Gabriel Adriano de Melo1, Gustavo Nahum Alvarez Ferreira2  
1,2 Aluno(a) de Graduação em Engenharia do Instituto Tecnológico de Aeronáutica – ITA.

[[1]](#footnote-1)

Aula 2 – Codificação de fonte

*Resumo*—Este trabalho explora a informação dentro de linguagens. Analisa-se não apenas a entropia de textos, mas também a eficiência de cada texto em representar a informação. Estabelecem-se alguns parâmetros para se medir a entropia condicional dos textos estudados assim como a previsibilidade de cada um.

Ao final, utiliza-se o conhecimento acumulado ao longo do trabalho para se medir a validade das traduções realizadas pelo algoritmo da Google.

*Palavras-chave*— Eficiência de linguagens, entropia, entropia condicional, previsibilidade de linguagens, teoria da informação, tradução.

# Introdução

N

O contexto da teoria de telecomunicações, encontra-se o conceito de entropia da informação. A entropia é uma medida da incerteza de uma variável aleatória, e está presente desde em cálculos de previsão sobre variáveis aleatórias, com base no conhecimento de outras variáveis aleatórias, até no cálculo da quantidade de bits necessários para representar uma variável aleatória, caso se atribuam rótulos binários para cada valor possível de X. Neste relatório, discutem-se os resultados e conclusões acerca da análise da informação contida em textos escritos em idiomas distintos. Os estudos envolveram comparações de eficiência de idiomas no que se refere ao armazenamento digital de textos escritos nesses idiomas, bem como análises de previsão de caracteres, com base no conhecimento de caracteres adjacentes e, por fim, um estudo da injeção ou depleção de informação existentes no processo de tradução de um texto de um idioma para outro ao usar-se o serviço de tradução online *Google translate*.

# Descrição do algoritmo

Os integrantes do grupo codificaram um algoritmo para o cálculo da entropia média e da entropia condicional. A linguagem escolhida para as linhas de código foi o C++, preferência que se justifica tanto por sua rapidez como pela familiaridade dos integrantes com a linguagem. Já a estrutura de dados adotada para armazenar as informações extraídas dos textos analisados foi a de mapa (*map*): para o cálculo da entropia média, produziu-se um mapa usando-se os diferentes caracteres que constavam no texto como chaves (*keys*), e as suas respectivas probabilidades de aparecerem no texto como os valores mapeados (*mapped values*), armazenados em tipo *double*; já para o cálculo das entropias condicionais, estruturou-se um mapa de pares (*map of pairs*), em cujo mapa havia um par contendo duas variáveis do tipo *char*, a primeira delas referente à letra que, no cálculo da probabilidade condicional, é a suposta já ter ocorrido no texto, e a segunda variável *char* é referente à letra que, na probabilidade condicional, é a desejosa de se conhecer a probabilidade de ocorrência, dado que a anterior foi verificada; finalmente, o segundo membro do mapa de pares era a frequência em que a sequência de caracteres em questão constava no texto.

O código apresenta um *loop* *while* principal, que itera sobre cada caractere do texto, e atualiza as informações de frequência por caractere nos mapas de cálculo de entropia média e entropia condicional. Nesta etapa do código, houve cautela, em particular, com os casos de início e término de leitura do texto (em que a entropia condicional não estaria propriamente definida).

Para prosseguir com o cálculo da probabilidade condicional, fez-se uso da seguinte igualdade:



Assim, a frequência em que os pares de caracteres Y (caractere anterior) e X (próximo caractere) aparecem no texto é, então, dividida pela frequência em que o Y aparece no texto, obtendo-se o valor da probabilidade condicional de que o próximo caractere seja X, dado que o imediatamente anterior foi Y. Essa probabilidade é armazenada no campo *double* do mapa de pares.

