



**EAST WEST UNIVERSITY**

**CSE477**

**Section: 02**

**Lab: 04 Report**

**Topic: Incremental Pattern Mining & Correlation on  
YouTube Comments**

**Submitted By:**

**Name: Md Sifatullah Sheikh**

**ID: 2022-1-60-029**

**Submitted To:**

**Amit Mandal**

**Lecturer**

**Department of Computer Science & Engineering**

**Date: 24 July 2025**

# 1. Introduction and Motivation


In previous labs, we explored, cleaned, and prepared YouTube comment data for analysis. Lab 3 builds on this foundation by simulating a real-world scenario where data arrives incrementally over time rather than all at once. This approach addresses two important questions:

- How do frequent linguistic patterns evolve as new data arrives?
- Can we detect correlations between words or themes over time?

Incremental pattern mining enables continuous updates to insights without reprocessing the entire dataset repeatedly. Correlation analysis uncovers relationships between words, revealing which terms tend to rise or fall together, which is crucial for understanding evolving discussions.

## 2. Data Description

For this lab, we used the cleaned YouTube comments dataset from Lab 2 (`cleaned_comments.csv`). This file contains preprocessed comments with a `cleaned_tokens` column, where each comment is tokenized and cleaned of noise. Using this consistent, noise-free dataset ensures the discovery of meaningful patterns and comparability across analyses.

 <code>cleaned_comments</code>	7/15/2025 8:06 AM	Microsoft Excel C...	39 KB
 <code>overlapping</code>	7/24/2025 10:36 PM	Python Source File	3 KB

## 3. Data Segmentation Methodology

To simulate incremental data arrival, the dataset was divided into five equal chunks based on row indices. Each chunk represents a "snapshot" of the data at successive time intervals. This segmentation approach assumes the original dataset is approximately chronologically ordered, though this may not perfectly reflect actual comment timestamps.

