**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: .png, .jpeg, .txt, .pdf, .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 2, utilizando a IDE do *Visual Studio Code*. O código encontra-se disponível no *Github* descrito no rodapé do relatório.

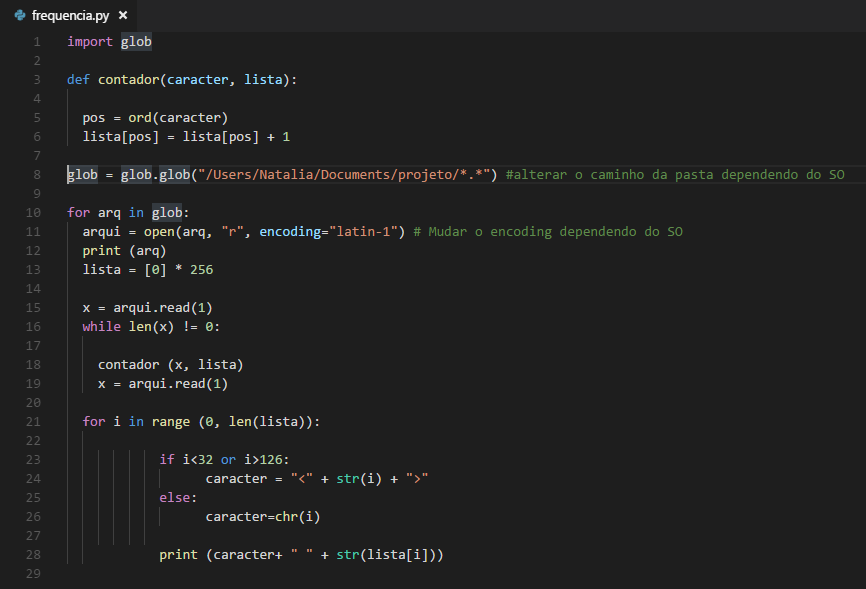


Figura 1: Contador de frequência de Bytes

Para o código da Figura 1 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.

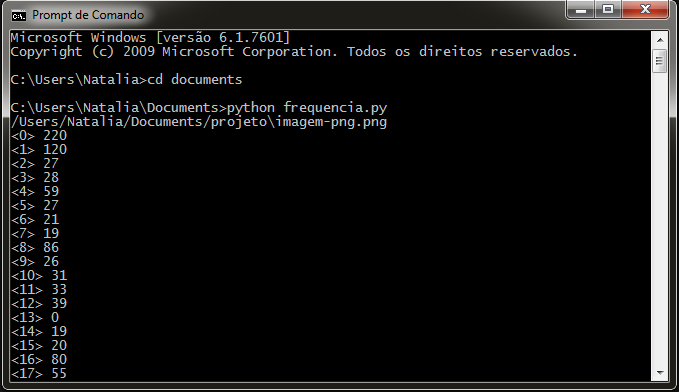


Figura 2: Exemplo de saída no console

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

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

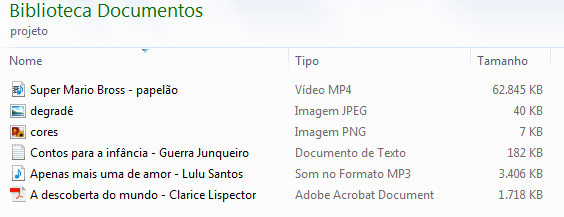


Figura 3: Arquivos nos formatos selecionados

Para cada arquivo foram utilizados os padrões de compactação abaixo descritos nas Figuras 4 e 5.

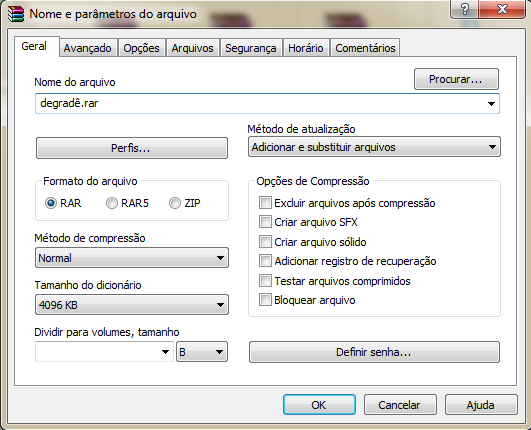


Figura 4: Especificações de compactação *.rar*

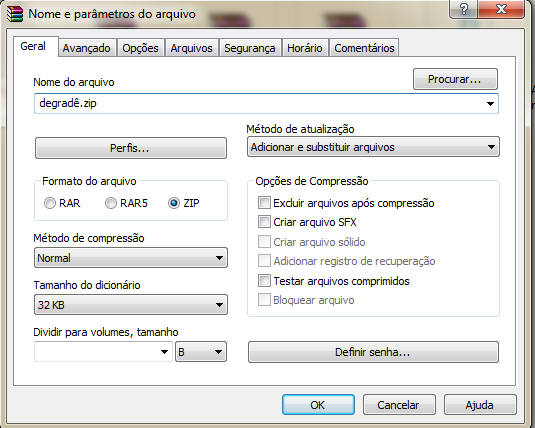


Figura 5: Especificações de compactação *.zip*

Todos os arquivos foram compactados, conforme a Figura 5.



Figura 6: Arquivos originais e compactados

**Fase 3 (três) – Testes:**