Congreso Internacional en Ingeniería Electrónica. Mem. Electro 2015, Vol. 37, pp 353-357, Chihuahua, Chih. México
http://depi.itchihuahua.edu.mx/display/memorias_electro/MemoriaElectro2015
ISSN 1405-2172

DESCIFRADO DE TEXTOS EN ESPAÑOL UTILIZANDO ANÁLISIS DE FRECUENCIAS INCLUYENDO UNIGRAMAS, BIGRAMAS Y TRIGRAMAS

Bárbara Emma Sánchez Rinza, .Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, 14 Sur y Avenida San Claudio,
brinza@hotmail.com

RESUMEN.

En este trabajo se presenta un criptosistema basado para el idioma español. Todos los idiomas tienen unas palabras más comunes que otras para poder hacer las conexiones de las frases u oraciones. En Español tenemos varias palabra como las preposiciones, los artículos que son palabra de una, dos o tres letras que comúnmente utilizamos más. En este trabajo se utiliza estas reglas ortográficas de nuestro idioma, para poder descifrar información.

El descifrado por frecuencias silábico es un algoritmo basado en las reglas gramaticales en el español, lo cual se realiza haciendo un estudio estadístico de las palabras de una, dos y de tres letras más comunes en el español. Es decir escogemos un tema y un texto de aproximadamente 10,000 palabras .y sacamos sus frecuencias de esas palabras de una de dos y de tres letras más usadas en el español, lo que llamaremos unigrama, bigrama y trigrama y comparando con el texto cifrado que tiene que ser del mismo tema. Para el proceso del cifrado, utilizaremos el algoritmo de Vigenére para codificar anteriormente el texto cifrado y posteriormente por esta técnica se decodificara [1].

ABSTRACT.

In this work for the Spanish language based cryptosystem we are presented.

All languages have some more common than others words to make connections phrases or sentences. In Spanish we have several words like prepositions, articles that are word of one, two or three letters most commonly used. In this paper these orthographic rules of our language is used to decrypt the data.

The deciphered syllabic frequencies is based on the grammar rules in Spanish algorithm, which is done by a statistical study of the words of one, two and three most common letters in Spanish. That is choose a topic and a text of about 10,000 words .and we get frequencies of these words in a two- and three letters more used in Spanish, which call unigram, bigram and trigram and comparing with the ciphertext that has to be the same issue. For the encryption process, we use the above algorithm to encode Vigenere ciphertext and subsequently decode this technique [1].

1. INTRODUCCIÓN

La palabra criptografía es un término que describe todas las técnicas que permiten cifrar mensajes o hacerlos inteligibles sin recurrir a una acción específica.

La criptografía se basa en la aritmética: En el caso de un texto, consiste en transformar las letras que conforman el mensaje en una serie de números (en forma de bits ya que los equipos informáticos usan el sistema binario) y luego realizar cálculos con estos números para:

- modificarlos y hacerlos incomprensibles. El resultado de esta modificación (el mensaje cifrado) se llama texto cifrado, en contraste con el mensaje inicial, llamado texto simple.
- asegurarse de que el receptor pueda descifrarlos.

El hecho de codificar un mensaje para que sea secreto se llama cifrado. El método inverso, que consiste en recuperar el mensaje original, se llama descifrado.

El criptoanálisis consiste en la reconstrucción de un mensaje cifrado en texto simple utilizando métodos matemáticos. Por lo tanto, todos los criptosistemas deben ser resistentes a los métodos de criptoanálisis. Cuando un método de criptoanálisis permite descifrar un mensaje cifrado mediante el uso de un criptosistema, decimos que el algoritmo de cifrado ha sido decodificado.

2. REALIZACIÓN DEL CRIPTOSISTEMA

Para el desarrollo del algoritmo se utilizó java para implementar dicho sistema. Primero tomamos un texto de entrenamiento de más de 10 000 palabras de un tema en específico y sacamos las frecuencias de los unigramas, bigramas y trigramas. Posteriormente del mismo tema se cifra por el algoritmo de vigenére. De la misma manera sacaremos las frecuencias del texto cifrado. Después de tener las dos tablas de frecuencia del texto de entrenamiento y del texto cifrado, sustituiremos primero por monosílabas bisílabas y trisílabas. Posteriormente se utiliza un procesador de palabras que nos arroje el porcentaje de parentesco por letras y por palabras entre el texto original y el texto descifrado.

