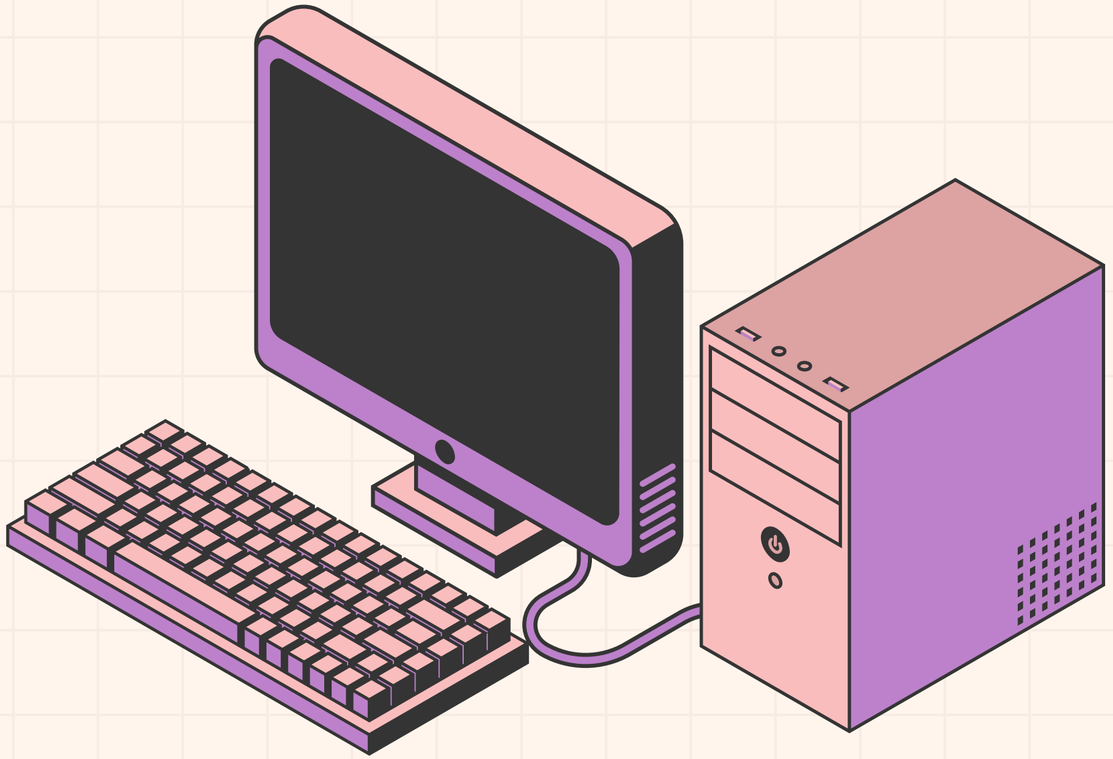
INFORMATION MODELS FOR PREDICTION

Guilherme Amorim 107162
José Gameiro 108840
Tomás Victal 109018

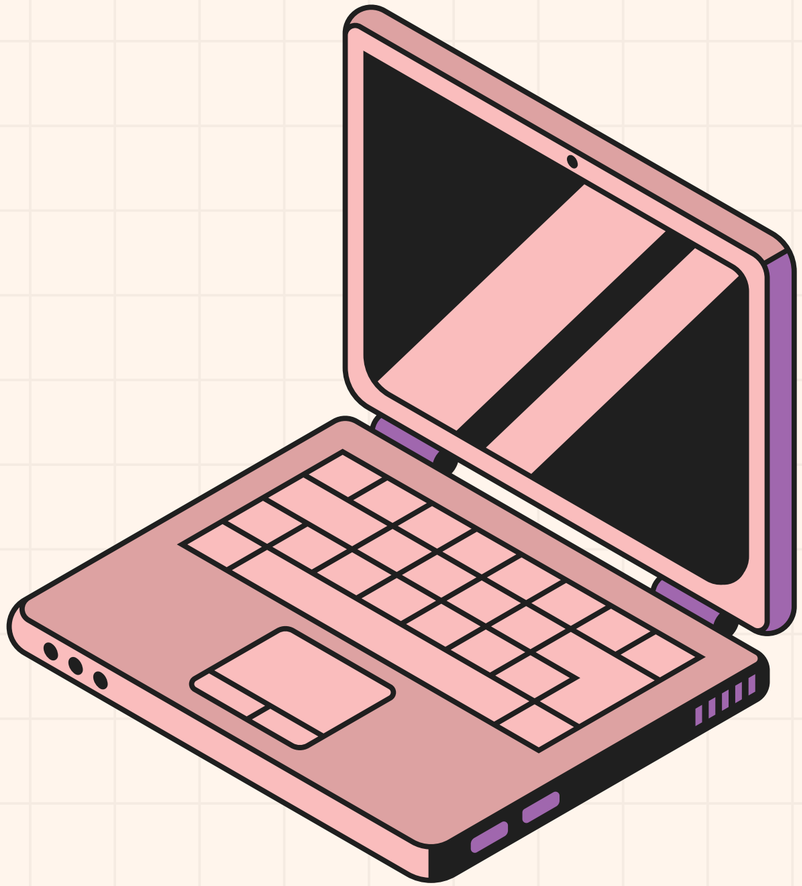


INTRODUCTION

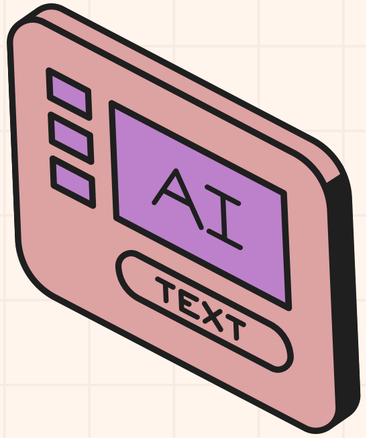
This project consists on the development of two main components:

- **fcm:** a program that measures the information content of text provided using a learned finite-context model

