UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS DE CRATEÚS

Trabalho 01

Criptoanálise por Análise de Frequência • 2023.1 • 10/04/2022

Organização das pessoas

- O trabalho será realizado em equipe, as quais devem ser formadas por **seis** membros.
- Cada equipe deve se dividir em duas, em que **três** pessoas serão responsáveis por **cifrar** e as outras três serão responsáveis por **decifrar**.

Realização do trabalho

1ª etapa.

- A equipe deve escolher um idioma (português ou inglês)
- Uma parte da equipe usa uma cifra de substituição com **palavra-chave** para cifrar um texto com aproximadamente **1000** palavras.
- A equipe deve estudar/pesquisar como usar uma cifra de substituição com palavra-chave.
- Poderá ser desenvolvido um algoritmo em qualquer linguagem para o processo de cifração

2ª etapa.

- A outra parte da equipe deve fazer o papel de criptoanalistas e tentar decifrar o texto por análise de frequência.

Observações sobre criptoanálise

- A técnica de análise de frequência não é uma receita infalível, pois as tabelas de frequências representam uma média.
- Para tentar valer a tabelas de frequência de cada idioma, a equipe que fará a cifração deve escolher texto do cotidiano, como matéria de jornais do cotidiano. Não devem usar texto técnicos.
- Um exemplo de criptoanálise feito por Simon Singh em o Livro dos Código está anexo.

Orientações para a entrega

- A entrega deverá ser feita por meio de um seminário em sala de aula em 26/04/2023.
- Apenas os arquivos .java devem ser enviados.
- O conjunto dos arquivos .java deve ser compactado em formato .zip
- Obs.: Em caso de dificuldade para resolver a lista, solicitem ajuda ao monitor da disciplina.

Criptoanalisando um texto cifrado

PCQ VMJYPD LBYK LYSO KBXBJXWXV BXV ZCJPO EYPD KBXBJYUXJ LBJOO KCPK. CP LBO LBCMKXPV XPV IYJKL PYDBL, QBOP KBO BXV OPVOV LBO LXRO CI SX'XJMI, KBO JCKO XPV EYKKOV LBO DJCMPV ZOICJO BYS, KXUYPD: 'DJOXL EYPD, ICJ X LBCMKXPV XPV CPO PYDBLK Y BXNO ZOOP JOACMPLYPD LC UCM LBO IXZROK CI FXKL XDOK XPV LBO RODOPVK CI XPAYOPL EYPDK. SXU Y SXEO KC ZCRV XK LC AJXNO X IXNCMJ CI UCMJ SXGOKLU?'

OFYRCDMO, LXROK IJCS LBO LBCMKXPV XPV CPO PYDBLK

Imagine que inteceptamos esta mensagem cifrada. O desafio é decifrá-la. Nós sabemos que o texto original está em inglês, e que foi misturado de acordo com uma cifra de substituição monoalfabética, mas não temos idéia de qual seja a chave. Pesquisar todas as chaves possíveis não é prático, por isso devemos aplicar a análise de freqüências. O que se segue é um guia, passo a passo, para criptoanalisar o texto cifrado, mas se você se sentir confiante pode ignorá-lo e tentar sua própria criptoanálise independente.

A reação imediata de qualquer criptoanalista ao ver um texto cifrado desse tipo é analisar a freqüência com que ocorrem todas as letras, o que resulta na Tabela 2. Não é surpreendente que a freqüência das letras varie. A pergunta é: será que podemos identificar o que representa qualquer uma delas, baseado em suas freqüências? O texto cifrado é relativamente curto; assim, não podemos aplicar a análise de freqüências de um modo simples e direto. Seria ingenuidade presumir que a letra mais comum nesse texto cifrado, o **O**, representasse a letra mais comum no idioma inglês, o **e**, ou que a oitava letra mais freqüente no texto cifrado, **Y**, representasse o **h**, que é a oitava letra mais freqüente no inglês. Uma aplicação incondicional da análise de freqüências levaria a uma mistura de letras sem sentido. Por exemplo, a primeira palavra **PCQ** seria decifrada como **aov**.