Em seguida, calculam-se as probabilidades individuais de cada caractere estar presente no texto. Para facilitar o cálculo desta etapa, criou-se a variável inteira “*numberOfChar*” (número de caracteres) ainda no trecho do código de leitura dos caracteres do texto; essa variável foi inicializada com o valor 0, e a cada caractere que era lido no texto, incrementava-se essa variável em uma unidade. Na etapa presente do código, então, dividiu-se a frequência de cada caractere distinto do texto pelo valor dessa variável, tendo-se obtido, assim, a probabilidade individual de cada caractere aparecer no texto.

Finalmente, criaram-se as funções “*getEntropy*” e “*getConditionalEntropy*”, responsáveis pelos cálculos da entropia média e da entropia condicional, respectivamente. Assim, essas funções recebiam como argumento o mapa de probabilidades, no caso de “*getEntropy*”, e ambos o mapa de probabilidades como o de probabilidades condicionais, no caso de “*getConditionalEntropy*”. Essencialmente, essas funções consistiam de código que implementasse as seguintes expressões:

Tabela II

Intervalos de tempo transcorridos na compactação e descompactação dos arquivos

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Idioma da Bíblia | Tamanho do arquivo original (KB) | Tempo de compactação (s) | Tamanho do arquivo compactado(KB) | Tempo de descom-pactação (s) |
| Alemão | 4038 | 22,637 | 1304 | 8,116 |
| Coreano | 4355 | 4,378 | 1000 | 9,716 |
| Espanhol (1909) | 3870 | 14,943 | 1248 | 23,009 |
| Espanhol (2004) | 3887 | 13,983 | 1241 | 7,837 |
| Filipino | 4780 | 17,363 | 1297 | 15,376 |
| Francês | 4145 | 7,288 | 1324 | 8,991 |
| Hebraico | 3674 | 3,773 | 853 | 6,937 |
| Inglês (1769) | 4100 | 13,272 | 1270 | 8,495 |
| Inglês (1901) | 4087 | 5,368 | 1261 | 10,289 |
| Inglês (1995) | 4035 | 6,498 | 1338 | 20,276 |
| Italiano | 4098 | 9,976 | 1346 | 8,176 |
| Latim | 3359 | 4,806 | 1147 | 8,38 |
| Maori | 4375 | 7,368 | 1190 | 7,945 |
| Polonês | 3942 | 5,768 | 1313 | 11,033 |
| Português | 3829 | 5,558 | 1262 | 7,928 |
| Português (Brasil) | 3865 | 5,095 | 1263 | 10,49 |



no caso da função “*getEntropy*” e:



no caso da função “*getConditionalEntropy*”. Essas funções retonam suas respectivas entropias em tamanho numérico *double*, e são chamadas pela função *main* para que tenham os seus valores impressos por “*printf*”.

# Desenvolvimento

Os códigos de compactação e descompactação produzidos foram, então, aplicados a uma coleção de arquivos de texto em versão .txt, que eram referentes aos mesmos livros presentes no Velho Testamento da Bíblia, porém escritos em línguas variadas. Na tabela I, listam-se os resultados obtidos nesse processo.

Tabela I

Tamanhos dos arquivos antes e após a compactação

|  |  |  |  |
| --- | --- | --- | --- |
| Idioma da Bíblia | Tamanho do arquivo original (KB) | Tamanho do arquivo compactado (KB) | Porcentagem de compactação |
| Alemão | 4038 | 1304 | 67,7% |
| Coreano | 4355 | 1000 | 77,0% |
| Espanhol (1909) | 3870 | 1248 | 67,8% |
| Espanhol (2004) | 3887 | 1241 | 68,1% |
| Filipino | 4780 | 1297 | 71,5% |
| Francês | 4145 | 1324 | 68,1% |
| Hebraico | 3674 | 853 | 76,8% |
| Inglês (1769) | 4100 | 1270 | 69,0% |
| Inglês (1901) | 4087 | 1261 | 69,1% |
| Inglês (1995) | 4035 | 1338 | 66,8% |
| Italiano | 4098 | 1346 | 67,2% |
| Latim | 3359 | 1147 | 65,9% |
| Maori | 4375 | 1190 | 72,8% |
| Polonês | 3942 | 1313 | 66,7% |
| Português | 3829 | 1262 | 67,0% |
| Português (Brasil) | 3865 | 1263 | 67,3% |

Tabela II. Análise dos intervalos de tempo de compactação e descompactação dos arquivos analisados.