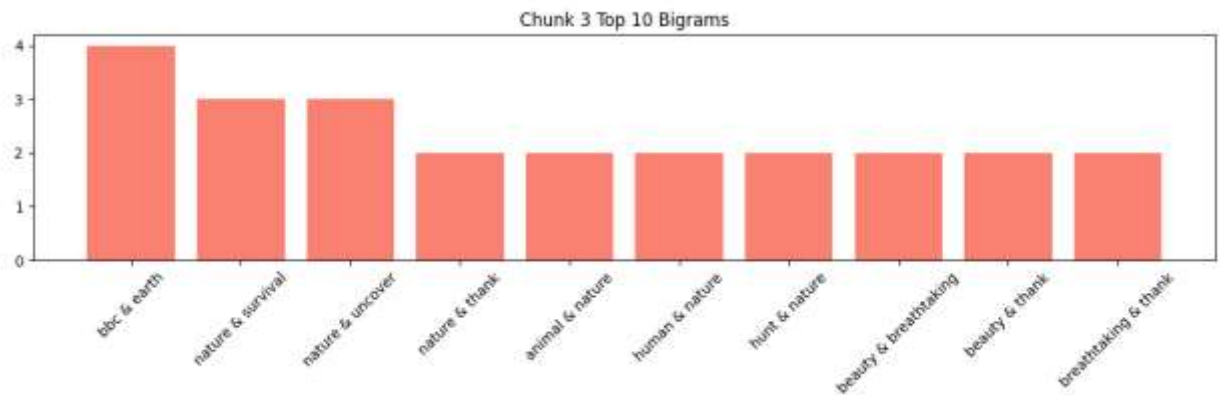
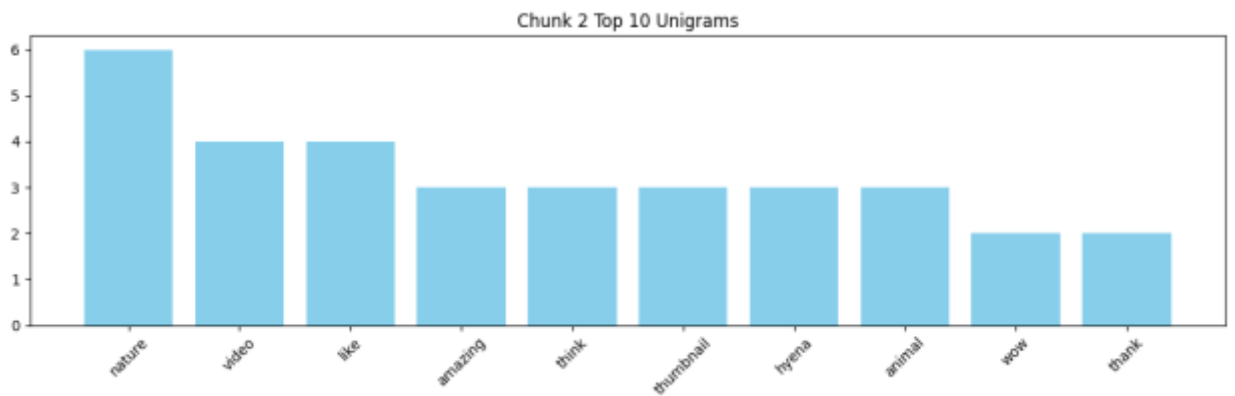
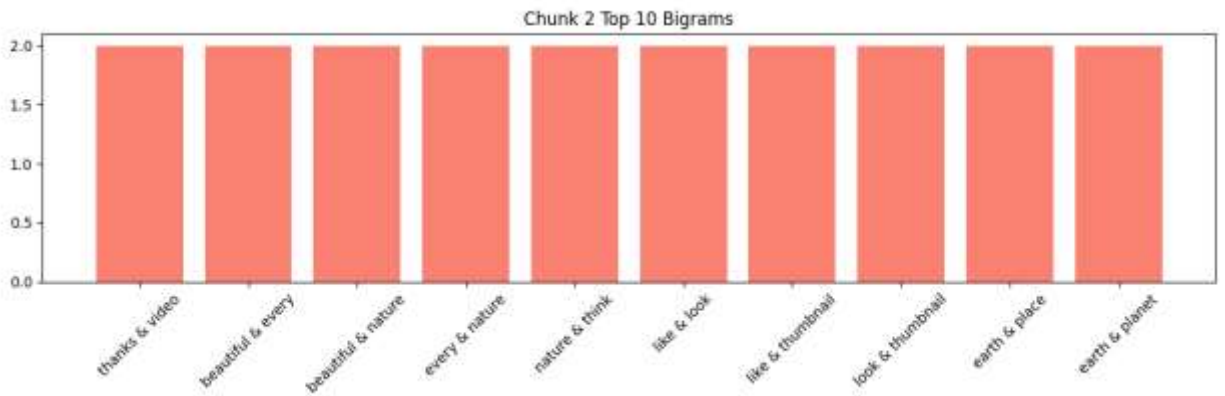
- Chunk 1: Rows 1–20%
- Chunk 2: Rows 21–40%
- Chunk 3: Rows 41–60%
- Chunk 4: Rows 61–80%
- Chunk 5: Rows 81–100%

