TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO GIỮA KỲ MÔN**

**XÁC SUẤT THỐNG KÊ ỨNG DỤNG CHO**

**CÔNG NGHỆ THÔNG TIN**

**.**

*Người hướng dẫn*: **Th.S NGUYỄN QUỐC BÌNH**

*Người thực hiện*: **NGUYỄN TRẦN QUANG HUY - 52000668**

Lớp **: 20050201**

Khoá  **: 24**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2022**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO GIỮA KỲ MÔN**

**XÁC SUẤT THỐNG KÊ ỨNG DỤNG CHO**

**CÔNG NGHỆ THÔNG TIN**

**.**

Người hướng dẫn: **ThS NGUYỄN QUỐC BÌNH**

Người thực hiện: **NGUYỄN TRẦN QUANG HUY - 52000668**

Lớp **: 20050201**

Khoá  **: 24**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2022**

LỜI CẢM ƠN

Em xin gửi lời cảm ơn đến thầy Th.S Nguyễn Quốc Bình là giảng viên hướng dẫn em để em có thể hoàn thành bài báo cáo này xin chân thành cảm ơn thầy.

TÓM TẮT

Trong bài báo cáo này chúng ta sẽ cũng nhau tìm hiểu về mã hóa, một lĩnh vực khá là thú vị và cũng rất quan trọng trong thời đại ngày nay. Chúng ta sẽ đi qua các khái niệm cơ bản khi ta bắt đầu tìm hiểu về mã hóa, mã hóa bằng Monoalphabetic Substitution Cipher, phá mã bằng Frequency Analysis. Thực hành và nhận xét về các giải thuật trên.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc101471423)

[TÓM TẮT ii](#_Toc101471424)

[MỤC LỤC 1](#_Toc101471425)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 2](#_Toc101471426)

[CHƯƠNG 1 – TỔNG QUAN VỀ MÃ HÓA VÀ GIẢI MÃ. 3](#_Toc101471427)

[CHƯƠNG 2 – GIẢI THUẬT MONOALPHABETIC SUBSTITUTION CIPHER 5](#_Toc101471428)

[CHƯƠNG 3 – GIẢI THUẬT FREQUENCY ANALYSIS 6](#_Toc101471429)

[CHƯƠNG 4 – THỰC HÀNH 6](#_Toc101471430)

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

[Hình 2.1: Kiến trúc FTP 7](#_Toc387689394)

**DANH MỤC BẢNG**

[Bảng 3.1 Ví dụ cho chèn bảng 8](#_Toc387689363)

CHƯƠNG 1 – TỔNG QUAN VỀ MÃ HÓA VÀ GIẢI MÃ.

Mã hóa (encryption) là phương pháp để biến thông tin (phim ảnh, văn bản, hình ảnh...) từ định dạng bình thường sang dạng thông tin không thể hiểu được. Ta có thể hiểu đơn giản là việc 2 người giao tiếp với nhau thông qua một phương tiện giao tiếp là thư tay và không muốn trong quá trình truyền thư có bị người khác đọc được thì vì yêu cầu như vậy sẽ sinh ra một quy ước ngầm với nhau mà quy tắc này là sẽ ẩn thông điệp của lá thư đi bằng một cách nào đó mà chỉ 2 người trên mới có thể hiểu được bức thư đó nói gì đó là sơ khai của mã hóa.

Giải mã (decryption) là phương pháp để đưa từ dạng thông tin đã được mã hóa về dạng thông tin ban đầu, quá trình ngược của mã hóa. Thì ví dụ ở trên cũng đã nêu ra được lý do tại sao có mã hóa rồi thì khi mà người gửi mã hóa thư gửi cho người nhận thì khi đến tay người nhận quá trình mà người nhận dùng những quy tắt ngầm đã thống nhất trước với nhau để có thể đọc lại bức thư quá trình này được gọi là giải mã ngược lại với mã hóa ở trên.

Plaintext là thông điệp ban đầu (muốn gửi) chưa được mã hóa (có kí hiệu là P).

Ciphertext là thông điệp lúc sau đã được mã hóa xong (có kí hiệu là C).

Key (gọi là khóa) là chuổi các ký tự được tạo ra và được dùng trong hàm mã hóa plaintext tạo ciphertext và giải mã để nhận lại được plaintext.

Hệ thống mật mã đối xứng (Symmetric Key Cryptography) có chổ còn có tên gọi khác là mã khóa bí mật (Sycret Key Crytosystem) là các thuật toán khóa đối xứng là một lớp các thuật toán mật mã hóa trong đó các khóa dùng cho việc mật mã hóa và giải mã có quan hệ rõ ràng với nhau. Một lưu ý trong hệ thống này là khóa này phải giữ bí mật với mọi người khác. Cho nên vậy mới có tên gọi là đối xứng.