Tabela I. Comparação de tamanhos de arquivos antés e após a compactação, e a respectiva porcentagem de compactação. Os conteúdos dos arquivos foram preservados, de modo que os arquivos descompactados apresentaram tamanhos iguais aos arquivos originais.

A tabela I permite que as porcentagens de compactação das bíblias sejam apreciadas, e nota-se que essas porcentagens são elevadas, com média de 69,3% e desvio padrão de 3,45% de compactação. Com isso, é possível perceber a eficiência considerável do algoritmo de Lempel-Ziv na compactação desses textos.

A tabela II mostra o intervalo de tempo decorrido durante cada um dos processos de compactação e descompactação, individualmente. Além das colunas referentes aos intervalos de tempo para compactação e descompactação, incluíram-se, também, colunas com os tamanhos dos arquivos original e compactado, para facilitar possíveis relações com os seus respectivos tempos.

Contudo, a análise da tabela permite avaliar que uma tal relação não é tão nítida: tomando-se as primeiras linhas da tabela, é possível verificar que o texto em alemão, que contém 4038KB, levou mais de 22 segundos para ser compactado; já o texto em coreano, com 4355KB, levou cerca de 4,4s em sua compactação.

Uma parte dessa discrepância pode ser entendida pela elevada oscilação de disponibilidade da CPU para o processamento do código estudado.

Uma vez que essa análise foi muito comprometida pela presença de ruídos dessa natureza, o grupo optou por fazer análises de correlação linear positiva dos dados presentes: a correlação calculada para os dados de tamanho original do arquivo e tempo de compactação foi de ; a correlação entre o tamanho do arquivo original e o tempo de descompactação foi de ; a correlação entre o tamanho do arquivo compactado e o tempo de compactação foi de ; por fim, a correlação entre o tamanho do arquivo compactado e o tempo de descompactação foi de .

Portanto, houve uma correlação linear positiva baixa ou quase inexistente entre os pares de parâmetros considerados. Isso pode ter ocorrido em função dos ruídos dos dados de tempo que foram medidos; ainda assim, o resultado obtido impossibilitou que se confirmasse a existência de uma forte correlação linear entre o tamanho do arquivo tomado e o tempo interveniente para qualquer um dos processos.

Para o desenvolvimento do trabalho é necessário responder a algumas perguntas previamente selecionadas e divididas em tópicos.

## Qual é a entropia dos arquivos compactados?

Tabela IV

Entropias dos arquivos originais

|  |  |  |  |
| --- | --- | --- | --- |
| Idioma da Bíblia | Tamanho do arquivo original (KB) | Entropia média | Entropia condicional |
| Alemão | 1304 | 4,522951 | 3,271865 |
| Coreano | 1000 | 5,023910 | 2,971844 |
| Espanhol (1909) | 1248 | 4,538020 | 3,254991 |
| Espanhol (2004) | 1241 | 4,491513 | 3,245227 |
| Filipino | 1297 | 3,977094 | 2,926025 |
| Francês | 1324 | 4,521425 | 3,337303 |
| Hebraico | 853 | 3,339732 | 2,168272 |
| Inglês (1769) | 1270 | 4,388496 | 3,280042 |
| Inglês (1901) | 1261 | 4,350252 | 3,304698 |
| Inglês (1995) | 1338 | 4,510906 | 3,479508 |
| Italiano | 1346 | 4,477818 | 3,269902 |
| Latim | 1147 | 4,169905 | 3,407692 |
| Maori | 1190 | 3,771488 | 2,689717 |
| Polonês | 1313 | 4,897213 | 3,304324 |
| Português | 1262 | 4,499726 | 3,256847 |
| Português (Brasil) | 1263 | 4,446478 | 3,261015 |

O valor da entropia dos arquivos compactados, bem como sua entropia condicional, puderam ser calculados por meio do código produzido na primeira aula experimental, e seus valores estão listados na tabela III.

