Aplicaciones matriciales a criptografía

Juan Gabriel Triana Laverde

Universidad Nacional de Colombia
Ciencias, Matemáticas
Bogotá, Colombia
2011

Aplicaciones matriciales a criptografía

Juan Gabriel Triana Laverde

Tesis o trabajo de grado presentada(o) como requisito parcial para optar al título de:
Magister en Matemática aplicada

Director(a):
Ph.D. Jorge Mauricio Ruiz Vera

Universidad Nacional de Colombia
Ciencias, Matemáticas
Bogotá, Colombia
2011

Resumen

Desde los inicios de la escritura se ha visto la necesidad de transmitir mensajes de manera que sean ocultos para aquellos que no sean el destinatario. Las técnicas de cifrar mensajes han sido desarrolladas desde tiempos remotos, lo cual permitió la proliferación de diversos métodos, algunos de ellos aun no han logrado ser totalmente desmantelados, como es el caso del cifrado por sustitución monoalfabética. En este trabajo se establece un método, basado en la descomposición en valores singulares y herramientas computacionales, que permite mejorar los resultados obtenidos con otros métodos de descifrado basados en el análisis de frecuencias.

Palabras Clave: Algebra Lineal Numérica, criptografía, Procesamiento de texto..

Abstract

From the beggining of writing we have seen the need to hide and transmit mes- sages in a way that they will be hidden to everybody but the receiver of the message. Techniques to encrypt messages have been developed from a long time ago, which al- lowed the proliferation for diverse methods, some of them have not been completely dismantled, like the monoalphabetic substitution encryption. In this work we are looking to establish a method, based on the singular value decomposition and computational tools, to improve the results obtained with other decoding methods.

Keywords: Numerical Linear Algebra, cryptography, Text Processing.

Contenido

1. Introducción

Existen diversos métodos para encriptar o cifrar un mensaje, *Giovanni Battista della Porta*(1535-1615) en su texto de cuatro volúmenes *furtivis literarum notis-vulgo de ziferis*[1] clasifica las técnicas de cifrado en tres grupos, que constituyen la criptografía clásica:

- Cifrado por trasposición
- Cifrado por sustitución
- Cifrado por sustitución por símbolos

En el cifrado por trasposición los caracteres cambian de posición pero mantienen su rol.
En el cifrado por sustitución los caracteres mantienen su posición, pero cambian de rol.
El cifrado de sustitución por símbolos, es un cifrado por sustitución en el cual el alfabeto original del mensaje no coincide con el alfabeto de cifrado, en esta técnica se usan alfabetos extraños para cifrar el mensaje, este procedimiento de cifrado fue utilizado por la orden de los Templarios quienes extraían los símbolos de un objeto conocido como la estrella de las ocho beatitudes, la cual fue utilizada como emblema de la orden.

En la actualidad los métodos de cifrado son mucho más sofisticados, entre ellos el más utilizado es el algoritmo RSA, creado por *Martin Gardner* publicado en 1977 en la revista *Scientific American*, basado en números primos de gran magnitud [[1],p.104], el cual emplea el modelo de clave pública. La confiabilidad que ofrece el algoritmo RSA permitió a *Phil Zimmerman*, en 1991, desarrollar el PGP (Pretty Good Privacy) que es un algoritmo de cifrado que funciona fácilmente en computadores domésticos. PGP utiliza conceptos de criptografía clásica y los combina con el algoritmo RSA.

Los métodos clásicos de cifrado siguen vigentes aún en nuestra era, en la que la tecnología es la principal herramienta, por esta razón llama la atención que el cifrado por sustitución no

[1]Esta obra resume el conocimiento criptográfico de la época

haya sido completamente desmantelado aún. En particular, la sustitución monoalfabética es una de las versiones más simples de los métodos de sustitución, sin embargo si tenemos un mensaje cifrado por sustitución monoalfabética en un alfabeto de n caracteres, tendremos que los posibles mensajes descifrados son $n!$.

El alfabeto que utilizamos en nuestro idioma consta de $n = 27$ caracteres (incluyendo la letra ñ), por tanto existen 27! posibles mensajes. Por supuesto no consideraremos todos los posibles mensajes, ya que almacenar, procesar y analizar 27! (del orden de 10^{28}) mensajes no es viable.

En el proceso de criptoanálisis de un mensaje, el punto de partida usual es el análisis de frecuencias, una técnica que permite relacionar caracteres a partir de la frecuencia con que aparecen en el mensaje, sin embargo el proceso de descifrado suele realizarse por comparación directa entre los resultados del análisis y tablas de frecuencias existentes para cada idioma, lo cual restringe dichos métodos a garantizar resultados satisfactorios cuando los mensajes son de gran longitud, debido a la ley de los grandes números. El principal problema de este tipo de técnicas de ataque a un mensaje cifrado es el no considerar las diferencias entre los caracteres del alfabeto, ya que las consonantes y vocales no son distinguibles.

En este trabajo se propone un algoritmo para descifrar mensajes cifrados por sustitución monoalfabética, dicho algoritmo se basa en los pilares del álgebra lineal con la que se establece un criterio para distinguir los caracteres que representan vocales de aquellos que representan consonantes en el mensaje cifrado, con lo cual se mejorará la precisión en el descifrado del mensaje. Para tal fin, este trabajo se dividió de tal manera que el lector podrá encontrar en el capítulo 2 los desarrollos a partir del análisis de frecuencias, seguido de las herramientas de álgebra lineal que permiten sustentar teóricamente la efectividad del algoritmo propuesto; también se incluyen los criterios de clasificación desarrollados. Las consideraciones adicionales que permiten mejorar la precisión del algoritmo propuesto son tratatadas con detalle en el capítulo 3, de este modo se proponé el algoritmo de descifrado. En el capítulo 4 se comparan los resultados obtenidos sobre un mensaje que se ataca utilizando aritmética modular, el análisis de frecuencias directamente y el método propuesto.

Las conclusiones de este trabajo se encuentran en el capítulo 5, seguidas por los apéndices que permitirán al lector conocer un poco más acerca de los métodos de cifrado por sustitución monoalfabética (Apéndice A), y de las herramientas de álgebra lineal empleadas (Apéndice B).

2. Descifrando la realidad

El proceso de conversión de un mensaje en texto plano a un texto cifrado, o viceversa, matemáticamente puede ser visto como una transformación entre dos espacios de caracteres, en particular dicha transformación puede ser una modificación de símbolos o un reordenamiento en la manera en la cual se presentan los caracteres. El proceso de construcción de dicha transformación, implica el diseño de algoritmos que permitan modificar la forma en que se presenta el mensaje, dependiendo de la función del algoritmo, puede ser un algoritmo de cifrado o de descifrado.

La principal garantía de que el mensaje se transmite de forma segura (incomprensible para cualquier persona que no sea el destinatario) es que el algoritmo de cifrado solo sea conocido por el destinatario y el emisario, sin embargo, debido a que el cifrado es de manera algorítmica, resulta probable la construcción de un algoritmo inverso que permita, dado un texto cifrado, descifrar un mensaje. Si dicho algoritmo inverso existe, se dice que el sistema criptográfico es *reversible*, es decir puedo a partir del texto plano construir un texto cifrado, y a partir de un texto cifrado reconstruir un texto plano, razón por la cual algunos autores dicen que el sistema es de dos vías.

En el caso de los mensajes cifrados por sustitución, sabemos que son altamente susceptibles al análisis de frecuencias [[1], p.39], dicho análisis puede ser considerado como el estudio de la frecuencia con la que aparecen los símbolos en los mensajes. El análisis de frecuencias fue desarrollado por el sabio Al-kindi (801-873), ha sido utilizado en diversas etapas de la historia, mostrando resultados bastante buenos, tanto así que muchos métodos de cifrado fueron creados con el fin de burlar este tipo de ataque.

La efectividad del análisis de frecuencias radica en que distintas letras no aparecen con la misma frecuencia en un mensaje, esto hace que algunas de ellas destaquen por su abundancia (e,a), y otras por su escasez (k,x), por ejemplo las vocales que, pese a ser solo 5, ocupan casi la mitad de la longitud de un texto en español. Este tipo de análisis nos permite determinar

que caracteres son más frecuentes en un mensaje, más aún que combinaciones de caracteres suelen verse con más frecuencia que otras, lo cual permite a un criptoanalista inferir sobre la correspondencia de los caracteres en el texto cifrado y en el texto plano.

El manejo de la información de un análisis de frecuencias puede ser algo complejo, ya que la cantidad de datos que se obtienen al calcular la frecuencia de aparición de dos caracteres consecutivos (bigramas), para cada carácter del alfabeto es elevada, por esta razón se almacenará en una matriz, ya que las matrices son una forma eficiente de almacenar información. Desde el punto de vista matemático, esto nos permite utilizar propiedades de algebra lineal, mientras que desde el punto de vista computacional nos ofrece la posibilidad de almacenar de manera óptima.

2.1. Matriz de frecuencias

El proceso de construcción de la matriz que nos permite visualizar un mensaje como un objeto matemático, se realiza de la siguiente manera:

Paso 1 Se enumeran las letras del alfabeto de cifrado.

Paso 2 Se asigna la i-ésima fila para la i-ésima letra del alfabeto.

Paso 3 Se asigna la j-ésima columna para la j-ésima letra del alfabeto.

Paso 4 En la posición i, j se ubica el número de veces en que la letra i-ésima es seguida por la letra j-ésima.

Paso 5 Se asume que el último carácter del mensaje es seguido por el primer carácter del mensaje.

Debemos tener en cuenta que la construcción de la matriz de frecuencias nos permite:

- Conocer el número de ocasiones en que el carácter i-ésimo sigue al caracter j-ésimo.
- Extraer el número de ocasiones en que un carácter aparece en el mensaje, basta multiplicar la matriz de frecuencias por un vector de 1's.
- Saber inmediatamente si un carácter aparece en el mensaje ya que, de no aparecer tendrá asociada una fila y columna nula.

- Conocer el número de caracteres utilizados en el mensaje.