Hệ thống mật mã không đối xứng (Asymmetric Key Cryptography) là một dạng mật mã hóa cho phép người sử dụng trao đổi các thông tin mật mà không cần phải trao đổi các khóa chung bí mật trước đó. Có thể được hiểu đơn giản là nâng cấp hơn của mã hóa đối xứng vì mã hóa đối xứng khi bị lộ khóa ra là sẽ lộ ra plaintext còn trong hệ thống này thì chúng ta sẽ có 2 khóa 1 khóa công khai dùng để mã hóa và khóa này được công bố rộng rãi cho mọi người và khóa còn lại là khóa bí mật chỉ những ai có khóa này mới có thể giải mã thông điệp từ khóa công khai. Vậy cho nên người ta mới gọi nó là bất đối xứng.

CHƯƠNG 2 – GIẢI THUẬT MONOALPHABETIC SUBSTITUTION CIPHER

Substitution Cipher được coi là một trong những mật mã đơn giản nhất. Cách thức mà nó hoặc động là dựa vào thay thế từng chữ cái của một văn bản với một chữ cái khác hoặc có biến thể khác là thay thế chữ thành một ký ngẫu nhiên (ký tự đặc biệt, …). Số trường hợp có thể xảy ra khi ta sử dụng bảng chữ cái tiếng anh (26 chữ cái) là 26! – 1 trường hợp có thể xảy ra (-1 là vì có 1 hoán vị mà ở đó ciphertext sẽ trùng hợp plaintext).

Monoalphabetic substitution cipher (mật mã thay thế monoalphabetic) được biết đến như là một mật mã thay thế đơn giản, dựa vào một bảng được định trước. Có thể hiểu là mỗi chữ cái sẽ được thay thế bằng một chữ cái khác cố định. Suy ra là nếu như ‘a’ được mã hóa thành ‘h’ thì tất cả những chữ ‘a’ mà ta thấy trong văn bản sẽ được đổi thành ‘h’ trong văn bản đã mã hóa.

Ví dụ: ta có plaintext như sau “Nguyen Tran Quang Huy”

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

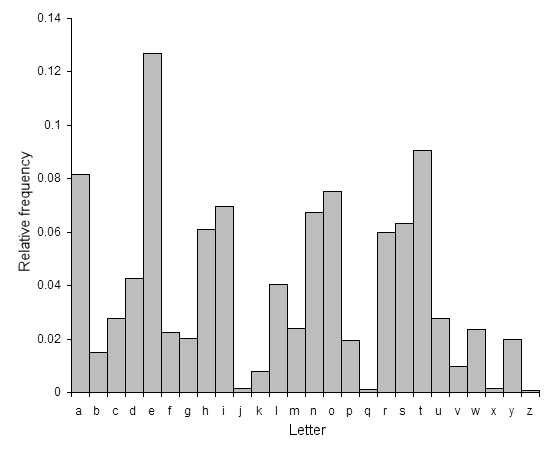
Sau khi mã hóa thì ta nhận được “Xhulox Czsx Nusxh Rul”

Biến thể phổ biến nhất mà nhắc đến Monoalphabetic Substitution Cipher đều phải biết đó chính là mật mã Caesar. Mật mã này mã hóa plaintext bằng cách thay thế mỗi ký tự trong plaintext bằng một ký tự đứng sau đó n đơn vị (a 🡪 d, b 🡪e, …, z 🡪 c). (ví dụ: plaintext “Nguyen Tran Quang Huy” với n = 3 sẽ thành “Qjxbhq Wudq Txdqj Kxb”)

Trên thực tế thì có rất nhiều mật mã thay thế monoalphabetic khác nhau, có thể nói là vô số vì mỗi chữ cái có thể được thay thế bằng bất kỳ cái gì (như biểu tượng, ký tự, …) không phải chỉ là một chữ cái khác mà thôi.

Ở quá khứ thì người ta sử dụng chúng rất nhiều (chúng được áp dụng trong quân sự, thông tin quốc gia, …) và thời điểm lúc đó rất khó để mà phá được mật mã như vậy. Nhưng theo thời gian với sự phát triển chóng mặt của máy tính và linh kiện phần cứng thì việc mà một máy tính có thể thực hiện hàng triệu triệu phép thế trong một thời gian rất là ngắn là chuyện rất bình thường rồi cho nên thật tế ngày nay có thể nói là không còn ai sử dụng thuật toán như vầy để đi mã hóa cả. Nhưng không thể nói vậy mà phủ nhận hoàn toàn được bởi vì suy cho cùng thì nó là nền tảng để phát triển những thuật toán mã hóa mạnh mẽ hơn.

CHƯƠNG 3 – GIẢI THUẬT FREQUENCY ANALYSIS

Đầu tiên ta có thể phải xác nhận với nhau là số lần xuất hiện của các chữ cái trong plaintext là khác nhau (không đồng đều) dựa vào đó là nền tảng để ta có thể phá được Ciphertext.

