Lab Report: **04**

Subject: CSE - 477

Section: 1

Submitted By:

Syed Musayedul Hussain

ID: 2021-2-60-015

Submitted to:

Amit Mandal
Lecturer

Department of Computer Science and Engineering ,East West University

This report presents an analysis of a cleaned comments dataset to identify and interpret frequent linguistic patterns across different segments of the data. The dataset cleaned_comments.csv was loaded, and the cleaned_tokens column was converted from string representations to Python lists. Data was split into 5 equal chunks for temporal or segment-wise analysis.

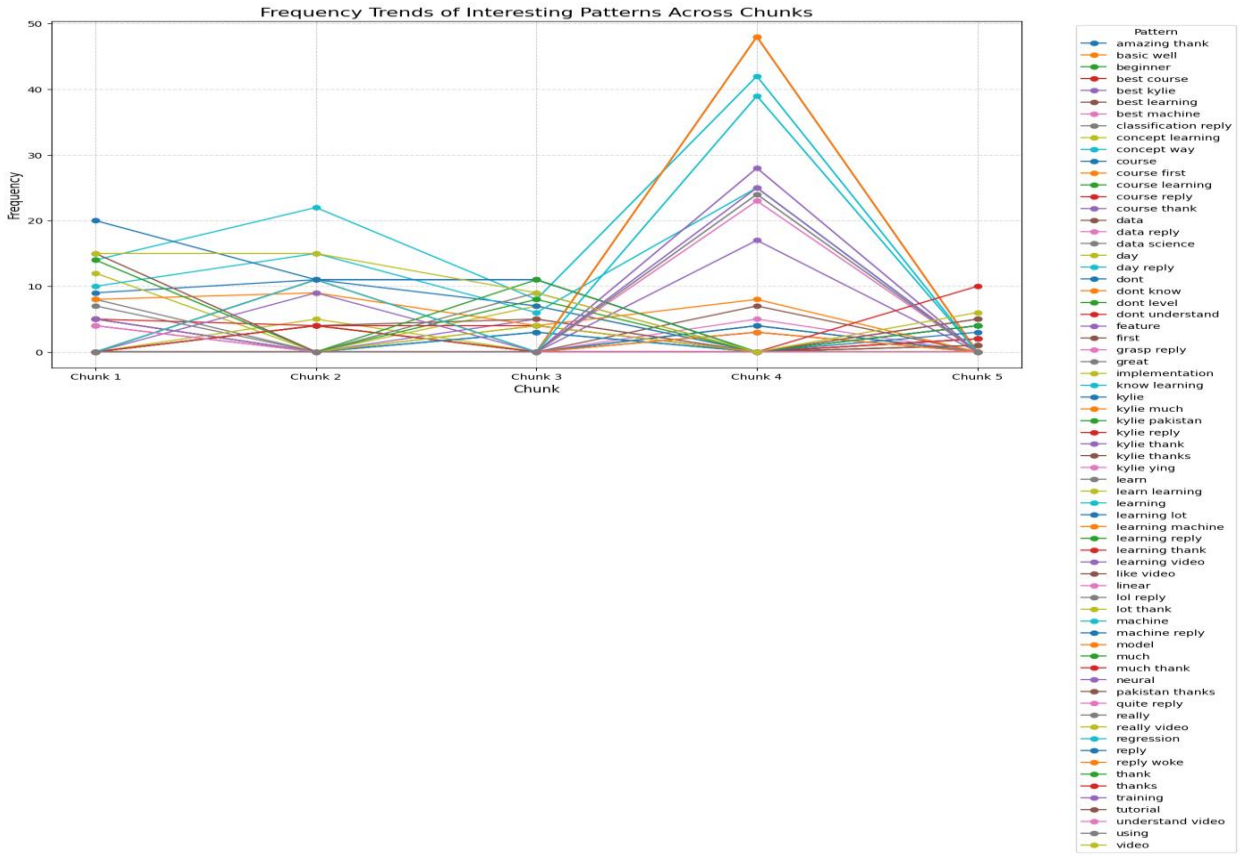
1. Unigram and Bigram Analysis

A. Unigrams (single words) and bigrams (unordered pairs of words) were counted for each chunk.
B. Top recurring unigrams included: learning, videomachine, course, thank, data, model, and regression.
C. Frequent bigrams often paired learning-related terms: (learning, machine), (like, video), (course, thank), (data, science), etc.

Notable Observations:

i. **Chunk 1 & 2:** Learning-related engagement (learning, machine, video) dominated.
ii. **Chunk 4:** Strong technical focus — high frequencies of data, model, regression, neural, and training.
iii. **Chunk 5:** More basic and beginner-oriented terms appeared (beginner, first, tutorial).