Ejemplo 1. *Tomando como ejemplo la palabra* decada *en el alfabeto* $\{a, b, c, d, e\}$ *podemos construir la siguiente tabla, que llamaremos tabla de frecuencias, en la cual se cuenta el número de veces que el caracter ubicado en la parte izquierda es seguido del caracter en la parte superior, de esta tabla se extrae la matriz de frecuencias, en la cual debemos tener en cuenta una condición adicional, la última letra del mensaje es seguida por la primera letra del mensaje.*

	a	*b*	*c*	*d*	*e*
a	*0*	*0*	*0*	*2*	*0*
b	*0*	*0*	*0*	*0*	*0*
c	*1*	*0*	*0*	*0*	*0*
d	*1*	*0*	*0*	*0*	*1*
e	*0*	*0*	*1*	*0*	*0*

En este caso la última letra es a *y la primera es* d*, por tanto se toma como si la letra* a *estuviese seguida de la letra* d*, por esta razón en la tabla anterior aparece que la letra a es seguida por la letra d en dos ocasiones.*

En la primera fila el caracter a, con lo cual se entiende que es el primer caracter, b esta en la segunda fila, se entiende que es el segundo caracter. Enumerar los caracteres es equivalente a construir la tabla de frecuencias.
A partir de la tabla de frecuencias anterior se visualiza con claridad la matriz de frecuencias del mensaje, que será:

$$A = \begin{bmatrix} 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Tomemos el vector $e = (1\ 1\ 1\ 1\ 1)^t$*, luego*

$$Ae = b$$

$$\begin{bmatrix} 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 1 \\ 2 \\ 1 \end{bmatrix}$$

El resultado es $b = (2\ 0\ 1\ 2\ 1)^t$ *que es el vector en el cual se almacena la información de cuantas veces aparece cada caracter en el mensaje. La componente i-ésima del vector b representa las veces que el caracter i-ésimo aparece en el mensaje; en nuestro ejemplo el mensaje era la palabra* decada*, del vector b se concluye:*

- *El primer caracter (a) aparece 2 veces en la palabra*
- *El segundo caracter (b) aparece 0 veces en la palabra*
- *El tercer caracter (c) aparece 1 vez en la palabra*
- *El cuarto caracter (d) aparece 2 veces en la palabra*
- *El quinto caracter (e) aparece 1 veces en la palabra*

La matriz de frecuencias es una representación de un mensaje a través de una matriz de números naturales, lo cual facilita la comprensión desde el punto de vista del algebra lineal, no obstante debemos ofrecer mayor facilidad en el análisis criptográfico y eficiencia desde el punto de vista computacional, por esta razón debemos extraer la información más relevante de la matriz de frecuencias, de este modo podemos establecer una estrategia de trabajo.

Inicialmente podemos pensar en aplicar un método de descomposición sobre la matriz de frecuencias, así podemos reducir los tiempos de computo y además facilitar el análisis sobre la representación del mensaje codificado; solo falta decidir que método de descomposición matricial es conveniente.

2.2. Descomposición en Valores Singulares

El proceso de descomposición en valores singulares [ver teorema 4], conocida como SVD, puede ser utilizado sobre la matriz de frecuencias, la razón para usar esta descomposición es la amplia gama de propiedades y facilidades que ofrece, entre ellas el permitir descomponer la matriz de frecuencias en 3 matrices, las cuales establecen bases para el rango y el espacio nulo de la matriz de frecuencias, y sus complementos ortogonales [ver teorema 5]. El concepto de base implica tener una noción de independencia, esto nos será de utilidad para poder hablar del alfabeto como un conjunto formado por dos clases diferentes de elementos, vocales y consonantes.

Sea A la matriz de frecuencias del mensaje. Al aplicar la descomposición en valores singulares obtenemos:

$$A = X\Sigma Y^T \tag{2-1}$$

donde

- X y Y son matrices unitarias
- Σ es una matriz diagonal que contiene los valores singulares σ_i de la matriz A ordenados de mayor a menor

En este caso, como las componentes de la matriz de frecuencias son reales, podemos decir que X y Y son matrices ortogonales.

Recordemos que los valores singulares de la matriz A son las raíces cuadradas de los valores propios de la matriz A^tA, debido a que A^tA es simétrica definida positiva, tenemos que los valores singulares son reales (ver teorema 3).

La estrategia de trabajo que se propondrá se basa en los valores y vectores singulares de la matriz de frecuencias del mensaje, para poder determinarlos aplicaremos la descomposición matricial SVD ya que permite:

- La posibilidad de trabajar incluso si la matriz de frecuencias tiene filas o columnas nulas (es decir, si algún carácter del alfabeto no se encuentra en el mensaje).

- El no restringir el trabajo a matrices cuadradas (es decir, cuando el alfabeto del mensaje cifrado y el alfabeto del mensaje descifrado no tienen el mismo tamaño).
- La facilidad para determinar el rango de la matriz de frecuencias.
- La posibilidad de aproximar la matriz de frecuencias utilizando matrices de rango menor.

El poder aproximar la matriz de frecuecias, en términos de matrices de rango menor nos será de gran utilidad para realizar un criptoanálisis, ya que no debemos utilizar toda la matriz de frecuencias, basta con escoger correctamente las matrices de rango menor que permitan establecer un criterio.

Considerando la matriz de frecuencias A asociada al mensaje, y aplicando la descomposición SVD obtenemos

$$A = X\Sigma Y^T$$

donde

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ & \ddots & \ddots & \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix}$$

Luego, realizando algunos cálculos obtenemos la siguiente descomposición

$$\Sigma = \sum_{i=1}^{n} \psi_i$$

Donde, cada ψ_i tiene el tamaño de Σ y se define como.

$$\psi_i(k,j) = \begin{cases} \sigma_i & \text{si } i = k, i = j \\ 0 & \text{En otro caso} \end{cases}$$

Considerando las matrices X, Y con columnas $[X_1\ X_2\ \cdots\ X_n]$, y $[Y_1\ Y_2\ \cdots\ Y_n]$ respectivamente, podemos reescribir la descomposición en valores singulares como:

$$A = X\Sigma Y^t$$
$$A = \sum_{i=1}^{n} X\psi_i Y^t$$
$$A = \sum_{i=1}^{n} X_i\sigma_i Y_i^t$$

De donde obtenemos una expresión alternativa de la matriz de frecuencias en términos de sus valores singulares y vectores singulares. Como los valores singulares de la matriz de frecuencias son reales, obtenemos:

$$A = \sum_{i=1}^{n} \sigma_i X_i Y_i^t \tag{2-2}$$

Los vectores X_i y Y_i se denominan vectores singulares a izquierda y derecha, respectivamente.

En la ecuación 2-2 se observa la manera en la que podemos escribir la matriz de frecuencias A en términos de los valores singulares y los vectores singulares. Cabe destacar que si un valor singular σ_k es nulo, no aportará información a la matriz de frecuencias; partiendo de este hecho podemos decir que si un valor singular es muy cercano a cero, no aportará una cantidad significativa de información por esta razón al considerarlos nulos, no se presentará gran pérdida de información.

En particular, recordando que los valores singulares aparecen ordenados de mayor a menor, podemos suponer que los últimos valores singulares son cercanos a cero, por tanto no representan una parte significativa dentro de la matriz y por lo que pueden ser obviados. Siguiendo esta idea, si suponemos que el primer valor singular σ_1 es de gran magnitud en comparación con los demás σ_i podemos decir que la mayor parte de la información consignada en la matriz A recae en el valor singular σ_1 y por ende los demás valores singulares pueden ser obviados; así obtenemos una aproximación de la matriz de frecuencias A en términos del primer valor singular σ_1, la cual se conoce como aproximación de rango 1 de la matriz A.

2.2.1. Aproximación de Rango 1

Suponiendo que el valor singular σ_1 es de gran magnitud en comparación con los demás valores singulares, los cuales supondremos son cercanos a cero, obtenemos:

$$A \approx \sigma_1 X_1 Y_1^T \tag{2-3}$$

La ecuación anterior se denomina *Aproximación de Rango 1.*

Esta aproximación de la matriz A nos permite trabajar con mayor facilidad ya que no debemos considerar todos los términos de la aproximación para aplicar propiedades de algebra lineal. Cabe destacar que la matriz de frecuencias es una matriz positiva (matriz cuyas entradas son mayores o iguales que 0), por esta razón podemos decir que los vectores singulares asociados a la aproximación de Rango 1 poseen todas sus componentes positivas (se deduce del teorema de Perron Frobennius [ver teorema 7]).

Intrínsecamente, al tomar la aproximación de rango 1, estamos asumiendo que el alfabeto considerado es un conjunto sin particiones, es decir, catalogamos todos los caracteres como si fuesen del mismo tipo (no se distinguen vocales ni consonantes)[ver teorema 5]. No obstante, pese a esta limitación podemos extraer información interesante, por ejemplo la frecuencia con la que aparece un caracter en el mensaje, para ello basta considerar el vector

$$e = [1\ 1 \cdots 1]^t$$

El cual se multiplica a la matriz de frecuencias

$$Ae = b$$

Obteniendo como resultado el vector b, en cuyas filas se encuentra el número de veces que aparece cada carácter.

La aproximación de rango 1 nos permite determinar un orden entre los caracteres que aparecen en un mensaje, dicho orden se puede establecer ordenando los caracteres de mayor a menor aparición en el mensaje, este orden varia respecto al idioma en el cual se escriba el mensaje, ya que por ejemplo, en español es muy poco probable que aparezca la letra w, mientras que inglés es frecuente verla en las preguntas. También tenemos en español la aparición del símbolo ñ que no está presente en el idioma inglés.

Debemos tener en cuenta que la construcción de una tabla de frecuencias de aparición de caracteres no es tarea simple ya que puede ser susceptible de interpretación[1]. Para evitar el mayor número de ambigüedades, los encargados de diseñar la tabla de frecuencias consideran textos de longitudes exageradas, para así reducir el sesgamiento.

Los porcentajes de aparición de cada carácter, en el lenguaje español son los siguientes:

A	11,96 %	B	0,92 %	C	2,92 %
D	6,87 %	E	16,78 %	F	0,52 %
G	0,73 %	H	0,89 %	I	4,15 %
J	0,30 %	K	0,01 %	L	8,37 %
M	2,12 %	N	7,01 %	Ñ	0.29 %
O	8,69 %	P	2,77 %	Q	1,53 %
R	4,94 %	S	7,88 %	T	3.31 %
U	4.80 %	V	0.39 %	W	0.01 %
X	0.06 %	Y	1.54 %	Z	0.15 %

Tabla 2-1.: Tomado de [[1], p.39]

El cálculo de la frecuencia de aparición de las letras en una lengua es difícil y está sujeto a condicionamientos. Se cuenta la frecuencia de las letras de un texto arbitrariamente largo, pero en los resultados influyen varios parámetros, incluso algunos autores consideran el espacio como un carácter.

Al contar con la tabla de frecuencias de aparición de caracteres, y la aproximación de Rango 1, el criptoanalista podrá intentar realizar una sustitución directa debido a la frecuencia que muestre cada carácter. Supongamos que tenemos un mensaje cifrado, si continuamos con la idea descrita, aplicando la aproximación de rango 1, el criptoanalista lograra obtener el carácter de mayor frecuencia, si se sabe (o se cree) que el mensaje esta en español, resulta razonable relacionar el carácter más frecuente del mensaje con la letra *e*, siguiendo con esta idea se relacionaría el segundo caracter más frecuente del mensaje con la letra *a*, y así sucesivamente para cada caracter.

