**Contagem de frequência de Bytes em diversos tipos de arquivo**

**CEFET/RJ**– Centro Federal de Educação Tecnológica Celso Suckow da Fonseca

**Disciplina de Organização e Estrutura de Arquivos**

Natália Nunes Vieira

Resumo: Este estudo tem como objetivo realizar uma análise sobre métodos de compressão de arquivos, onde serão analisadas as frequências de bytes ocorrentes em um arquivo antes e depois dos mesmos serem compactados como *.zip* ou *.rar*.

**Introdução**

O trabalho consiste em elaborar um código que seja capaz de contar a frequência de bytes de um arquivo, depois compactar estes arquivos e realizar uma nova contagem de frequência. Identificar o grau de compressão de cada um deles e redigir uma documentação com o que foi analisado. Para isso, o desenvolvimento foi divido em três fases:

**Fase 1 (um):**

Desenvolver um algoritmo que seja capaz de ler cada Byte de um arquivo e contar a frequência em que eles ocorrem.

O algoritmo precisa:

* Ler uma pasta com diversos tipos de arquivo;
* Ler cada arquivo da pasta individualmente;
* Ler cada Byte do arquivo;
* Identificar qual Byte foi lido;
* Incrementar uma lista com a frequência que cada Byte aparece no arquivo;
* Escrever a lista na tela.

Linguagem: Python

**Fase 2 (dois):**

Compactar os arquivos utilizando formatos já conhecidos.

Será preciso:

* Escolher pelo menos dois formatos de compactação;
* Compactar cada um dos arquivos;
* Identificar o tamanho anterior e posterior à compactação.

Formato dos arquivos: *.txt, .pdf,* .*jpeg,* .*png*, *.mp3, .mp4.*

Formatos para compactação: *.rar* e *.zip*

**Fase 3 (três):**

Comparar a frequência de Bytes dos arquivos antes e depois de serem compactados.

Será preciso:

* Realizar a contagem da frequência de Bytes do arquivo antes da compactação;
* Realizar a contagem da frequência de Bytes do arquivo após a compactação;
* Identificar qual a diferença da contagem entre os dois testes;
* Redigir um documento com o que for identificado entre os arquivos.

**Desenvolvimento**

**Fase 1 (um) - Implementação:**

O algoritmo foi desenvolvido na linguagem de programação Python, utilizando a IDE do *Visual Studio Code*. O código encontra-se disponível no *Github* descrito no rodapé do relatório.

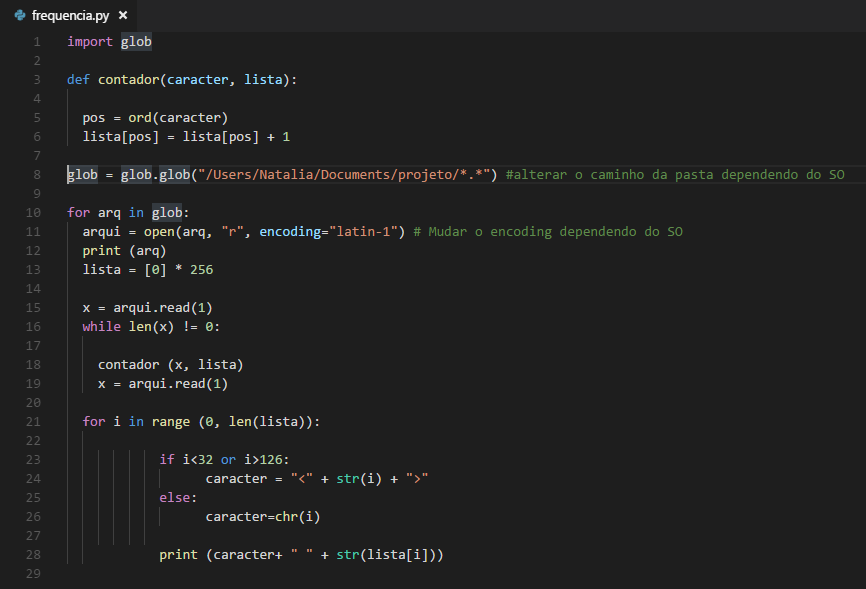


Figura 1: Contador de frequência de Bytes

Para o código da Figura 1: Contador de frequência de Bytes é importante ressaltar que, para que os testes ocorram com precisão, é necessário alterar o caminho da pasta onde estarão os arquivos para teste, e também verificar o *encoding* utilizado por cada sistema operacional.

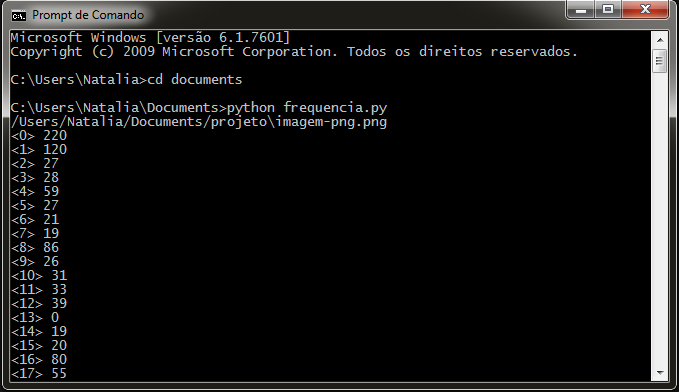


Figura 2: Exemplo de saída no console

Para que fosse gerada uma identificação para cada caractere, utilizamos a tabela ASC II, que pode ser encontrada no GitHub informado no rodapé deste documento.

**Fase 2 (dois) - Formatos:**

Para a realização dos testes, foram escolhidos os formatos de arquivo *.txt*, .*pdf*, *.jpeg*, *.png*, *.mp3*, *.mp4*. E os formatos de compactação .*zip* e .*rar*.

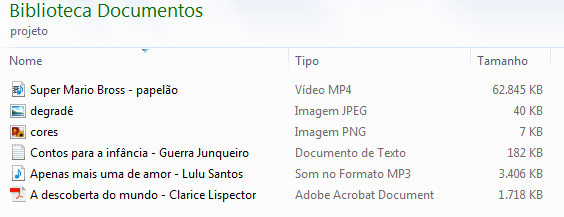


Figura 3: Arquivos nos formatos selecionados

Para cada arquivo foram utilizados os padrões de compactação abaixo descritos na Figura 4: Especificações de compactação *.rar* e na Figura 5: Especificações de compactação *.zip*.

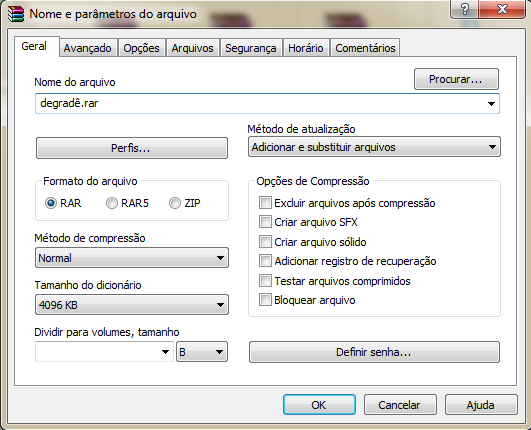


Figura 4: Especificações de compactação *.rar*

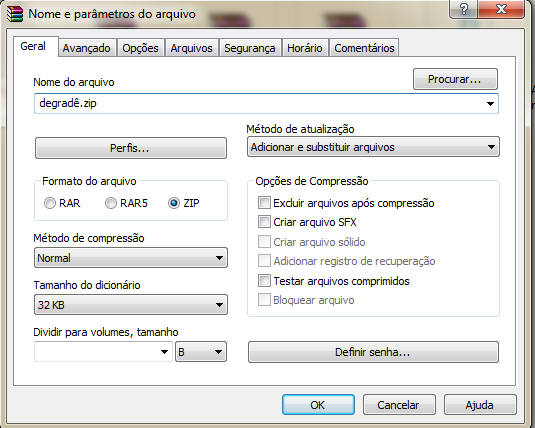


Figura 5: Especificações de compactação *.zip*

Todos os arquivos foram compactados, conforme a Figura 6: Arquivos originais e compactados.



Figura 6: Arquivos originais e compactados

**Fase 3 (três) – Testes:**

Para a realização dos testes foi utilizada uma máquina com a seguinte configuração: processador *Core i5*, 8GB de memória RAM e um SSD de 120 GB. Os arquivos utilizados para execução dos testes encontram-se disponíveis na pasta do drive informada no rodapé deste relatório.

Abaixo, os resultados identificados em cada teste.

* .*txt*:

O arquivo utilizado para testar o formato .*txt* é o livro *Contos para Infância*, do autor Guerra Junqueiro e foi obtido através do Projeto Gutemberg.