2. Pattern Frequency Trends

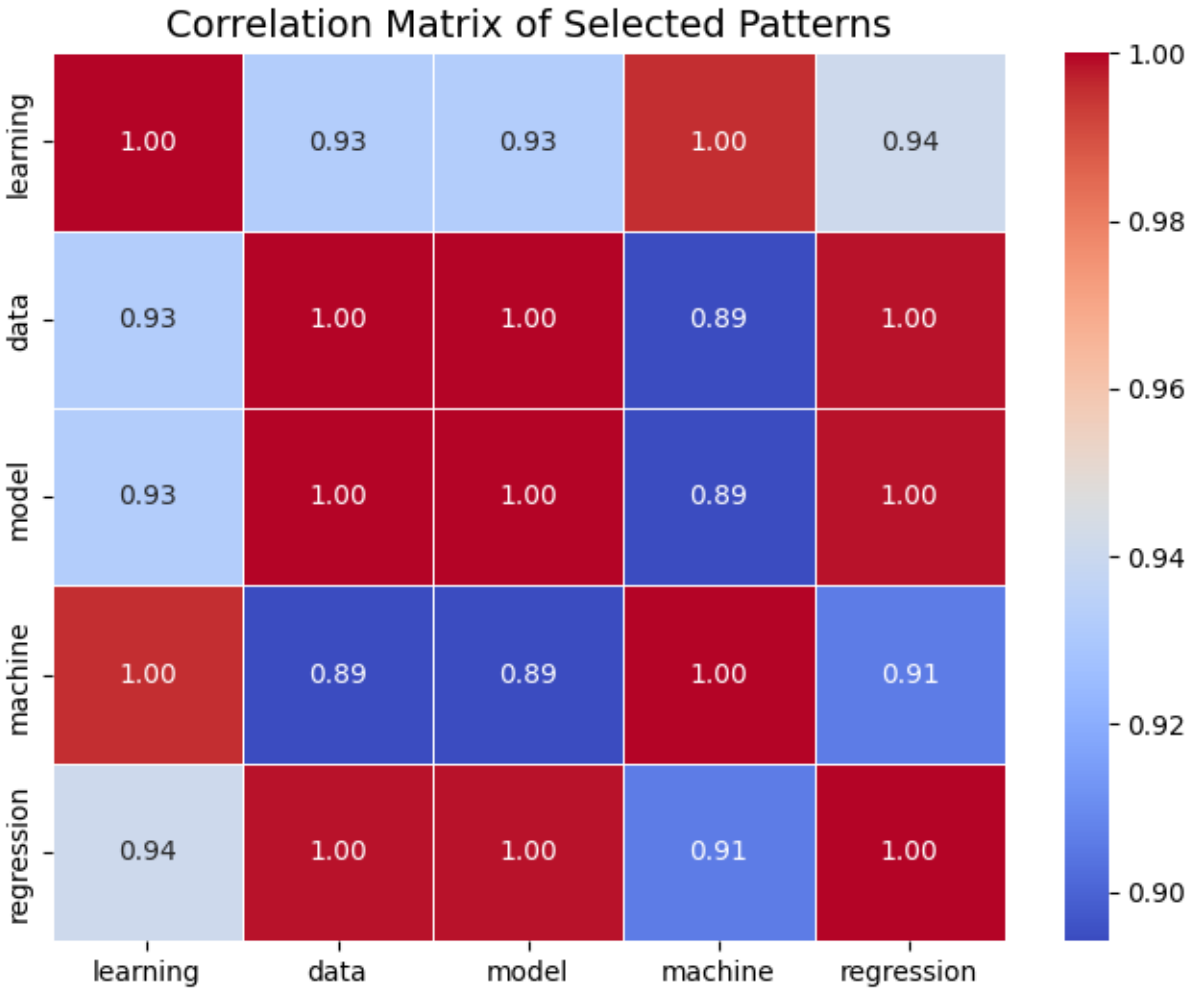


A. A table was created with counts for all top unigrams and bigrams across chunks.
B. Certain patterns showed significant variation across chunks — notably: learning, data, model, machine, regression.
C. These patterns spiked sharply in Chunk 4, suggesting a period of heavy technical discussion or content.

3. Interesting Pattern Selection

Selected Patterns for Correlation Analysis:

	learning	data	model	machine	regression
Chunk 1	14	0	0	10	0
Chunk 2	22	11	11	15	11
Chunk 3	8	0	0	6	0
Chunk 4	42	48	48	25	39
Chunk 5	0	0	0	0	0



A. Criteria: present in some but not all chunks, frequency variation, and at least one significant jump.
B. Selected patterns: learning, data, model, machine, regression.
C. Correlation among these terms was likely strong, reflecting a thematic shift toward data science topics in specific segments.

4. Hypothesis

A. The dataset might reflect evolving audience interest or curriculum progression:
    a. Early: learner motivation, feedback, and general terms.
    b. Middle to late: deeper technical topics as complexity of material increased.
B. The spike in technical terms in Chunk 4 may align with lessons/videos on model training and regression.

From a decision-making perspective, these findings can guide targeted content strategies:

i. For early audiences**,** content that is motivational and accessible encourages engagement.
ii. For advanced learners**,** technical depth sustains interest and promotes topic mastery.
iii. Monitoring such vocabulary trends can help content creators align material with the audience's evolving skill level and maintain balanced engagement.

The strong correlation among terms like *"learning"*, *"machine"*, *"data"*, and *"regression"* in technical-heavy segments further validates the thematic clustering of advanced topics. This correlation can be used to identify high-value knowledge areas for deeper exploration.