En palabras de Al-Kindi

[1] En este tipo de labores, los signos de puntuación son ignorados

"Una manera de resolver un mensaje cifrado, si sabemos en qué lengua está escrito, es encontrar un texto llano escrito en la misma lengua, suficientemente largo, y luego contar cuantas veces aparece cada letra. A letra que aparece con más frecuencia la llamamos "primera", a la siguiente en frecuencia la llamaremos "segunda"....y así hasta que hayamos cubierto todas las letras que aparecen en nuestro texto. Luego observamos el texto cifrado que queremos resolver y clasificamos sus símbolos de la misma manera. Encontramos el símbolo que aparece con mayor frecuencia y lo sustituimos por la "primera"de la nuestro texto, hacemos lo mismo con la "segunda", y así sucesivamente, hasta que hayamos cubierto todos los símbolos del criptograma que queremos resolver".[[3],p.125]

Esta técnica no siempre será efectiva, pero resulta ser un muy buen punto de partida ya que, es posible que algunos caracteres sean descifrados. No obstante, extraer información de la matriz de fecuencias a partir de una aproximación de rango 1 nos limita a considerar todo el alfabeto como un conjunto sin particiones, en el cual solo consideramos la frecuencia de aparición de cada caracter. La matriz de frecuencias permite determinar cuántas veces un carácter es seguido por otro, lo cual es de gran importancia ya que en español hay bastantes ideas que se deben considerar y contar con esa información será de gran apoyo, pues permite restringir las posibles soluciones. Entre las tantas ideas que permite brindar la frecuencia de aparición de un carácter seguido de otro están:

- Existen combinaciones de dos caracteres iguales que se presentan con frecuencia en el lenguaje español.(ll,rr,ee).
- Existen combinaciones de caracteres que en español no se presentan ($\tilde{n}t$,fg).
- Existen reglas lingüísticas que permiten establecer, hasta cierto punto, que tipo de carácter sigue a otro (esta idea se profundizará más adelante).
- Si se identifica algún caracter en el mensaje descifrado, de seguro se podra intuir algo acerca del carácter que más veces le sigue.

Estas ideas podrían desarrollarse con mayor facilidad si, de alguna manera, se lograse identificar en el mensaje cifrado cuales caracteres corresponden a vocales y cuales corresponden a consonantes. No obstante, esto no es posible determinarlo solo con la aproximación de rango 1, para ello se debe establecer un nivel mayor de precisión en la apoximación de la matriz de frecuencias, además de establecer ciertas reglas del lenguaje natural.

2.2.2. Aproximación de Rango 2

Retomemos la idea de escribir la matriz A en la forma de la ecuación (2)

Esta vez supondremos que los dos primeros valores son grandes en magnitud en comparación a los demás, de esta manera obtenemos:

$$A = \sigma_1 X_1 Y_1^T + \sigma_2 X_2 Y_2^T$$

Debido a que en esta ocasión consideramos el segundo valor singular de la matriz de frecuencias, no podemos garantizar que los vectores propios asociados tengan siempre términos positivos, ya que el Teorema de Perron Frobenius [ver teorema 6], establece una condición para el mayor valor singular y su correspondiente vector singular asociado. Debemos tener en cuenta que la aplicación de la aproximación de rango 2, sugiere que nuestro lenguaje de codificación puede ser visto como dos conjuntos que lo particionan [ver teorema 5](dos conjuntos disjuntos que al unirse forman todo el alfabeto).

Estableciendo como punto de partida la aproximación de rango 2, se pueden sugerir diversas particiones para el alfabeto del mensaje, sin embargo debemos escoger la más objetiva ya que nuestro interés es poder establecer un proceso algorítmico, por esta razón la partición del alfabeto se hará en dos conjuntos:

- Letras del alfabeto que son vocales
- Letras del alfabeto que son consonantes

Ninguna letra es consonate y vocal a la vez, por tanto los conjuntos son disjuntos.
Al unir el conjunto de las vocales con el de las consonantes obtenemos todo el alfabeto. Con lo cual se garantiza la escogencia realizada es una partición del alfabeto.

Esta idea lingüística, pese a ser muy sencilla, puede parecer algo de poca utilidad en el desarrollo algebraico que hemos realizado, no obstante podemos empalmar la idea propuesta con lo desarrollado hasta el momento, para ello utilizamos el vector $e = (1\ 1 \cdots\ 1)^t$ que hemos mencionado anteriormente.

Construyamos los vectores V y C, correspondientes a vocales y consonantes, de la siguiente manera

$$V(i) = \begin{cases} 1 & si \quad \text{la i-ésima letra del alfabeto es vocal} \\ 0 & si \quad \text{la i-ésima letra del alfabeto es consonante} \end{cases}$$

$$C(i) = \begin{cases} 1 & si \quad \text{la i-ésima letra del alfabeto es consonante} \\ 0 & si \quad \text{la i-ésima letra del alfabeto es vocal} \end{cases}$$

De este modo podemos enlazar con el trabajo que hemos realizado ya que el vector de unos que hemos denominado e puede expresarse como

$$e = V + C$$

Esta construcción de V y C nos permite extraer información de la matriz de frecuencias de manera mucho más ágil, para ello basta observar las siguientes propiedades:

$$\begin{aligned} C^tAV &= \text{Número de ocasiones en que una consonante es seguida de una vocal} \\ V^tAV &= \text{Número de ocasiones en que una vocal es seguida de una vocal} \\ V^tAC &= \text{Número de ocasiones en que una vocal es seguida de una consonante} \\ C^tAC &= \text{Número de ocasiones en que una consonante es seguida de una consonante} \\ V^tAe &= \text{Número de vocales} \\ C^tAe &= \text{Número de consonantes} \end{aligned}$$

Resulta evidente que $V^TAe + C^TAe$ es el número de letras del alfabeto que son usadas.

Ejemplo 2. *Considere el texto "cada decada", y determine:*

- *El número de veces en que una consonante es seguida por una vocal*
- *El número de veces en que una vocal es seguida por una vocal*
- *El número de veces en que una vocal es seguida por una consonante*
- *El número de veces en que una consonante es seguida por una consonante*
- *El número de vocales*

- *El número de consonantes*
- *El número total de caracteres*

El mensaje propuesto tiene por alfabeto $\Lambda = \{a, b, c, d, e\}$.

Para poder determinar todo lo anterior, debemos primero concentrarnos en la matriz de frecuencias asociada al mensaje, la cual es dada por

$$\begin{bmatrix} 0 & 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

cabe destacar que la primera fila corresponde a la primera letra del alfabeto Λ*, la segunda fila corresponde a la segunda letra del alfabeto y así sucesivamente.*

En el alfabeto Λ *el primer y el quinto caracter son vocales (a,e respectivamente), mientras el segundo, tercer y cuarto caracter son consonantes (b,c,d respectivamente), por tanto obtenemos los siguientes vectores*

$$V = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \qquad C = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \qquad e = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Por tanto, tendremos:

$C^tAV = 5$ *ocasiones en que una consonante es seguida de una vocal*

$V^tAV = 0$ *ocasiones en que una vocal es seguida de una vocal*

$V^tAC = 5$ *ocasiones en que una vocal es seguida de una consonante*

$C^tAC = 0$ *ocasiones en que una consonante es seguida de una consonante*

$V^tAe = 5$ *vocales*

$C^tAe = 5$ *consonantes*

$V^tAe + C^tAe = 10$ *caracteres utilizados en el mensaje*

2.3. Construcción del criterio de clasificación

Por supuesto, el método puede ser susceptible de error, por ello debemos considerar un caso en el cual el criterio no decida o simplemente no resulte ser lo suficientemente claro como para dar un veredicto; pensando en ello retomamos la aproximación de rango 2

$$A = \sigma_1 X_1 Y_1^T + \sigma_2 X_2 Y_2^T$$

A partir de la cual construimos los siguientes vectores que representaran los casos en que, según el criterio, el carácter es una vocal o una consonante, o el criterio no decide:

$$C(i) = \begin{cases} 1 & si \quad X_2(i) > 0 \text{ y } Y_2(i) < 0 \\ 0 & \text{en otro caso} \end{cases}$$

$$V(i) = \begin{cases} 1 & si \quad X_2(i) < 0 \text{ y } Y_2(i) > 0 \\ 0 & \text{en otro caso} \end{cases}$$

$$N(i) = \begin{cases} 1 & si \quad X_2(i)Y_2(i) > 0 \\ 0 & \text{en otro caso} \end{cases}$$

La forma en que se definen los vectores es la siguiente:

- Si $X_2(i)$ es positivo y $Y_2(i)$ es negativo, el carácter es consonante
- Si $X_2(i)$ es negativo y $Y_2(i)$ es positivo, el carácter es vocal
- Si $X_2(i)$ y $Y_2(i)$ tienen el mismo signo, el criterio no decide sobre el i-ésimo carácter

El funcionamiento sigue siendo el mismo, ya que

$$e = (1\ 1 \cdots\ 1)^T = N + C + V$$

Cabe destacar que se utiliza como criterio de asignación el cambio de signo de los vectores singulares, por esta razón el método no se establece con los vectores singulares asociados al radio espectral de la matriz, ya que el radio espectral tiene asociados vectores con entradas positivas [ver teorema 7], con lo cual las componentes de los vectores no cambiaran de signo y, según la construcción, al no presentarse cambio de signo el criterio no decide.

Ejemplo 3. *Consideremos el fámoso poema "Nocturno"[[10],p.121][2], compuesto por José Asunción Silva (1865-1898) publicado en la revista "Lectura Para Todos", en la ciudad de Cartagena en 1894.*

Los resultados al analizar los caracteres partiendo de la matriz de frecuencias asociadas al poema fueron:

	V	C	N		V	C	N
A	1	0	0	N	0	1	0
B	0	1	0	O	1	0	0
C	0	1	0	P	0	1	0
D	0	1	0	Q	0	1	0
E	1	0	0	R	0	1	0
F	0	0	1	S	0	1	0
G	0	1	0	T	0	0	1
H	0	0	1	U	1	0	0
I	0	0	1	V	0	1	0
J	0	1	0	W	0	0	0
K	0	0	0	X	0	0	0
L	0	1	0	Y	0	1	0
M	0	1	0	Z	0	1	0

Tabla 2-2.: Asignación de caracteres correspondiente al poema "Nocturno".

El valor numérico corresponde al valor que toma cada vector en la fila correspondiente a cada caracter.

*Tomando los resultados obtenidos en **2-2** podemos decir:*

- *Para 4 caracteres el criterio no decide.*
- *Los caracteres para los cuales los vectores V,C,N son nulos son aquellos que no aparecen en el poema.*
- *Los caracteres asignados en V y N fueron asignados correctamente.*

[2] Un fragmento de dicho poema puede ser visto en el reverso del billete de 5000 pesos colombianos, en el frente se encuentra su autor José Asunción Silva.

3. Consideraciones adicionales

3.1. Regla vfc

Las aproximaciones de Rango 1 y de Rango 2 nos permiten establecer un criterio de asignación de caracteres, con los cuales estamos en condiciones de proponer posibles mensajes descifrados.

Sin embargo, para poder restringir el conjunto de posibles soluciones debemos ubicar una restricción lingüística que nos permitirá ajustarnos más a la realidad, ya que cada cada idioma posee sus propias reglas.

En particular existe una regla, conocida como *regla vfc*, la cual enuncia que es más frecuente que las consonantes sean seguidas por vocales y no que las vocales sean seguidas por vocales.

En el momento en que se aplica la aproximación de Rango 2 debemos analizar la partición del alfabeto que se propone, ya que no sabemos que corresponde a vocal ni que corresponde a consonante, es necesario verificar que la partición propuesta cumpla la regla vfc pues de no ser así la clasificación como vocal o consonante será errónea.

