Cifra de César e a frequência das letras: Uma solução estatística

Caesar cipher and letter frequency: A statistical solution

Matheus da Silva Santos

Instituto Federal do Mato Grosso (IFMT) - Pontes e Lacerda, MT – Brazil

silva.matheus1@estudante.ifmt.edu.br

Abstract. *This work explores the use of letter frequency analysis in solving substitution ciphers, using the Caesar cipher as the basis of the study. Using a combination of literary texts and musical lyrics, an algorithm was developed to normalize characters and calculate the frequencies of each letter. From the comparison between the frequency observed in a cipher text and the expected frequencies, it was possible to decipher the message by statistical approximation. The results demonstrate that statistical analysis is an efficient and practical solution for breaking simple ciphers.*

Keywords: Caesar Cipher, Programming, Statistics.

Resumo. *Este trabalho explora o uso da análise de frequência de letras na resolução de cifras de substituição, utilizando a cifra de César como base do estudo. Utilizando uma combinação de textos literários e letras musicais, foi desenvolvido um algoritmo para normalizar caracteres e calcular as frequências de cada letra. A partir da comparação entre a frequência observada em um texto cifrado e as frequências esperadas, foi possível decifrar a mensagem por aproximação estatística. Os resultados demonstram que a análise estatística é uma solução eficiente e prática para quebrar cifras simples.*

Palavras-chave: Cifra de César, Programação, Estatística..

1. Introdução

Técnicas para ocultar uma mensagem e transmiti-la de maneira secreta têm sido usadas há muitos anos, sendo algumas bem inventivas.[Costa & Figueiredo, 2010] Considerado um dos primeiros e mais simples métodos de cifragem, a ‘cifra de César’ ou ‘cifra de troca’, consiste em substituir um caractere por outro usando uma regra, que deve ser de conhecimento do emissor e do receptor, para que seja decifrada no recebimento.

2. Objetivo

O objetivo deste trabalho é explorar, de forma simples, o funcionamento de um dos métodos clássicos de criptografia e sua base matemática. Mostrando como as tecnologias disponíveis hoje podem resolver desafios que, em determinado momento da história, foram considerados seguros em termos de comunicação cifrada.

3. Materiais e Métodos

Para a elaboração deste documento, foi necessário realizar a extração de texto de diferentes tipos de produções. Foram utilizadas quatro fontes, totalizando 3.200 páginas. Essas fontes incluem três livros e uma coletânea de letras musicais do grupo Racionais

MC 's. Os livros são: Anna Karenina de Liev Tolstói, Dom Casmurro de Machado de Assis, e Novo Diccionario da Língua Portuguesa de Cândido de Figueiredo. Para realizar essa coleta, foi desenvolvido um algoritmo em Python que extrai o texto da fonte indicada, realiza a normalização dos caracteres e calcula com que frequência cada letra é utilizada nos textos base.

Para representar matematicamente a cifra, as letras precisam ser convertidas em números: A = 0, B = 1, C = 2, ..., Z = 25. A substituição é fixa e realizada por um valor N, de modo que a cifragem é calculada como (X + N), onde X representa a letra a ser substituída. Se o valor da soma ultrapassar o intervalo de 0 a 25, subtrai-se 26, o processo de decifração é similar, sendo necessário conhecer o valor de N para realizar (X - N).

4. Resultados

Com base na análise do material de referência foi obtida uma tabela de frequência.(Figura 1)

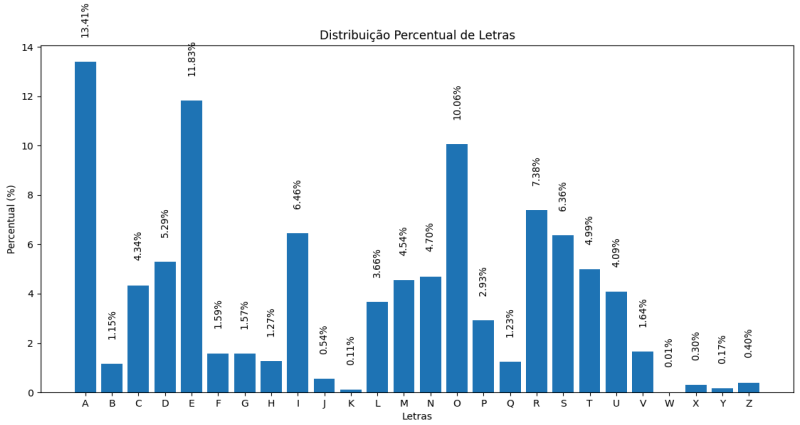

Figura 1.

