

Descriptografia de textos binários criptografados em Cifra de César e Cifra de Substituição

Victor Emmanuel Susko Guimarães
Cristiano Augusto Dias Mafuz

I. Introdução

Criptografia é o termo dado ao conjunto das técnicas empregadas para a codificação de um texto, visando torná-lo ininteligível para indivíduos que desconhecem o processo de descodificação (chamado de descriptografia). Nesse aspecto, a computação teve como uma de suas bases a criptografia e descriptografia de mensagens, a exemplo da máquina Enigma, dos anos 1920, utilizada para codificar mensagens alemãs durante o período do nazismo. Nos dias atuais, as criptografias possuem aplicações na segurança de dados de incontáveis sistemas da informação, desde dados bancários até credenciais em aplicativos e websites e, por isso, a descriptografia é recorrentemente utilizada de maneira benéfica para garantir que o algoritmo de codificação é confiável e assegura as informações dos usuários.

Uma das criptografias famosas é a Cifra de César, que consiste em rotacionar as letras do alfabeto, seguindo um deslocamento fixo e circular do alfabeto original, que pode ser feito da esquerda para a direita e vice-versa. Assim, cada letra do alfabeto rotacionada possui uma letra correspondente no alfabeto original, e o valor do deslocamento é chamado de chave, pois a partir dele é possível obter a mensagem decifrada. A figura 1 mostra o exemplo de um deslocamento de 3 letras da direita para a esquerda.

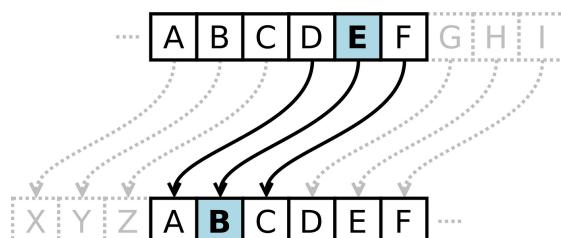


Figura 1 - Representação da Cifra de César

Outra criptografia conhecida é a Cifra de Substituição, que similarmente à Cifra de César, possui como base a desfiguração do alfabeto para a criação de uma cifra. Entretanto, esta diverge em relação à anterior porque não existe um padrão de ordenamento do novo alfabeto, ou seja, o alfabeto gerado não segue um deslocamento fixo, logo, cada letra possui uma correspondente aleatória. Dessa maneira, a Cifra de César possui 26 possíveis chaves,

enquanto a Cifra de Substituição possui $26!$ (número próximo à 4×10^{26}) possíveis chaves. A Figura 2 exemplifica uma Cifra de Substituição de uma parte do alfabeto.

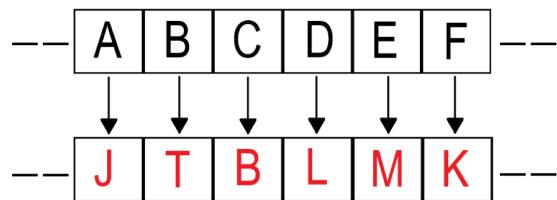


Figura 2 - Representação da Cifra de Substituição

Desse modo, este trabalho tem como objetivo demonstrar, utilizando algoritmos na linguagem Python de programação, a descriptografia de dois textos binários, sendo o primeiro codificado em Cifra de César e o segundo codificado em Cifra de Substituição.

Das disposições do trabalho, a seção II possui o referencial teórico usado na solução das cifras, a seção III aborda quais os procedimentos utilizados para a obtenção das mensagens, bem como a métrica de avaliação dos resultados, a seção IV traz os resultados dos testes e, por fim, a seção V mostra uma breve síntese geral de todo o conteúdo apresentado, juntamente com uma análise crítica dos resultados.

II. Referencial Teórico

Na descriptografia da Cifra de César, o algoritmo realiza todos os deslocamentos possíveis do alfabeto, e, a cada deslocamento, uma métrica é utilizada para encontrar o deslocamento mais provável. Já para a Cifra de Substituição, como esta possui uma quantidade massiva de chaves possíveis, caso fosse proposto um algoritmo que encontre a chave exata da cifra por força bruta, isso acarretaria um tempo de execução do programa muito longo. Nesse sentido, a proposta seria criar um algoritmo que encontre a chave de maior probabilidade de acerto, e, para tal, a métrica foi utilizada novamente para encontrar a chave mais provável.