Biểu đồ tần suất các ký tự trong tiếng Anh

Để dễ hình dung hơn ta vào thẳng ví dụ luôn:

Ta có một plaintext: “THIS IS A PROPER SAMPLE FOR ENGLISH TEXT. THE FREQUENCIES OF LETTERS IN THIS SAMPLE IS NOT UNIFORM AND VARY FOR DIFFERENT CHARACTERS. IN GENERAL THE MOST FREQUENT LETTER IS FOLLOWED BY A SECOND GROUP. IF WE TAKE A CLOSER LOOK WE WILL NOTICE THAT FOR BIGRAMS AND TRIGRAMS THE NONUNIFORM IS EVEN MORE.”

Sau khi mã hóa: “YKHLBA JCZ SVIJ JZB TZVHI JCZ VHJ DR IZXKHLBA VSS RDHEI DR YVJV LBXSKYLBA YLALJVS IFZZXC CVI LEFHDNZY EVBLRDSY JCZ FHLEVHT HZVIDB RDH JCLI CVI WZZB JCZ VYNZBJ DR ELXHDZSZXJHDBLXI JCZ XDEFSZQLJT DR JCZ RKBXJLDBI JCVJ XVB BDP WZ FZHRDHEZY WT JCZ EVXCLBZ CVI HLIZB YHVEVJLXVSST VI V HXXIKSJ DR JCLI HZXZBJ YZNZXDFEZBJ LB JZXCBDSDAT EVBT DR JCZ XLFCZH ITIJZEIJCVJ PZHZ DBXZ XDBILYXHZYIZKHZ VHZBDP WHZVMVWSZ”

Phân tích tần số ta thấy đoạn mã trên bao gồm 338 chữ, thống kế tần xuất như sau:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |
| 5 | 24 | 19 | 23 | 12 | 7 | 0 | 24 | 21 | 29 | 6 | 21 | 1 | 3 | 0 | 3 | 1 | 11 | 14 | 8 | 0 | 27 | 5 | 17 | 12 | 45 |

Quan sát Z là chữ mã có tần suất lớn hơn hẳn các chữ cái còn lại nên rút ra: e 🡪 Z. Quan sát những chữ mã có tần suất cao tiếp theo fj = 29, fv = 27. Đồng thời chú ý đến bộ ba jcz có tần suất cao, dễ thấy fjcz = 8 , t 🡪 j, h 🡪 c (suy luận jcz chính là từ bản rõ the). Tương tự như vậy ta sẽ so sánh tần số với bảng dưới đây.

Text

Description automatically generated with low confidence

Và để nâng cao độ chính xác thì ta nên thêm các cặp chữ thường xuất hiện như th, an, …

CHƯƠNG 4 – THỰC HÀNH

|  |
| --- |
| *def* **Encryption**(input):  *# assign output*      output = input  *# template*      alphabet = list("*abcdefghijklmnopqrstuvwxyz*")      substitutionAlphabet = list("*--------------------------*")  *# create substitution alphabet*      for i in range(**len**(alphabet)):          index = random.randint(MIN, MAX) % 26          while substitutionAlphabet[index] != "*-*":              index = random.randint(MIN, MAX) % 26          substitutionAlphabet[index] = alphabet[i]  *# store key*  **writeToFile**("*./key.txt*", "".**join**(alphabet) + "*\n*" + "".**join**(substitutionAlphabet))  *# encryption*      output = **myReplace**(input, alphabet, substitutionAlphabet)  **print**("*source* ", alphabet)  **print**("*destination*", substitutionAlphabet)      return output |
| Encryption function |

Hàm Encryption sẽ nhận vào đối số là một chuỗi sao đó tạo ra key ngẫu nhiên và mã hóa input đầu vào.

Graphical user interface

Description automatically generated with medium confidence

Kết quả với độ dài plaintext là < 200 chữ

**TÀI LIỆU THAM KHẢO**

**Online**

1. [Mã hóa – Wikipedia tiếng Việt](https://vi.wikipedia.org/wiki/M%C3%A3_h%C3%B3a)
2. [Thuật toán khóa đối xứng – Wikipedia tiếng Việt](https://vi.wikipedia.org/wiki/Thu%E1%BA%ADt_to%C3%A1n_kh%C3%B3a_%C4%91%E1%BB%91i_x%E1%BB%A9ng)
3. [Mật mã hóa khóa công khai – Wikipedia tiếng Việt](https://vi.wikipedia.org/wiki/M%E1%BA%ADt_m%C3%A3_h%C3%B3a_kh%C3%B3a_c%C3%B4ng_khai)
4. [Phân tích tần suất – Wikipedia tiếng Việt](https://vi.wikipedia.org/wiki/Ph%C3%A2n_t%C3%ADch_t%E1%BA%A7n_su%E1%BA%A5t)