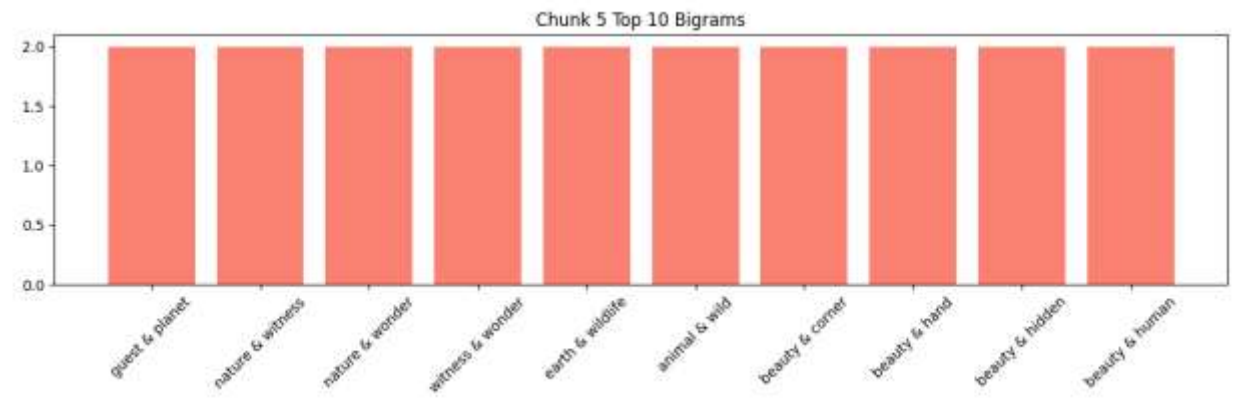
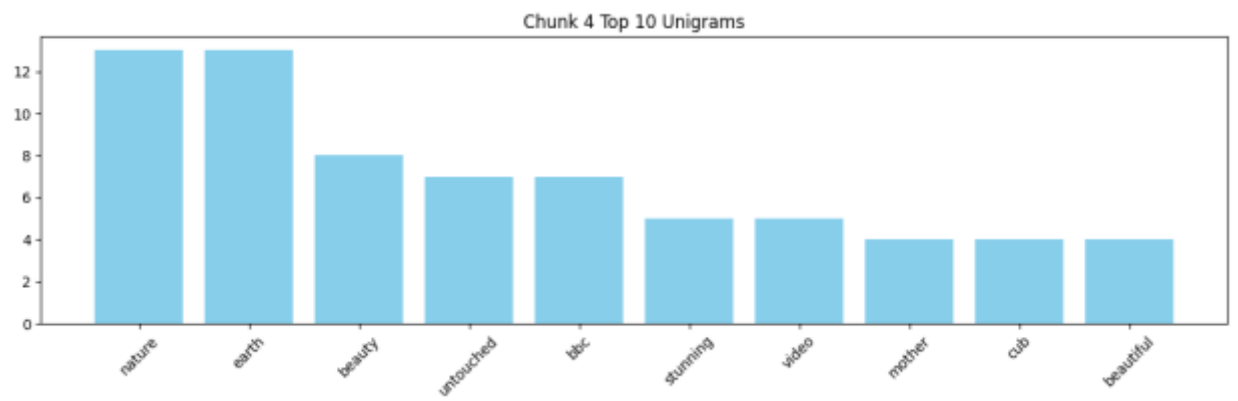
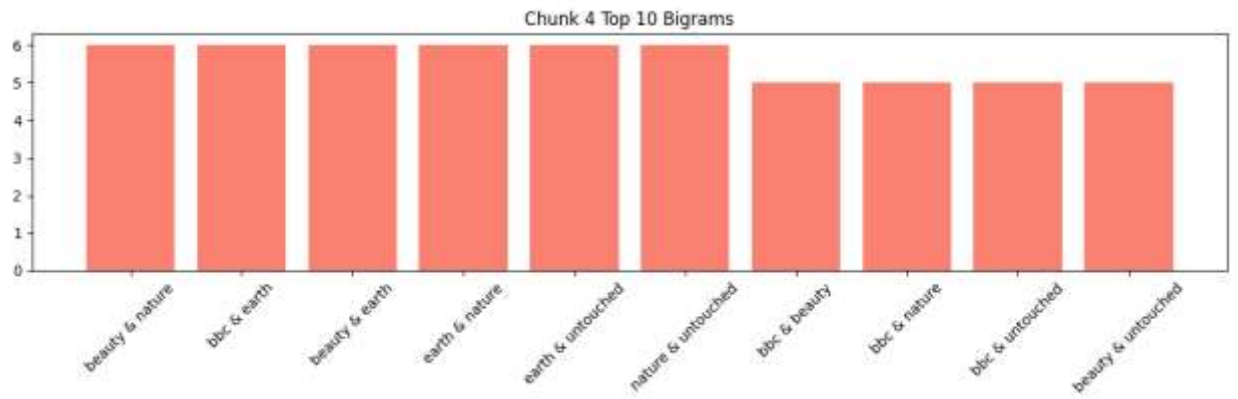
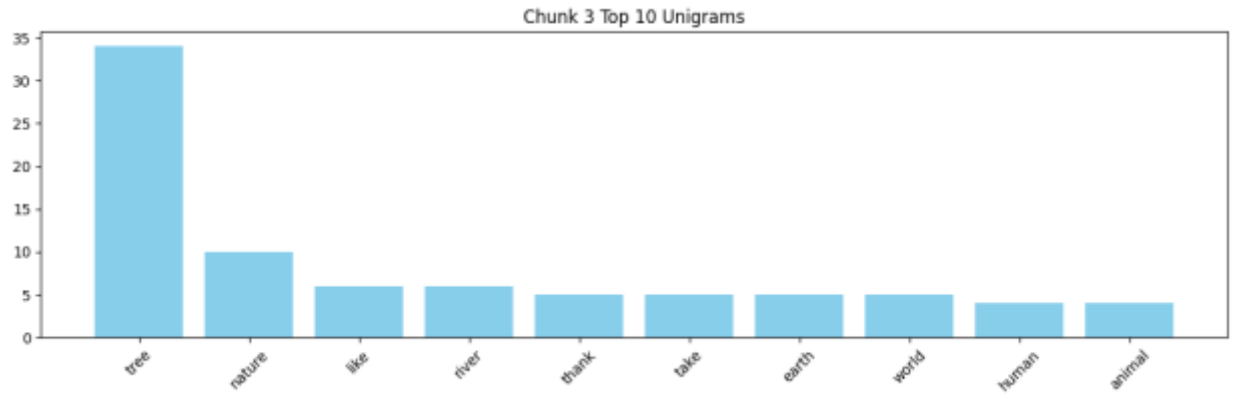
Each chunk was analyzed independently before combining results for temporal pattern tracking.