Para la implementación del descifrado de un texto mediante el análisis de frecuencias de palabras mono, bi, y trisilabicas. La idea es analizar el texto cifrado de entrada para conocer la frecuencia de cada una de las letras, la frecuencia de las palabras de una letra (unigramas), palabras de dos letras (bigramas) y palabras de tres letras (trigramas). Del texto de

entrenamiento. Una vez que se tienen estas frecuencias, ya conocemos los elementos más frecuentes en el texto cifrado. Finalmente se procede a sustituir estos elementos por los más frecuentes en el idioma español.

Diagrama de Caso de Uso

En la Figura 1 se muestra en diagrama de caso de uso en el cual el usuario interactúa con el sistema.

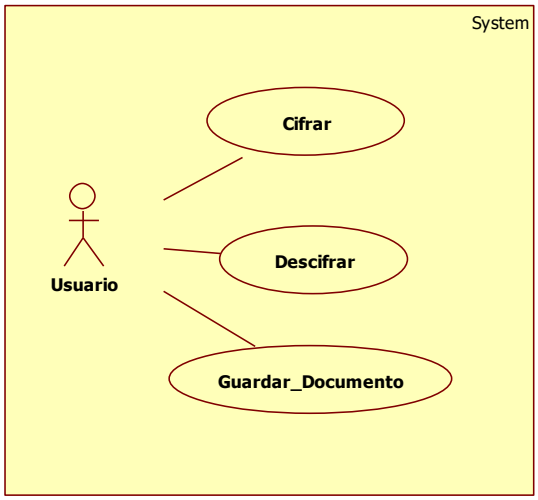

Figura 1. Caso de uso

A continuación se hace la descripción de cada caso de uso:

Descripción de cada caso de uso

Caso de uso	Descripción
Cifrar	Cifrado del documento original, puede ser cifrado por el usuario o dado de forma exterior.
Descifrar	Descifrado del documento obtenido. EL usuario debe cargar el documento de forma manual.
Guardar_Documento	El usuario puede guardar el documento descifrado. Debe seleccionarse la ubicación donde se guarda.

2.1. Diagrama de clases

En la figura 2 se muestra el diagrama de clases del sistema. La clase principal es “Cifrado”, la cual realiza las operaciones de cifrado (documento de entrenamiento) y descifrado (documento proporcionado por el usuario). Las clases “Lista” y “Nodo” se Utilizan para la elaboración de las tablas de frecuencias. Las clases “Practica1” e “Interface” se encargan de la visualización de la interfaz.

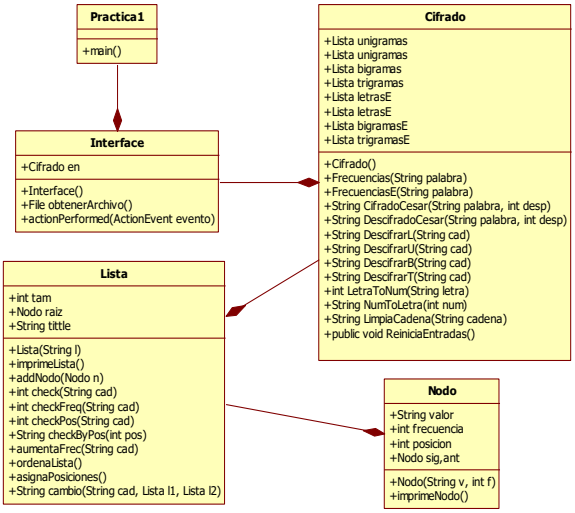

Figura 2. Diagrama de clases

2.2. Diagrama de actividades

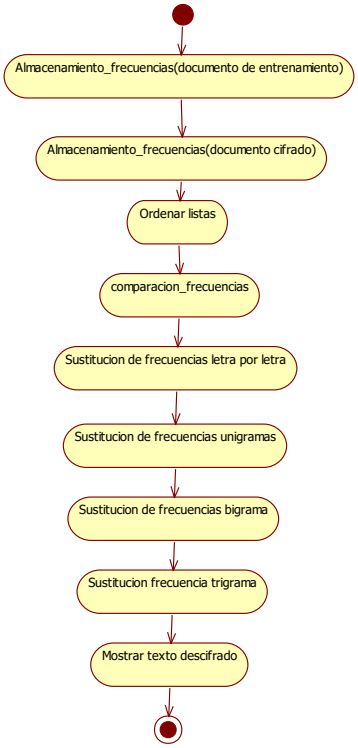

Figura 3. Diagrama de actividades

2.3. implementación.