III. Metodologia

A métrica utilizada consiste em uma janela deslizante, que irá percorrer o texto analisando a probabilidade de todos conjuntos de quatro letras (chamados *quadgrams*) do texto estarem próximos do conjunto de *quadgrams* da base de dados da métrica. Tais dados, por sua vez, são resultado da contagem dos *quadgrams* de um texto escrito na língua inglesa e, dessa forma os *quadgrams* do texto cifrado são constantemente comparados com os *quadgrams* dos dados, gerando um número negativo que determina a probabilidade de a chave encontrada ser a correta. O número em questão trata-se do logaritmo, em base 10, da soma de todos os logaritmos das probabilidades dos *quadgrams*, e quanto mais próximo de 0, mais provável, com base nas comparações realizadas, é a chave encontrada. Por exemplo, se o texto fosse a palavra *ATTACK*, a métrica faria o seguinte cálculo:

$$\log(p(\text{ATTACK})) = \log(p(\text{ATTA})) + \log(p(\text{TTAC})) + \log(p(\text{TACK}))$$

Para a Cifra de César, o programa inicia-se com a leitura do texto binário, que será armazenado em uma variável do tipo *string*. Depois, todos os espaços do arquivo lido serão retirados e, em seguida, é feita a conversão de todos números para letras, de acordo com a tabela ASCII. Uma vez que as letras são obtidas, inicia-se um laço de repetição com a utilização da métrica 26 vezes, pois são todas as possibilidades de tradução da Cifra de César. Por fim, o último passo é a seleção, entre todas as chaves geradas, daquela que a métrica indicou possuir maior valor. A figura 3 mostra um fluxograma da ordem de execução das operações do programa.