- **generator:** a text generator that creates text following depending on a model created;



OUR IMPLEMENTATION



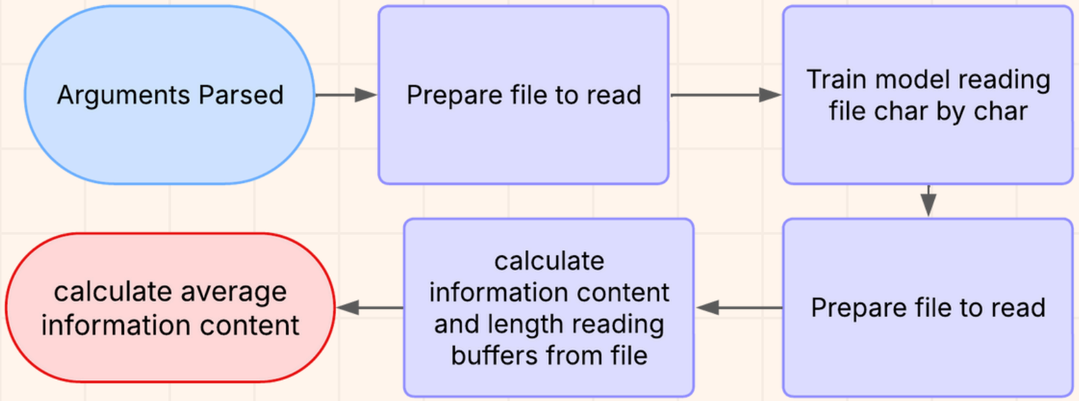
File Reader
- Open File() - Read Char() - Read Buff() - Read Word()

Finite Context Model
- new() - train_char() - compute_probability() - calculate_information_content() - sample_next_char() - get_k()

