Codificação de Fonte (Outubro 2017)

Lucas Gonçalves da Conceição Nogueira, *Lucas da Silva Jorge*

*Palavras chave*— Lempel-ziv, compactação, descompactação, algoritmo, textos, idiomas e resultados.

# Apresentaçao e descrição do Algoritmo

O algoritmo é caracterizado por não causar perda de informações em sua compactação, sendo assim, a partir de textos compactados, é possível obter o texto original.

Para a compactação, primeiramente, armazena-se todos os caracteres em um dicionário com índices correspondentes ao que será transmitido. O arquivo é percorrido adicionando-se sequências as menores sequências que não estão presentes no dicionário, adicionando seus endereços e codificações. Tal procedimento é seguido até o fim do texto. Vale notar que com o intuito de diminuir a complexidade do algoritmo, o tamanho do dicionário foi limitado, sendo o mesmo reiniciado após atingir o tamanho máximo.

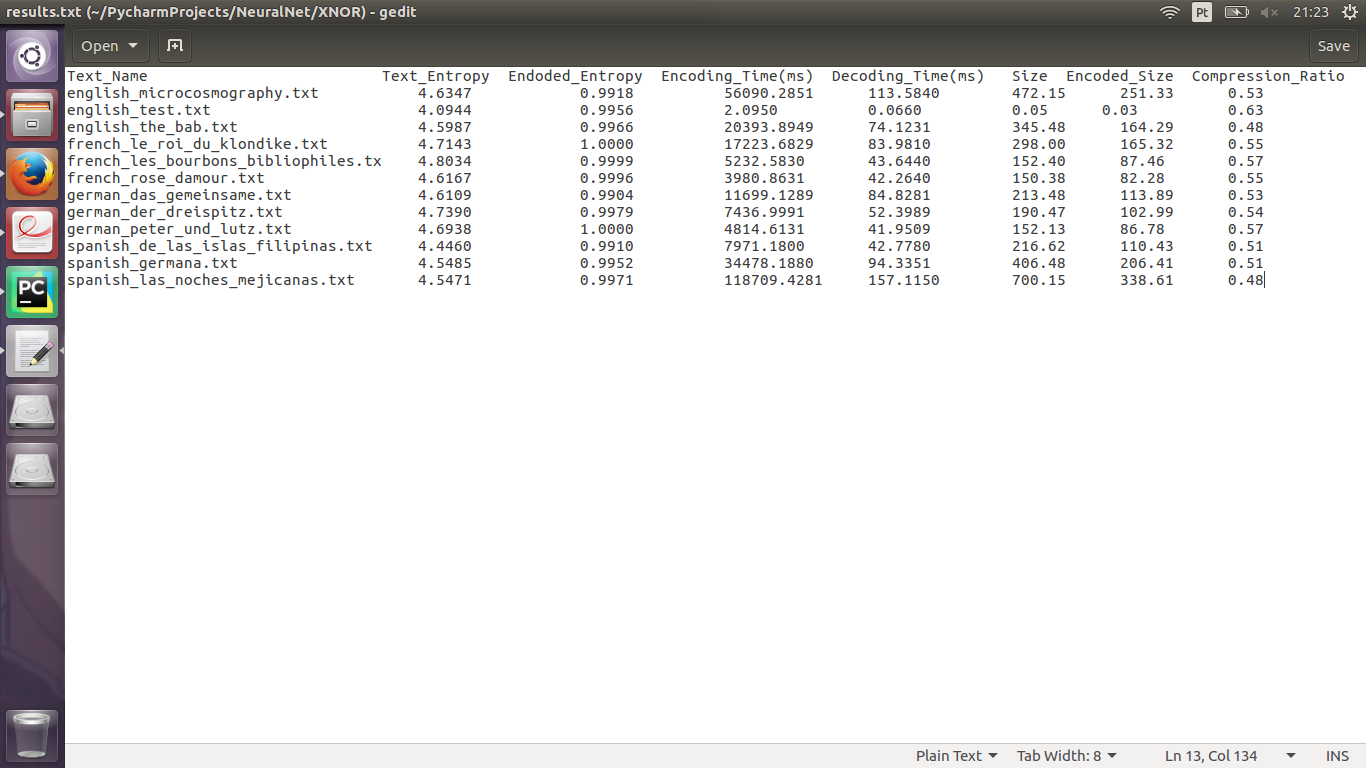
Para o algoritmo de descompactação, primeiramente, é identificado o alfabeto correspondente. Após isso, são lidas determinadas quantidades de bits, de acordo com o endereço do dicionário. Seguindo esse procedimento, ao final, o texto descompactado é retornado.