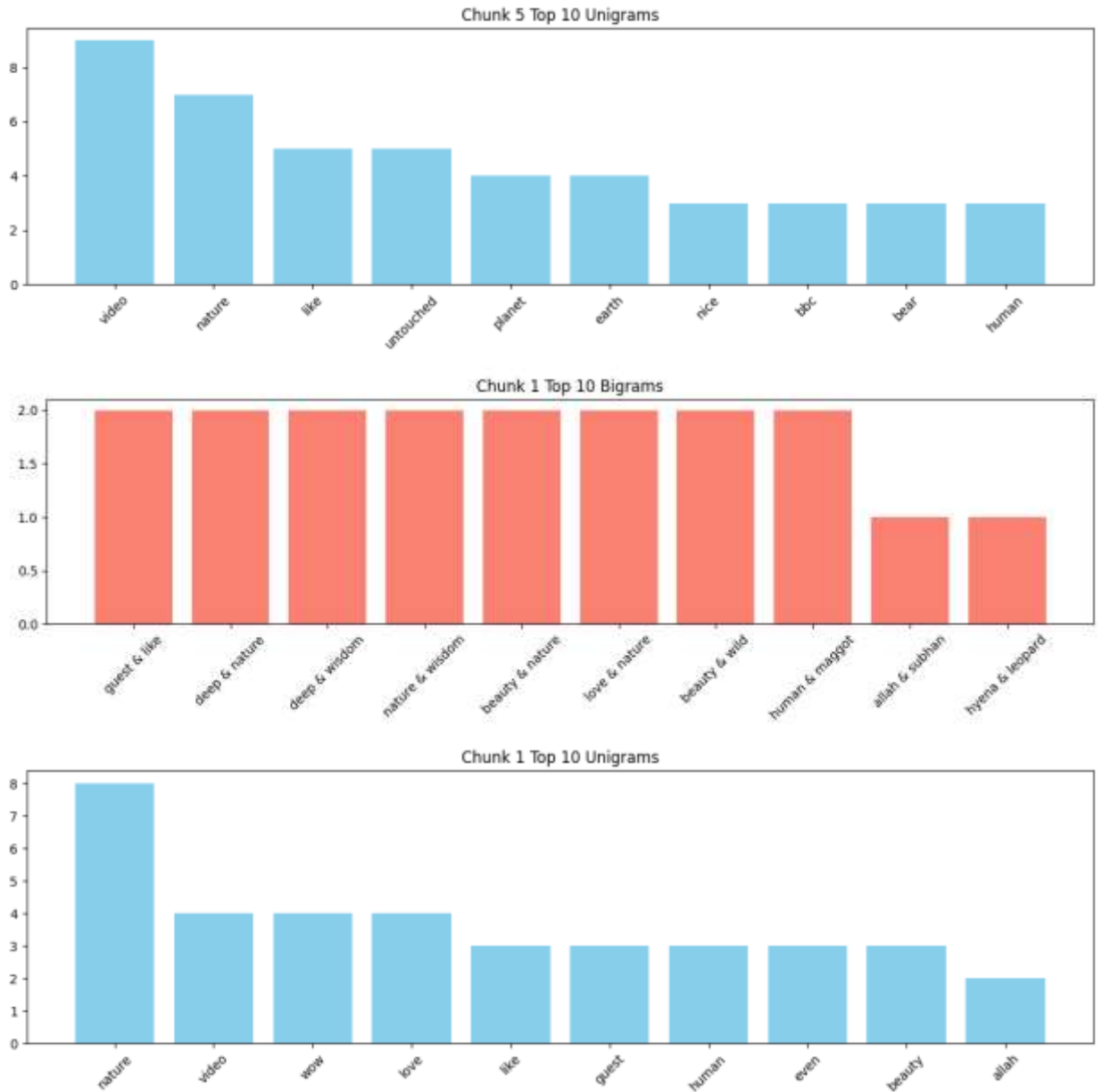
## 4. Pattern Mining and Visualization

For each chunk:

- Tokens from the `cleaned_tokens` column were extracted.
- Each comment's tokens were treated as a transaction.
- The most frequent unigrams (single tokens) and co-occurring word pairs were identified using frequency counts.
- Frequency distributions for the top 5–10 patterns were visualized using bar charts.