Contudo, podemos começar voltando nossa atenção apenas para as três letras que aparecem mais de trinta vezes no texto cifrado, ou seja: **O**, **X** e **P**. É razoavelmente seguro supor que as letras mais comuns no texto cifrado provavelmente representem as letras mais comuns do alfabeto inglês, mas não estejam necessariamente na ordem correta. Em outras palavras, não

temos certeza de que O = e, X = t, e que P = a, mas podemos tentar a suposição de que:

O = e, t ou a, X = e, t ou a, P = e, t ou a.

Tabela 2 Análise de freqüência da mensagem cifrada

Letra	Freqüência		Letra	Freqüência	
	Ocorrência	Porcentagem		Ocorrência	Porcentagem
A	3	0,9	N	3	0,9
B	25	7,4	O	38	11,2
C	27	8,0	P	31	9,2
D	14	4,1	Q	2	0,6
E	5	1,5	R	6	1,8
F	2	0,6	S	7	2,1
G	1	0,3	T	0	0,0
H	0	0,0	U	6	1,8
I	11	3,3	V	18	5,3
J	18	5,3	W	1	0,3
K	26	7,7	X	34	10,1
L	25	7,4	Y	19	5,6
M	11	3,3	Z	5	1,5

De modo a prosseguir com confiança e determinar a identidade das letras mais comuns, O, X e P, precisamos empregar uma forma mais sutil de análise de freqüências. No lugar de simplesmente contar a freqüência com que aparecem as três letras, devemos voltar nossa atenção para com que freqüência elas aparecem ao lado das outras letras. Por exemplo, será que a letra O aparece antes ou depois de várias letras ou teria ela a tendência a ficar ao lado de algumas letras em especial? A resposta a esta pergunta nos dará uma boa indicação de se O representa uma vogal ou uma consoante. Se O for uma vogal, ela aparecerá antes e depois da maioria das letras, mas se for uma consoante ela tenderá a evitar a maioria das letras. Por exemplo, a letra e pode aparecer antes e depois de todas as outras letras, mas a letra t é raramente vista antes ou depois de b, d, g, j, k, m, q, ou v.

A tabela a seguir pega as três letras mais freqüentes no texto cifrado O, X e P e faz uma lista da freqüência com que cada uma aparece antes ou depois de cada letra. Por exemplo, o O aparece antes do A somente em uma ocasião, mas nunca aparece depois, dando o total de 1 no primeiro espaço. A letra O é vizi-

nha da maioria das letras, e existem somente sete que ela evita completamente, o que é representado pelos sete zeros na fileira do O. A letra X é igualmente sociável, porque ela também fica ao lado da maioria das letras e evita apenas oito delas. Contudo a letra P é muito menos amistosa. Ela tende a ficar ao lado de apenas umas poucas letras e evita 15 delas. A evidência sugere que O e X representam vogais, enquanto P é uma consoante.

	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
O	1	9	0	3	1	1	1	0	1	4	6	0	1	2	2	8	0	4	1	0	0	3	0	1	1	2
X	0	7	0	1	1	1	1	0	2	4	6	3	0	3	1	9	0	2	4	0	3	3	2	0	0	1
P	1	0	5	6	0	0	0	0	0	1	1	2	2	0	8	0	0	0	0	0	0	11	0	9	9	0

Agora devemos nos perguntar que vogais são representadas por O e X. Provavelmente se trata de **e** e **a**, as duas vogais mais populares do idioma inglês, mas será que O = **e** e X = a, ou é O = **a** e X = **e**? Um detalhe interessante no texto cifrado é que a combinação OO aparece duas vezes, enquanto XX não aparece nenhuma vez. E como as letras **ee** aparecem juntas com muito maior freqüência do que **aa** num texto em inglês, é provável que O = **e** e X = **a**.