# Resultados Obtidos

Com o intuito de validar e avaliar a eficiência do algoritmo apresentado, foram utilizados 3 textos, sendo cada um em 4 idiomas diferentes. Os resultados obtidos são representados nas tabelas seguintes.

TABELA I

RESULTADOS OBTIDOS – L=1000



OBS: tamanhos em KB.

Como se pode observar, analisando a tabela 1, o arquivo descompactado, após a compactação, corresponde ao arquivo original, conforme o esperado. Tal resultado pode ser interprestado como um dos critérios de validação do projeto.

Outro resultado calculado pelo algoritmo é o cálculo da entropia dos textos compactados, considerando os mesmos uma sequência de bits.

Analisando os mesmos resultados para um limitantes de memória diferentes, foi possível obter os seguintes resultados:

Como se pode observar, foram testados diversos textos agrupados por idioma, afim de tentar se estabelecer argumentos para as análises que serão propostas relacionados ao idioma considerado.

TABELA II

RESULTADOS OBTIDOS L=10.000

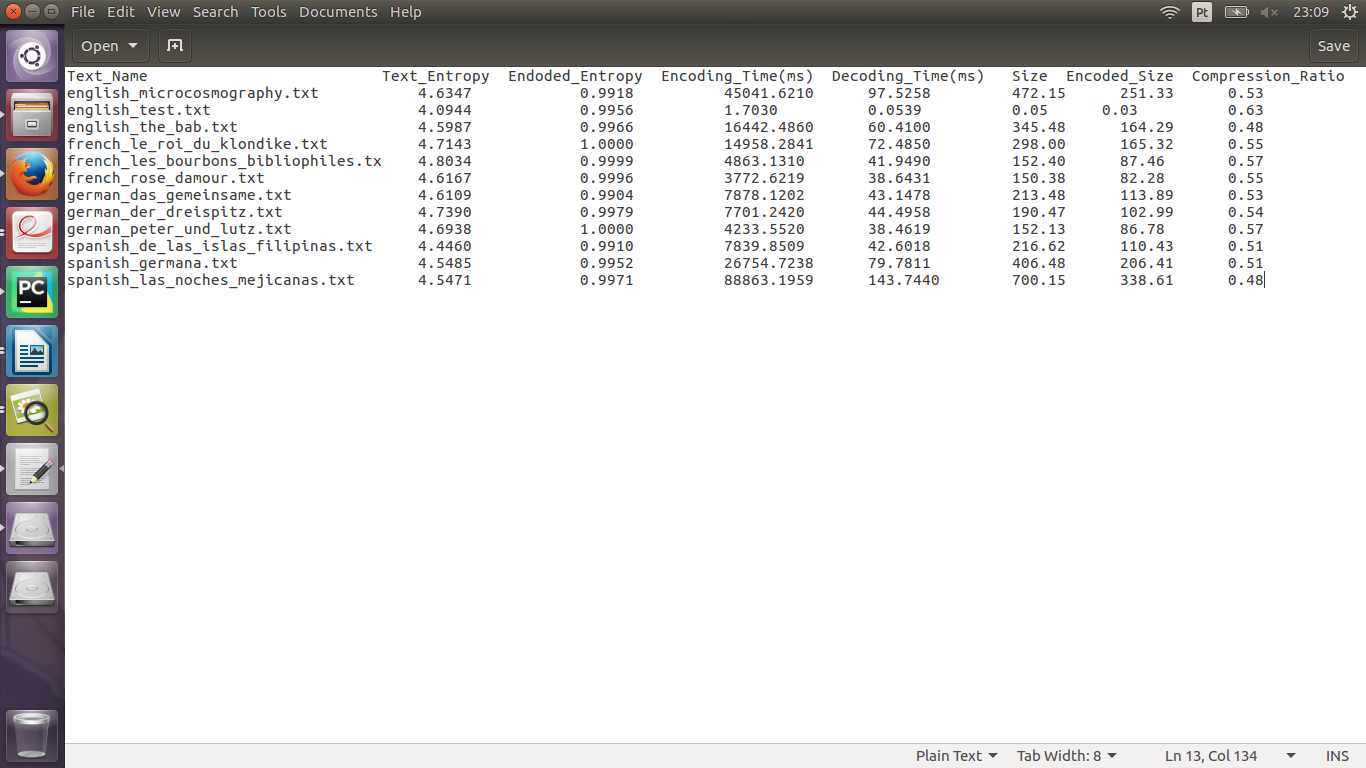


TABELA III

RESULTADOS OBTIDOS L=100.000

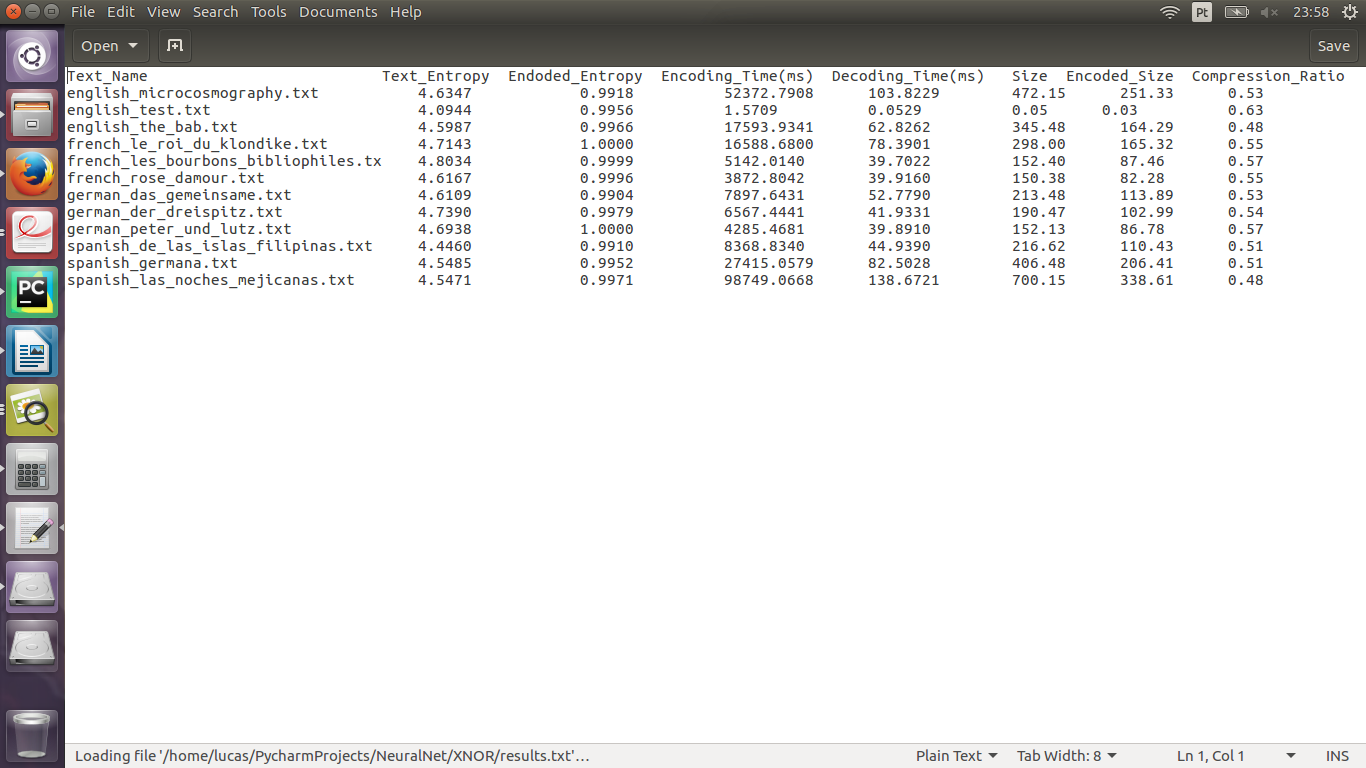
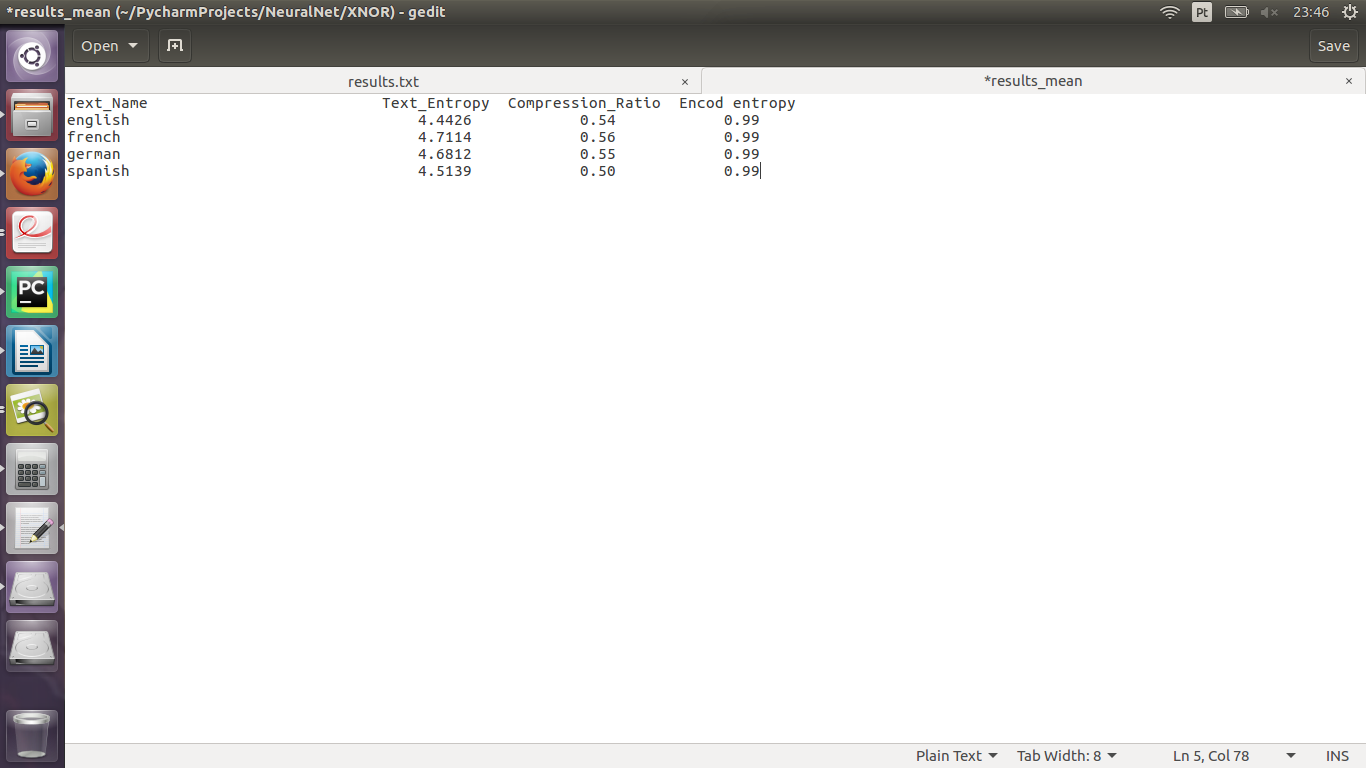