Matemáticamente, la condición *vfc* se puede escribir como:

$$\frac{\text{Número de vocales seguidas por vocal}}{\text{Número de vocales}} < \frac{\text{Número de consonantes seguidas por vocal}}{\text{Número de consonantes}}$$

En la aproximación de rango 2 se propuso una partición que permitía crear los vectores V,C en los cuales se establecía si un caracter corresponde a una vocal, a una consonante. Utilizando estas definiciones, podemos escribir la regla vfc en términos de dichos vectores.

$$\frac{V^t A V}{V^t A (V + C)} < \frac{C^t A V}{C^t A (V + C)}$$

De donde se obtiene

$$[V^tAV][C^tAC] < [V^tAC][C^tAV] \tag{3-1}$$

Teorema 1. *La partición del alfabeto en vocales y consonantes propuesta en la aproximación de Rango 2, satisface la regla* vfc

Demostración. Establecemos inicialmente la partición de Rango 2 de la matriz de frecuencias, luego

$$A = \sigma_1 X_1 Y_1^t + \sigma_2 X_2 Y_2^t$$

Debemos probar que la ecuación (3-1) se cumple, para ello tomamos la aproximación de Rango 2 y reemplazando, obtenemos:

$$\begin{aligned}
V^tAV &= [V^t(\sigma_1 X_1 Y_1^t + \sigma_2 X_2 Y_2^t)V] \\
C^tAC &= [C^t(\sigma_1 X_1 Y_1^t + \sigma_2 X_2 Y_2^t)C] \\
V^tAC &= [V^t(\sigma_1 X_1 Y_1^t + \sigma_2 X_2 Y_2^t)C] \\
C^tAV &= [C^t(\sigma_1 X_1 Y_1^t + \sigma_2 X_2 Y_2^t)V].
\end{aligned}$$

De lo anterior se observa que

$$\begin{aligned}
[V^tAV][C^tAC] &= [V^t(\sigma_1 X_1 Y_1^t + \sigma_2 X_2 Y_2^t)V]C^t(\sigma_1 X_1 Y_1^t + \sigma_2 X_2 Y_2^t)C] \\
&= [V^t(\sigma_1 X_1 Y_1^t + \sigma_2 X_2 Y_2^t)VC^t(\sigma_1 X_1 Y_1^t + \sigma_2 X_2 Y_2^t)C].
\end{aligned}$$

En la ecuación anterior aparece el término VC^t, el cual es el producto punto entre los vectores C y V. Sin embargo, este producto es nulo ya que ningún caracter es vocal y consonante a la vez. Por tanto, la expresión

$$[V^tAV][C^tAC] < [V^tAC][C^tAV]$$

equivalente a

$$[V^tAV][C^tAC] - [V^tAC][C^tAV] < 0$$

Se transforma en:

$$-[V^tAC][C^tAV]$$

Al reemplazar la aproximación de rango 2 se convierte en

$$-[V^t(\sigma_1X_1Y_1^t+\sigma_2X_2Y_2^t)C][C^t(\sigma_1X_1Y_1^t+\sigma_2X_2Y_2^t)V]$$

continuando con los cálculos, se tiene que:

$$\begin{aligned}
-[V^tAC][C^tAV] &= \\
-[V^t(\sigma_1X_1Y_1^t+\sigma_2X_2Y_2^t)C][C^t(\sigma_1X_1Y_1^t+\sigma_2X_2Y_2^t)V] &= \\
-[V^t(\sigma_1X_1Y_1^t+\sigma_2X_2Y_2^t)CC^t(\sigma_1X_1Y_1^t+\sigma_2X_2Y_2^t)V] &= \\
-[V^t(\sigma_1X_1Y_1^t+\sigma_2X_2Y_2^t)\|C\|(\sigma_1X_1Y_1^t+\sigma_2X_2Y_2^t)V]. &
\end{aligned}$$

Distribuyendo en la ecuación se deduce

$$-\sigma_1^2\|C\|(V^tX_1Y_1^tX_1Y_1^tV)-\sigma_2^2\|C\|(V^tX_2Y_2^tX_2Y_2^tV)-\sigma_1\sigma_2\|C\|(V^tX_1Y_1^tX_2Y_2^tV+V^tX_2Y_2^tX_1Y_1^tV)$$

Como cada X_i y Y_j son vectores columnas, los productos $X_2Y_2^t$ y $X_1Y_1^t$ son escalares, por tanto la expresión anterior se convierte en:

$$\begin{aligned}
-\sigma_1^2\|C\|(X_1Y_1^t)^2(V^tV)-\sigma_2^2\|C\|(X_2Y_2^t)^2(V^tV)-\sigma_1\sigma_2\|C\|(V^t(2(X_2Y_2^t)(X_1Y_1^t)V) &= \\
-\sigma_1^2\|C\|(X_1Y_1^t)^2\|V\|-\sigma_2^2\|C\|(X_2Y_2^t)^2\|V\|-\sigma_1\sigma_2(2(X_2Y_2^t)(X_1Y_1^t))\|C\|\|V\| &= \\
\|C\|\|V\|\left[-\sigma_1^2(X_1Y_1^t)^2-\sigma_2^2(X_2Y_2^t)^2-2\sigma_1\sigma_2(X_2Y_2^t)(X_1Y_1^t)\right] &= \\
-\|C\|\|V\|\left[\sigma_1(X_1Y_1^t)+\sigma_2(X_2Y_2^t)\right]^2 &
\end{aligned}$$

Dado que $\|C\|\geq 0$, $\|V\|\geq 0$ y $\left[\sigma_1(X_1Y_1^t)+\sigma_2(X_2Y_2^t)\right]^2\geq 0$, obtenemos

$$-\|C\|\|V\|\left[\sigma_1(X_1Y_1^t)+\sigma_2(X_2Y_2^t)\right]^2\leq 0$$

Para finalizar, basta recordar que los conjuntos C y V son no vacíos, lo cual hace que $\|C\|,\|V\|>0$.

Por tanto, la partición propuesta en la aproximación de Rango 2, satisface la regla vfc □

Como la partición propuesta satisface la regla *vfc*, entonces podemos aplicar la partición del alfabeto en vocales y consonantes en la construcción de nuestro método de descifrado.

3.2. Bigramas

El proceso de clasificación de caracteres obtenido de la aproximación de rango 2, combinado con las aproximaciones de rango 1, nos permite establecer la ubicacion de algunos caracteres en el mensaje, sin embargo para poder aumentar la precisión, se sugiere utilizar el concepto de k-grama, el cual definiremos como una cadena de caracteres de longitud k.

Similar al estudio de la tabla de frecuencias de aparición de caracteres de cada lenguaje, se puede establecer los k-gramas más frecuentes en cada idioma. Cabe destacar que la frecuencia de aparición de un caracter es el caso especial en el cual se estudia la frecuencia de aparición de 1-gramas.

La matriz de frecuencia nos permite extraer más información, para ello nos restringiremos al uso de algunos 2-gramas o bigramas [[6],p.115-125], con el fin de mejorar la precisión del descifrado. No obstante, se corren algunos riesgos ya que la longitud del mensaje cifrado juega un papel vital en la efectividad del descifrado.

Utilizar 3-gramas, mejoraría aún más la precisión del descifrado, luego aplicar 4-gramas mejoraría aún más la precisión; sin embargo, el uso de k-gramas muestra menor mejora a medida que se consideran cadenas con valores de k en aumento, es decir, el aumento de precisión obtenido con 2-gramas es superior al obtenido al considerar luego 3-gramas y este a su vez será superior al obtenido al considerar luego 4-gramas.

En pocas palabras, considerar más alla de los bigramas (2-gramas) implicaría mayor costo computacional y mayores consideraciones teóricas para obtener una mejora muy leve. Claro está, se consideraran solo algunos bigramas asociados al carácter de mayor aparición en el idioma correspondiente, ya que se reduce la probabilidad de error, puesto que el carácter más frecuente es normalmente el más probable de ubicar en un mensaje cifrado. El uso de algunos bigramas no implica el considerar más información sobre el mensaje, ya que se pueden extraer a partir de la matriz de frecuencias asociada.

3.3. Entropía

En 1949 Claude Shannon, en su publicación *Communication Theory of Secrecy Systems*[[9], p.45], definió la *entropía* de un mensaje como el número de bits (unidad de información) necesarios para representarlo bajo una codificación óptima. En pocas palabras, la entropía

es una medida de incertidumbre con la cual se busca conocer a priori la seguridad de la infromación. Por esta razón, al aplicar la definición de entropía sobre un mensaje cifrado, tenemos una variable aleatoria X que representará a los posibles mensajes de texto plano que se pueden obtener.

Definicion 1. *Sea W un evento que ocurre con probabilidad $P(W)$, definimos la* ***información*** *obtenida por la ocurrencia del evento como*

$$-\log_2 P(W)$$

Esta definición nos indica que un evento con poca probabilidad de ocurrir aportará gran cantidad de información, mientras un evento casi seguro no aportará información considerable.

Definicion 2. *Sea X una variable aleatoria discreta que puede tomar los valores x_1,x_2,...,x_n, sea p el vector de probabilidades $p = (P(X = x_1), \ldots, P(X = x_n))$. Definimos la* ***entropía de Shannon*** *como:*

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Ejemplo 4. *Determinar la entropía de una variable aleatoria con distribución Bernoulli*

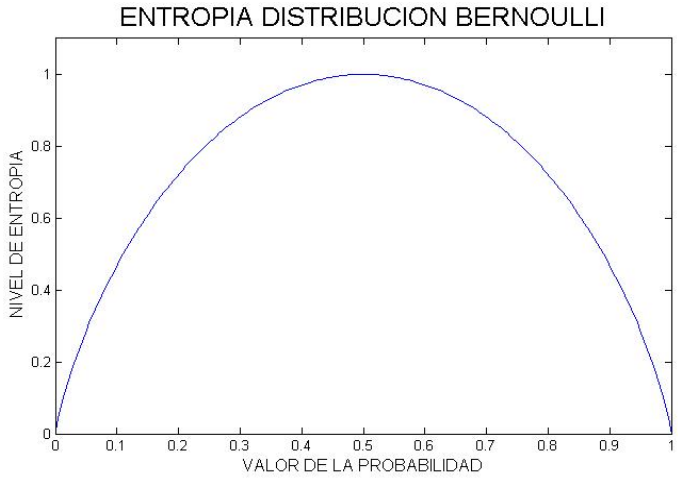

Figura 3-1.: Entropía de una variable aleatoria bernoulli

A partir de la definición, y el ejemplo anterior, podemos realizar las siguientes observaciones:

- Se asume que $0(\log(0)) = 0$, con el fin de garantizar continuidad.
- Entre más probable resulte ser un valor, menos información recibimos por su aparición
- Un evento seguro ($p_i = 1$ o $p_i = 0$) no aporta información, lo cual se concluye como caso límite del item anterior.
- El caso en que se consideran dos eventos equiprobables arroja como resultado un valor de entropía 1, el cual se escoge como unidad de medida.
- Cuando los eventos tienden a ser equiprobables, la entropía tiende a aumentar (debido a la continuidad de la entropía).
- En la definición de entropía se considera el logaritmo base 2 ($\log_2$), precisamente para que el caso de dos eventos equiprobables resulte ser la unidad.