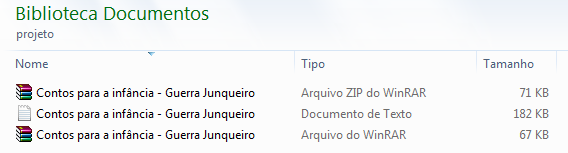


Figura 7: Arquivo .txt original e compactados

Os testes foram realizados para os três formatos de compactação do arquivo, conforme Figura 8: Testes para arquivo .*txt* original e compactado.

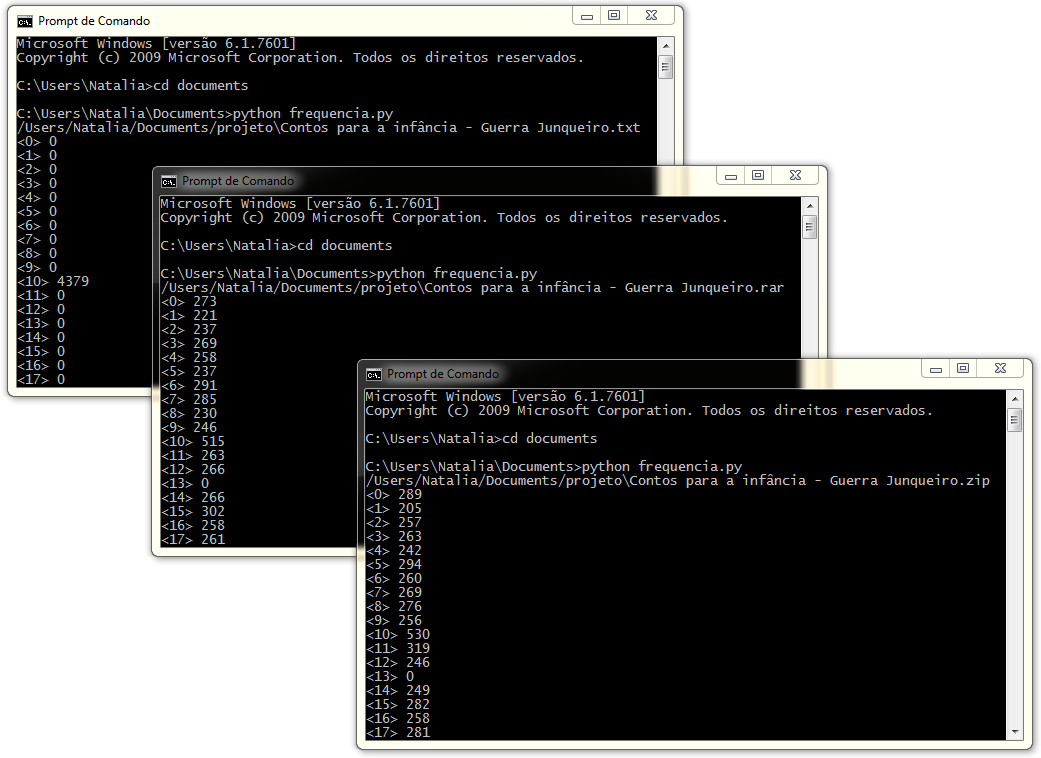


Figura 8: Testes para arquivo .*txt* original e compactado

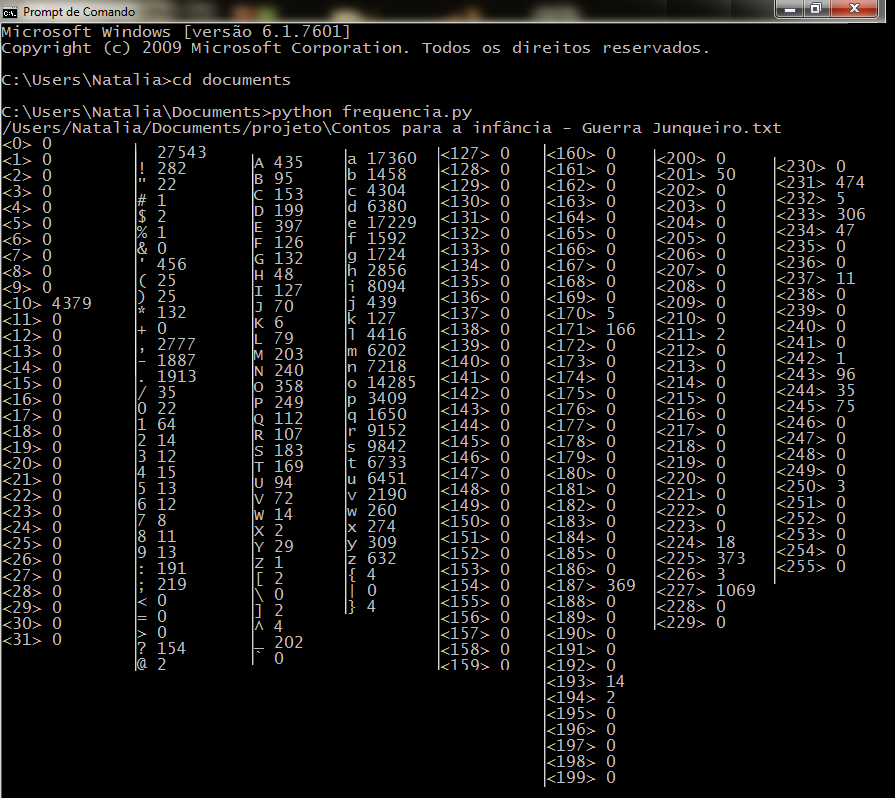


Figura 9: Resultado da contagem de Bytes do arquivo *.txt*

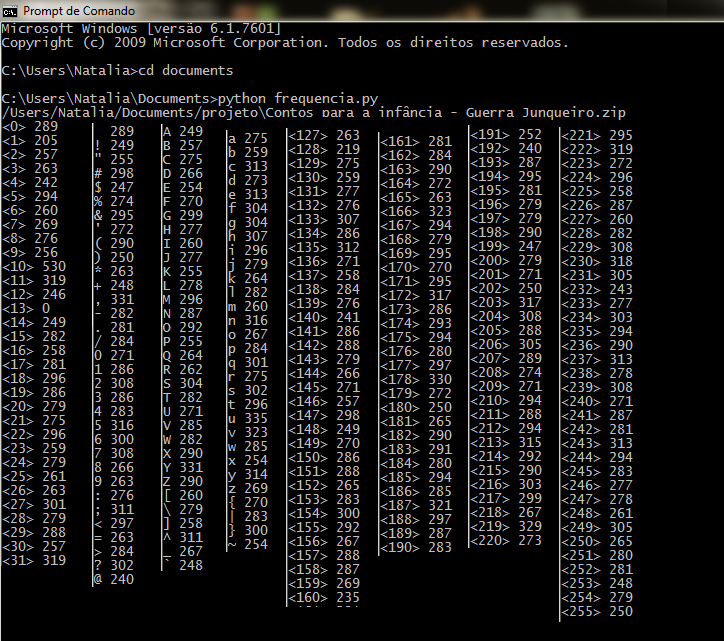


Figura 10: Resultado da contagem de bytes para o arquivo *.txt* compactado em formato *.zip*

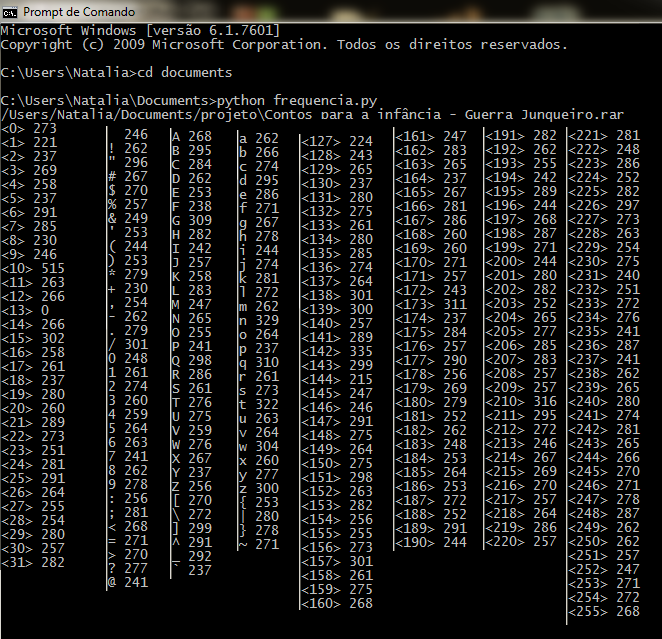


Figura 11: Resultado da contagem de bytes para o arquivo *.txt* compactado em formato *.rar*

Na Figura 7: Arquivo .txt original e compactados, podemos perceber a diferença de tamanho para os arquivos compactados, enquanto a compactação em .*zip* reduziu o arquivo de 182KB para 71KB, a compactação do .*rar* reduziu o mesmo arquivo para 67KB. Para a análise da frequência de bytes, do arquivo .*txt* utilizamos, após identificação dos caracteres que apareceram com maior frequência, os caracteres de *espaço* e ‘*a*’ para comparação entre os três tipos de arquivo.