Neste ponto conseguimos identificar com confiança duas das letras do texto cifrado. Nossa conclusão de que X = **a** é apoiada pelo fato de que o X aparece sozinho no texto cifrado, e o a é uma das duas únicas palavras do inglês formadas por uma única letra. Só uma outra letra aparece sozinha no texto cifrado e esta é o Y. Isto torna altamente provável que ela represente a outra palavra inglesa de uma letra só que é o **i**. Focalizar a atenção em palavras de uma só letra é um truque padrão da criptoanálise, e eu o incluí na lista de dicas criptoanalíticas do Apêndice B. Esse truque em particular funciona apenas porque este texto cifrado ainda mantém os espaços entre as palavras. Freqüentemente, um criptógrafo remove todos os espaços para tornar mais difícil aos interceptadores inimigos a decodificação da mensagem.

Embora tenhamos espaços entre as palavras, o truque seguinte também deve funcionar nos casos em que o texto cifrado foi unido numa linha contínua de caracteres. Esse truque nos permite identificar a letra **h** depois que tenhamos identificado a letra **e**. No idioma inglês a letra **h** aparece com freqüência antes da letra **e** (como em **the**, **then**, **they**, etc.), mas raramente ele ocorre depois do

e. A tabela a seguir mostra com que freqüência o **O**, que pensamos representar o **e**, aparece antes e depois de todas as letras do texto cifrado. A tabela sugere que o **B** representa o **h**, porque ele aparece antes do **O** em nove ocasiões, mas nunca depois dele. Nenhuma outra letra na tabela tem esse relacionamento assimétrico com o **O**.

	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
depois do O	1	0	0	1	0	1	0	0	1	0	4	0	0	0	2	5	0	0	0	0	0	2	0	1	0	0
antes do O	0	9	0	2	1	0	1	0	0	4	2	0	1	2	2	3	0	4	1	0	0	1	0	0	1	2

Cada letra do idioma inglês tem sua própria personalidade, que lhe é única e que inclui sua freqüência e relacionamento com as outras letras. É esta personalidade que nos permite estabelecer a verdadeira identidade de uma letra, mesmo quando ela foi disfarçada pela substituição monoalfabética.

Até aqui já foram identificadas com confiança quatro letras, **O** = **e**, **X** = **a**, **Y** = **i**, e **B** = **h**, assim podemos começar a substituir algumas das letras do texto cifrado por suas equivalentes no texto original. Eu devo manter a convenção de deixar as letras do texto cifrado em maiúsculas enquanto as letras do texto decodificado ficam em minúsculas. Isso vai nos ajudar a distinguir as letras que ainda precisamos identificar daquelas que já são conhecidas.

PCQ VMJiPD LhiK LiSe KhahJaWaV haV ZCJPe EiPD KhahJiUaJ LhJee KCPK. CP Lhe LhCMKaPV aPV IiJKL PiDhL, QheP Khe haV ePVeV Lhe LaRe CI Sa'aJMI, Khe JCKe aPV EiKKev Lhe DJCMPV ZeICJe hiS, KaUiPD: 'DJeaL EiPD, ICJ a LhCMKaPV aPV CPe PiDhLK i haNe ZeeP JeACMPLiPD LC UCM Lhe IaZReK CI FaKL aDeK aPV Lhe ReDePVK CI aPAiePL EiPDK. SaU i SaEe KC ZCRV aK LC AJaNe a IaNCMJ CI UCMJ SaGeKLU?'

eFiRCDMe, LaReK IJCS Lhe LhCMKaPV aPV CPe PiDhLK

Este passo simples nos ajuda a identificar várias outras letras, porque podemos adivinhar algumas das palavras do texto cifrado. Por exemplo, as palavras de três letras mais comuns no ingles são **the** e **and**, e elas são relativamente fáceis

de localizar — **Lhe**, que aparece seis vezes e **aPV**, que aparece cinco vezes. Portanto, **L** provavelmente representa **t**, enquanto **P** provavelmente representa **n** e **V** representa **d**. Podemos agora substituir essas letras no texto cifrado por seus verdadeiros valores:

