Finite-Context Models

Text Prediction and Generation

107637 André Oliveira
107849 Alexandre Cotorobai
124467 Francisco Ferreira

Table of contents

Introduction

This project consists on the development of two main components:

- **fcm:** Measures the information content of text provided using a learned finite-context model;
- **generator:** Text generator that relies on an already created model or trains one with a given input.

01

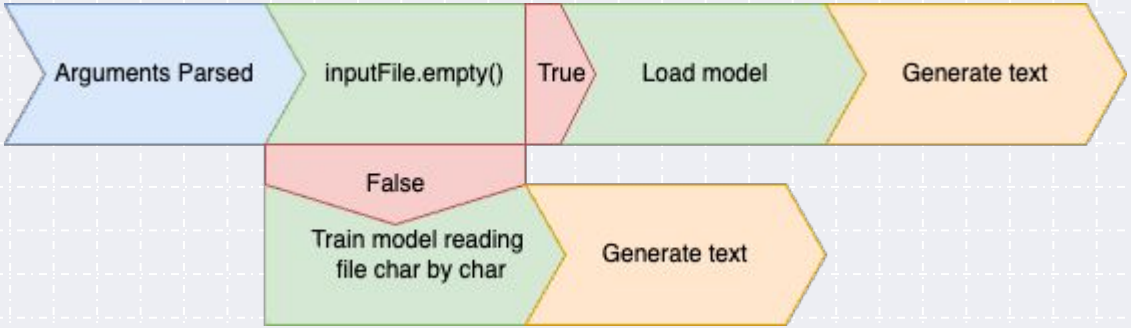
Implementation

...

Implementation

➔ FCM

➔ Generator

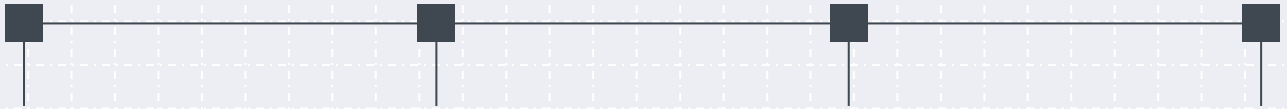


02

Methodology

...

Our Approach to the problem



Minimize Bits per Second

Lead to unintended Losses

Average Word Length Context

Good for textual data, but not effective on DNA

Double Model Approach

Might lead to a loop in the lower order model

Levenshtein Distance

Allowed to generate previously unseen words

03

Results

...

BPS Comparison Experiments

In our experiments, we executed the finite-context model (FCM) twice:

- ➔ Original Text Analysis
- ➔ Generated Text Analysis

Sequences that we will present:

- ➔ Sequence1 (DNA)
- ➔ Sequence2 (Portuguese Literary Text)

Sequence 1: DNA Data

Higher context order (k)

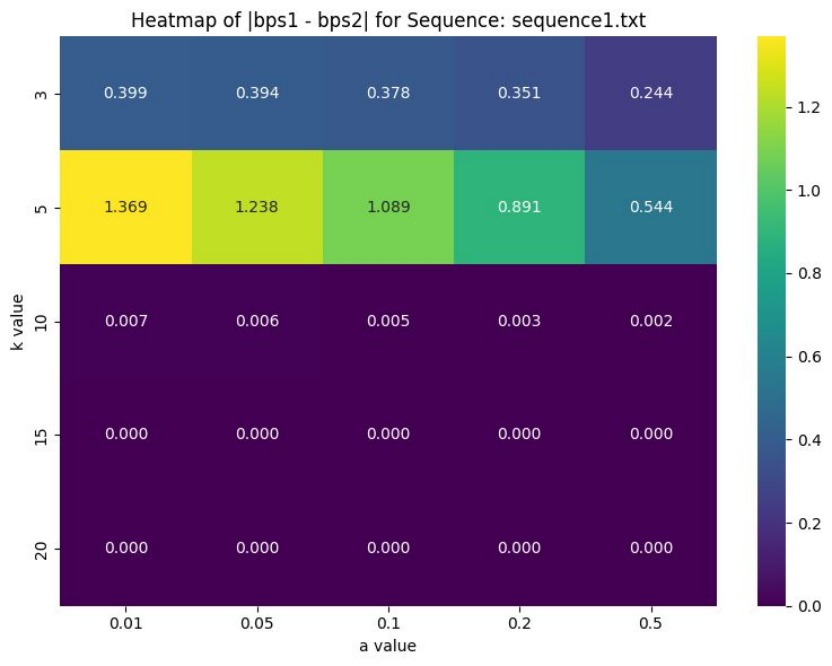
➔ Decrease in BPS difference

➔ Approach zero (k >= 10)