Utilizando o seguinte texto cifrado com uma troca de N = 3: “Xpd qrlwh, txdqgr Udlpxqgr txlv vxuudu Edudqgãr, Shgur wrprx dv gruhv gr qhjulqkr h urodudp qd oxwd pdlv vhqvdflrqdo d txh dv duhldv gr fdlv mdpdlv dvvlvwludp. Udlpxqgr hud pdlv dowr h pdlv yhokr. Sruép Shgur Edod, r fdehor orlur yrdqgr, d flfdwulc yhuphokd qr urvwr, hud gh xpd djlolgdgh hvsdqwrvd h ghvgh hvvh gld Udlpxqgr ghlarx qãr vó d fkhild grv Fdslwãhv gd Duhld, frpr r suósulr duhdo. Hqjdmrx whpsrv ghsrlv qxp qdylr.”, foi obtido um gráfico de frequência.(Figura 2)

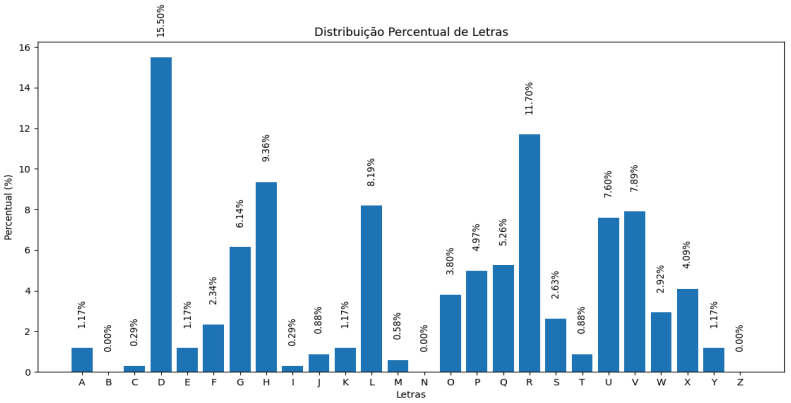

Figura 2.

Ao fazer a comparação dos gráficos e substituições podemos chegar ao resultado da remoção da cifra por aproximação estatística, concluindo que a mensagem cifrada era: "UMA NOITE QUANDO RAIMUNDO QUIS SURRAR BARANDAO PEDRO TOMOU AS DORES DO NEGRINHO E ROLARAM NA LUTA MAIS SENSACIONAL A QUE AS AREIAS DO CAIS JAMAIS ASSISTIRAM RAIMUNDO ERA MAIS ALTO E MAIS VELHO POREM PEDRO BALA O CABELO LOIRO VOANDO A CICATRIZ VERMELHA NO ROSTO ERA DE UMA AGILIDADE ESPANTOSA E DESDE ESSE DIA RAIMUNDO DEIXOU NAO SO A CHEFIA DOS CAPITAES DA AREIA COMO O PROPRIO AREAL ENGAJOU TEMPOS DEPOIS NUM NAVIO"

5. Considerações Finais

Os resultados obtidos comprovam que a análise estatística da frequência das letras é uma solução simples e poderosa para a resolução de cifras de substituição, o desenvolvimento de um algoritmo para a realização dessa análise se demonstrou eficiente para a realização deste tipo de estudo. Devido a sua fragilidade, esse tipo de criptografia se tornou obsoleto, mas pode ser usado como exemplo prático para aplicação de conceitos matemáticos, podendo ser uma ótima ferramenta em sala de aula.

Referências

Paar, C. and Pelzl, J. (2010) Understanding Cryptography: A Textbook for Students and Practitioners, Springer, Germany.

Gregório, M. B. dos S., Barvinski, C. A., Odakura, V., e Sanabria, L. (2014) "CriptoMat1: Ensinando Matemática Utilizando Conceitos de Criptografia - Cifra de César e César Estendida."

Costa, C., & Figueiredo, L. M. (2010). *Introdução à Criptografia*.Fundação CECIERJ, Rio de Janeiro.