TABELA IV

RESULTADOS DE CADA IDIOMA L=1.000



# Análise

Com base na comparação entre as tabelas de resultados obtidas e nos tamanhos de dicionário utilizados, foi possível chegar as seguintes conclusões e análises:

Tendo em vista que o algoritmo de compressão tem como responsabilidade retirar a máxima quantidade de texto, de maneira a se conservar o que se deseja, é natural que haja uma diminuição de entropia em relação ao texto original, pois após a compactação se tem a mesma quantidade de informação representada de maneira otimizada. Analisando-se a tabela I, pode-se notar essa diminuição presente em todos os testes.

Além disso, pode-se notar que a entropia dos arquivos compactados tendem a 1. Tal fato já era esperado tendo em vista que tal resultado é o indicativo de um texto formado por uma linguagem binária.

Tendo em vista que a entropia e redundancia são inversamente proporcionais, conforme explicado, os resultados corroboram com o esperado pela teoria.

Com base no algoritmo de Lempel-ziv, pode-se notar que com o aumento do tamanho do dicionário é de se esperar que a eficiência de compactação aumente junto, pois quanto maior L maior a quantidade de simbolos no dicionário com codificação correspondente. Seguindo essa mesma linha, quanto maior for o valor de L menor tende a ser o texto compactado, pelo fato do processo possuir uma taxa de compressão menor.

Quanto à variação do tempo de processamento para os diversos valores de L, pode-se observar que o maior valor de L ocasiona um menor tempo de processamento, resultado esperado tendo em vista que para tamanhos limitados de dicionário o mesmo acaba sendo preenchido diversas vezes, levando a um maior consumo de tempo computacional. Para L muito grande, aumentar L ainda mais não altera muito o resultado, já que o tamanho do dicionário jamais atingirá o limite.

Fazendo-se a comparação entre os compactadores comerciais e o implementado, pode-se notar uma taxa de compactação do comercial, em geral, um pouco melhor que a do algoritmo implementado. Para um mesmo texto foi obtida uma taxa de compactação de 55,3% utilizando o algoritmo, equanto que para o compactadro Winrar foi obtida uma taxa de 30,4% (L=100.000). Analisando a literatura, o Winrar possui uma taxa média de compressão de 30% enquanto o algoritmo possui de 50%.

Quanto à influência do idioma na eficiência do algoritmo, pôde-se notar que quanto maior a entropia do idioma menor a taxa de compactação. Tal resultado já era esperado, tendo em vista que idiomas com maiores entropias menos redundantes. Assim, é de se esperar que o algoritmo de compactação cause uma diferença de tamanho entre arquivo compactado e descompactado maior. Assim, naturalmente, idiomas com menor entropia possuem informação menos compactada em seu dialeto e por consequência possuem uma compressão mais significativa com o uso do algoritmo, causando assim uma diminuição na taxa de conversão.

OBS: Foram encontradas diferentes definições para taxa de compactação. Neste relatório foi considerada como taxa de compactação a razão entre o tamanho do arquivo compactado sobre o tamanho do arquivo original.

# Conclusão

## Conforme pode-se comprovar através da comparação do algorítmo de Lempel-Ziv utilizado com algorítmos comerciais como 7-Zip, Winar e Winzip, a eficiência da compactação obtida foi menor (maior Compression Ratio), o que é esperado, já que o algorítmo usado é simplista, ainda atendendo todos os requisitos para o código: transmissão de mais informação em menos espaço com zero perda de informação.