Percebemos que, para o arquivo não compactado, o caractere de *espaço* apareceu 27543 vezes, para 289 vezes no formato .*zip* e 246 vezes no formato .*rar*; para o caractere ‘*a*’, a ocorrência identificada no arquivo original foi de 17360 vezes, para 275 vezes no .*zip* e 262 vezes no .*rar*.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***.txt*** | ***.zip*** | ***.rar*** | ***Frequência de Bytes: Arquivo original/Compactado*** | |
| ***.zip*** | ***.rar*** |
| ***espaço*** | 27543 | 289 | 246 | 95,30 | 111,96 |
| ***a*** | 17360 | 275 | 262 | 63,13 | 66,26 |

Sendo assim, é possível identificar que, após a compactação do arquivo, tanto pela frequência de bytes, quanto pelo tamanho dos arquivos, para um formato .*txt*, o formato de compactação .*rar* compacta melhor que o .*zip*.

* .*pdf*:

Para os testes com arquivo pdf foi selecionado o livro *A Descoberta do Mundo*, de autora Clarice Lispector. O arquivo foi disponibilizado pela equipe do *LeLivros*, para uso parcial em pesquisas e estudos acadêmicos.

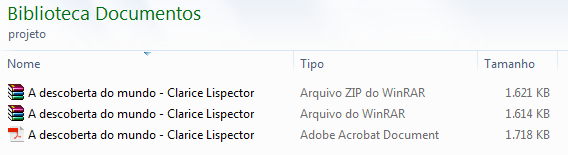


Figura 12: Arquivo *.pdf* original e compactados

Os testes foram realizados para os três formatos do arquivo, conforme Figura 13: Compactação de arquivo .*pdf*.

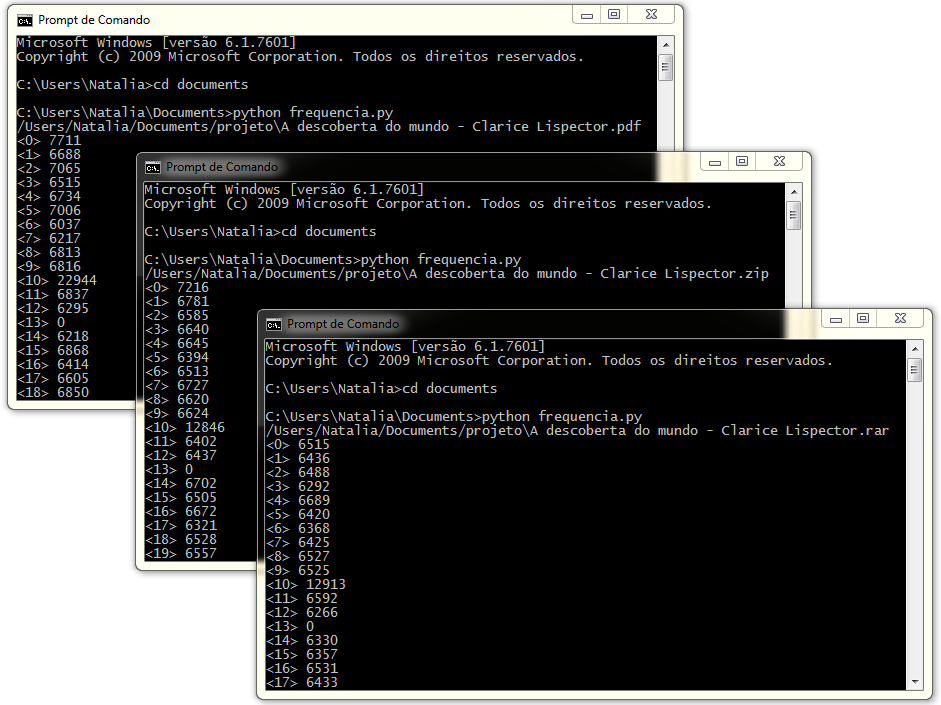


Figura 13: Compactação de arquivo .*pdf*

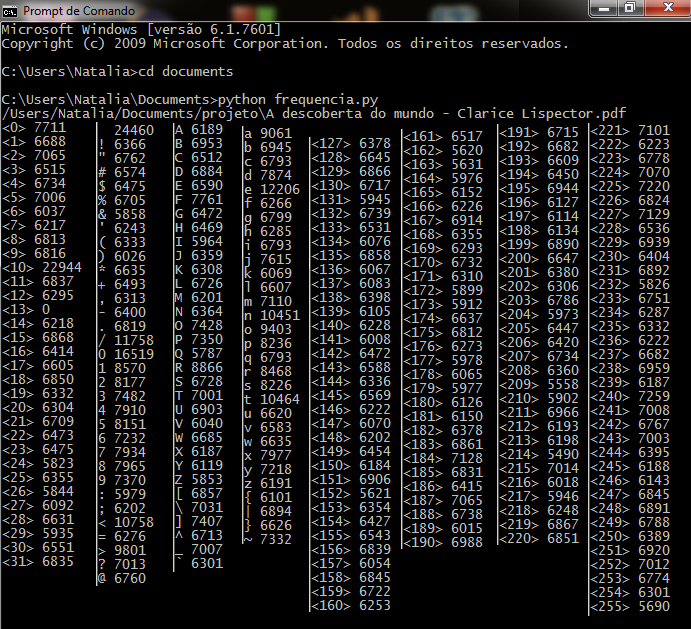


Figura 14: Resultado da contagem de bytes para o arquivo .*pdf*

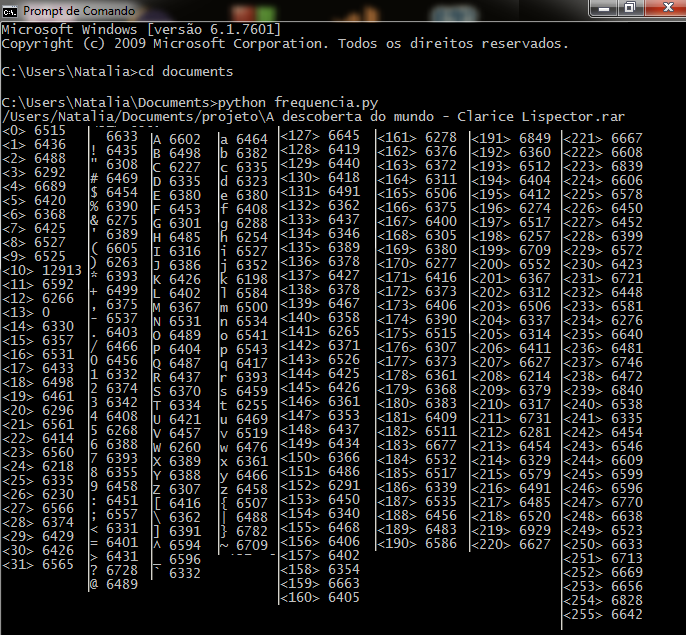


Figura 15: Resultado da contagem de bytes para o arquivo .*pdf* compactado em formato .*rar*

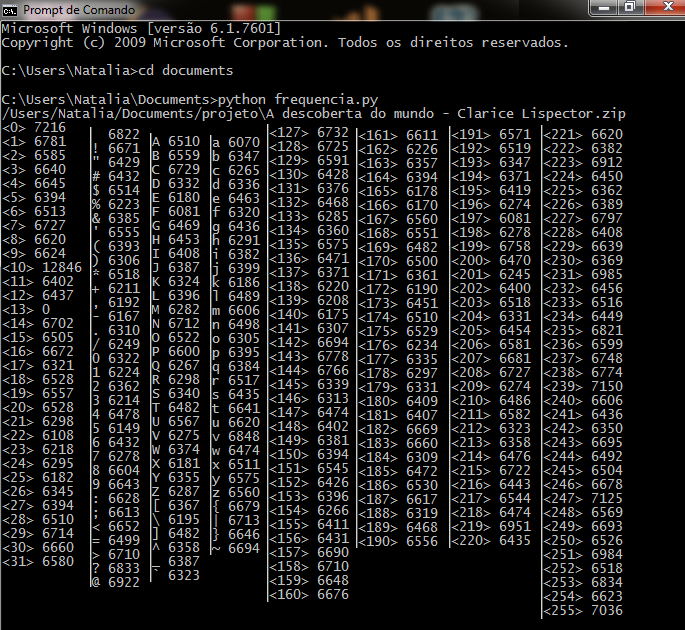


Figura 16: Resultado da contagem de bytes para o arquivo .*pdf* compactado em formato .*zip*

A Figura 12: Arquivo *.pdf* original e compactados, nos permite identificar a redução de tamanho para os arquivos .*pdf*, compactados em .*zip* e .*rar*. Ao comparar os valores que em .*pdf* foram de 1718 KB, em .*zip* teve uma redução para 1621KB e 1614KB em .*rar*.