Teorema 2. *$H(X) \leq \log_2 n$, además $H(x) = \log_2 n$ si $P_i = \frac{1}{n}$, para cada i*

Demostración. Para probar que $H(X) \leq \log_2 n$, basta demostrar que $\log_2 n$ es el valor máximo, para ello consideremos el siguiente problema de optimización

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

sujeto a la restricción

$$g(p_1, \ldots, p_n) = \sum_{i=1}^{n} p_i = 1$$

Este problema puede ser visto como un problema de multiplicadores de Lagrange. Luego:

$$\nabla H(p_1, \ldots, p_n) = \lambda \nabla g(p_1, \ldots, p_n)$$

Sin pérdida de generalidad, consideremos el término i-ésimo de cada gradiente, así obtenemos:

$$
\begin{aligned}
-\left(p_i \log_2 p_i\right)' &= \lambda (p_i)' \\
-\left(p_i \frac{\ln p_i}{\ln 2}\right)' &= \lambda (p_i)' \\
-\left(\log_2 p_i + \frac{1}{\ln 2}\right) &= \lambda \\
-\frac{\ln p_i + 1}{\ln 2} &= \lambda \\
-\ln p_i &= \lambda \ln 2 + 1 \\
\ln\left(\frac{1}{p_i}\right) &= \lambda \ln 2 + 1 \\
\frac{1}{p_i} &= e^{\lambda \ln 2} e \\
\frac{1}{p_i} &= 2^{\lambda} e \\
p_i &= \frac{1}{2^{\lambda} e} \qquad (*)
\end{aligned}
$$

Recordemos que $\sum_{i=1}^{n} p_i = 1$, luego

$$
\begin{aligned}
\sum_{i=1}^{n} p_i &= \sum_{i=1}^{n} \frac{1}{2^{\lambda} e} \\
1 &= \frac{n}{2^{\lambda} e} \\
n &= 2^{\lambda} e \\
\lambda &= \log_2\left(\frac{n}{e}\right) \qquad (**)
\end{aligned}
$$

Por tanto, reemplazando (**) en (*), obtenemos

$$p_i = \frac{1}{n}$$

De donde se concluye que si tomamos $p_i = \frac{1}{n}$, para cada i (eventos equiprobables) el valor de la entropía es máximo, basta ahora calcular dicho valor. Para ello tomamos

$$H(x) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

El cual será calculado como:

$$H(x) = -\sum_{i=1}^{n} \left(\frac{1}{n}\right) \log_2 \frac{1}{n}$$

Por propiedades de logaritmo $-\log_2\left(\frac{1}{n}\right) = \log_2 n$, por tanto:

$$H(x) = \sum_{i=1}^{n} \left(\frac{1}{n}\right) \log_2 n$$

Calculando la sumatoria, obtenemos

$$H(x) = n\left(\frac{1}{n}\right) \log_2 n$$

De donde se concluye

$$H(x) = \log_2 n$$

□

El teorema anterior, nos permite reconocer bajo que condiciones la entropía alcanza su valor máximo, no obstante debemos recordar que un evento con probabilidad alta de ocurrencia no brinda gran información, sin embargo un evento con poca probabilidad de ocurrencia brindará bastante información.

La entropía nos proporciona una forma de medir la cantidad de información que esperamos obtener de un evento, de este modo podemos estimar la incertidumbre que proporciona una variable aleatoria. Entre mayor sea la entropía, mayor será el valor de la incertidumbre.

4. Resultados

A continuación se presenta el algoritmo de descifrado construido a partir de los conceptos analizados a lo largo de este trabajo:

P1 Inicialmente debemos tomar el mensaje cifrado para aplicar una transformación inyectiva entre la galería de símbolos usados y un conjunto de símbolos más familiares, esto con el fin de trabajar el mensaje con mayor facilidad utilizando un computador. Este proceso se lleva a cabo si el alfabeto de cifrado no es el cotidiano.

P2 A partir del mensaje cifrado, realizamos la construcción de la matriz de frecuencias correspondiente, en la cual se condensara la información de los caracteres utilizados en el mensaje.

P3 Se aplica la aproximación de Rango 1

P4 Se propone como solución la sustitución de texto plano basándonos en la tabla de frecuencias de cada carácter (depende del idioma)

P5 Se aplica la aproximación de Rango 2

P6 Se propone la partición del alfabeto, en nuestro caso será en vocales y consonantes

P7 Se verifica que la partición propuesta cumpla la regla *vfc*

P8 Se establece un criterio combinando las aproximaciones de Rango 1 y de Rango 2, y el uso de algunos bigramas.

P9 Se propone como solución el resultado obtenido.

Este procedimiento descrito nos permitirá tener un acercamiento al objetivo de descifrar un mensaje. El resultado final será un criterio sobre cada carácter, en el cual se aclara si, según

el algoritmo, se trata de una consonante, una vocal o si no es posible determinarlo. También podremos ir más allá y establecer, con cierta precisión, que carácter del mensaje cifrado corresponde a cada carácter en el mensaje real, de este modo es posible postular posibles soluciones que, aunque no sean exactas siempre, ayudarán en el proceso de descifrado del mensaje.

Con el fin de establecer la efectividad del método propuesto, supondremos la siguiente situación: El usuario tiene un mensaje cifrado del cual desea conocer su contenido, posiblemente el cifrado sea por sustitución, de ser así será susceptible al análisis de frecuencias y podrá ser descifrado, para ello aplicaremos los siguientes métodos de descifrado:

- Aritmética modular
- Sustitución directa a partir de la tabla de frecuencias
- El Algoritmo propuesto

Ejemplo 5. *Descifrar el siguiente mensaje:*

"Oz xrvmxrz vh vo xlmqfmgl wv xlmlxrnrvmglh hrhgvnzgrxznvmgv vhgifxgfizwlh lygvmrwlh nvwrzmgv oz lyhviezxrlm wv kzgilmvh ivtfozivh, wv izalmznrvmglh b wv vckvirnvmgzxrlm, vm znyrglh vhkvxrurxlh wv olh xfzovh hv tvmvizm kivtfmgzh, hv xlmhgifbvm srklgvhrh, hv wvwfxvm kirmxrkrlh b hv vozylizm ovbvh tvmvizovh b vhjfvnzh nvglwrxznvmgv litzmrazwlh. Oz xrvmxrz fgroraz wruvivmgvh nvglwlh, b gvxmrxzh kziz oz zwjfrhrxrlm b litzmrazxrlm wv xlmlxrnrvmglh hlyiv oz vhgifxgfiz wv fm xlmqfmgl wv svxslh hfurxrvmgvnvmgv lyqvgrelh b zxxvhryovh z ezirlh lyhviezwlivh, zwvnzh wv yzhzihv vm fm xirgvirl wv eviwzw b fmz xliivxxrlm kvinzmvmgv.

Oz zkorxzxrlm wv vhlh nvglwlh b xlmlxrnrvmglh yfhxz oz tvmvizxrlm wv nzh xlmlxrnrvmgl lyqvgrel vm ulinz wv kivwrxxrlmvh xlmxivgzh, xfzmgrgzgrezh b xlnkilyzyovh, ivuvirwzh z svxslh lyhviezyovh kzhzwlh, kivhvmgvh, b ufgfilh. Xlm uivxfvmxrz vhzh kivwrxxrlmvh kfvwvm ulinfozihv nvwrzmgv izalmznrvmglh b vhgifxgfizihv xlnl ivtozh l ovbvh tvmvizovh, jfv wzm xfvmgz wvo xlnkligznrvmgl wv fm hrhgvnz, b kivwrxvm xlnl zxgfziz wrxsl hrhgvnz vm wvgvinrmzwzh xrixfmhgzmxrzh. Zotfmlh wvhxfyirnrvmglh xrvmgrurxlh kfvwvm ivhfogzi xlmgizirlh zo hvmgrwl

xlnfm. Vqvnkolh wv vhgl hlm oz gvlirz zglnrxz l oz nvxzmrxz xfzmgrxz jfv wvhzurzm mlxrlmvh xlnfmvh hlyiv oz nzgvirz. Nfxszh xlmxvkxrlmvh rmgfrgrezh wv oz mzgfizovaz szm hrwl gizmhulinzwzh z kzigri wv szoozatlh xrvmgrurxlh xlnl vo nlernrvmgl wv gizhozxrlm wv oz grviiz zoivwvwli wvo hlo"

A continuación se presentan los resultados obtenidos al aplicar cada uno de los distintos métodos de descifrado mencionados.

4.1. Resultado al descifrar utilizando Aritmética Modular

"Xi gaevgai eq ex guvzovpu fe guvugawaevpuq qaqpewipagiwevpe eqprogporifuq uhpevafuq wefaivpe xi uhqernigauv fe tipruveq recoxireq, fe rijuviwaevpuq k fe elterawevpigauv, ev iwhapuq eqtegadaguq fe xuq goixeq qe ceveriv trecovpiq, qe guvqprokev batupeqaq, qe fefogev travgatauq k qe exihuriv xekeq ceverixeq k eqsoewiq wepufagiwevpe urcivajifuq. Xi gaevgai opaxaji faderevpeq wepufuq, k pegvagiq tiri xi ifsoaqagauv k urcivajigauv fe guvugawaevpuq quhre xi eqprogpori fe ov guvzovpu fe begbuq qodagaevpewevpe uhzepanuq k iggeqahxeq i nirauq uhqernifureq, ifewiq fe hiqirqe ev ov graperau fe nerfif k ovi gurreggauv terwivevpe. Xi itxagigauv fe equq wepufuq k guvugawaevpuq hoqgi xi ceverigauv fe wiq guvugawaevpu uhzepanu ev durwi fe trefaggauveq guvgrepiq, goivpapipaniq k guwtruhihxeq, rederafiq i begbuq uhqernihxeq tiqifuq, treqevpeq, k doporuq. Guv dregoevgai eqiq trefaggauveq toefev durwoxirqe wefaivpe rijuviwaevpuq k eqprogporirqe guwu recxiq u xekeq ceverixeq, soe fiv goevpi fex guwturpiwaevpu fe ov qaqpewi, k trefagev guwu igpoiri fagbu qaqpewi ev feperwavifiq gargovqpivgaiq. Ixcovuq feqgohrawaevpuq gaevpadaguq toefev reqoxpir guvprirauq ix qevpafu guwov. Ezewtxuq fe eqpu quv xi peurai ipuwagi u xi wegivagi goivpagi soe feqidaiv vugauveq guwoveq quhre xi wiperai. Wogbiq guvgetgauveq avpoapaniq fe xi viporixeji biv qafu privqdurwifiq i tirpar fe bixxijcuq gaevpadaguq guwu ex wunawaevpu fe priqxigauv fe xi paerri ixrefefur fex qux"