Smoothing Factor (α)

➔ Higher impact on lower k

➔ Lower impact on higher k



Sequence 2: Literary Text

Context order (k)

➔ Depending on the smoothing factor also increases for lower k values

➔ Increases (k >= 10)

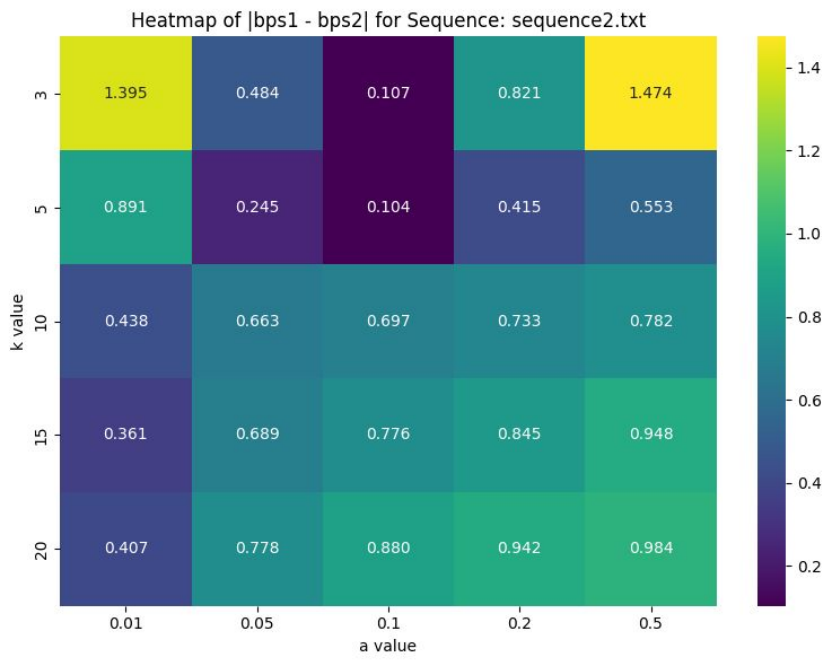
Smoothing Factor (α)

➔ For most cases, lower BPS difference for lower values

Optimal Values

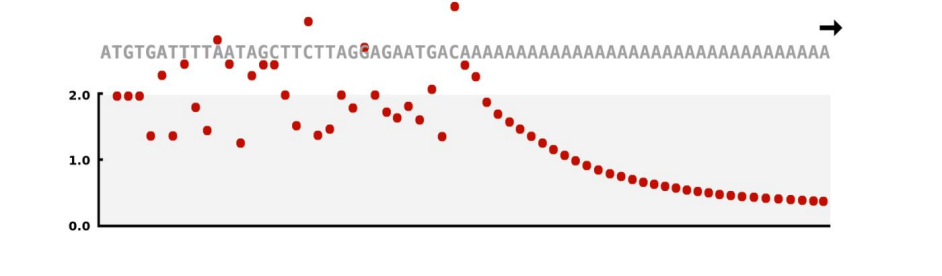
➔ (5 ≤ k < 10)