> nCQ dMJinD thiK tiSe KhahJaWad had ZCJne EinD KhahJiUaJ thJee KCnK. Cn the thCMKand and liJKt niDht, Qhen Khe had ended the taRe CI Sa'aJMI, Khe JCKe and EiKKed the DJCMnd ZeICJe hiS, KaUinD: 'DJeat EinD, ICJ a thCMKand and Cne niDhtK i haNe Zeen JeACMntinD tC UCM the laZReK CI FaKt aDeK and the ReDendK CI anAient EinDK. SaU i SaEe KC ZCRd aK tC AJaNe a laNCMJ CI UCMJ SaGeKtU?'
>
> eFiRCDMe, taReK IJCS the thCMKand and Cne niDhtK

Depois que algumas letras já foram determinadas, a criptoanálise avança muito rapidamente. Por exemplo, a palavra no começo da segunda frase é **Cn**. Toda palavra tem uma vogal, de modo que **C** deve ser uma vogal. Existem apenas duas vogais que ainda não indentificamos, **u** e **o**; o **u** não se encaixa, portanto o **C** deve representar o **o**. Também temos a palavra **Khe**, o que implica que o **K** deve representar ou **o t** ou o **s**. Mas nós já sabemos que **L** = **t**, assim se torna claro que o **K** = **s**. Tendo identificado essas duas letras, nós as inserimos no texto cifrado e aparece então a frase **thoMsand and one niDhts**. Um palpite razoável é de que se trata do título **thousand and one nights** (Mil e uma noites), e parece provável que a última linha esteja nos dizendo que se trata de uma passagem de *Tales from the Thousand and One Nights*. Isto implica que **M** = **u**, **I** = **f**, **j** = **r**, **D** = **g**, **R** = **l**, e **S** = **m**.

Podíamos continuar tentando identificar as outras letras por palpite, mas no lugar disso vamos dar uma olhada no que sabemos sobre o alfabeto original e o alfabeto cifrado. Esses dois alfabetos formam a chave e foram usados pelo criptógrafo de modo a fazer a substituição da mensagem misturada. Ao identificar a verdadeira identidade das letras no texto cifrado, nós efetivamente estivemos descobrindo os detalhes do alfabeto cifrado. Um resumo de nossas conquistas, até agora, é dado pelos alfabetos, normal e cifrado, a seguir.

Alfabeto original	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
Alfabeto cifrado	X – – V O I D B Y – – R S P C – – J K L M – – – – –

Examinando o alfabeto cifrado parcial, podemos completar a criptoanálise. A seqüência **VOIDBY** do alfabeto cifrado sugere que o criptógrafo escolheu uma frase-chave como base para o seu código. Um trabalho de suposição é suficiente para sugerir que a frase-chave pode ser **A VOID BY GEORGES PEREC**, que é reduzida a **AVOIDBYGERSPC** depois da remoção dos espaços e das repetições. Depois as letras continuam em ordem alfabética, omitindo-se qualquer uma que tenha aparecido na frase-chave. Neste caso em particular o criptógrafo teve o cuidado incomum de não começar a frase-chave no início do alfabeto cifrado, e sim três letras depois do começo. Isto é possível porque a frase-chave começa com a letra **A** e o criptógrafo queria evitar codificar **a** como **A**. Finalmente, tendo determinado o alfabeto cifrado completo, podemos decodificar todo o texto cifrado e a criptoanálise está completa.

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

Now during this time Shahrazad had borne King Shahriyar three sons. On the thousand and first night, when she had ended the tale of Ma'aruf, she rose and kissed the ground before him, saying: 'Great King, for a thousand and one nights I have been recounting to you the fables of past ages and the legends of ancient kings. May I make so bold as to crave a favour of your majesty?'

Epilogue, *Tales from the Thousand and One Nights*