4.2. Resultado al descifrar utilizando directamente la tabla de frecuencias

"Ma inerina es em iorztrlo ce ioroinunerlos snsleualniauerle esldtiltdacos oblerncos uecnarle ma obsedqainor ce paldores deytmades, ce daforaunerlos g ce expednuerlainor, er aubnlos espeinvnios ce mos itames se yeredar pdeytrlas, se iorsldtger hnpolesns, se cectier pdnrinpnos g se emabodar meges yeredames g esjteuas uelocniauerle odyarnfacos. Ma inerina tlnmnfa cnvederles uelocos, g leirnias pada ma acjtnsninor g odyarnfainor ce ioroinunerlos sobde ma esldtiltda ce tr iorztrlo ce heihos stvninerleuerle obzelnqos g aiiesnbmes a qadnos obsedqacodes, aceuas ce basadse er tr idnledno ce qedcac g tra ioddeiinor peduarerle. Ma apmniainor ce esos uelocos g ioroinunerlos btsia ma yeredainor ce uas ioroinunerlo obzelnqo er vodua ce pdecniinores ioridelas, itarlnlalnqas g ioupdobabmes, devednCas a heihos obsedqabmes pasacos, pdeserles, g vtltdos. Ior vdeiterina esas pdecniinores ptecer vodutmadse uecnarle daforaunerlos g esldtiltdadse iouo deymas o meges yeredames, jte car iterla cem ioupodlaunerlo ce tr snsleua, g pdecnier iouo ailtada cniho snsleua er celedunracas inditrslarinas. Amytros cesitbdnunerlos inerlnvnios ptecer destmlad iorldadnos am serlnco ioutr. Ezeupmos ce eslo sor ma leodna alounia o ma ueiarnia itarlnia jte cesavnar roinores ioutres sobde ma ualedna. Utihas ioriepinores nrltnlnqas ce ma raltdamefa har snco ldarsvoduacas a padlnd ce hammafyos inerlnvnios iouo em uoqnunerlo ce ldasmainor ce ma lnedda amdececod cem som"

4.3. Resultado al descifrar con el método propuesto

"Pa ciencia es ep conjunto de conocimientos sistematicamente estluctulados ogtenidos mediante pa ogselhacion de batlones lequpales, de lazonamientos v de ewbelimentacion, en amgitos esbeciyicos de pos cuapes se qenelan blequntas, se constluven fibotesis, se deducen blincibios v se epagolan peves qenelapes v esxuemas metodicamente olqanizados. Pa ciencia utipiza diyelentes metodos, v tecnicas bala pa adxuisicion v olqanizacion de conocimientos sogle pa estluctula de un conjunto de fecfos suyicientemente ogjetihos v accesigpes a halios ogselhadoles, ademas de

gasalse en un clitelio de heldad v una colleccion belmanente. Pa abpicacion de esos metodos v conocimientos gusca pa qenelacion de mas conocimiento ogjetiho en yolma de bledicciones concletas, cuantitatihas v comblogagpes, leyelidas a fecfos ogselhagpes basados, blesentes, v yutulos. Con ylecuencia esas bledicciones bueden yolmupalse mediante lazonamientos v estluctulalse como leqpas o peves qenelapes, xue dan cuenta dep comboltamiento de un sistema, v bledicen como actuala dicfo sistema en detelminadas cilcunstancias. Apqunos descuglimientos cientiyicos bueden lesuptal contlalios ap sentido comun. Ejembpos de esto son pa teolia atomica o pa mecanica cuantica xue desayian nociones comunes sogle pa matelia. Mucfas concebciones intuitihas de pa natulapeza fan sido tlansyolmadas a baltil de fappazqos cientiyicos como ep mohimiento de tlaspacion de pa tiella aplededol dep sop”

5. Conclusiones

El método de descifrado que se propusó en este trabajo inicia con la descomposición en valores singulares de la matriz de frecuencias, con la que se establece un método de clasificación que permite determinar si un caracter representa una consonante o una vocal, los resultados del criterio sobre un texto pueden ser vistos en la tabla (**2-2**). La posibilidad de distinguir vocales de consonantes permite el uso de bigramas, de este modo se obtiene una mejora en la precisión del método sin necesidad de realizar más definiciones ya que esta información se extrae de la matriz de frecuencias.

El mensaje propuesto en el ejemplo 5, será utilizado para comparar los distintos métodos de descifrado, dicho mensaje consta de 1230 caracteres, 199 espacios y 18 símbolos de puntuación, presenta un nivel de entropía de 3,986, este valor es bastante cercano al valor máximo de entropía que está alrededor de 4,75. De esta observación podemos concluir que el nivel de incertidumbre sobre el mensaje es alto.

El resultado obtenido al comparar directamente con la tabla de frecuencias (**2-1**) no muestra un buen rendimiento, debido a la longitud del mensaje, pues la cantidad de caracteres utilizados en el mensaje es pequeña y no garantiza la convergencia de las frecuencias de aparición en el mensaje a las mostradas en la tabla de frecuencias.

En el caso de la aritmética modular, solo se utiliza una clave para descifrar el mensaje, sin embargo en los cifrados por sustitución cada pareja representa en sí una clave, por lo cual es de esperarse que los resultados no sean satisfactorios, a menos que el cifrado sea afín.

Los resultados obtenidos con el método propuesto, muestran una clara mejoría en la precisión, respecto a los métodos basados directamente en el análisis de frecuencias, ya que el combinar aproximaciones algebraicas y el uso de bigramas permite obtener resultados satisfactorios para mensajes cuya longitud es corta (por ejemplo mensajes que no superan una página de longitud).

A. Anexo: Cifrados por sustitución monoalfabética

Los cifrados por sustitución se caracterizan por que los caracteres mantienen su posición, pero cambian de rol, el proceso de cifrado se lleva a cabo al realizar una correspondencia entre caracteres, la clave del cifrado es la forma en que dicha correspondencia es efectuada.

Los métodos de sustitución pueden ser polialfabéticos o monoalfabéticos:

- En la sustitución monoalfabética solo se cuenta con un único alfabeto, lo cual convierte el proceso de cifrado en una transformación del texto plano en un mensaje, en el cual se sustituyen los caracteres del mensaje por caracteres en el alfabeto de crifrado.
- En los métodos polialfabéticos el cifrado es realizado con más de un alfabeto (puede ser el mismo alfabeto utilizado más de una vez), lo cual da la posibilidad de cambiar el modo de sustituir durante el cifrado.

En particular nos concentraremos en los cifrados monoalfabéticos ya que son el objeto de estudio de este trabajo. La sustitución monoalfabética suele ser denominada sustitución simple ya que, en el mensaje original, se sustituyen los caracteres (o grupos de caracteres), de una manera establecida, o simplemente se sustituyen por otros caracteres dentro del alfabeto; al finalizar la sustitución obtenemos el mensaje cifrado. Este proceso de cifrado puede ser de tres tipos:

- Monográmico
- Poligrámico
- Tomográmico.

A.1. cifrado monográmico

En esta técnica de cifrado se toma como alfabeto de cifrado el alfabeto del mensaje original, los carácteres son reemplazados por otros de manera inyectiva, razón por la cual:

- Los representantes de cada caracter en el mensaje cifrado mantienen su posición.
- Los caracteres en el mensaje cambian de rol
- La longitud del mensaje se mantiene
- El número de veces que aparece un caracter se mantiene

Para realizar el cifrado de un mensaje, debemos inicialmente establecer el protocolo de cifrado, es decir, la relación que mantendrán los caracteres del mensaje original con los caracteres en el mensaje cifrado.

Tomemos como lenguaje de cifrado el mismo lenguaje del texto plano, de este modo el mensaje parecera algo sin sentido, y consideremos diversas reglas de cifrado.

A.1.1. Cifrado afín

Los procesos de cifrado en los cuales el alfabeto de cifrado consiste en desplazar cierta cantidad de caracteres, son enmarcados bajo el nombre de *cifrado afín*. Los cifrados afines son basados en aritmética modular, sin embargo, en un alfabeto de n caracteres tendremos solo n posibles cifrados afines, esto hace que un ataque por fuerza bruta sea factible.

En general, asignando valores entre 0 y $n-1$ para un alfabeto de n caracteres, con una clave de cifrado dada, el cifrado afín será de la siguiente forma:

$$C \equiv M + clave \ (\text{mód } n)$$

Donde M es el caracter en el mensaje, C es el correspondiente caracter cifrado

Para poder obtener el mensaje original basta correr el alfabeto hacia la izquierda el número que indique la clave. A continuación veremos algunos ejemplos del cifrado afín

Ejemplo 6. *Cifrar el mensaje: "Una noche, una noche toda llena de perfumes, de murmullos y de músicas de alas".*

- *Desplazando UNA letra el alfabeto.*
  Con lo cual la relación entre los caractres será:

  $$a\ b\ c\ d\ e\ f\ g\ h\ i\ j\ k\ l\ m\ n\ o\ p\ q\ r\ s\ t\ u\ v\ w\ x\ y\ z$$
  $$b\ c\ d\ e\ f\ g\ h\ i\ j\ k\ l\ m\ n\ o\ p\ q\ r\ s\ t\ u\ v\ w\ x\ y\ z\ a$$

  Luego, el mensaje y el mensaje codificado serán:

  ```
  Una noche,
  una noche toda llena de perfumes, de murmullos y de musicas de alas
  Vob opdif,
  vob opdif upeb mmfob ef qfsgvnft, ef nvsnvmmpt z ef nvtjdbt ef bmbt
  ```

- *Desplazando DOS letras el alfabeto*
  Con lo cual la relación entre los caractres será:

  $$a\ b\ c\ d\ e\ f\ g\ h\ i\ j\ k\ l\ m\ n\ o\ p\ q\ r\ s\ t\ u\ v\ w\ x\ y\ z$$
  $$c\ d\ e\ f\ g\ h\ i\ j\ k\ l\ m\ n\ o\ p\ q\ r\ s\ t\ u\ v\ w\ x\ y\ z\ a\ b$$

  Luego, el mensaje y el mensaje codificado serán:

  ```
  Una noche,
  una noche toda llena de perfumes, de murmullos y de musicas de alas
  Wpc pqejg,
  wpc pqejg vqfc nngpc fg rgthwogu, fg owtownnqu a fg owukecu fg cncu
  ```

- *Desplazando QUINCE letras el alfabeto*
  Con lo cual la relación entre los caractres será:

  $$a\ b\ c\ d\ e\ f\ g\ h\ i\ j\ k\ l\ m\ n\ o\ p\ q\ r\ s\ t\ u\ v\ w\ x\ y\ z$$
  $$p\ q\ r\ s\ t\ u\ v\ w\ x\ y\ z\ a\ b\ c\ d\ e\ f\ g\ h\ i\ j\ k\ l\ m\ n\ o$$

Luego, el mensaje y el mensaje codificado serán:

```
Una noche,
una noche toda llena de perfumes, de murmullos y de musicas de alas
Jcp cdrwt,
jcp cdrwt idsp aatcp st etgujbth, st bjgbjaadh n st bjhxrph st paph
```

El cifrado afín más famoso es aquel en el que se desplazan 3 caracteres, conocido como el *cifrado César* (en honor a Julio César)[1], esta cifra fue empleada para enviar mensajes a sus tropas.

- Utilizando el cifrado César obtenemos

```
Una noche,
una noche toda llena de perfumes, de murmullos y de musicas de alas
Wpc pqejg,
wpc pqejg vqfc nngpc fg rgthwogu, fg owtownnqu a fg owukecu fg cncu
```

A.1.2. Cifras hebreas

Los hebreos desarrollaron procesos de cifrado, uno de ellos es conocido como el *atbash* utilizado en las sagradas escrituras[[3],p.123]. Este proceso consiste en asociar cada letra del alfabeto con su antipoda, es decir, el primer caracter se intercambia con el último, el segundo se intercambia con el penúltimo, y así sucesivamente.