Para analisar a frequência de Bytes, os caracteres escolhidos foram o numeral ‘*0*’ e a letra ‘*e*’, onde o ‘*0’* aparecia 16519 vezes no arquivo .*pdf* e passou a aparecer 6322 vezes no .*zip* e 6456 no .*rar*; e a letra ‘*e’* aparecia 12206 vezes no arquivo .*pdf* e reduziu para 6463 aparecimentos no .*zip* e 6380 no .*rar*.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***.pdf*** | ***.zip*** | ***.rar*** | ***Frequência de Bytes: Arquivo original/Compactado*** | |
| ***.zip*** | ***.rar*** |
| ***0*** | 16519 | 6322 | 6456 | 2,61 | 2,56 |
| ***e*** | 12206 | 6463 | 6380 | 1,89 | 1,91 |

Após análise do tamanho dos arquivos e da frequência dos bytes ‘*0*’ e ‘*e*’ pudemos perceber que, apesar do arquivo original ter sido compacto, o formato de arquivo .*pdf* não teve uma redução significativa em nenhum dos dois formatos.

* .*jpg*:

A imagem abaixo utilizada para os testes em .*jpg* foi obtida através do *Google Imagens* e está disponível para download no *Pinterest*.

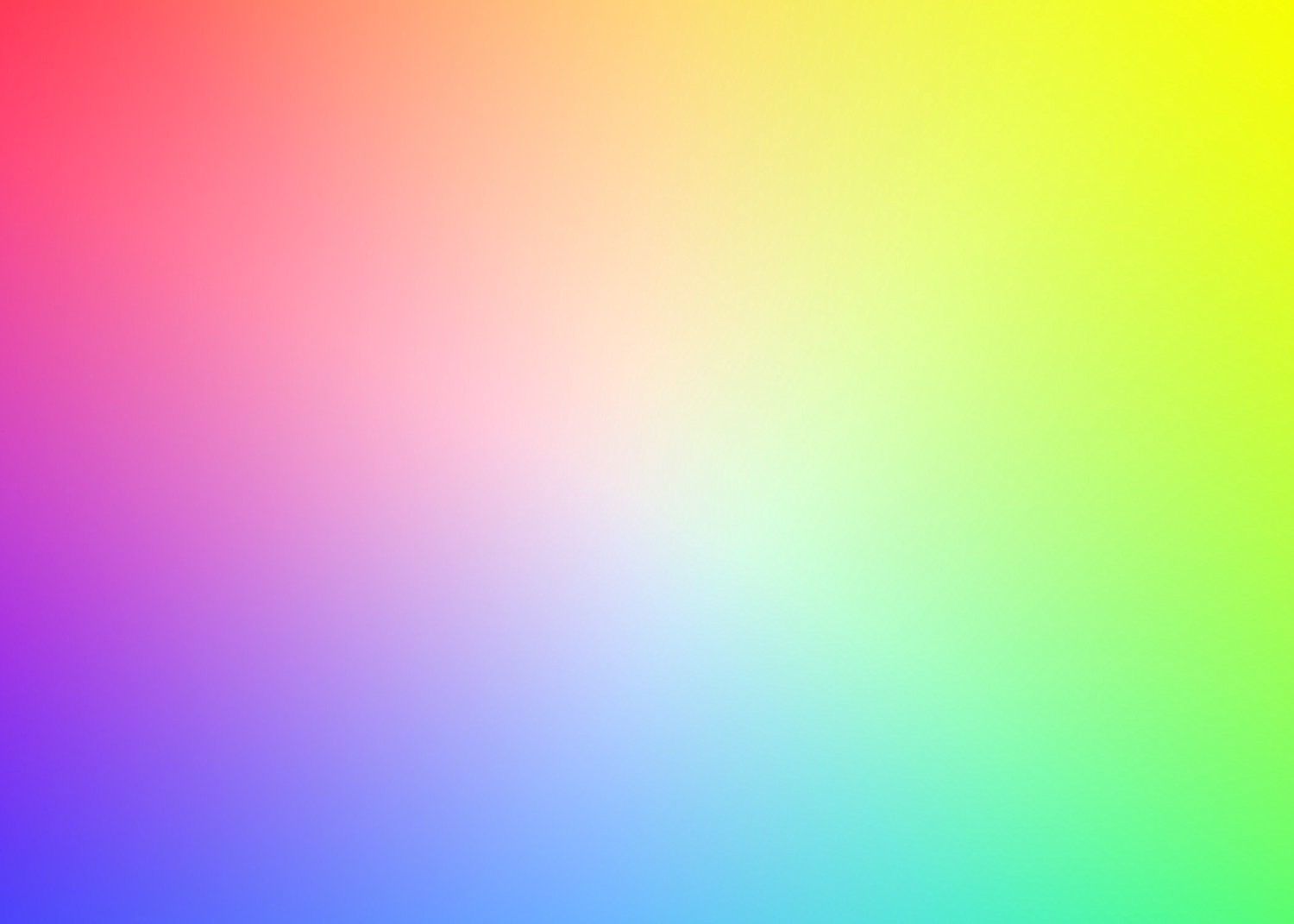


Figura 17: Imagem utilizada para o teste de contagem de frequência de Byte no formato .*jpg*

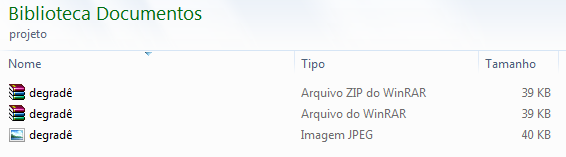


Figura 18: Arquivo .*jpg* original e compactados

Os testes foram realizados para os três formatos do arquivo, conforme Figura 19: Compactação de arquivo .*jpg*.