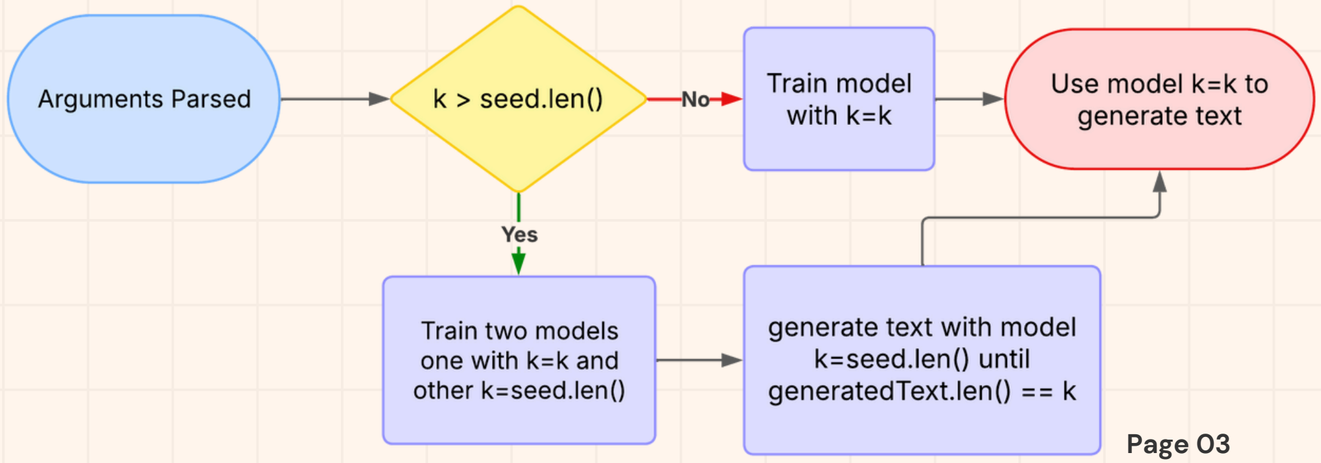
Text Generator
- generate_text() - generate_text_words()

Chart Generator
- new() - compute_probibility() - train_chart() - draw_chart()