- Tomando el alfabeto de cifrado como el alfabeto original en orden inverso.
  Con lo cual la relación entre los caracteres será:

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
z y x w v u t s r q p o n m l k j i h g f e d c b a

Luego, el mensaje y el mensaje codificado serán:

[1] El historiador *Suetonio* documento este hecho en la obra *Vida de los Césares*

```
Una noche,
una noche toda llena de perfumes, de murmullos y de musicas de alas
Fmz mlxsv,
fmz mlxsv glwz oovmz wv kviufnvh, wv nfinfoolh b wv nfhrxzh wv zozh
```

Entre las cifras hebreas destaca el *Albam*, el cual consiste en correr el alfabeto 13 posiciones hacia adelante, esto lo hace equivalente a un cifrado afín con clave 13, esta cifra es utilizada por algunos foristas en la web, donde es conocida como ROT13.

- Cifrando el fragmento del nocturno con el método albam obtenemos

```
Una noche,
una noche toda llena de perfumes, de murmullos y de musicas de alas
Han abpur,
han abpur gbqn yyran qr creshzrf, qr zhezhyybf l qr zhfvpnf qr nynf
```

A.1.3. La cifra del Kamasutra

El famoso libro Kamasutra, es en realidad un manual de conocimientos recopilados para que una mujer se convierta en una gran esposa, este libro fue escrito por Vatsyayana alrededor del siglo IV a.c.

El texto recomienda 64 habilidades y cualidades que toda buena esposa debe tener, entre ellas el arte de la escritura secreta conocido como *mlecchita-vikalpa* (la número 45), la cual permite mantener en secreto detalles de su relación.

El método de cifrado consiste en dividir el alfabeto por la mitad, y establecer relaciones entre las dos mitades, mediante la creación de parejas las cuales pueden ser incluso de manera aleatoria.

Ejemplo 7. *Diseñar un sistema de cifrado utilizando la cifra del Kamasutra*

Una forma de emparejar es escribir el alfabeto en dos filas, pero ordenado de arriba abajo, así:

$$\begin{matrix} a & c & e & g & i & k & m & o & q & s & u & w & y \\ b & d & f & h & j & l & n & p & r & t & v & x & z \end{matrix}$$

De este modo, los caracteres de la primera fila serán intercambiados por los de la segunda fila y viceversa. Luego para el fragmento del nocturno, tendremos:

```
Una noche,
una noche toda llena de perfumes, de murmullos y de musicas de alas
Vmb mpdgf,
vmb mpdgf spcb kkfmb cf ofqevnft, cf nvqnvkkpt z cf nvtjdbt cf bkbt
```

A.1.4. cifra con palabra clave

Continuando la idea de establecer parejas de caracteres para realizar sustitución simple, podemos pensar en un método rápido que permita establecer un cifrado fácil de realizar, pero difícil de descifrar, con la ventaja de poder ofrecer una forma de descifrado sencilla una vez conocida la clave.

El método consiste en acordar una palabra clave, en la cual no se repitan caracteres, en la parte superior se ubica el alfabeto ene l orden usual, en la parte inferior se inicia con la palabra clave y luego se continua el alfabeto con las letras que faltan.

Ejemplo 8. *Establecer un método de cifrado utilizando la palabra clave "cifrado".*

$$\begin{array}{c} a\ b\ c\ d\ e\ f\ g\ h\ i\ j\ k\ l\ m\ n\ o\ p\ q\ r\ s\ t\ u\ v\ w\ x\ y\ z \\ c\ i\ f\ r\ a\ d\ o\ p\ q\ s\ t\ u\ v\ w\ x\ y\ z\ b\ e\ g\ h\ j\ k\ l\ m\ n \end{array}$$

El resultado de dicha cifra se muestra a continuación

```
Una noche,
una noche toda llena de perfumes, de murmullos y de musicas de alas
Hwc wxfpa,
hwc wxfpa gxrc uuawc ra yabdhvae, ra vhbvhuuxe m ra vheqfce ra cuce
```

Este método permite al usuario transmitir el secreto con solo una palabra clave, así el receptor podrá fácilmente descifrar el mensaje. Sin embargo un espía del mensaje tendrá la dura tarea de descifrar la forma en que se realizo el cifrado, sin embargo cada emparejamiento es una clave, por tanto si el alfabeto tiene n caracteres, tendrá $n!$ posibles mensajes.

A.2. cifrado poligrámico

En este método de cifrado por sustitución no se sustituyen los caracteres uno por uno, sino que cada caracter puede ser representado por grupos de caracteres(pares, tríos, etc.). La principal diferencia con respecto al cifrado monográmico es que no se mantiene la frecuencia de aparición de caracteres.

Ejemplo 9. *Cifrar el mensaje: "Una noche, una noche toda llena de perfumes, de murmullos y de músicas de alas".*

Supongamos que tenemos el siguiente protocolo de cifrado

a	1 , 27 , 53	n	14 , 40 , 66
b	2 , 28 , 54	o	15 , 41 , 67
c	3 , 29 , 55	p	16 , 42 ,68
d	4 , 30 , 56	q	17 , 43 , 69
e	5 , 31 , 57	r	18 , 44 ,70
f	2 , 28 , 58	s	19 , 45 , 71
g	2 , 28 , 59	t	20 , 46 , 72
h	2 , 28 , 60	u	21 , 47 , 73
i	2 , 28 , 61	v	22 , 48 , 74
j	2 , 28 , 62	w	23 , 49 , 75
k	2 , 28 , 63	x	27 , 50 , 76
l	2 , 28 , 64	y	28 , 51 , 77
m	2 , 28 , 65	z	29 , 52 , 78

De este modo, nuestro mensaje

"Una noche, una noche toda llena de perfumes, de murmullos y de músicas de alas".

Se convierte en:

21 14 1 40 15 3 8 5 47 66 27 14 41 29 34 31 20 67 4 53 12 38 57 66 1 30 5 68 5 18 6 21 39 31 19 56 31 39 21 70 39 73 38 64 15 45 25 56 5 13 47 45 35 55 1 19 56 31 27 12 53 71

Si queremos podemos agregar 0 a la izquierda de los números menores que 10, así todos los términos del lenguaje de llegada se compondran de dos caracteres, obteniendo de este modo:

21 14 01 40 15 03 08 05 47 66 27 14 41 29 34 31 20 67 04 53 12 38 57 66 01
30 05 68 05 18 06 21 39 31 19 56 31 39 21 70 39 73 38 64 15 45 25 56 05
13 47 45 35 55 01 19 56 31 27 12 53 71

En el mensaje cifrado podemos ver ciertas características

- Cada caracter tiene mas de un representante en el alfabeto de cifrado
- La frecuencia de aparición de cada caracter no se matiene

Basta observar que en el mensaje original la parte: "una noche, una noche", es sustituida por:
21 14 01 40 15 03 08 05 47 66 27 14 41 29 34 31

Las letras u,n,a tienes más de un representante en el mensaje. También se observa esto cuando se mira la codificación dada a la cadena "ll", la cual aparece dos veces en el mensaje con codificación diferente.

Esta técnica de cifrado parece ser infalible, sin embargo el lector se dará cuenta que los caracteres tienen asociados números, (a 1 , b 2 , ...), los números asignados a cada caracter dependen de la posición que el caracter ocupa en el alfabeto. En el mensaje cifrado se observan números que parecen no tener algún orden, la clave para descifrar el mensaje es la congruencia del número modulo 26 (el número de caracteres), así se obtiene un número de 1 a 26 (0) que corresponde al caracter que ocupa esa posición en el alfabeto.

A.3. Cifrado Tomográmico

Los sistemas de cifrado tomográmico permiten que cada caracter sea representado por un grupo de símbolos, donde cada símbolo de cifrado se obtiene a través de un método específico

Ejemplo 10. *Cifrar el mensaje: "Una noche, una noche toda llena de perfumes, de murmullos y de músicas de alas".*

Para cifar el mensaje utilizaremos una técnica conocida como el cifrado de Polibio. Para ello consideremos la siguiente tabla

	1	2	3	4	5
1	a	b	c	d	e
2	f	g	h	i	j
3	k/q	l	m	n	o
4	p	r	s	t	u
5	v	w	x	y	z

El alfabeto que trabajamos consta de 26 caracteres, los cuales se han ubicado en una tabla de 5 filas y 5 columnas dando un total de 25 casillas, por esta razón k y q se encuentran en la misma casilla ya que, además de ser poco frecuentes, tienen un parecido fonético.

El proceso de cifrado consiste en ubicar la fila y columna que ocupa cada caracter en la tabla, así por ejemplo al caracter n corresponde el número 34. De este modo, el mensaje

"Una noche, una noche toda llena de perfumes, de murmullos y de músicas de alas".

se transforma en:
45 34 11 34 35 13 23 15 45 34 11 34 35 13 23 15 44 35 14 11 32 32 15 34 11 14 15 41 15 42 21 45 33 15 43 14 15 33 45 42 33 45 32 32 35 43 54 14 15 33 45 43 24 13 11 43 14 15 11 32 11 43

Nota En los métodos anteriores no se incluyo el caracter ñ en el alfabeto, esto con el fin de facilitar la explicación de cada método.

B. Anexo: Algebra Lineal

Definicion 3. *Diremos que una matriz A es* ***Reducible*** *si existe alguna matriz de permutación P talque PAP^{-1} sea una matriz triangular por bloques.*

Definicion 4. *Diremos que una matriz A es* ***Irreducible*** *si no es reducible*

Definicion 5. *Sea A una matriz cuadrada A, diremos que A es:*

- ***Definida Positiva*** *si:*

  Para todo x vector columna no nulo, se satisface que:

  $$x^t Ax > 0$$

- ***Semidefinida Positiva*** *si:*

  Para todo x vector columna no nulo, se satisface que:

  $$x^t Ax \geq 0$$

Definicion 6. *Sea A una matriz cuadrada, definimos el* ***polinomio característico de*** *A como*

$$P_A(\lambda) = det(A - \lambda I) = 0$$

Donde la matriz I representa a la matriz identidad.

Definicion 7. *Sea $P_A(\lambda)$ el polinomio característico de la matriz A, talque*

$$P_A(\lambda) = det(A - \lambda I) = (\lambda - \lambda_1)^{a_1} + \cdots + (\lambda - \lambda_k)^{a_k}$$ [1]

Denominamos a las raíces del polinomio $P_A(\lambda)$ (es decir cada λ_i) ***Valores propios asociados a*** *A*

Definicion 8. *Sea $P_A(\lambda)$ el polinomio característico de la matriz A, talque*

$$P_A(\lambda) = (\lambda - \lambda_1)^{a_1} + \cdots + (\lambda - \lambda_k)^{a_k}$$

Los valores a_i se denominan ***Multiplicidad del valor propio*** *λ_i.*

Si $a_j = 1$, diremos que λ_j es ***raíz simple*** *del polinomio característico de A*

Definicion 9. *Se define el* ***radio espectral*** *de una matriz $A \in M_{n\times n}(\mathbb{R})$ como:*

$$\rho(A) = \sup_{i \leq n} (|\lambda_i|)$$

Definicion 10. *Se definen los* ***valores singulares*** *de la matriz A como las raíces cuadradas de los valores propios de la matriz A^tA.*

Es decir, si λ_i es valor propio de A^tA, $\sigma_i = \sqrt{\lambda_i}$ será valor singular de A

La definición de valor singular requiere lar raíces cuadradas de los valores propios de A^tA, por tanto es razonable preguntarnos si estos valores singulares pueden ser complejos, sin embargo el siguiente teorema nos muestra que los valores propios de A^tA son positivos y por ende los valores singulares de A serán reales.

