CRIPTOGRAFÍA Y COMPUTACIÓN

GRADO EN INGENIERÍA INFORMÁTICA

Cifrados Clásicos

PRÁCTICA 2

Autores

Roberto García Pérez

Vladislav Nikolov Vasilev

Víctor Alejandro Vargas Pérez

Fecha de Entrega

Granada, 10 de marzo de 2020

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

—

CURSO 2019-2020

1. Introducción

Esta práctica consiste en el descifrado de 5 textos que se han cifrado con técnicas clásicas. Se adjunta un directorio con los 5 textos descifrados, así como un fichero fuente decipher.py con la implementación de las funciones **decipher_caesar(text, k)** y **decipher_vigenere(text, key)**, utilizadas para obtener el resultado de descifrar un texto con César y Vigenère respectivamente, con una clave dada. Para obtener el índice de coincidencia y las frecuencias relativas, así como el descifrado de un texto por sustitución (dada una permutación) y por transposición (dado un n) se ha utilizado el fichero Funciones.py que forma parte del material de la práctica.

2. Texto 1

- **Cifrado: Permutación**
- **Clave:**

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

En primer lugar, como suponemos que el texto original está en español, podemos buscar la secuencia de tres caracteres que más se repite. Si lo hacemos vemos que la secuencia más común es **TÑW**, luego, como la secuencia de tres palabras que más se repite en español es **QUE**, podemos asociar Q con T, U con Ñ, y E con W.

Posteriormente, si hacemos un análisis de frecuencias relativas, vemos que el carácter que más se repite es la **M**, por tanto, podemos asociarla con la letra **A**, ya que la E ya ha sido asignada con la W. Siguiendo a las letras M y W, la letra con mayor frecuencia es la **U**, que la asociaremos con la **O**. Sabiendo cuatro de las cinco vocales, podemos probar a ver cuales son las secuencias de longitud dos que más se repiten. Si lo hacemos, dos de las que más se repiten son **HM** y **WH**, es decir, **Ha** y **eH**, con lo cual es evidente que las secuencias son **LA** y **EL**, por ello, asociaremos a la **H** con la **L**.

La siguientes letras en frecuencia son **N**, **S**, **P** y **I**, y las letras correspondientes en probabilidad en idioma español son **S**, **N**, **R**, **I**. La letra **N** aparece muchas veces consecutivamente (83), con lo cual, lo razonable es que se corresponda con la **R**. La **S** la podemos asignar con la **N** ya que **eS**, es una subsecuencia de tamaño dos que se repite mucho en el texto, y podemos deducir que se trata de la preposición **en**. Por último, si contamos el número de veces que aparece la secuencia **II** vemos que no aparece ninguna vez, mientras que la secuencia **PP** aparece 51 veces, luego lo lógico es que la letra **P** se corresponda con la **S**, y la **I** se corresponda con la **I**.

El resto de asignaciones se pueden deducir fácilmente si leemos el texto con las sustituciones anteriores.

3. Texto 2

- **Cifrado: Vigenère**
- **Clave: HABITUALMENTE**

Presuponiendo que el texto original estaba en español (como el resto de textos), lo primero que se ha hecho es calcular el índice de coincidencia del texto cifrado, el cuál ha resultado ser de aproximadamente 0.04. Teniendo en cuenta que el índice de coincidencia del español es de aproximadamente 0.07, esto solo puede significar que **el texto ha sido cifrado mediante cifrado de Vigenère**, ya que este tipo de cifrado reduce el índice de coincidencia original. Para descifrar el texto, necesitamos conocer dos cosas:

- El tamaño de la clave, k.
- La clave como tal.

Para atacar el cifrado se ha utilizado el método del **índice de coincidencia**. Primero se ha ido dividiendo el texto en particiones donde se cogen los caracteres de k en k, obteniendo por tanto k particiones, y se ha analizado el índice de coincidencia de cada una. Si el índice de coincidencia de todas las particiones es aproximadamente el mismo que el del texto original, entonces se sabe que la clave utilizada es de tamaño k. En caso contrario, se repite el proceso hasta obtener resultados satisfactorios.

En este caso, se ha ido probando con distintos valores de k, hasta dar que con k=13 el índice de coincidencia de todas las particiones era de aproximadamente 0.07. Por tanto, de aquí hemos determinado que **el tamaño de la clave es k=13**.