FINITE CONTEXT MODEL



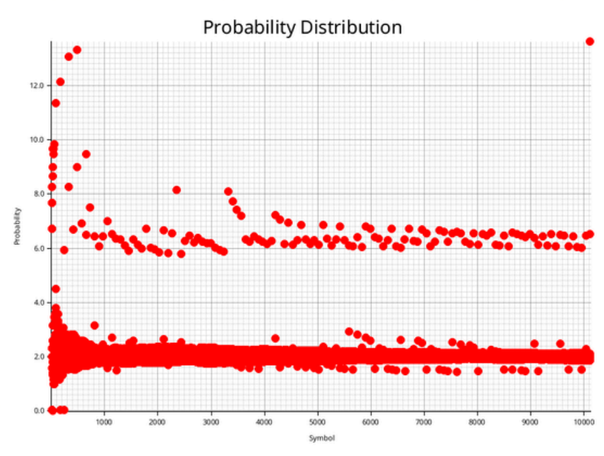
GENERATOR



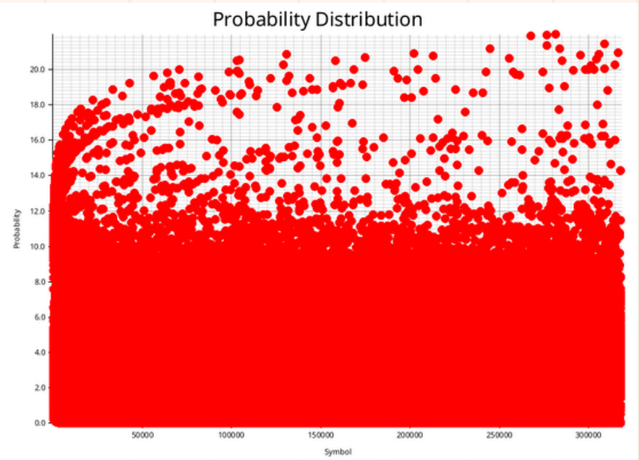
EXPERIMENTS

MULTIPLE SEQUENCES

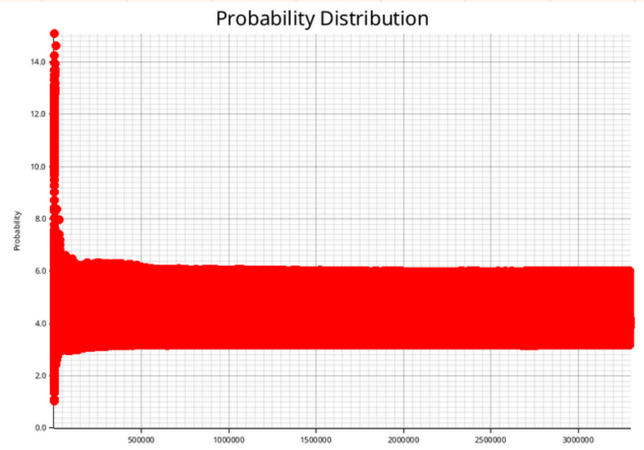
SEQUENCE 1 – 2.0437



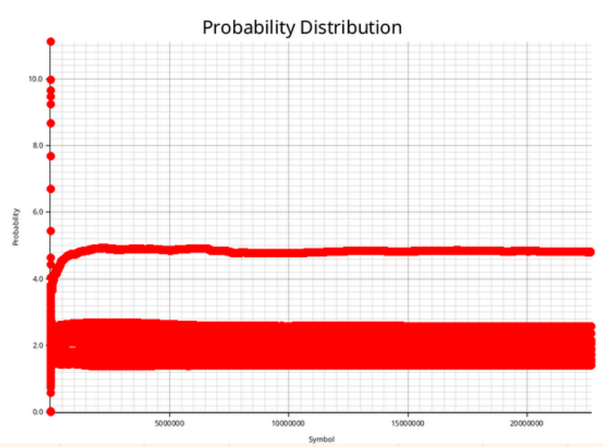
SEQUENCE 2 – 2.2099



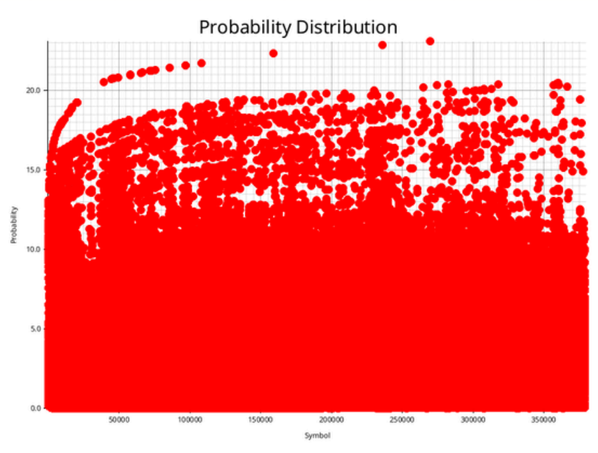
SEQUENCE 3 – 3.9966



SEQUENCE 4 – 1.8793

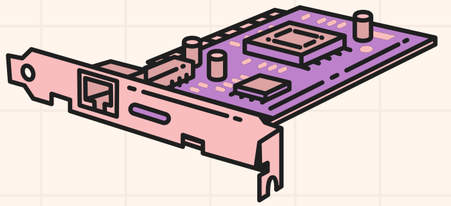


SEQUENCE 5 – 1.3215



EXPERIMENTS

CHANGING THE **K** AND **ALPHA** VALUES



Alpha = 0.01

K	Average Information Content
2	2.0528
4	2.0409
6	1,9257
8	1.7540
10	1.7267

k = 3

Alpha	Average Information Content
0.0001	2.0108
0.001	2.0143
0.01	2.0473
0.1	2.2908
1	3.4798
10	5.8183

SEQUENCE 1

TEXT GENERATED

CHARACTERS

Qualescerguantos e gerrando inadade-
parendos toda pro tam,
Que peça,
Não Eôo porfeinhega,
No pelo aquecinas reito do sena frio de.