Uma primeira análise da tabela III indica os seguintes valores para as estatísticas de cada parâmetro presente: média de 6,7177 para a entropia média, com um desvio padrão de 0,4480, e média de 1,1483 para a entropia condicional, com um desvio padrão de 0,9217. Ao comparar esses valores com os valores obtidos para os mesmos parâmetros na atividade experimental anterior, verificou-se que: na atividade anterior, a entropia foi de, aproximadamente, 4,5 e, logo, a entropia dos arquivos mostrou-se consideravelmente maior quando eles estão na forma compactada; já a entropia condicional possuía, em média, um valor da ordem de 3,5 no primeiro laboratório, sugerindo que textos compactados diminuem muito a entropia mútua.

Uma análise relacionando o tamanho dos textos compactados e suas respectivas entropias mostrou os seguintes resultados: o coeficiente de correlação linear calculado entre o tamanho do arquivo e a entropia média foi de 0,1694; já o coeficiente de correlação linear entre o tamanho do arquivo e a entropia condicional foi de -0,4594, o que é classificado como uma correlação linear negativa baixa. Dessa forma, o tamanho dos textos compactados não se mostrou ser um fator muito impactante no que diz respeito a suas respectivas entropias.

## Como as entropias dos arquivos compactados se comparam às entropias dos textos?

As entropias dos textos puderam ser obtidas, dessa vez, aplicando o código produzido na primeira atividade experimental, porém aos textos originais. Os resultados podem ser vistos na tabela IV.

A média da entropia apresentou o valor de 4,3704, e seu desvio padrão foi de 0,4053; já a média da entropia condicional apresentou o valor de 3,1518, e seu desvio padrão foi de 0,3271. Dessa forma, confirmou-se a tendência antecipada na análise anterior, a partir dos dados da primeira atividade experimental, de que a mesma informação, quando codificada através do algoritmo de compactação de Lempel-Ziv, possui uma quantidade de entropia média maior – seu valor médio foi calculado em 6,7177, substancialmente maior do que a média das entropias dos textos originais, 4,3704 –, ao passo que sua entropia condicional é bastante inferior à dos textos originais – a entropia condicional média dos textos compactados foi calculada em 1,1483, bem abaixo da média de entropia condicional dos textos não compactados, que é de 3,1518.

Tabela III

Entropias dos arquivos compactados

|  |  |  |  |
| --- | --- | --- | --- |
| Idioma da Bíblia | Tamanho do arquivo compactado (KB) | Entropia média | Entropia condicional |
| Alemão | 1304 | 7,036425 | 1,069312 |
| Coreano | 1000 | 6,073538 | 4,107575 |
| Espanhol (1909) | 1248 | 6,541842 | 0,814010 |
| Espanhol (2004) | 1241 | 7,400419 | 1,502718 |
| Filipino | 1297 | 6,605368 | 0,729289 |
| Francês | 1324 | 6,359770 | 0,669473 |
| Hebraico | 853 | 7,032909 | 1,381728 |
| Inglês (1769) | 1270 | 6,477110 | 0,240741 |
| Inglês (1901) | 1261 | 6,801624 | 0,757052 |
| Inglês (1995) | 1338 | 6,461093 | 0,261682 |
| Italiano | 1346 | 7,518313 | 1.885832 |
| Latim | 1147 | 6,692257 | 0,921599 |
| Maori | 1190 | 5,908819 | 0,293333 |
| Polonês | 1313 | 7,262663 | 1,361502 |
| Português | 1262 | 6,769306 | 1,457130 |
| Português (Brasil) | 1263 | 6,541153 | 0,920317 |

O fenônemo relacionado à entropia condicional pode ser entendido da seguinte forma: uma vez que a quantidade de dígitos binários

# Desafio

# Conclusão

Referências

1. [↑](#footnote-ref-1)