusing on specific events or time periods to uncover the underlying drivers of tweet activity.

Section 2: K-means Clustering on 20% of the Dataset

In the second part of our analysis, we randomly selected 20% of the dataset and focused on five numerical features related to emotions: 'valence_intensity', 'fear_intensity', 'anger_intensity', 'happiness_intensity', and 'sadness_intensity'. We applied k-means clustering with two different values of k (k=2 and k=3) and used PCA to reduce the dimensionality, allowing us to visualize the clusters in a lower-dimensional space.



Clustering and Visualization:

- **k=2:** The visualization showed two relatively distinct clusters with a clear boundary near PC1=0, and the points were evenly distributed along PC2 in the range of (-2, 2). This suggests that the two clusters might correspond to positive and negative sentiments.
- **k=3:** The clustering result displayed two relatively clear boundaries. Due to hardware limitations, I initially tried to cluster 10,000 samples but faced kernel crashes. However, when clustering 20% of the entire dataset, the clusters were more distinct, with boundaries near PC1=±0.17. This indicates the possibility of three sentiment categories: positive, negative, and neutral.

Silhouette Scores: We calculated the silhouette scores to statistically assess the clarity of the clustering:

- Silhouette Score for k=2: 0.41739173531089546
- Silhouette Score for k=3: 0.3240152541349422

These scores indicate relatively clear boundaries between clusters, with k=2 showing more distinct clustering.

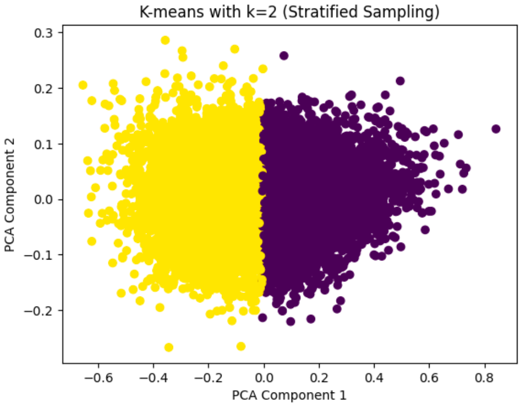
Assigning Remaining 80% Data: We then assigned the remaining 80% of the data to the clusters using cosine similarity and evaluated the clustering quality using the homogeneity score. This score measures the extent to which each cluster contains only a single class of data points:

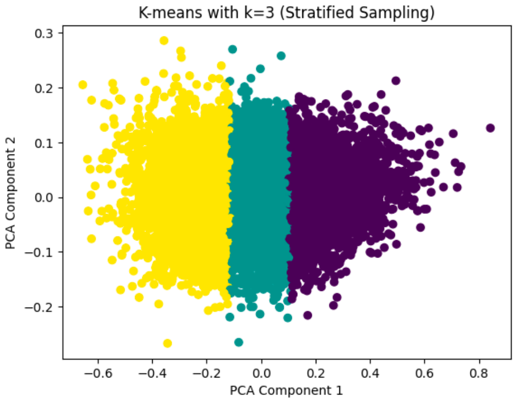
- Homogeneity Score for 20% data with k=3: 0.4208224836348746
- Homogeneity Score for 80% data with k=3: 0.42377200812340265

The results were nearly identical, indicating consistent clustering quality across both subsets of data.

Section 3: Stratified Sampling and Generalization

In the third part, we repeated the analysis using stratified sampling based on sentiment to evaluate the generalization and robustness of the results.





Stratified Sampling Results:

- Silhouette Score for k=2 (Stratified Sampling): 0.41738168428576405
- Silhouette Score for k=3 (Stratified Sampling): 0.32332622920185083

Homogeneity Scores:

- Homogeneity Score for 20% stratified data with k=3: 0.4122551099768931
- Homogeneity Score for 80% stratified data with k=3: 0.42546364069783355

Conclusion: The results from both random and stratified sampling were similar, indicating that the clustering results are robust and generalizable. The slight differences in silhouette and homogeneity scores suggest that both sampling methods perform similarly in terms of clustering quality. Stratified sampling offers a more balanced representation of sentiments, but the overall clustering performance remains consistent. This indicates that the chosen features and clustering method are effective in capturing the underlying sentiment structure in the dataset.