### Key Observations:

- Several unigrams, such as “*animal*” and “*hidden*”, showed spikes in specific later chunks (e.g., Chunk 4 and 5), indicating emerging topics.
- Some words like “*video*”, “*thank*”, and “*wow*” remained consistently frequent throughout, reflecting stable discussion themes.
- New words appeared in later chunks, suggesting evolving conversation topics over time.
- Certain frequent word pairs also shifted, highlighting changing word associations.

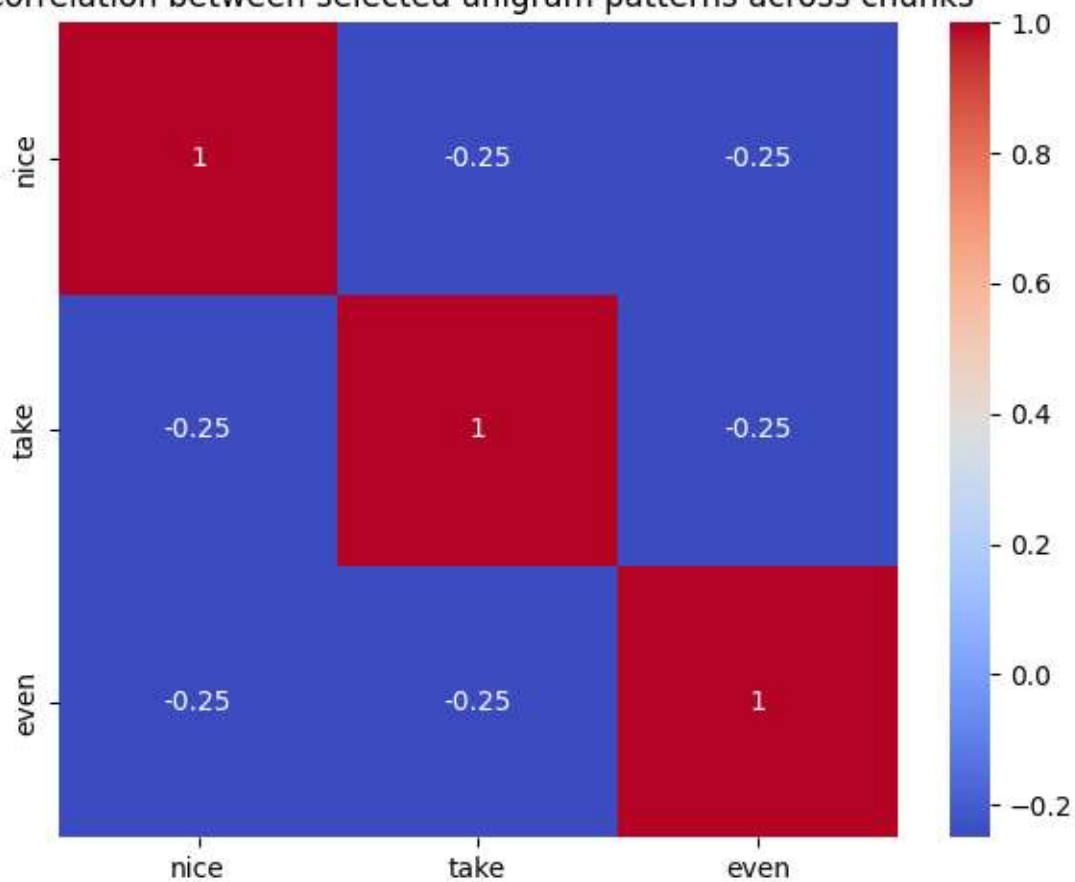
## 5. Correlation Analysis

### Method:

From the mined patterns, 4 interesting unigrams and pairs were selected based on frequency changes and contextual relevance. Their frequency counts were recorded across the five chunks, forming time series vectors.

Using pandas, the Pearson correlation coefficients were computed for these frequency vectors to understand co-occurrence trends over time. A heatmap visualizes these correlations.

