Assignment 1 - Copy models for data compression

Algorithmic Information Theory (2022/23)
Universidade de Aveiro

Martinho Tavares, 98262, martinho.tavares@ua.pt
Nuno Cunha, -, - Pedro Lima, -, -

March 28, 2023

1 Introduction

2 Work organization

3 Copy model

...

In order to evaluate whether the copy model can provide acceptable results, we can use a baseline below which we expect the model to report the file's entropy. We decided to use, as a baseline, the entropy considering each symbol's relative frequency in the entire file, which is given by:

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

With this value in mind, we evaluated the model as a whole with different values for its parameters, on different files. The files chosen for testing are present in the repository[1], and they have the following baselines:

- `chry.txt`: ...
- ...: ...

...

Throughout this section, the different program parameters are detailed, and their effect on the model's performance is studied.

[1] `https://github.com/NMPC27/TAI-G7-Lab1`

3.1 Pattern size

When choosing a pointer in the past from which to start copying, we need to look for an occurrence of the same k-sized pattern as the one we are currently on.

Thus, k is one of the parameters that affects program performance, where k is a positive integer. On one hand, a lower value of k

3.1.1 Results

3.2 Smoothing parameter alpha

3.3 Base probability distribution

3.4 Copy pointer repositioning strategy

3.5 Copy pointer reposition threshold

4 Text generator

5 Conclusion

6 References

References