Teoria da Informação

Trabalho Prático nº 2

CODEC não destrutivo para Texto

Introdução

Período de execução: 5 Semanas

Prazo de Entrega: Sexta-feira, 16 de dezembro 2020, 23h59

Formato de Entrega:
Relatório (em formato de artigo)
Código fonte
Instruções de instalação e execução

Objetivo: Pretende-se que o aluno desenvolva a sua capacidade de resolução de problemas e de integração de conhecimentos na área da Teoria da Informação.

Neste trabalho prático pretende-se que os alunos explorem os conceitos de teoria de informação, em particular no que respeita aos conceitos relativos à teoria da compressão, e proponham uma solução para a compressão eficiente não destrutiva de texto.

A cada grupo será proposto desenvolver uma solução de compressão para contexto de armazenamento de um conjunto de ficheiros de texto.

Importa referir que o foco do trabalho está na proposta da solução, na sua fundamentação e validação, não na implementação propriamente dita. Nesse sentido, os alunos podem usar código existente (por exemplo, publicado na internet) não havendo quaisquer restrições ou penalizações relativas ao uso de componentes existentes, desde que devidamente identificadas. A linguagem de programação poderá ser qualquer. Contudo, têm que implementar uma solução integrada e com o código suficientemente estruturado e documentado por forma a permitir aos docentes executar a solução e confirmar os resultados.

No final do trabalho, os alunos deverão entregar um relatório em formato de artigo IEEE, com um máximo de 6 páginas (*template* disponível em https://www.ieee.org/conferences/publishing/templates.html), contendo os seguintes elementos:
1 – Estado da arte do domínio específico;
2 – Descrição do(s) algoritmo(s) e fundamentação das opções;
3 – Análise e discussão de resultados no dataset original, comparando as diferentes soluções analisadas;
4 – Conclusões e trabalho futuro sugerido.

Dataset:

Os ficheiros a comprimir estão representados na tabela 1. Têm características diferentes, pelo que o melhor *codec* para um ficheiro pode não ser o melhor para os outros. Assim, deverá analisar criticamente a performance da(s) sua(s) solução(ões) com base nas diferentes características dos ficheiros.

Tabela 1 – Lista de ficheiros a codificar e descodificar

Nome	**Descrição**
bible.txt	Texto integral da bíblia (King James version)
jquery-3.6.0.js	Biblioteca jQuery, versão de desenvolvimento 3.6.0
finance.csv	Ficheiro CSV com informação do "annual enterprise survey" da Nova Zelândia (ano 2020)
random.txt	Coleção aleatória de caracteres

Para permitir acompanhar o trabalho e sugerir melhoramentos, a sequência de atividades deverá ser a seguinte:

Semana 1 a 2:

Explore o dataset "original" a comprimir com vista a caracterizar a sua distribuição estatística e potencial de compressão entrópica sem métodos adicionais (rever TP1).

Elabore um pequeno estado da arte (máximo 3 páginas no formato IEEE) de codecs lossless de texto, identificando os principais passos e códigos usados.

Esse estado da arte deverá versar pelo menos os seguintes aspectos: principais módulos dos codecs e transformadas aplicáveis no domínio do texto e principais códigos usados, bem como a suas combinações.

Como ponto de partida sugere-se a seguinte bibliografia:

- Gupta, A., Bansal, A., & Khanduja, V. (2017). Modern lossless compression techniques: Review, comparison and analysis. *Proceedings of the 2017 2nd IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT 2017*, *February*. https://doi.org/10.1109/ICECCT.2017.8117850
- Bell, T., Witten, I. H., & Cleary, J. G. (1989). Modeling for text compression. *ACM Computing Surveys (CSUR)*, *21*(4), 557–591. https://doi.org/10.1145/76894.76896
- Kodituwakku, S. R., & Amarasinghe, U. S. (2010). Comparison of Lossless Data Compression Algorithms. *Indian Journal of Computer Science and Engineering*, *1*(4), 416–425.
- S, S., & L, R. (2011). Text Compression Algorithms - a Comparative Study. *ICTACT Journal on Communication Technology*, *02*(04), 444–451. https://doi.org/10.21917/ijct.2011.0062

- Sharma, N., Kaur, J., Kaur, N., Sharma, N., Kaur, J., & Kaur, N. (2014). A Review on various Lossless Text Data Compression Techniques. *An International Journal of Engineering Sciences*, *2*((December), 58–63.

✍ Escreva um documento com a descrição e caracterização do problema e do estado da arte (formato artigo IEEE, máximo 3 páginas, templates: https://www.ieee.org/conferences/publishing/templates.html).
🖿 Submeta o documento no inforestudante até 26/11.

Semana 3: Explore mecanismos de transformação não destrutiva da sua fonte por forma a aumentar a redundância estatística da mesma. Analise os resultados em termos de entropia.

Semana 4 e 5: Investigue os diversos tipos de códigos assumindo fontes independentes e não-independentes (considere mecanismos que tirem partido da dependência estatística da fonte) e analise a eficiência de compressão obtida. Para o efeito deverá considerar o *overhead* em termos de armazenamento gerado pela utilização de cada código (por exemplo, a utilização de um código de Huffman poderá implicar o armazenamento da árvore binária). Compare os resultados de compressão dos diferentes métodos.

Entrega Final

✍ Entregue um documento com a descrição da solução proposta, fundamentando-a com as diversas experiências realizadas. Em particular, deverá entregar um relatório em formato de artigo IEEE com um máximo de 6 páginas (vide *template* em anexo) contendo os seguintes elementos:
1 – Descrição e caracterização do problema abordado
2 – Estado da arte do domínio específico
3 – Descrição do(s) algoritmo(s) selecionado(s) e fundamentação das opções
4 – Análise de resultados no dataset fornecido e discussão
5 – Conclusões e trabalho futuro sugerido

✍ Entregue um documento com as instruções para instalar e correr a solução.
📊 Prepare uma apresentação PowerPoint com duração máxima de 10 minutos.
🖿 Submeta o seu artigo e a apresentação na inforestudante. Inscreva-se para realizar a apresentação e defendê-la.