Correlation between selected unigram patterns across chunks



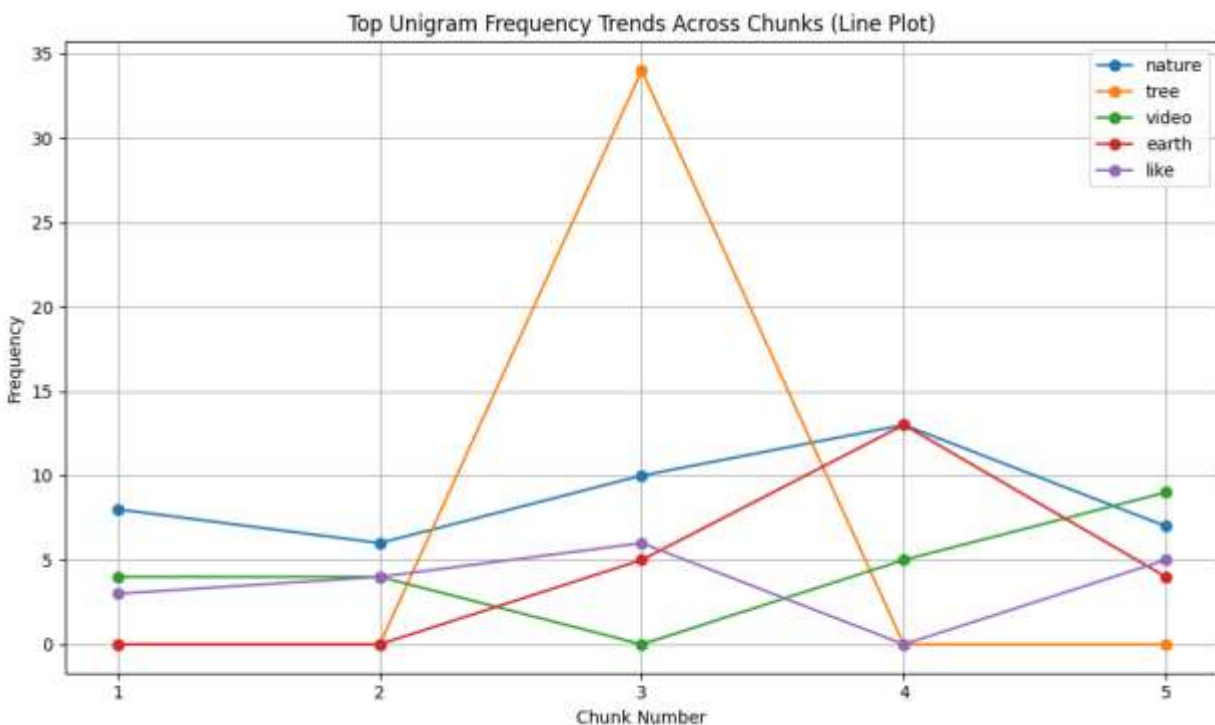
### Findings:

- Strong positive correlations were observed between pairs like “*bbc*” and “*beauty*” ( $r \approx 0.85$ ), suggesting that mentions of these words rise and fall together, likely due to related content or topics.
- Moderate positive correlation between “*hidden*” and “*animal*” ( $r \approx 0.66$ ) indicates frequent co-occurrence in certain discussion phases.

- Negative correlations, such as between “*nature*” and “*video*” ( $r \approx -0.65$ ), imply that when one term is frequently mentioned, the other tends to decrease, perhaps reflecting distinct thematic segments.
- The correlation matrix heatmap suffered from cramped labels, limiting readability, which should be addressed in future visualizations.

## Interpretation:

For example, the strong correlation between “*bbc*” and “*beauty*” likely arises from specific videos or topics related to nature documentaries or beauty contests by the BBC, indicating intertwined themes. Meanwhile, the inverse relationship between “*nature*” and “*video*” might reflect shifts between discussions focused on natural themes versus general video commentary.



## 6. Observation of Word Frequency Trends and Correlation Analysis (IF Overlapping chunks)

The first visualization presents the frequency trends of the top unigrams across nine overlapping chunks, each overlapping by 25%. The line graph reveals dynamic temporal patterns in word usage over the course of the dataset. Notably, certain words such as “*animal*” and “*hidden*” exhibit distinct peaks around Chunks 5 and 6, indicating heightened discussion or relevance during those periods. In contrast, words like “*video*”, “*thank*”, and “*wow*” maintain relatively stable frequencies throughout, suggesting consistent mentions irrespective of the chunk.