Teorema 3. *Sea A una matriz de tamaño $m \times n$, entonces los valores propios de A^tA son reales, mayores o iguales que 0*

[1] Esta descomposición es posible ya que según el teorema fundamental del Algebra, un polinomio de grado n tiene n raíces complejas

Demostración. La matriz A^tA es siempre simétrica ya que $(A^tA)^t = (A^t)(A^t)^t = A^tA$.
Por tanto sus valores propios son positivos.

Como A es de tamaño $m \times n$, entonces A^t tendrá tamaño $n \times m$, con lo cual A^tA será una matriz de tamaño $n \times n$.

Veamos ahora que A^tA tiene todos sus valores propios positivos, para ello probaremos que A^tA es semidefinida positiva. Sea x un vector columna, de tamaño n, no nulo, luego Ax será un vector columna.

Aplicando el producto entre vectores dado por $< X, Y >= X^tY$ (Producto Punto), obtenemos

$$\begin{aligned} < Ax, Ax > &= (Ax)^tAx \\ \|Ax\|^2 &= x^tA^tAx \end{aligned}$$

Luego $\|Ax\|^2 = x^t(A^tA)x$.
Como $\|Ax\| \geq 0$, entonces $x^t(A^tA)x \geq 0$.

Por tanto A^tA es semidefinida positiva, luego sus valores propios son mayores o iguales a cero. □

Teorema 4. *Sea A una matriz de tamaño $m \times n$, entonces puede ser representada como*

$$A = X\Sigma Y^t$$

Donde:

- *X es una matriz unitaria de tamaño $m \times n$*
- *Y es una matriz unitaria de tamaño $m \times n$*
- *Σ es una matriz diagonal, de tamaño $n \times n$, en cuya diagonal principal se encuentran los valores singulares de A ordenados de mayor a menor.*

Demostración. Recordemos que A^tA es simétrica, por tanto tiene valores propios reales. Sin pérdida de generalidad, consideremos $r \leq n$ el número de valores propios no nulos de A^tA (Es decir, A^tA es una matriz de rango r).

Sea $Y = [y_1 \ y_2 \ \dots \ y_n]$ una matriz conformada por los vectores propios normalizados de la matriz A, luego

$$Ay_i = \lambda_i y_i \qquad \text{para cada } i$$

Ahora, consideremos los vectores Ay_i, y veamos que son un conjunto de r vectores ortogonales, para ello, tomemos Ay_i, Ay_j con $i \neq j$ y verifiquemos que el producto interno es 0

$$\begin{aligned}
\langle Ay_j, Ay_i \rangle &= (Ay_j)^t Ay_i \\
&= y_j^t A^t A y_i \\
&= y_j^t (A^t A y_i) \\
&= y_j^t (\lambda_i y_i) \\
&= \lambda_i y_j^t y_i \\
&= 0
\end{aligned}$$

Por tanto Ay_j es ortogonal a Ay_i, para cada i, j tales que $i \neq j$, de donde se concluye que los vectores $Ay_1, Ay_2, \dots, Ay_n$ son r ortogonales (Resultan r ortogonales porque asumimos que existen r valores propios no nulos). No obstante, pese a ser ortogonales, no están normalizados, por ello consideremos

$$x_i = \frac{Ay_i}{\|Ay_i\|}$$

Veamos que valor toma $\|Ay_i\|$, para ello recordemos que y_i es vector propio de A^tA, luego

$$\begin{aligned}
A^t A y_i &= \lambda_i y_i \\
y_i^t A^t A y_i &= y_i^t \lambda_i y_i \\
(Ay_i)^t Ay_i &= \lambda_i y_i^t y_i \\
\|Ay_i\|^2 &= \lambda_i \|y_i^t y_i\|^2 \\
\|Ay_i\|^2 &= \lambda_i \\
\|Ay_i\| &= \sigma_i
\end{aligned}$$

Por tanto, tenemos que

$$x_i = \frac{Ay_i}{\sigma_i}$$

De donde se obtiene que los vectores $x_1, x_2, \dots x_r$ son ortogonales.

En general, tenemos lo siguiente:

$$\begin{aligned} Ay_i &= \frac{\sigma_i}{\sigma_i} Ay_i \\ &= \sigma_i \frac{Ay_i}{\sigma_i} \\ &= \sigma_i x_i \end{aligned}$$

Por tanto, $Ay_i = \sigma_i x_i$.
Para verlo matricialmente, tomemos Σ la matriz de valores singulares definida como

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ & \ddots & \ddots & \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix}$$

Donde $\sigma_i \leq \sigma_j$ para $i \geq j$. Cabe destacar que si un valor propio es nulo, entonces un valor singular también lo es.

Por tanto, matricialmente obtenemos

$$AY = X\Sigma$$

Cabe destacar que $X\Sigma = [\sigma_1 X_1 \ \dots \ \sigma_n X_n]$.
Como Y es una matriz unitaria, tenemos que $Y^t = Y^{-1}$. Luego

$$\begin{aligned} AY &= X\Sigma \\ AYY^t &= X\Sigma Y^t \\ A &= X\Sigma Y^t \end{aligned}$$

De donde concluimos que, $A = X\Sigma Y^t$

□

Teorema 5. *Sea $A = X\Sigma Y^t$, una descomposición en valores singulares de una matriz A de tamaño $m \times n$, sean $\sigma_1, \cdots, \sigma_r$ valores singulares no nulos de A. Entonces:*

1. *El rango de A es r*
2. *$\{x_1\ x_2\ \cdots\ x_r\}$ es una base ortonormal de $Rango(A)$*
3. *Si $r < n$, entonces $\{y_{r+1}\ y_{r+2}\ \cdots\ y_n\}$ es base ortonormal de $Nul(A)$*
4. *$\{y_1\ y_2\ \cdots\ y_r\}$ es una base ortonormal de $Rango(A^t)$*
5. *Si $r < n$, entonces $\{x_{r+1}\ x_{r+2}\ \cdots\ x_m\}$ es base ortonormal de $Nul(A^t)$*

Demostración. Recordemos que el rango de una matriz, coincide con el número de valores propios no nulos que contenga.

1. Por hipótesis conocemos la existencia de r valores singulares no nulos, los cuales son obtenidos a partir de los valores propios no nulos de A^tA, en pocas palabras A^tA tiene rango r.
   La matriz A (también A^t) tiene el mismo rango que A^tA, por tanto A tiene rango r, de donde se concluye que el rango de A es r.

2. En la prueba del teorema (4) de la descomposición SVD, vimos que los vectores $x_i = \frac{Av_i}{\sigma_i}$ son r vectores ortonormales, por tanto forman una base ortonormal del Rango(A).

3. Como $r < n$, tenemos que $\sigma_k = 0$ para $k > r$, por tanto para cada $k > r$ se cumple $Ay_k = \sigma_k x_k = 0$.

   Luego, como $\{y_{r+1} \dots y_n\}$ anulan a A, y son ortonormales, entonces forman una base de $Nul(A)$

4. Los vectores $\{y_1\ \dots y_r\}$ son una base del espacio complementario de $Nul(A)$, por tanto forman una base para $Nul(A)^\perp$, es decir, son base de $Ran(A^t)$

5. Sabemos que si $r < n$ entonces, el espacio generado por $\{x_{r+1} \dots x_n\}$ es complementario al espacio generado por $\{x_1 \dots x_r\}$ que es el rango de A, luego $\{x_{r+1} \dots x_n\}$ es base de $Rango(A)^\perp$, es decir, es base de $Nul(A^t)$.

   Nota El teorema de la dimensión es utilizado intrínsecamente, para relacionar las dimensiones de los espacion Nul y Rango.

□

Teorema 6. *(Perron-Frobenius) Sea A una matriz irreducible, talque cada componente es positiva (es decir $a_{i,j} \geq 0$ para cada i, j), Entonces.*

1. *El radio espectral $\rho(A)$ es positivo, y $\rho(A)$ es un valor propio*
2. *Existe x talque $Ax = \rho(A)x$, talque $x_i >= 0$ para cada i*
3. *$\rho(A)$ es una raíz simple del polinomio característico de A.*

Nota La condición de irreducibilidad sobre la matriz A impide la existencia de filas nulas en la matriz.

En particular, utilizaremos un resultado similar al teorema de Perron-Frobenius, en el cual nos restringimos al trabajo con matrices de entradas no negativas en las cuales podemos tener filas nulas. (Es posible considerar un caracter en el alfabeto que no aparezca en el mensaje).

Teorema 7. *Sea A matriz talque $a_{ij} \geq 0$ para cada i, j, Entonces*

1. *$\rho(A)$ es un valor propio*
2. *Existe x talque $Ax = \rho(A)x$, talque $x_i >= 0$ para cada i*

Demostración. Sea $E \in M_{nXn}(\mathbb{R})$ talque $E_{ij} = 1$ para cada i, j.
Tomemos $\varepsilon > 0$ talque $A_\varepsilon = A + \varepsilon E$. Luego $\rho(A) \leq \rho(A_\varepsilon)$.

Sea x_ε el vector propio asociado a A_ε.
Tomemos una sucesión $\varepsilon_n \to 0$ estrictamente decreciente, talque para cada término de la sucesión se construye A_{ε_n} y x_{ε_n}. Luego:

$$\rho(A_{\varepsilon_{i+1}}) \leq \rho(A_{\varepsilon_i}) \qquad \text{para cada } i$$

Entonces

$$\lim_{n\to\infty} \rho(A_n) = \lim_{\varepsilon\to 0} \rho(A + \varepsilon E) = \rho(A)$$
$$\lim_{n\to\infty} x_n = \lim_{\varepsilon\to 0} \rho(x + \varepsilon E) = x$$

Entonces, $Ax = \rho(A)x$ donde x es un vector propio talque $x_i > 0$ para cada i □

Bibliografía

[1] Gómez, Joan: *Matemáticos, espías y piratas informáticos* Barcelona: RBA Libros S.A, 2010

[2] Hanselman, Duane C: *Mastering Matlab* Pearson Prentice Hall, 2005

[3] Fernández, S: *La criptografía Clásica*. En: *Revista SIGMA* 24 (2004), p. 119–141

[4] Honn, T & Ston, S: *linear Algebra, SVD and cryptograms* (2002)

[5] Rodao, J.M: Piratas cibernéticos : cyberwars, seguridad informática e Internet Alfaomega, 2002

[6] Alvaréz, C & Carreiras, M & De Vega, M: *Estudio estadístico de la ortografía castellana*. En: *Cognitiva* 20 (1992), p. 107–125

[7] Hart, G: *To decode short cryptograms*. En: *Communications of the acm* (1994), p. 102–108

[8] Subiza, B: *Juegos Matriciales y su aplicación a la teoría de Perron-Frobenius*. En: *Estadística Española* 112-113 (1986), p. 31–43

[9] Stinson, D: *Cryptography: Theory and Practice*. Ontario: Chapman & Hall, 2006

[10] Orjuela, H: *La primera versión del nocturno de Silva*. En: *Thesaurus* 34 (1974), p. 118–128