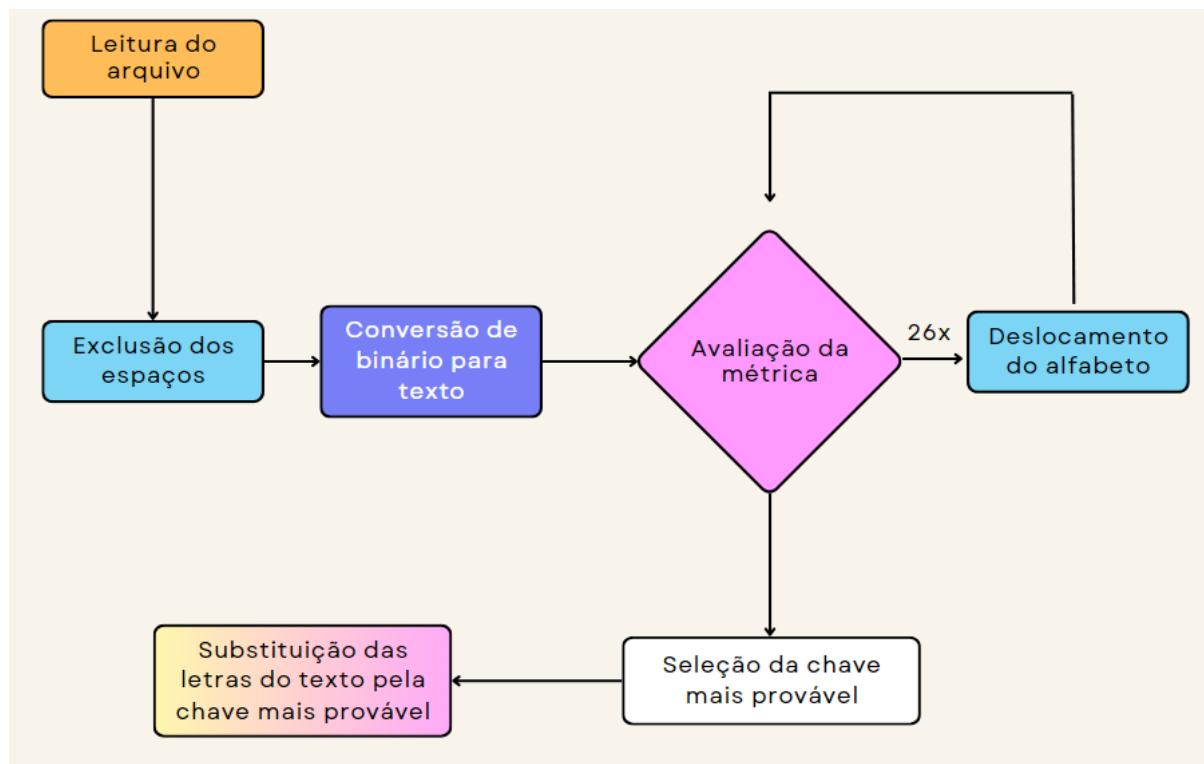


Figura 3 - Fluxograma Cifra de César

Uma lógica parecida foi utilizada na Cifra de Substituição, o algoritmo realiza a leitura do texto binário, retira os espaços entre caracteres e converte números em letras. Entretanto, a próxima etapa é a repetição de 20 mil testes da contagem dos *quadgrams* do texto e a substituição, por ordem de frequência, dos *quadgrams* do texto pelos *quadgrams* mais frequentes do banco de dados da métrica, gerando uma chave nova a cada iteração. Por fim, a chave é trocada no texto e a métrica fará outro cálculo. A figura 4.1 ilustra o fluxograma deste algoritmo

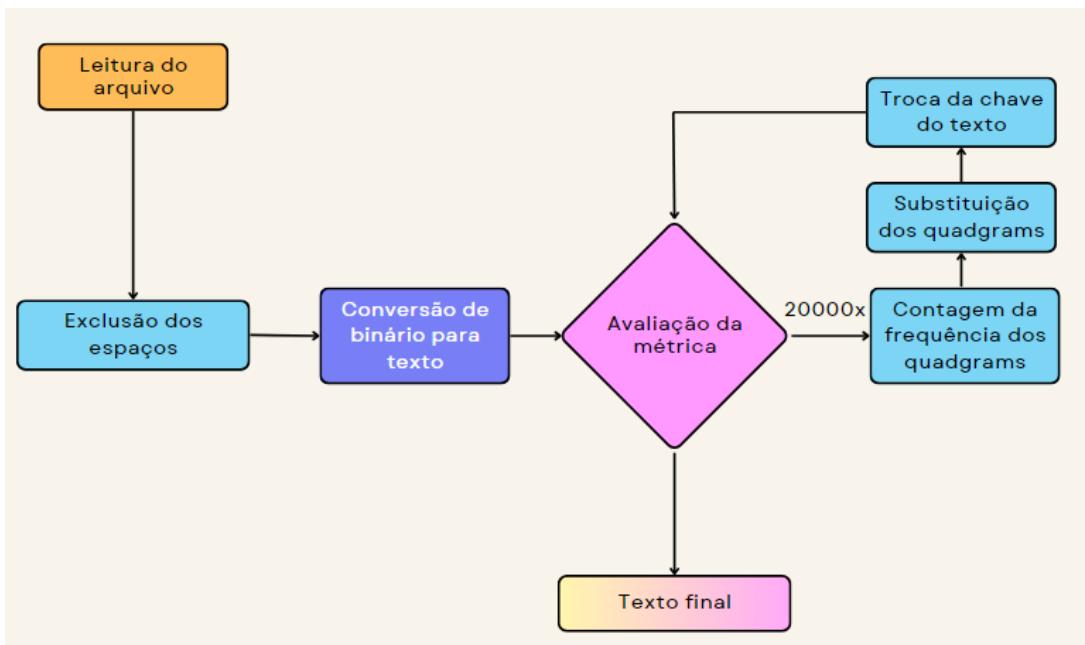


Figura 4.1 - Fluxograma Cifra de Substituição - algoritmo 1

Além disso, um algoritmo similar ao último também foi testado, também para encontrar a chave da Cifra de Substituição. A diferença entre eles é que este tenta substituir diretamente os *quadgrams* da métrica no texto e realiza os cálculos de avaliação a partir deles. A figura 4.2 mostra como seria o fluxo de execução desse outro algoritmo