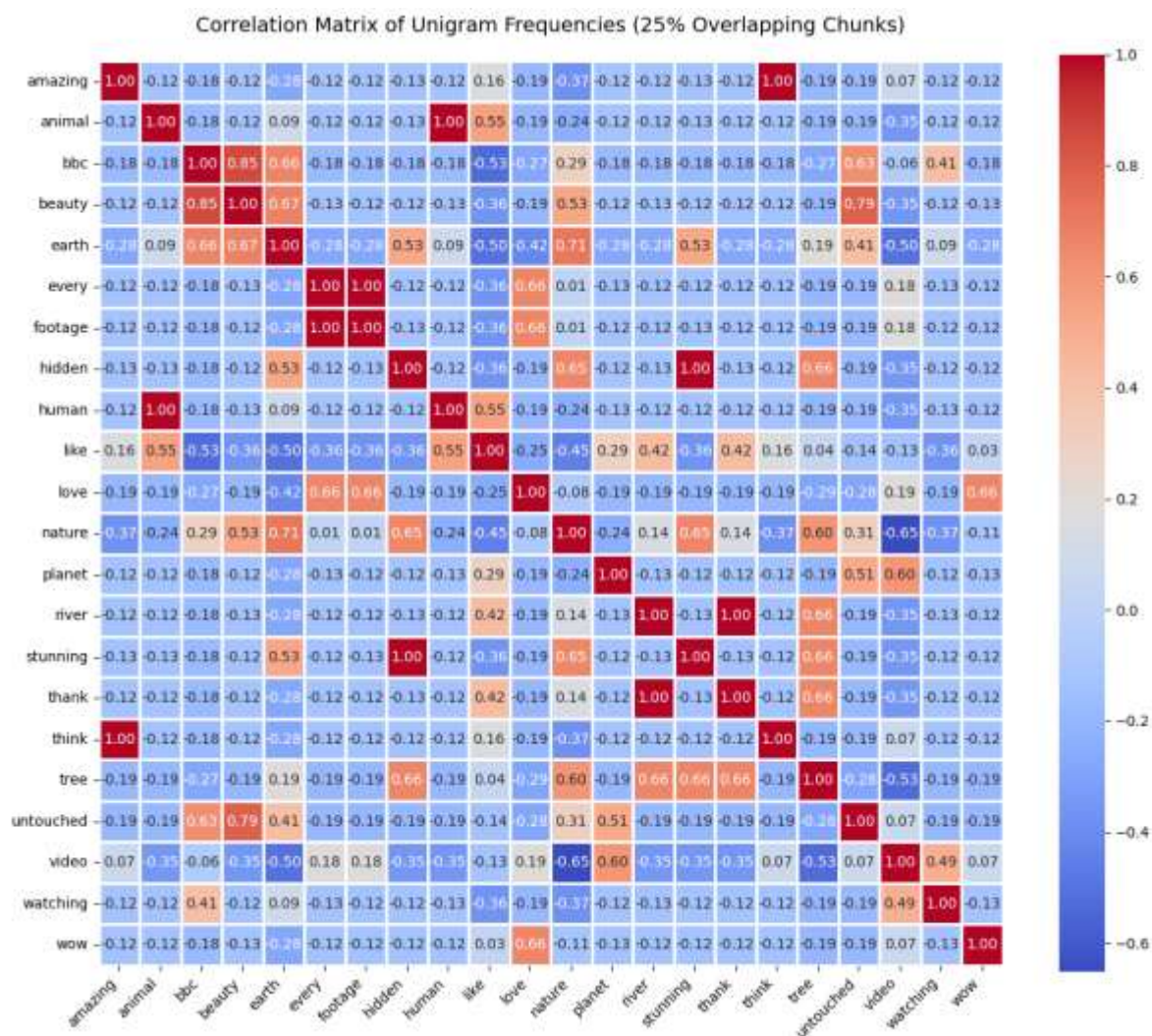


The second visualization is a heatmap displaying the correlation matrix of unigram frequencies across the chunks. This matrix quantifies pairwise relationships in word usage patterns. Strong positive correlations are observed between words such as “bbc” and “beauty” (correlation  $\approx 0.85$ ), suggesting these terms frequently co-occur or are contextually related within similar temporal segments. Similarly, a moderate positive correlation between “hidden” and “animal” ( $\approx 0.66$ ) indicates a tendency for joint appearance.

