Departamento de Ciência da Computação
Universidade de Brasília

Cifra de Vigenère:
implementação, decifragem e quebra

Nome: Gabriel Mendes Ciriatico Guimarães
Matrícula: 202033202

Implementação da cifragem

Para começar a cifragem, é necessário definir um alfabeto de caracteres possíveis de serem cifrados, bem como os números que os representam. Esses números devem ser uma sequência finita de inteiros. Caso algum caractere que esteja no texto a ser cifrado não esteja nesse alfabeto, ele não será cifrado, sendo mantido na mesma posição e forma do texto não cifrado.

Esse alfabeto é utilizado para filtrar o texto dado, com a criação de um texto temporário desprovido de qualquer caractere indecifrável, cujo tipo e posição são armazenados para posterior inserção no texto cifrado.

Supondo o texto abaixo para cifrar:

ALICE ANN MUNRO É UMA ESCRITORA CANADENSE DE CONTOS, CONSIDERADA UMA DAS PRINCIPAIS ESCRITORAS DA ATUALIDADE EM LÍNGUA INGLESA. FOI AGRACIADA COM O PRÊMIO NOBEL DA LITERATURA EM 2013.

Temos que ele se transforma em:

ALICEANNMUNROUMAESCRITORACANADENSEDECONTOSCONSIDERADAUMA DASPRINCIPAISESCRITORASDAATUALIDADEEMLÍNGUAINGLESAFOIAGRACIADACOM OPRMIONOBELDALITERATURAEM

O projeto desenvolvido considera apenas letras de A a Z, maiúsculas. Portanto, todo texto que é inserido como dado de entrada no cifrador tem suas letras transformadas em letras maiúsculas. Símbolos, números e letras diferentes (como letras acentuadas) são mantidos na sua forma original, já que não estão no dicionário de caracteres cifráveis.

Esse texto temporário é usado para a criação do *key stream*, onde a chave é repetida para ter o mesmo tamanho do texto cifrável. Usando a chave VIDAQUERIDA para o texto anterior, por exemplo, temos:

VIDAQUERIDAVIDAQUERIDAVIDAQUERIDAVIDAQUERIDAVIDAQUERIDAVIDAQ UERIDAVIDAQUERIDAVIDAQUERIDAVIDAQUERIDAVIDAQUERIDAVIDAQUERIDAVIDA QUERIDAVIDAQUERIDAVID

Com o *keystream* e o texto cifrável, passa-se para a cifragem propriamente dita, feita com base no cálculo do módulo, que define o caractere cifrado. O algoritmo de cifragem pode ser facilmente ilustrado através de uma tabela de Vegenère, onde temos como nome de cada coluna cada caractere do alfabeto, e na primeira célula de cada linha, também cada caractere do alfabeto, de forma ordenada.

Considerando um caractere qualquer na posição i no texto original, precisamos localizar a coluna ki e a linha cuja primeira célula seja o mesmo caractere daquele na posição i no *keystream*, na linha li. O caractere cifrado está na posição (ki, li). A cifragem precisa repetir esse processo para todo caractere entre 0 e n, onde n é o tamanho do texto.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
B	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
C	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B
D	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C
E	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D
F	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E
G	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F
H	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G
I	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H
J	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I
K	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J
L	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K
M	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L
N	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M
O	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N
P	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Q	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
R	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
S	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
T	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
U	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
V	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
W	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
X	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
Y	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Z	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y

Para realizar esse processo computacionalmente, sem recurso visual, é possível utilizar a operação de módulo. Para isso, transpomos o caractere na posição i do texto original em X posições, onde X é o número do caractere na posição i do *keystream*. Assim, num alfabeto A-Z, o caractere L (de número 11) cifrado com I (8) torna-se T (19), já que temos:

$$(11 + 8) \bmod 26 = 19$$

Com cada caractere cifrável cifrado, pega-se a tabela de caracteres que não puderam ser cifrados, colocando-os em suas posições originais. Assim, temos:

VTLCU URE UXNMW É XMQ YWTZLTJZD CQHEUMQSZ LH CEHXFA, FOIALDULEUI XMV LDS FLMEKLPVQV EIWVZBRRVA GA QNYRTLDVLH EC FÍRXCD IIOOEIU. JFQ DGMIFIQXE TWP O KZÊPIE HSSMO DV TLTULEKCUA ZU 2013.