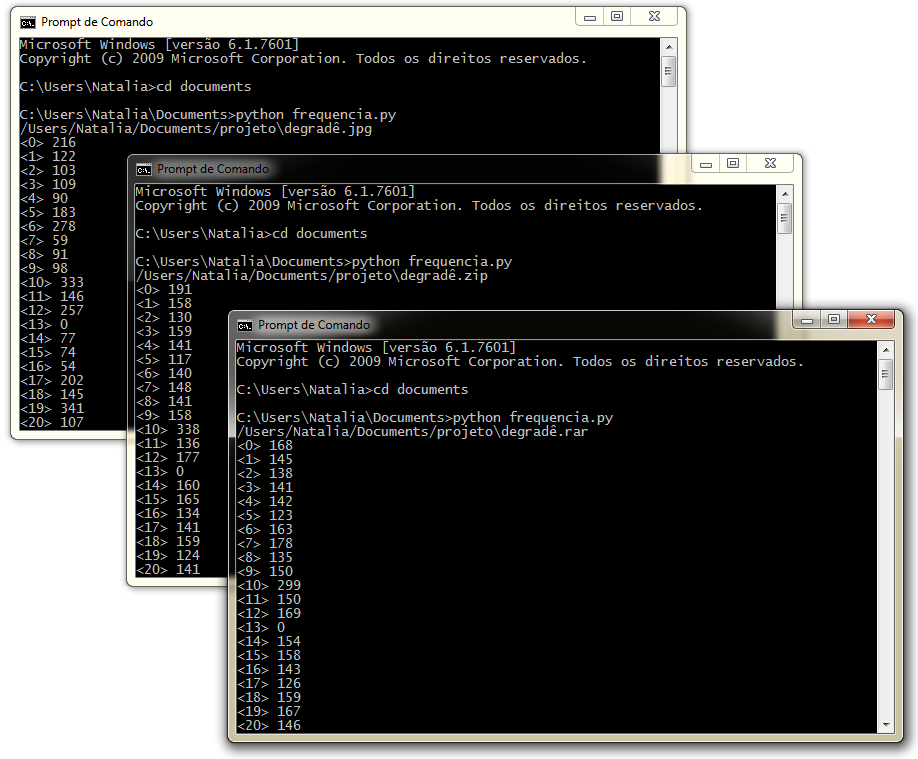


Figura 19: Compactação de arquivo .*jpg*

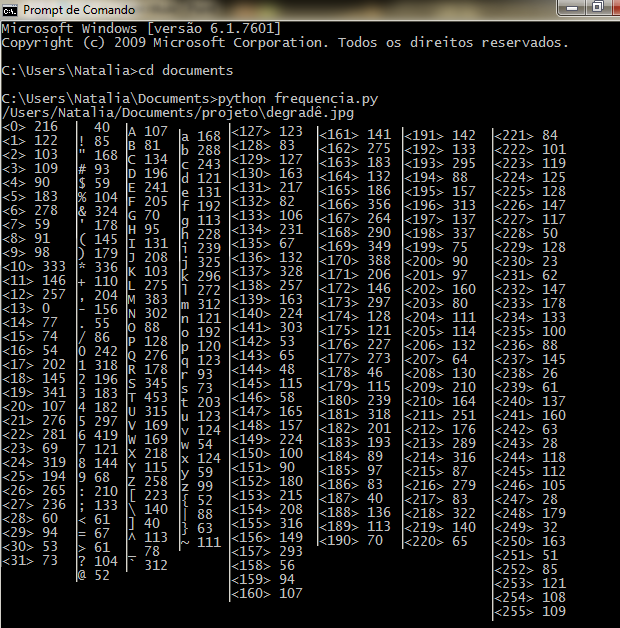


Figura 20: Resultado da contagem de bytes para o arquivo .*jpg*

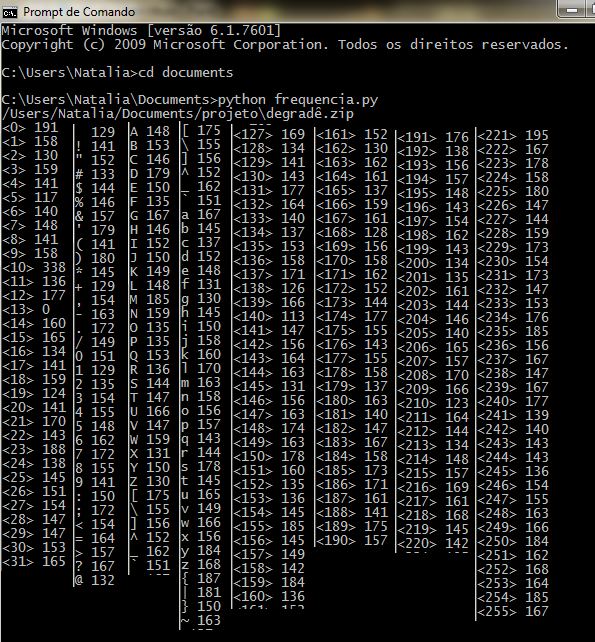


Figura 21: Resultado da contagem de bytes para o arquivo .*jpg* compactado em formato .*zip*

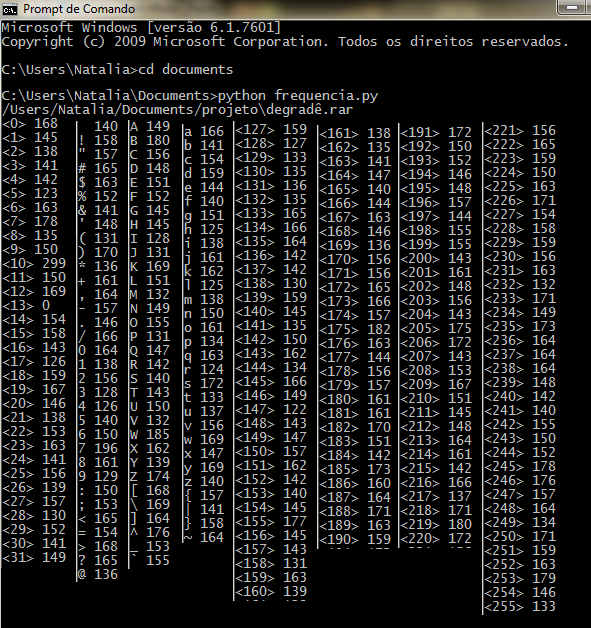


Figura 22: Resultado da contagem de bytes para o arquivo .*jpg* compactado em formato .*rar*

A Figura 18: Arquivo .*jpg* original e compactados nos mostra a redução de tamanho entre os formatos .*jpeg*, .*zip* e .*rar*, e, através dela, podemos concluir que a compactação identificada foi mínima, considerando que um arquivo que antes tinha 40KB passou a ter 39KB em ambos os formatos de compactação.

Na Figura 20: Resultado da contagem de bytes para o arquivo .*jpg*, Figura 21: Resultado da contagem de bytes para o arquivo .*jpg* compactado em formato .*zip* e Figura 22: Resultado da contagem de bytes para o arquivo .*jpg* compactado em formato .*rar*, foram destacados os caracteres encontrados em maior número no formato .*jpg*: ‘*6’* e ‘*m’*. Para o formato .*jpg*, eles apareceram 419 vezes e 312 vezes, respectivamente. Para o formato .*zip*, os caracteres passaram a aparecer 162 e 163 vezes, respectivamente. E, para o formato .*rar*, a aparição dos caracteres ocorreu em 150 e 138 vezes, respectivamente.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***.jpg*** | ***.zip*** | ***.rar*** | **Frequência de Bytes: Arquivo original/Compactado** | |
| ***.zip*** | ***.rar*** |
| **6** | 419 | 162 | 150 | 2,58 | 2,79 |
| **m** | 312 | 163 | 138 | 1,91 | 2,26 |

Com base nos valores identificados na Figura 18: Arquivo .*jpg* original e compactados e na Figura 20: Resultado da contagem de bytes para o arquivo .*jpg*, Figura 21: Resultado da contagem de bytes para o arquivo .*jpg* compactado em formato .*zip* e Figura 22: Resultado da contagem de bytes para o arquivo .*jpg* compactado em formato .*rar*, podemos concluir que, como o .*jpg* é um formato de fácil compactação por ser gerado em blocos, a imagem utilizada no teste já possuía um alto nível de compactação, considerando que a diferença entre os tamanhos do arquivo original e dos compactados foi de apenas 1KB e a redução das frequências é inferior a 3 vezes menor.

* .*png*:

A imagem abaixo foi utilizada para os testes, foi encontrada através do *Google Imagens* e está disponível no *Pinterest*.



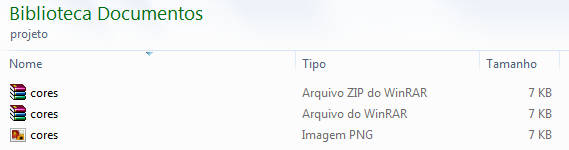


Figura 23: Arquivo .*png* original e compactados

Os testes foram realizados para os três formatos do arquivo, conforme Figura 24: Compactação de arquivo .*png*.

