**INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA**
**Engenharia Informática e de Computadores**

**Compressão e Codificação de Dados**
Inverno 2020/2021
Trabalho prático II

Este trabalho tem por objetivo explorar a ideia de, na compressão de ficheiros, usar dicionário inicial para aumentar a compressão, para os formatos DEFLATE (RFC 1951), LZ4 [1-2] e Zstandard (zstd) [3-4], variantes do algoritmo LZ77. Considere que o dicionário inicial está limitado a 32 kiB.

Com o gerador do trabalho anterior (exercício 3, usando, pelo menos, aproximação de 2.ª ordem) gerar conjunto de ficheiros.

Usando como dicionário o ficheiro (ou parte deste) utilizado no gerador, verificar o ganho de compressão em relação à não utilização de dicionário.

Analisar a possibilidade de uma sequência gerada a partir do dicionário ser utilizada como dicionário para codificar os ficheiros do conjunto gerado.

Analisar a utilização de dicionário diferente do utilizado para gerar os ficheiros a comprimir.

Analisar a dependência dos resultados anteriores em relação ao ficheiro utilizado no gerador.

Comparar a compressão com dicionário inicial com a compressão da concatenação do conjunto de ficheiros sem utilizar dicionário (verificar se há dependência da ordem de concatenação).

Analisar heurísticas para obter o dicionário a partir dos ficheiros que se pretendem comprimir e avaliar uma heurística com o conjunto inicial de ficheiros e com outro conjunto (Enron mail ou outro).

Usando o conjunto de ficheiros obter dicionário com Zstandard e comparar a compressão com e sem dicionário.

No caso de utilização de implementação sem suporte para dicionário inicial, para estimar o comprimento da sequência *f* comprimida com o dicionário *d* sugere-se a aproximação $|\text{comp}^{(d)}(f)| \cong |\text{comp}(d+f)| - |\text{comp}(d)|$, onde $|.|$ é o comprimento da sequência em *bytes*, comp(.) é o algoritmo de compressão para obtenção da sequência comprimida, $\text{comp}^{(d)}(.)$ é o algoritmo de compressão usando dicionário inicial *d* e *d* + *f* é a concatenação da sequência *f* depois de *d*.

O relatório deve incluir a descrição dos algoritmos usados e das aproximações consideradas e a análise crítica dos resultados obtidos. Deve também incluir informação suficiente para obter de modo independente os resultados apresentados.

[1] Y. Collet, LZ4 Block Format Description, 2019,
https://github.com/lz4/lz4/blob/dev/doc/lz4_Block_format.md, acedido em 19/12/2020.
[2] Y. Collet, LZ4 Frame Format Description, 16/8/2020,
https://github.com/lz4/lz4/blob/dev/doc/lz4_Frame_format.md, acedido em 19/12/2020.
[3] Zstandard - Fast real-time compression algorithm, http://www.zstd.net, acedido em 19/12/2020.
[4] Y. Collet, F. Handte, N. Terrell, 5 ways Facebook improved compression at scale with Zstandard, Facebook Engineering, Dec. 2018, https://engineering.fb.com/core-data/zstandard/, acedido em 19/12/2020.

19 de dezembro de 2020