Implementação da decifragem

O processo de decifragem segue o mesmo esquema do de cifragem, com exceção da função de cifra, que aqui é o inverso - isto é, no cálculo do módulo retoma-se a posição do caractere original subtraindo o valor do caractere cifrado com o valor do caractere correspondente no *keystream*.

Assim, em um primeiro momento o texto cifrado é filtrado com base no alfabeto dado, a partir do qual é gerado um *keystream* com o mesmo tamanho do texto cifrado. Cada caractere na posição i no texto cifrado e no *keystream* passa pela função de decifragem, resultando no caractere original.

Com o caractere cifrado T (19) e o caractere da chave I (8), obtemos o caractere decifrado L (11):

$$(19 - 8) \bmod 26 = 11$$

Quebra da cifra

A quebra da cifra é feita através da análise de frequência dos caracteres, exigindo para isso uma tabela de frequência de caracteres. Observe que isso implica que a pessoa saiba: o idioma em que está escrito o texto; e os caracteres cifráveis.

Para a quebra, existem 3 etapas principais: a identificação do tamanho provável da chave, o que é feito com a análise da ocorrência de bigramas repetidos; a análise de frequência de caracteres com base nos tamanhos adivinhados da chave; e a escolha de qual é a chave mais provável dentre as chaves de diferentes tamanhos encontradas.

O tamanho da chave pode ser encontrado com base nos bigramas que se repetem no texto. Se o bigrama “WP” aparece 2 vezes em um texto cifrado, por exemplo, é provável que a repetição exista no próprio texto original, com o tamanho do espaço entre as repetições sendo um fator do tamanho real da chave. Assim, se “WP” aparece na posição 10 e, depois, na posição 32, é possível que a chave tenha o tamanho de 1, 2, 4, 6 ou 12 caracteres.

É feita uma tabela com todas as repetições entre bigramas no texto decifrado, com todos os fatores possíveis. Os fatores que aparecem mais têm mais chance de serem o tamanho real da chave.

Com um tamanho de chave escolhido, passa-se para a análise de frequência. Supondo que escolhemos uma chave de tamanho k, então a cada k caracteres um mesmo caractere que cifra é usado no texto original. Fazendo a contagem de caracteres de k em k, começando em i (onde 0 <= i < k), podemos montar uma tabela de frequência de caracteres. Essa tabela pode ser comparada com a tabela de frequência de caracteres de alguma língua: a ordem de caracteres que resultar em maior semelhança com a tabela de frequência real revelará o caractere na posição i da chave.

A semelhança entre as tabelas de frequência pode ser observada de forma visual através de gráficos. No entanto, para tirar essa responsabilidade do usuário, recorro ao cálculo da similaridade de cosseno, que permite mensurar quão semelhantes são duas sequências numéricas. Assim, transformo as porcentagem em sequências numéricas e gero p tabelas de frequência com base na tabela de frequência de caracteres cifrados, onde p é o tamanho do alfabeto cifrável. Essas sequências variam a posição das letras de 1 em 1, sendo comparadas à sequência da tabela real de frequência.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

O cálculo é feito com:

A tabela com maior similaridade por cosseno revela qual é o caractere na posição i da chave, que é o caractere que inicia a sequência dada.

Observe que a análise de tamanhos possíveis de chave pode resultar em mais de 1 possível tamanho. Assim, temos n possíveis chaves, com tamanho variável. Para que o usuário não tenha que ver se cada possível chave decifra o texto corretamente, isso é feito pelo quebrador de cifra.

O quebrador verifica cada chave possível, checando em um dicionário se cada palavra do texto decifrado existe naquela língua. É feito, então, para cada chave, o cálculo da proporção de palavras que existem. A chave que tiver a maior proporção tem mais chances de ser a chave correta, portanto é ela a selecionada.

Esse processo permite tirar do usuário a tarefa de selecionar tamanho de chave e, depois, verificar qual das possíveis chaves retorna o texto com mais sentido para a leitura.