Project Assignment:
Motif Discovery & Repeated Substring Mining in Real Data

Objective

In this project, you will analyze a real sequence dataset and extract:

- **Motifs**: short patterns occurring many times.
- **Highly repeated substrings**: long substrings that repeat frequently or maximally.

You may choose any real dataset (DNA, RNA, proteins, EEG sequences after symbolization, temperature logs, event sequences, etc.).

Task 1: Motif Extraction

Find motifs of lengths k (for different values of k) that occur frequently.

- Report: motif, its frequency, and all occurrence positions.
- Consider overlapping occurrences.

Task 2: Highly Repeated Substrings

Compute and report the frequency of:

- The longest repeated substring.
- The most frequent repeated substring.
- Top-k repeated substrings of different lengths.
- All maximal repeats of length $\geq L$ (for different values of L).

Consider

- Exact motifs.
- Abelian motifs.
- Small edit distance motifs.
- Cartesian tree motifs.
- Order preserving motifs.
- Parametrized motifs.
- Motifs with swaps.
- Comparing motifs in two different, but related, datasets.

Deliverables

A. Report

- Dataset description.

- Algorithms used (KMP / Trie / AC / Suffix structures / ...).
- Runtime & memory complexity.
- Results: motifs and repeated substrings.
- Tables, figures, and interpretation.

B. Code A fully working implementation with clear instructions.

C. Reproducibility Provide the dataset and outputs.

Grading is based on the following:

Algorithmic correctness
Efficiency & complexity analysis
Quality of results
Presentation & clarity of report
Code quality

Good luck, and enjoy exploring real data with string algorithms!