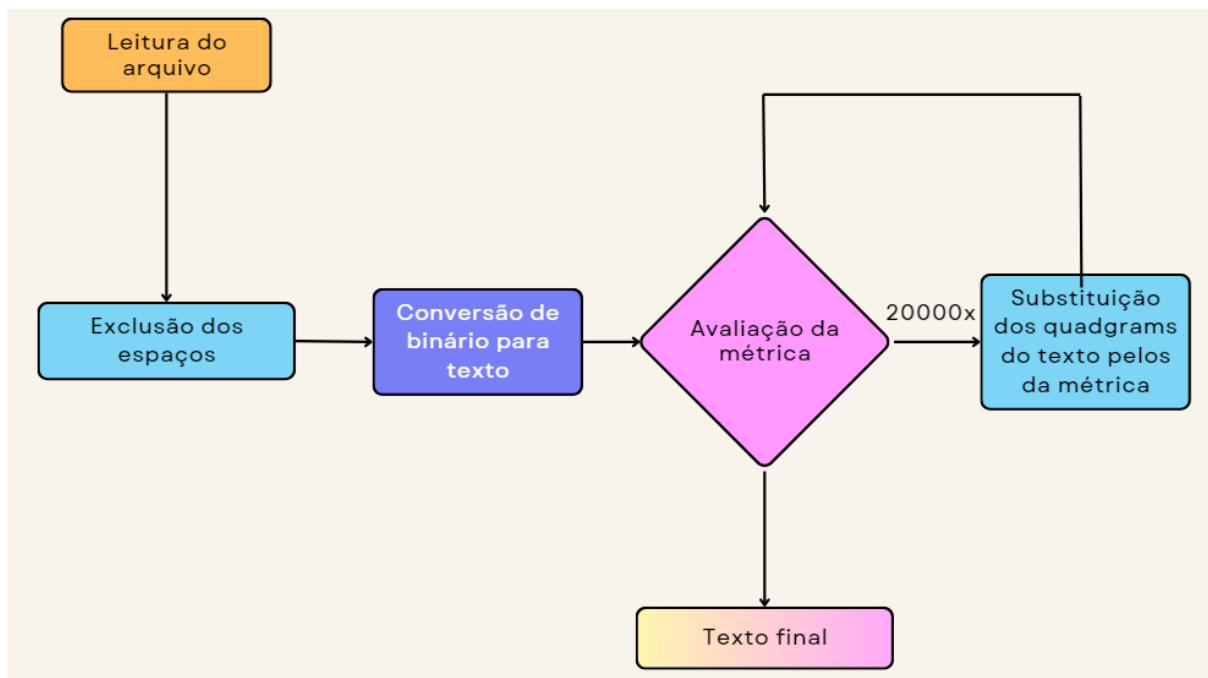


Figura 4.1 - Fluxograma Cifra de Substituição - algoritmo 2

IV. Resultados

Para a Cifra de César, o algoritmo consegue encontrar a chave correta e descriptografar a mensagem na maioria dos casos. A figura 4.1 mostra a saída do texto criptografado em Cifra de César e, posteriormente, são realizados os cálculos dos deslocamentos. A figura 4.2 mostra o restante dos cálculos de deslocamento e a saída do texto descriptografado. É possível notar que a chave selecionada foi a de deslocamento 15, porque era a que possuía maior valor absoluto segundo a métrica.

```
XI WPH QTTC HPXS IMPHGDDBN XH P WJBQAXCV PCS RMPGPRITG QJXASXCV TMETGXCTRT INTGT XH ETGWPEN CD QTIITG STBDCHGPXDC DU INT UDAAN DU WUBPC RCRXTXH IMPC IWDH SHIPCI XBPVT D
U DJG IXCN LDGAS ID BT XI JCSTGHRDGH DJG GHEDCHXXAXIN ID STPA BDGT ZXCSAN LXIW DCT PCDIWTG PCS ID EGTHGKT PCS RWITGXW INT EPAT QAJT SDI INT DCAN WDBT LT WPKT TKTG ZCDLC
Deslocamento 0: -8.016175542490819
Deslocamento 1: -8.066486521347132
Deslocamento 2: -7.557463050507031
Deslocamento 3: -8.571571198166586
Deslocamento 4: -7.709279892641594
Deslocamento 5: -7.737684741420081
Deslocamento 6: -8.797155131153549
Deslocamento 7: -8.968548954043982
Deslocamento 8: -7.754703971840475
Deslocamento 9: -8.448875442781791
```