K = 2

WORDS

armas Nas brandamente,
conduzidos adereça.
arremessa, apelida
resistirei Tomai consolar-te!
Oceano;
deixamos

K = 5

armas se no bélica elegantes,
Por que saíam,
Onde a pintura pôr a armas, porque a
tua frio,
A prata um remédio cerca areno,
Para ajudasse
De áspero, sábio e forte Artabro, e as-
pereza;

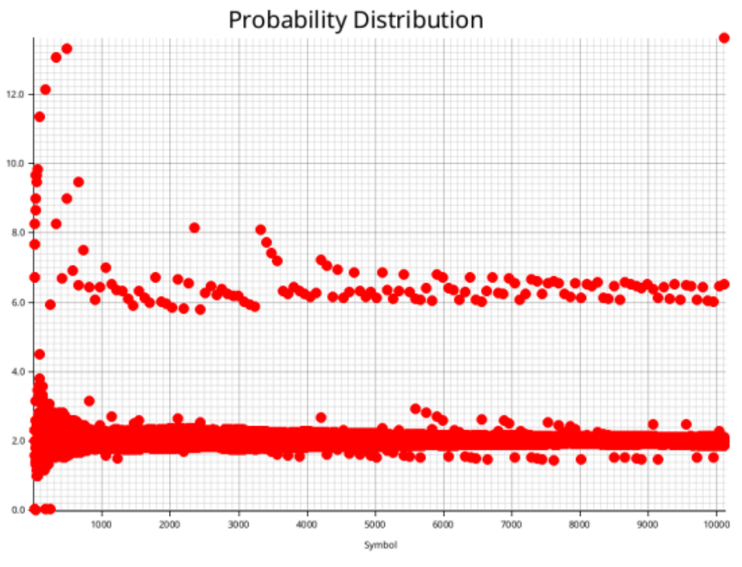
CHARACTERS

armas em hospitais, nos Baco na
Massília touro,
esquecerão Abrantes,
troque Umas, enseada
26
sagaz "Porém Canace, penhor Gueos
causaram adornado,