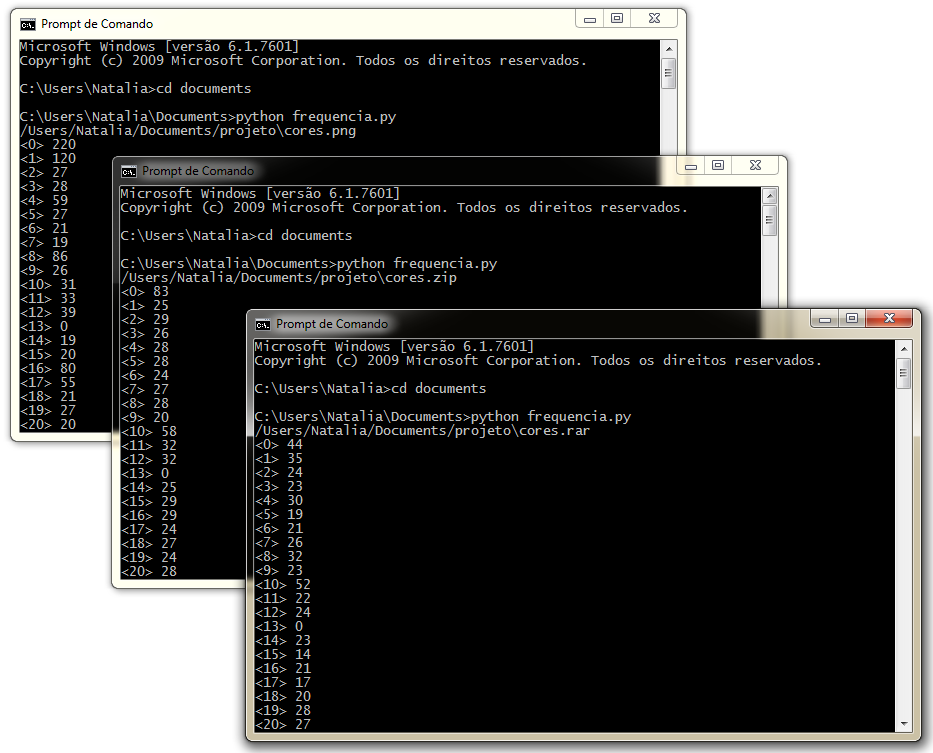


Figura 24: Compactação de arquivo .*png*

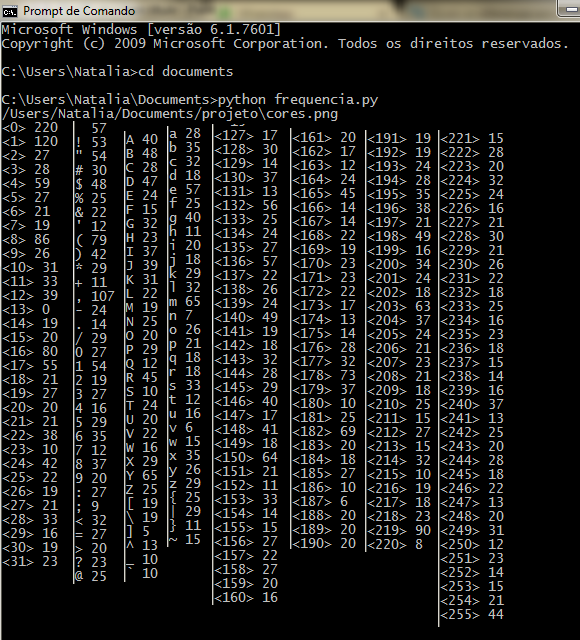


Figura 25: Resultado da contagem de bytes para o arquivo .*png*

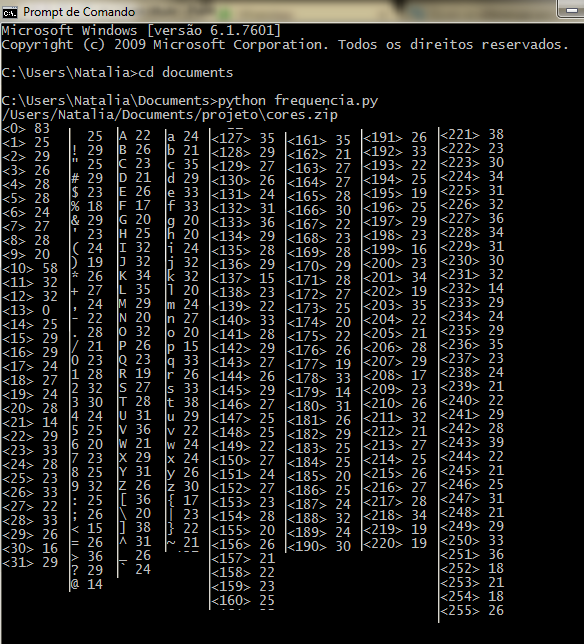


Figura 26: Resultado da contagem de bytes para o arquivo .*png* compactado em formato .*zip*

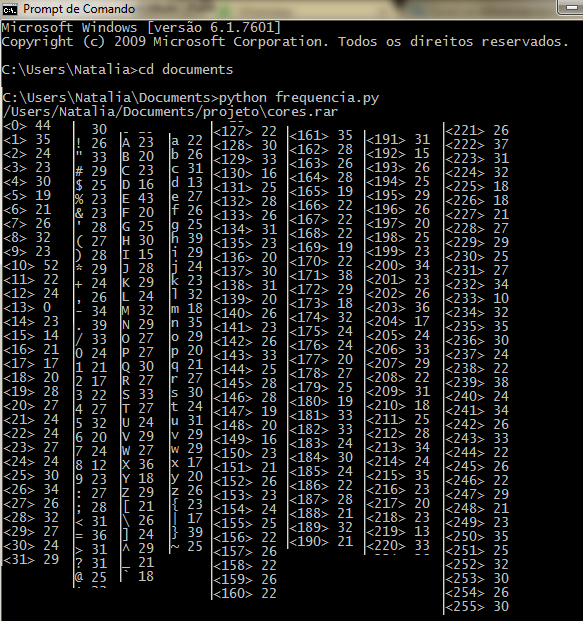


Figura 27: Resultado da contagem de bytes para o arquivo .*png* compactado em formato .*rar*

Na Figura 23: Arquivo .*png* original e compactados é possível perceber que quando o arquivo .*png* foi compactado não houve redução de tamanho e o arquivo que possuía 7KB se manteve com os 7KB no formato .*zip* e no formato .*rar*, isto não necessariamente se deve ao formato possuir baixo poder de compressão e sim aos pixels da imagem, que, por ser muito colorida, minimiza os bytes que podem ser suprimidos.

Na Figura 25: Resultado da contagem de bytes para o arquivo .*png*, Figura 26: Resultado da contagem de bytes para o arquivo .*png* compactado em formato .*zip* e Figura 27: Resultado da contagem de bytes para o arquivo .*png* compactado em formato .*rar* escolhemos os caracteres *‘,*’ e ‘*m’*, que foram identificados com maior frequência no formato .*png*, para realizarmos a análise. No formato .*png* a *‘,*’ apareceu 107 vezes, para 24 no .*zip* e 26 no .*rar*; já o caractere ‘*m’*, a frequência encontrada foi de 65 aparições no arquivo .*png* para 24 no .*zip* e 18 no .*rar*.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***.png*** | ***.zip*** | ***.rar*** | **Frequência de Bytes: Arquivo original/Compactado** | |
| ***.zip*** | ***.rar*** |
| ***,*** | 107 | 24 | 26 | 4,46 | 4,11 |
| ***m*** | 65 | 24 | 18 | 2,7 | 3,61 |

De acordo com o que foi identificado para o arquivo png, podemos concluir que, apesar da frequência de Bytes ter diminuído em média 4 vezes nos caracteres analisados para ambos os formatos, o tamanho dos arquivos se manteve constante, sendo assim, ainda que a compressão não tenha sido capaz de reduzir o tamanho da imagem a frequência de Bytes de alterou.

* .*mp3*:

O arquivo .*mp3* utilizado para a análise da contagem de frequência foi a música *Apenas mais uma de amor*, do Lulu Santos.

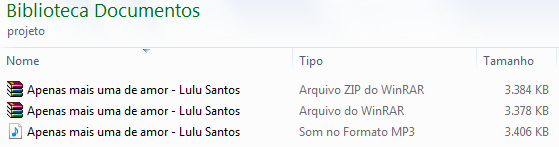


Figura 28: Arquivo .mp3 original e compactados

Os testes foram realizados para os três formatos do arquivo, conforme Figura 29: Compactação de arquivo .*mp3*.