Figura 4.1 - Texto criptografado - Cifra de César

```
Deslocamento 10: -8.91543300509316
Deslocamento 11: -7.723623617802799
Deslocamento 12: -8.85894871214405
Deslocamento 13: -9.288954170170642
Deslocamento 14: -7.888202149826091
Deslocamento 15: -4.084369746913945
Deslocamento 16: -8.55110385648157
Deslocamento 17: -8.551165385100074
Deslocamento 18: -8.763116790116406
Deslocamento 19: -7.8239754628295595
Deslocamento 20: -8.81343413899402
Deslocamento 21: -7.792014296437976
Deslocamento 22: -8.238932526254967
Deslocamento 23: -8.702572845866198
Deslocamento 24: -8.303436019340669
Deslocamento 25: -8.370307626847284

Deslocamento correto --> 15
IT HAS BEEN SAID THAT ASTRONOMY IS A HUMBLING AND CHARACTER BUILDING EXPERIENCE THERE IS PERHAPS NO BETTER DEMONSTRATION OF THE FOLLY OF HUMAN CONCEITS THAN THIS DISTANT IMAGE O
F OUR TINY WORLD TO ME IT UNDERSCORES OUR RESPONSIBILITY TO DEAL MORE KINDLY WITH ONE ANOTHER AND TO PRESERVE AND CHERISH THE PALE BLUE DOT THE ONLY HOME WE HAVE EVER KNOWN
```

Figura 4.2 - Texto descriptografado - Cifra de César

Na Cifra de Substituição, os algoritmos realizam várias trocas da chave do texto, no entanto, a possível chave, na maioria das vezes, não é encontrada. Isso deve-se ao fato de que mesmo com a tentativa de aumentar as chances de acerto dos *quadgrams* seguindo a ordem de frequência, como no algoritmo 1, ou até substituindo as possibilidades de *quadgrams*, como no algoritmo 2, as possibilidades de chaves ainda são muito grandes. A figura 5.1 mostra o texto criptografado em Cifra de Substituição, a figura 5.2 mostra as diferentes tentativas de decodificação do texto com o primeiro algoritmo e a figura 5.3 mostra as tentativas de decodificação do texto do segundo algoritmo da Cifra de Substituição.

VCO ATD HO SOTMD VCO OPJIKWITVMZ WG VCO SWMQMM NM VCO ATD HO VCIOTVOZ VCO SWZVNUNZR WG WUJ MIJOSNM NZ VCTV XKOTY HNUKA TIQM XOTINZR IOMNUISO CUZRID JOWJKO TZA ZTVNIMZM HNUKA XO JTMZO VW TSV NZ VCONI KWH SWZVITSVDA JTOLUANSOM TZM HNUKA CTBO MOOZ VCO KTMV RTMO WG CQUTZ OZKIRCVOZQZV UZVNK VCO INMO WG T BNMMWZTID ZOH SUKVUO VCTV WZSO TRTNZ OQXITSOM VCO S WMQNS JOIMIOSNVO T JOIMIOSNVO NZ HCNSC HO TIO WZO GNVNZR ZONVCOI TXWBZ ZWJ XOKWJ XUV HNVNCZ