Una vez que se ha determinado el tamaño de la clave, se ha atacado cada partición de forma separada, ya que cada una está cifrada utilizando un cifrado César, el cual es fácil de romper. En cada una de ellas se ha obtenido una tabla con las frecuencias de las letras y se ha mirado cuáles son las más frecuentes. Teniendo en cuenta que en el español las letras más frecuentes son la E y la A, si miramos cuáles son las más frecuentes en cada partición y comprobamos si tienen la misma distancia entre sí se puede determinar fácilmente qué sustitución se ha hecho. Una vez que se ha hecho todo el análisis, se ha descubierto que la clave utilizada para cifrar el texto ha sido **HABITUALMENTE**.

Partición 0: A → H	Partición 1: A → A	Partición 2: A → B	Partición 3: A → I
Partición 4: A → T	Partición 5: A → U	Partición 6: A → A	Partición 7: A → L
Partición 8: A → M	Partición 9: A → E	Partición 10: A → N	Partición 11: A → T
Partición 12: A → E			

4. Texto 3

- **Cifrado: Transposición, escítalo**
- **Clave(n): 25**

En primer lugar, se han obtenido las **frecuencias relativas** del texto, que son **casi idénticas a las del español**. De esto se deduce que se ha **cifrado por transposición** (se mantienen los símbolos del texto original, pero intercambiados de sitio). El cifrado que conocemos dentro de este tipo es el escítalo, por lo que se ha utilizado la función **descifra_transposicion(texto, n)**, con diferentes n (saltos) hasta que se ha obtenido un texto con sentido. Concretamente, la n es 25.

5. Texto 4

- **Cifrado: Sustitución por traslación (César)**
- **Clave: k=18**

El **índice de coincidencia** es el del español (0.07 aproximadamente). Por su parte, las **frecuencias relativas** no coinciden con las del español, pues son las siguientes:

(12.825, 'V')	(12.597, 'R')	(9.537, 'G')	(9.44, 'K')	(7.389, 'E')	(7.324, 'Z')
(6.119, 'J')	(5.208, 'U')	(4.589, 'C')	(4.069, 'L')	(3.873, 'D')	(3.71, 'T')
(2.994, 'M')	(1.79, 'H')	(1.204, 'S')	(1.139, 'W')	(1.009, 'X')	(0.976, 'P')
(0.878, 'N')	(0.846, 'F')	(0.748, 'Y')	(0.651, 'Q')	(0.553, 'I')	(0.292, 'A')
(0.227, 'O')	(0.0, 'Ñ')	(0.0, 'B')			

Todo esto nos indica que se ha debido cifrar por **sustitución monoalfabética**. Supongamos que se ha utilizado **traslación**, de forma que cada letra se cambia por la que esté k posiciones más adelante.

Si nos basamos en las frecuencias del español, en primer lugar diríamos que E se ha cifrado como V, por lo que k=18. En tal caso, la A se habría cifrado como R, la O como G y la S como K. Por lo tanto, hasta el momento habría correspondencia entre las frecuencias del texto y las frecuencias de sus correspondientes letras en el español. Por este motivo, probamos a descifrar con k=18, y efectivamente se obtiene un texto con sentido.

6. Texto 5

- **Cifrado: Sustitución por traslación (César)**
- **Clave: k=4**

Las frecuencias relativas del texto son las siguientes:

(25.636, 'E')	(7.872, 'I')	(6.848, 'W')	(6.438, 'V')	(6.028, 'O')	(5.677, 'Q')
(4.828, 'S')	(4.682, 'G')	(4.126, 'X')	(3.745, 'Y')	(3.658, 'H')	(3.482, 'M')
(3.219, 'P')	(2.897, 'T')	(1.785, 'F')	(1.668, 'K')	(1.346, 'Z')	(1.287, 'L')
(1.112, 'C')	(1.024, 'U')	(0.936, 'J')	(0.79, 'N')	(0.556, 'D')	(0.321, 'R')
(0.029, 'B')	(0.0, 'Ñ')	(0.0, 'A')			

En primer lugar, supongamos que se ha cifrado por **traslación**. En tal caso, dado que el orden de estas frecuencias no coincide con el del español, pongamos que la A se ha cifrado como E (no podría ser la E como E pues entonces k=0: no hay cifrado). En ese caso, k=4.

Por consiguiente, la E se cifraría como I, la O como S, la S como W y la R como V. Si bien el orden no coincidiría exactamente con el del español, es similar (estas letras se mantienen entre las más usadas), por lo que es razonable probar a descifrar con k=4. Efectivamente, el resultado obtenido tiene sentido.