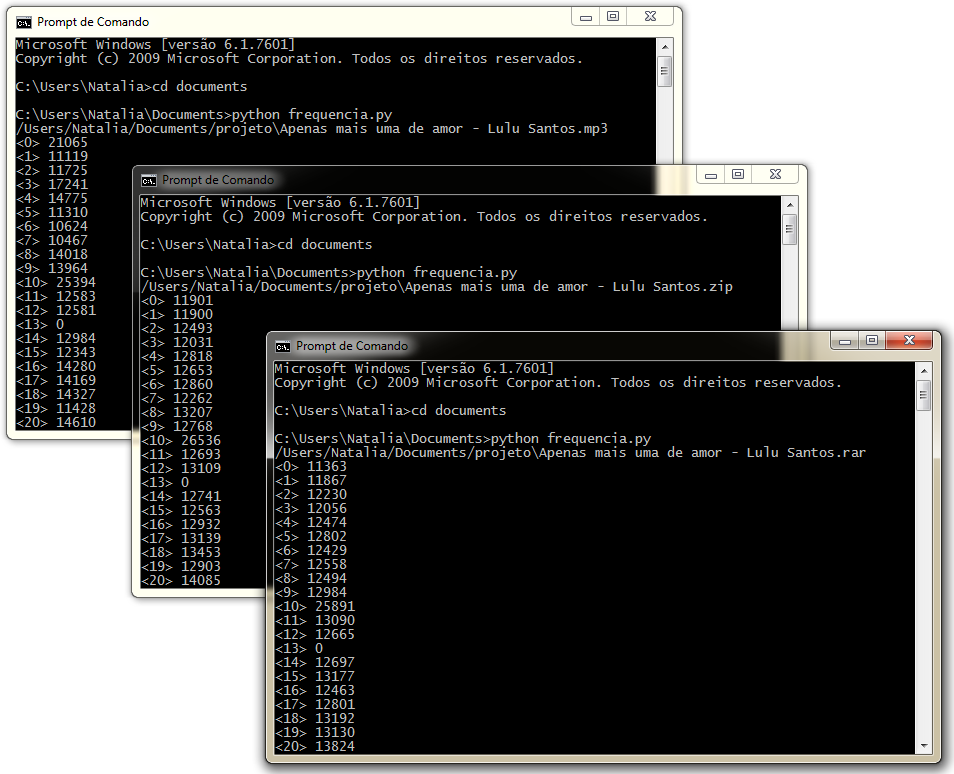


Figura 29: Compactação de arquivo .*mp3*

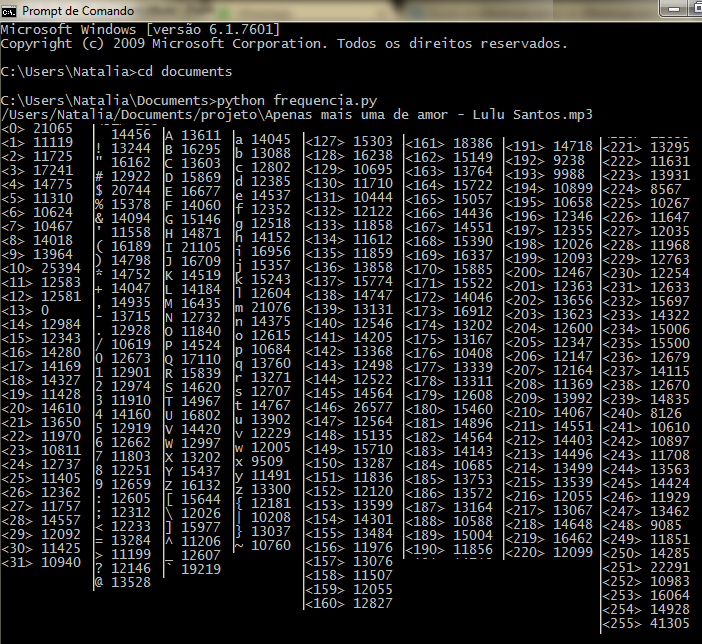


Figura 30: Resultado da contagem de bytes para o arquivo .*mp3*

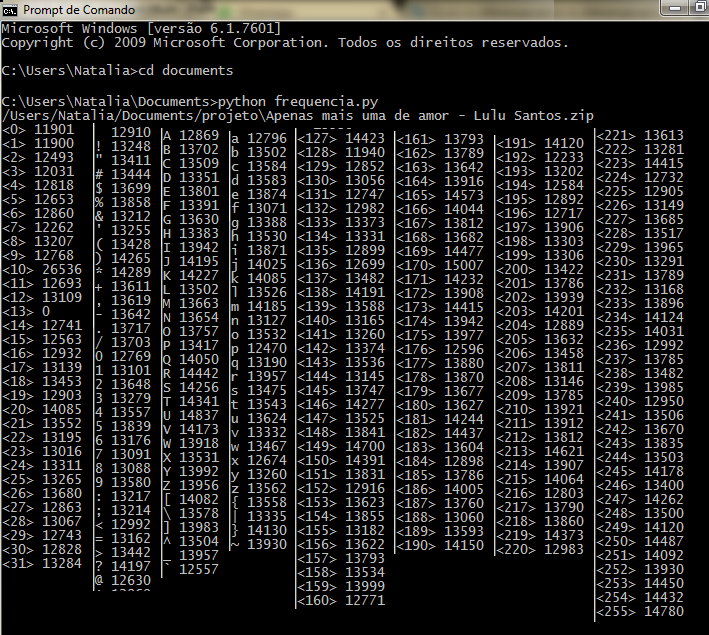


Figura 31: Resultado da contagem de bytes para o arquivo .mp3 compactado em formato .*zip*

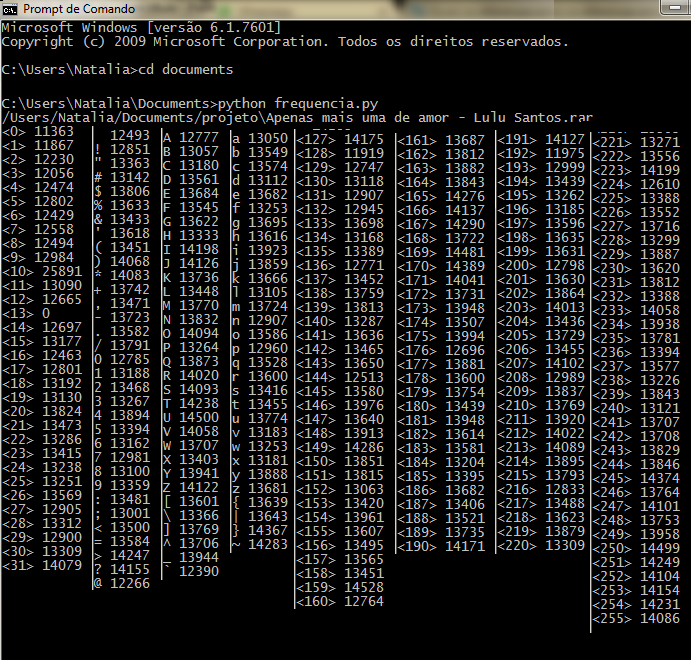


Figura 32: Resultado da contagem de bytes para o arquivo .mp3 compactado em formato .*rar*

Para a compressão de um arquivo .*mp3*, identificamos através da Figura 28: Arquivo .mp3 original e compactados, que apesar do formato já ser compactado, houve compressão e diminuição do tamanho dos arquivos, que originalmente tinha 3406KB e passaram a ter 3384KB no formato .*zip* e 3378KB no formato .*rar*. Ainda assim, o valor reduzido é quase insignificante para o tamanho total do arquivo.

Nos testes de frequência de bytes, identificados na Figura 30: Resultado da contagem de bytes para o arquivo .*mp3*, Figura 31: Resultado da contagem de bytes para o arquivo .mp3 compactado em formato .*zip* e Figura 32: Resultado da contagem de bytes para o arquivo .mp3 compactado em formato .*rar*, foram escolhidos os caracteres ‘*Q’* e ‘*m’*, que tiveram uma redução significativa se comparada à redução de tamanho total dos arquivos. Para o caractere ‘*Q’*, que aparecia 17110 vezes no arquivo original, passou a aparecer 14050 vezes no arquivo .*zip* e 13873 vezes no .*rar*. Para o caractere ‘*m’*, a incidência de 21076 vezes no arquivo principal caiu para 14185 vezes no .*zip* e 13724 vezes no .*rar*.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***.png*** | ***.zip*** | ***.rar*** | **Frequência de Bytes: Arquivo original/Compactado** | |
| ***.zip*** | ***.rar*** |
| ***Q*** | 17110 | 14050 | 13873 | 1,22 | 1,23 |
| ***m*** | 21076 | 14185 | 13724 | 1,48 | 1,53 |

Na tabela acima é possível perceber que, apesar do arquivo ter sido comprimido, a diferença de tamanho é extremamente insignificante, considerando que os caracteres analisados não atingiram redução de nem 50% da frequência.

* .*mp4*:

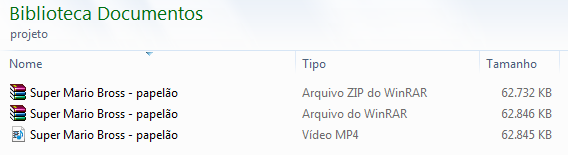


Figura 33: Arquivo .*mp4* original e compactados

Os testes foram realizados para os três formatos do arquivo, conforme Figura 34: Compactação de arquivo .*mp4*.