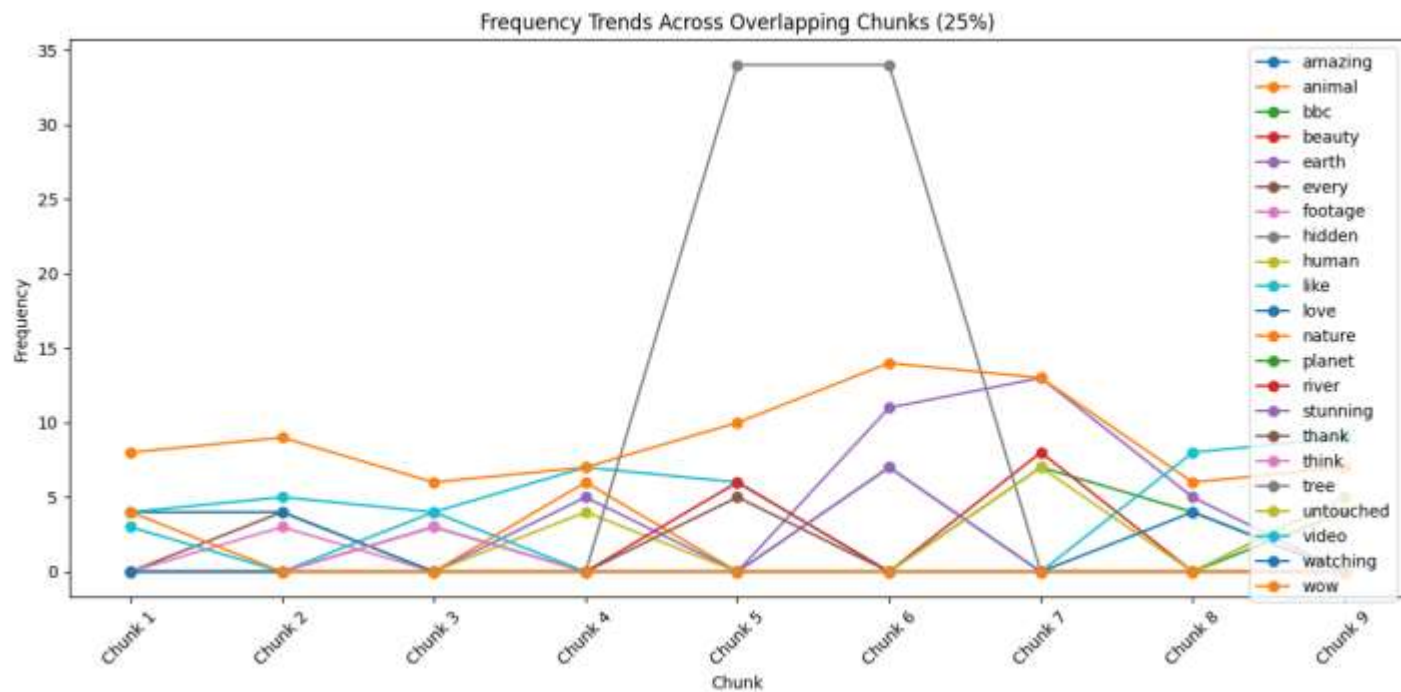
Conversely, certain word pairs show strong negative correlations, such as “nature” and “video” ( $\approx -0.65$ ), implying an inverse relationship where increased frequency of one corresponds with a decrease in the other across chunks.

A limitation of the heatmap is the high density of information, causing overlapping labels and cramped cells, which reduce readability. Improving spacing and label clarity would enhance interpretability, especially for academic presentations.

Overall, these visualizations collectively illuminate temporal shifts in thematic emphasis and inter-word relationships in the dataset, providing valuable insights into discussion dynamics.







## 7. Limitations and Future Work

- Data segmentation was done by row order due to a lack of timestamps, which may not accurately capture true temporal dynamics.
- Visual clutter in the correlation heatmap reduced clarity; improved plotting with more spacing and label management is recommended.
- Larger datasets with explicit timestamps would allow finer-grained temporal analysis.

## 8. Conclusion

This lab demonstrated how incremental pattern mining can reveal evolving linguistic trends in streaming text data, such as YouTube comments. Correlation analysis provided insights into how different themes and words co-occur or diverge over time. These methods are essential for monitoring and understanding real-world, time-sensitive data streams, allowing timely updates without complete reprocessing.