Figura 5.1 - Texto criptografado - Cifra de Substituição

Media da métrica: -7.894048060713551
 Texto Decifrado:
 JQH DRM VH THRUH JQH HMBlEPRJOEA EK JQH TEUCEU OU JQH DRM VH JOPHRJHA JQH TEAOJOAOS EK EYP UBHTOHU OA JQRJ GHRN VEPID RPCU GHROPS PHUEYPTH QYASPM BHEBII RAD ARJOEAU VEYID GH BPEAH JE RTJ EA JQHP IEV TEAJRPTJDH BPHZYDOTHU RAD VEYID QRFH UHHA JQH IRUJ SRUB EK QYCRRA HALOSQJHACHA YAOJ JQH POUH EK R FOUEARPM AHV TYLJYHP JQRJ EATH RSROA HCGPRTHU JQH T EUOT BHPUBHTJOEH R BHPUBHTJOEH OA VOJQO VH RPH EAH KOJJOS AHOJHP RGEFH AEP GHIEV GYJ VOJQOA
 Média da métrica: -7.894048060713551
 Texto Decifrado:
 JQH DRM VH THRUH JQH HMBlEPRJOEA EK JQH TEUCEU OU JQH DRM VH JOPHRJHA JQH TEAOJOAOS EK EYP UBHTOHU OA JQRJ GHRN VEPID RPCU GHROPS PHUEYPTH QYASPM BHEBII RAD ARJOEAU VEYID GH BPEAH JE RTJ EA JQHP IEV TEAJRPTJDH BPHZYDOTHU RAD VEYID QRFH UHHA JQH IRUJ SRUB EK QYCRRA HALOSQJHACHA YAOJ JQH POUH EK R FOUEARPM AHV TYLJYHP JQRJ EATH RSROA HCGPRTHU JQH T EUOT BHPUBHTJOEH R BHPUBHTJOEH OA VOJQO VH RPH EAH KOJJOS AHOJHP RGEFH AEP GHIEV GYJ VOJQOA
 Média da métrica: -7.741019354696085
 Texto Decifrado:
 JQH DRM VH THRUH JQH HMBlEPRJOEA EK JQH TEUCEU OU JQH DRM VH JOPHRJHA JQH TEAOJOAOS EK EYP UBHTOHU OA JQRJ GHRN VEPID RPCU GHROPS PHUEYPTH QYASPM BHEBII RAD ARJOEAU VEYID GH BPEAH JE RTJ EA JQHP IEV TEAJRPTJDH BPHZYDOTHU RAD VEYID QRFH UHHA JQH IRUJ SRUB EK QYCRRA HALOSQJHACHA YAOJ JQH POUH EK R FOUEARPM AHV TYLJYHP JQRJ EATH RSROA HCGPRTHU JQH T EUOT BHPUBHTJOEH R BHPUBHTJOEH OA VOJQO VH RPH EAH KOJJOS AHOJHP RGEFH AEP GHIEV GYJ VOJQOA
 Média da métrica: -7.741019354696085
 Texto Decifrado:
 JQH DRM VH THRUH JQH HMBlEPRJOEA EK JQH TEUCEU OU JQH DRM VH JOPHRJHA JQH TEAOJOAOS EK EYP UBHTOHU OA JQRJ GHRN VEPID RPCU GHROPS PHUEYPTH QYASPM BHEBII RAD ARJOEAU VEYID GH BPEAH JE RTJ EA JQHP IEV TEAJRPTJDH BPHZYDOTHU RAD VEYID QRFH UHHA JQH IRUJ SRUB EK QYCRRA HALOSQJHACHA YAOJ JQH POUH EK R FOUEARPM AHV TYLJYHP JQRJ EATH RSROA HCGPRTHU JQH T EUOT BHPUBHTJOEH R BHPUBHTJOEH OA VOJQO VH RPH EAH KOJJOS AHOJHP RGEFH AEP GHIEV GYJ VOJQOA
 Média da métrica: -7.741019354696085