WORDS

EXPERIMENTS

TRAIN A MODEL WITH TEXT GENERATED



K = 4 ALPHA = 0.1

A.I.C. = 2.7458

ORIGINAL SEQUENCE 1

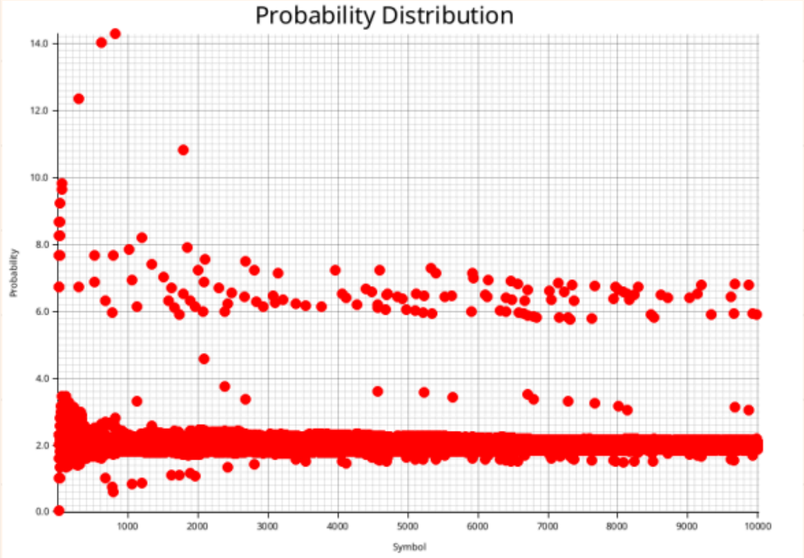
1ST SEQUENCE GENERATED

K = 4 ALPHA = 0.1

PRIOR = ACTG

CHARACTERS = 10000

A.I.C. = 2,6582



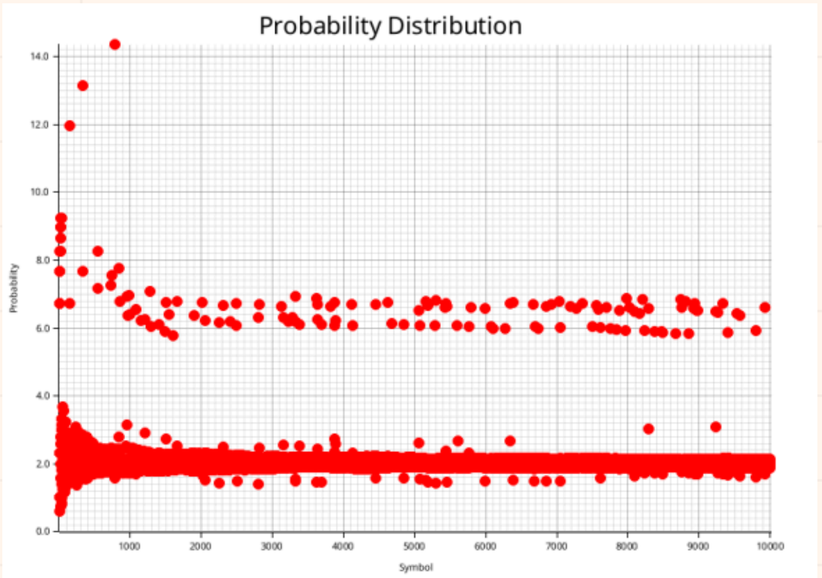
2ST SEQUENCE GENERATED

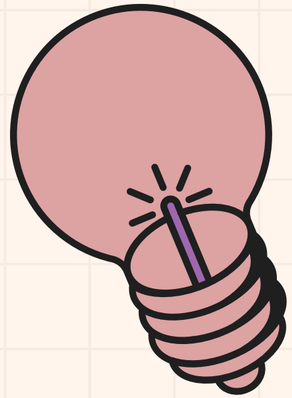
K = 4 ALPHA = 0.1

PRIOR = ACTG

CHARACTERS = 10000

A.I.C. = 2.5703

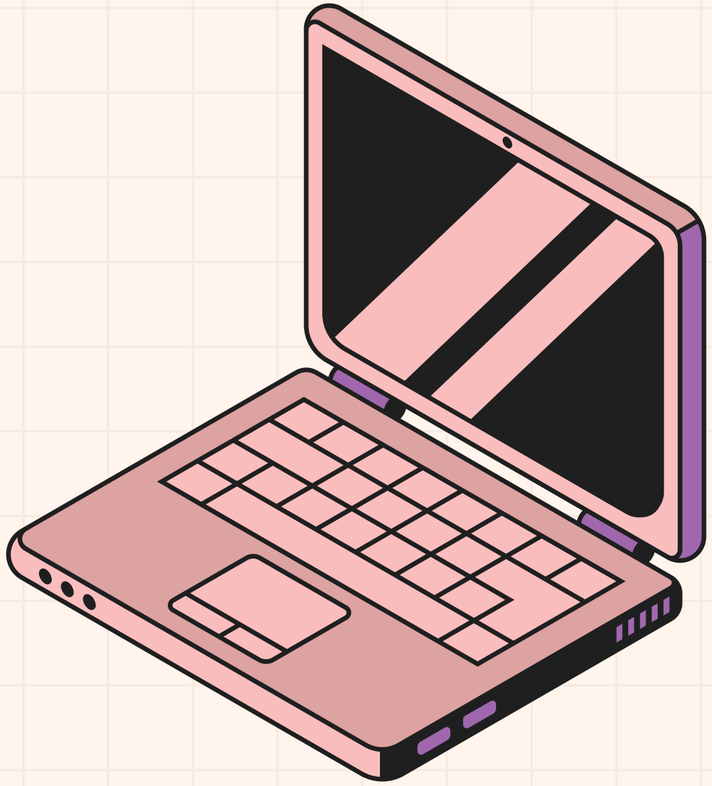


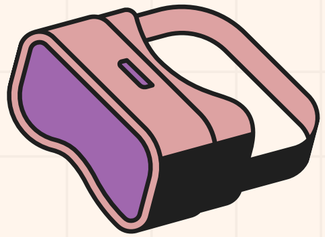


CONCLUSIONS

The experiments made were important to get some conclusions about the programs developed:

- Increasing the **context size (k)**, makes predictions more certain.
- Higher values of the **smoothing parameter (alpha)** leads to greater entropy.
- **Word-based models** can produce more coherent text when applied to natural language sequences.
- **Training a model on generated sequences** results in an increase in predictability over multiple iterations.





THANK YOU