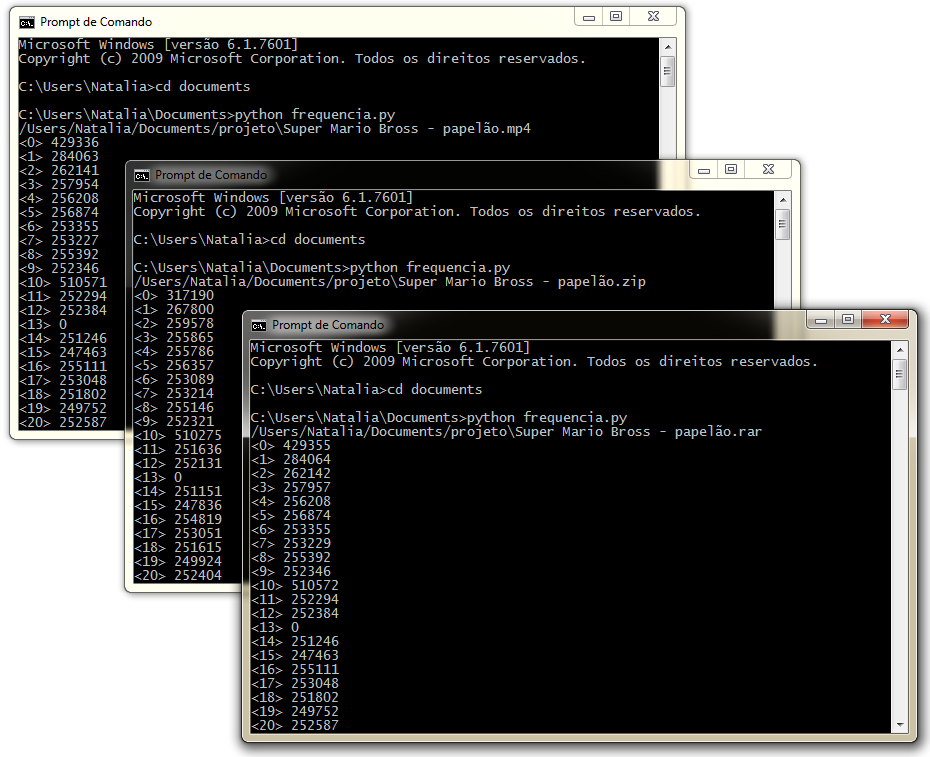


Figura 34: Compactação de arquivo .*mp4*

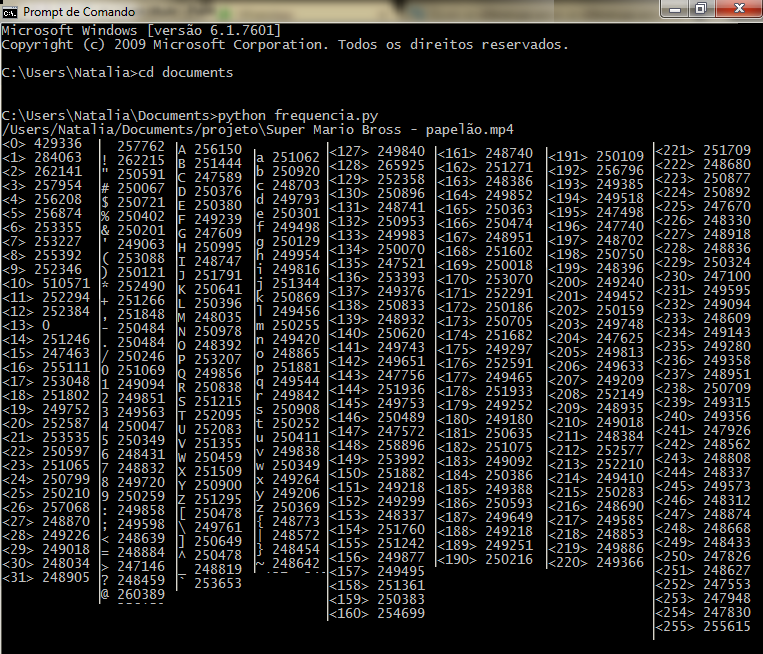


Figura 35: Resultado da contagem de bytes para o arquivo .*mp4*

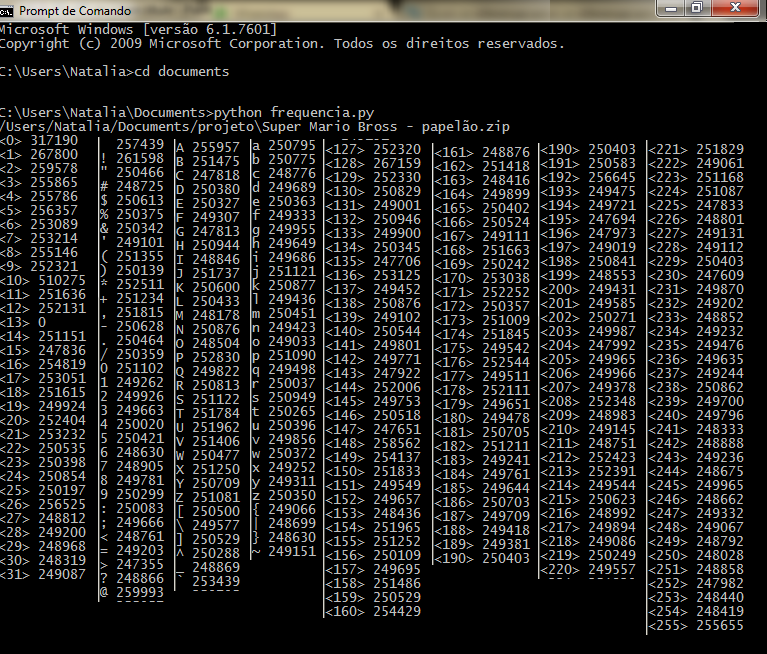


Figura 36: Resultado da contagem de bytes para o arquivo .*mp4* compactado em formato .*zip*

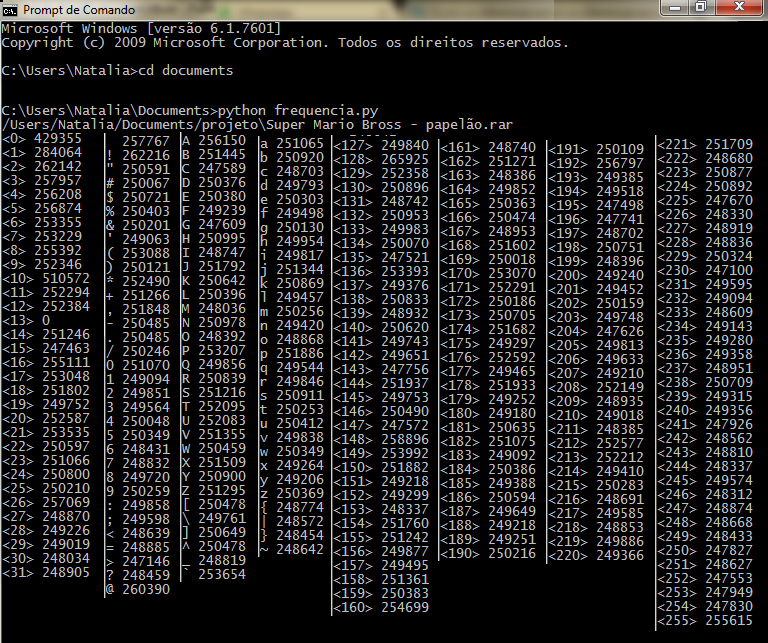


Figura 37: Resultado da contagem de bytes para o arquivo .*mp4* compactado em formato .*rar*

Com o auxílio da Figura 33: Arquivo .*mp4* original e compactados é possível perceber que, apesar da compressão do arquivo .*mp4* para o formato .*zip* ter gerado uma compressão mínima, a compressão do mesmo para o formato .*rar*, além de não ter reduzido seu tamanho, o tornou 1KB maior do que no formato original.

Na análise da frequência de bytes, foram utilizados os caracteres *‘!*’ e *‘@*’, que, como podemos ver na Figura 35: Resultado da contagem de bytes para o arquivo .*mp4*, Figura 36: Resultado da contagem de bytes para o arquivo .*mp4* compactado em formato .*zip* e Figura 37: Resultado da contagem de bytes para o arquivo .*mp4* compactado em formato .*rar*, possuíam 262215 e 260389 aparições no arquivo principal, respectivamente; estes valores foram reduzidos para 261598 e 259993 aparições para os caracteres no arquivo .*zip* e sofreram um aumento de um byte para cada caractere, que passaram a ter frequência de 262216 e 260390 vezes, respectivamente no arquivo .*rar*.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***.png*** | ***.zip*** | ***.rar*** | **Frequência de Bytes: Arquivo original/Compactado** | |
| ***.zip*** | ***.rar*** |
| ***!*** | 262215 | 261598 | 262216 | 1 | 0,99 |
| ***@*** | 260389 | 259993 | 260390 | 1 | 0,99 |

Com o auxílio da tabela acima, podemos verificar que, pelo fato de o formato de arquivo .*mp4* ser naturalmente comprimido, as diferenças geradas na frequência dos bytes em cada arquivo se torna insignificante tanto para a diminuição, quanto para o aumento de frequência de cada caractere.

**Conclusão**

Após análise da frequência dos bytes nos formatos de arquivo .*txt, .pdf, .jpg, .png, .mp3, .mp4* e suas compressões para os formatos .*zip* e .*rar*, pudemos perceber que nem todo formato de arquivo é suficientemente compactável, e que nem sempre a frequência dos bytes se torna menor em um arquivo após sua compactação. Identificamos também que, para a maioria dos formatos, nos testes realizados, a compactação para .*rar* foi superior, tornando-a melhor no quesito compressão, porém, para o único formato (.*mp4*) onde o .*rar* não teve maior índice de compressão, ele também tornou o arquivo maior que o formato original.

Também foi identificada uma semelhança para as compactações de imagens, onde em ambos os formatos (.*jpg* e .*png*) as compactações tiveram o mesmo tamanho (em .*jpg* a redução de 1KB e em .*png* sem reduções).

Considerando que só foram realizados testes com um arquivo de cada formato, para que a pesquisa tenha maior garantia é recomendado realizar testes com mais de um arquivo do mesmo formato, e também, como futuros trabalhos, seria interessante o estudo sobre outros formatos menos utilizados, tanto para arquivos quanto para formatos de compactação, ou até mesmo a comparação entre arquivos de mesmo tipo, seja ele texto, imagem, áudio ou vídeo.