Figura 5.2 - Textos gerados - Cifra de Substituição - algoritmo 1

Texto Decifrado:
 JOK ZXT FK SKQOK JOK KEWMQDIAHN WI JOK SWMQMNO AQ JOK ZXT FK JOKSKOKHN WI WND QBKSAKQ AH JOKX YGOKU FMDGZ XDQD YKODAHN DKQNDISK ONHNDT BKMBGK XHZ HXQAWHQ FNGZ YK BDWIK JW XSJ WH JOKAD GMF SWMTDXSKJZ BDKPNZASQK XHZ FNGZ OXUK QKHK JOK QKQJ XQOB WI OMWQ KHGAVQJHKMHQ NHQAG JOK DAQK WI X LAQAHXDT HKF SNGJNDK JOKJ WHSK XXAH KMYDXSKQ JOK S WQMAS BKDQBSKJALK X BKDQBSKJALK AH FOASO FK XOK WHK JAJJAHN HKAJOKD XWILK HND YKGMF YND FAJOAH
 -8.892897318600674
 Texto Decifrado:
 AWV IPL UV RVPFV AWV KHMUPYPAENT MN AWV RMFSMF EF AWV IPL UV AMYPPAVT AWV RMTAETCTX MN MCY FHREVF ET AMPA DJVPG UMQJ PYSF DVPYETX YVFCYRV WCTXYL HMHDV PTI TPAEMTF UMCJL DV HYMTV AM PRA MT AWVEY JMU RMTAYPRAVIE PTI UMCJL WPOV FWFT AWV JPFM XPHF MN WSPIT VTIXEWAVTSVTA CTAEJ AWV YEFV MN P DEEFEMTPY TVU RCJACVW AWPA MTRV PXPET VSDYPRVF AWV R MFSEF HVYFHVRAEV P HVYFHVRAEV T UMERW UV PVV MTV NEAAETX TVEAWWY PMOVW TMV DVJUW DCA UEAMET
 -7.842874308207148
 Texto Decifrado:
 BMP QCR YP UPCJP BMP POGIXKCBFIZ ID BMP UJHJCJ F3 BMP QCR YP BWPQPCBZ BMP UIZBZFZAFZT ID IAX JGPUPFZ F2 BWCB LKPC EYXQ OXQJ L PCXFZT XPJIAKUP WAZTXR GPIGKP CZQ ZCBFIZJ YIAK LP GXIZP BI CUB IZ BMPFX KIY UIZBCUBQZ QXMAPQFUPZ CZQ YIAKQ WCSP JRPZ BMP KCBJ TCG ID WANCH PZKFTWBPZPZ ABZFK BMP XFJZP ID C SFJFIZXOR ZPY UAKBAPW BWCB IZUP CTCFZ PHLXCUPJ BMP U IJHFU GPXQGPUBFSP C GPXQGPUBFSP Z YWFUW YP CXP IZP DFBBFZT ZPFBWPK CLISP ZIX LKPCY LAB YFBWFZ
 -9.028919947973545
 Texto Decifrado:
 BFO MVC GO AQVRO BFO OJUHMDIVBXOP DK BFO ADRLDR XR BFO MVC GO BFNOVBOP BFO ADPBPXPKP DK DYN RHQAXOR XP BFVB ZNQVS GDMM VNLR ZONVPE NORDYNAO FYPENC HODIMO VPM PVBODPR GDYM ZO HNDPO BD VAB DP BFOXN WDQ ADPBNVABOM HNQDMYKAOR VPM GDYMM FVDO ROOF BFO WRB EVRH DK FYLPV OPXKFBOPLOPB YPBXW BFO NXRD DK V JXRXDPVNC PGM AYMBWNO BFVW DPAO VEVXP OLZNVAOR BFO A DRXLA HONRHOABXOJ V HONRHOABXOJ XP GXFAF GO VNO DPO IXBBXP E POXBPN VZDJO PDN ZOMDG ZYB GXBPX
 -7.960735397172237

Figura 5.2 - Textos gerados - Cifra de Substituição - algoritmo 2

V. Conclusão

A Cifra de César e a Cifra de Substituição constituem duas formas de criptografia de dados, sendo que a primeira possui um nível de dificuldade de descriptografia menor em relação à segunda. A métrica utilizada levou em consideração a semelhança dos *quadgrams* do texto decifrado com os *quadgrams* de palavras da língua inglesa, e o método de deslocamento do alfabeto funcionou precisamente na obtenção da chave correta para a Cifra de César. Já a Cifra de Substituição não obteve o mesmo êxito da anterior, pois devido à imensa quantidade de possíveis chaves, os algoritmos carecem de uma lógica mais assertiva e que tenha uma precisão maior a cada nova chave gerada.

Por fim, este trabalho possibilitou a análise de duas diferentes criptografias, e seguindo a lógica de seus respectivos paradigmas, foram propostos algoritmos que manipulam os textos de acordo com a semelhança dos conjuntos quaternários de letras desses com outros da língua inglesa.