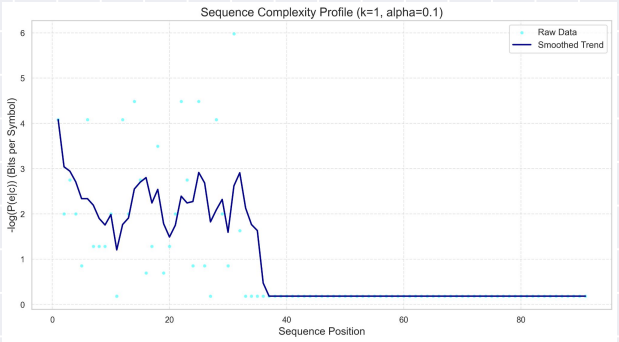
➔ 0.05 ≤ α ≤ 0.1

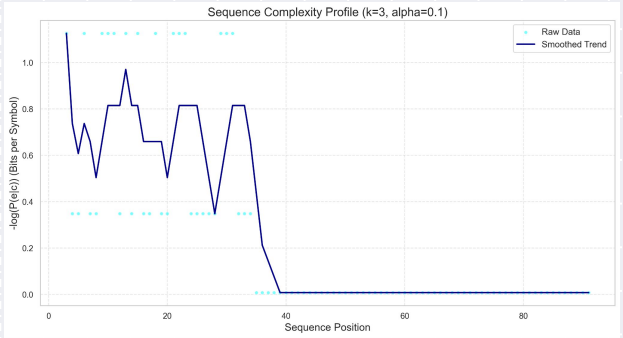


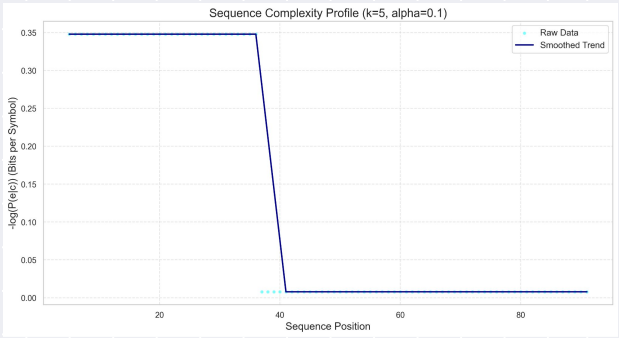
Complexity Profiling

Complexity Profiling validated with the class example



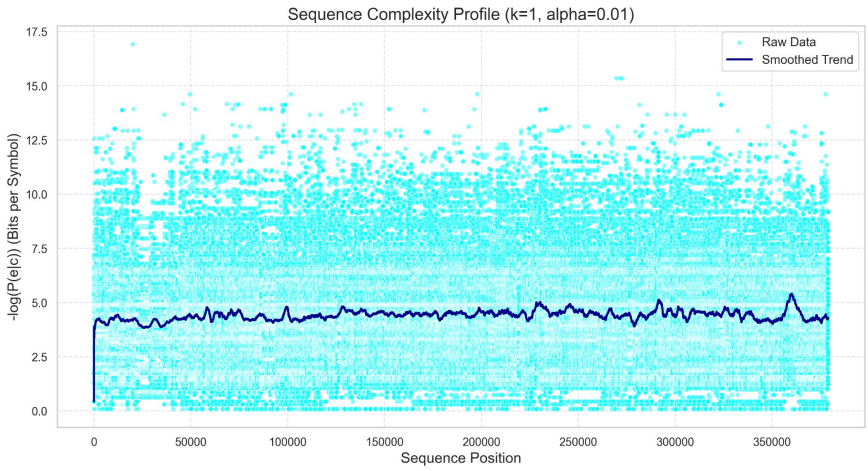




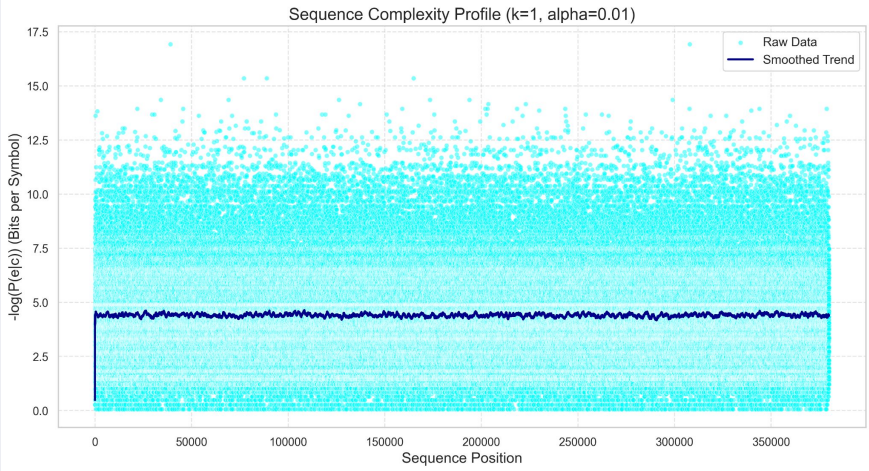


Complexity Profiling

Complexity Profile of sequence 5, k=1

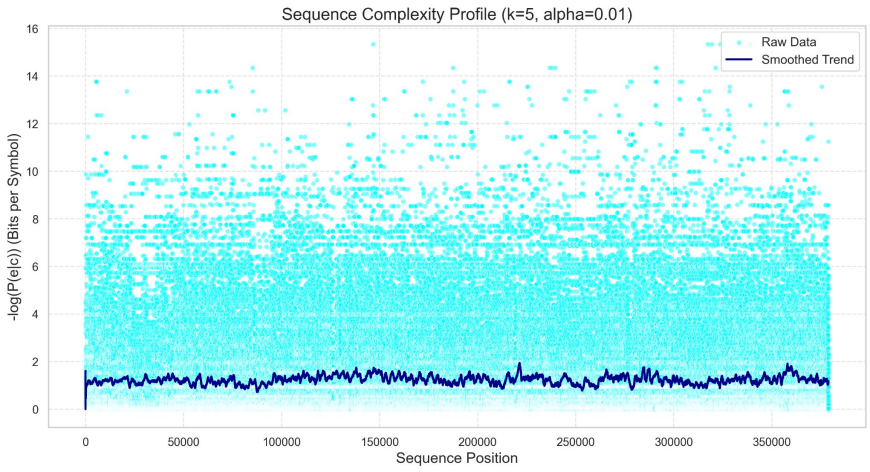


Complexity Profile of generated text from sequence 5, k=1

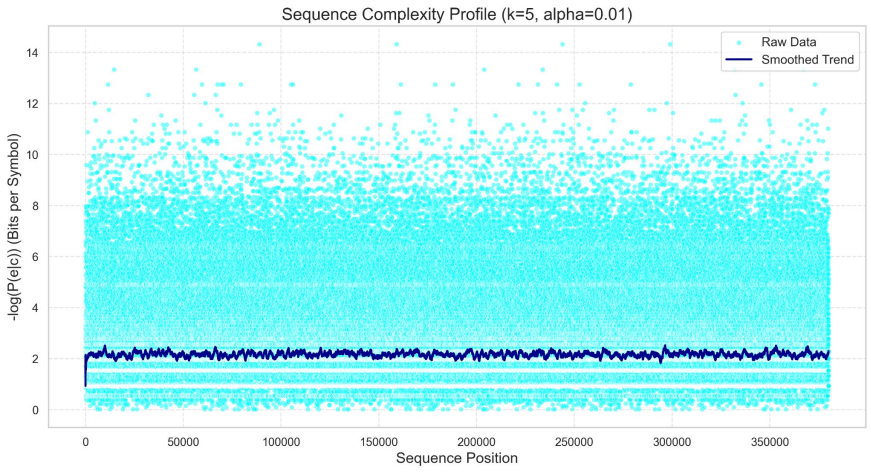


Complexity Profiling

Complexity Profile of sequence 5, k=5



Complexity Profile of generated text from sequence 5, k=5



Entropy Interpretation and Predictability

	Shannon Entropy	Conditional Entropy	Redundancy
K = 1	1.9652	1.9198	0.0231
K = 15	1.9652	0.1356	0.9310

Comparison with Zstandard Compression

Sequence	Size (bits)	Symbols	ZStandard BPS	FCM BPS
Sequence 1	28072	10126	2.7723	2.04688
Sequence 2	1065056	318185	3.3473	2.73718
Sequence 3	13837896	3295751	4.1987	4.0508
Sequence 4	48725896	22668225	2.1495	1.90792
Sequence 5	651056	378930	1.7181	1.85937

Conclusion

- The choice of context length (*k*) and smoothing factor (*α*) directly impacts predictive accuracy and computational efficiency;
- In highly structured sequences, such as DNA, long contexts ($k \geq 10$) significantly reduce predictive uncertainty;
- In literary texts, moderate values of *k* (between 5 and 10) and *α* (0.05 to 0.1) yield better performance, avoiding overgeneralization;
- Small variations in *Bits per Symbol* (BPS) between original and generated texts indicate good model adaptation to the original text patterns;

Thanks!

Do you have any questions?