Primero, para el almacenamiento de las frecuencias se crearon 4 listas (unigramas, bigramas y trigramas), llamaremos primero la frecuencia de las letras eso es en general, el idioma español tiene letras más utilizadas que otras como se puede ver en la figura 4, tenemos el texto claro o texto de entrenamiento y

ISSN 1405-2172

el texto cifrado, el programa para decodificar hará la sustitución de lista de texto claro en la lista del texto cifrado y guardara.

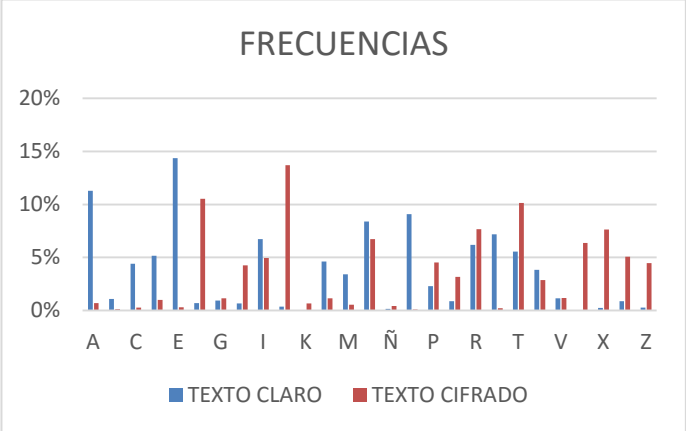

Figura 4 Frecuencias de una sola letra del texto claro o texto de entrenamiento y el texto cifrado

En la figura 5 tenemos la gráfica de los bigramas es decir las palabras de una dos letra que son más frecuentes en el español del texto de entrenamiento.

Figura 5.- Frecuencia del bigrama del texto de entrenamiento.

La palabras que tenemos en el bigrama de la figura 5 se sustituirán por la lista de los bigramas que se tienen en el texto cifrado, figura 6.

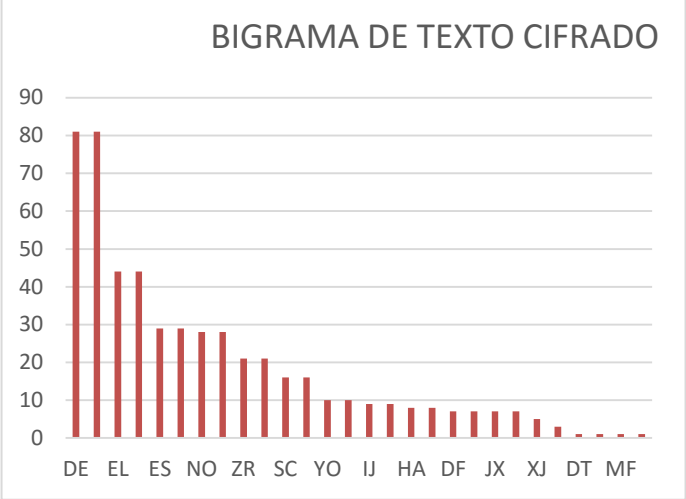

Figura6. Bigrana del texto cifrado.

A continuación se dará el trigrama del texto entrenamiento de 10 000 palabras que se ve en la figura 7 y se sustituirá en el texto que se ha ido decodificando por la lista de la figura 8 que es el trigrama del texto cifrado.

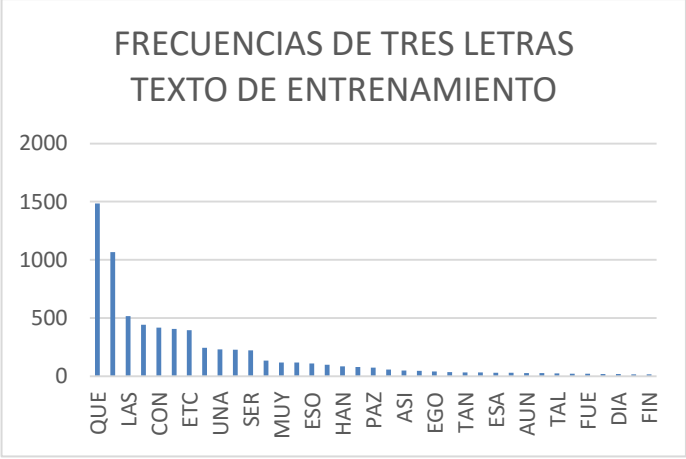

Figura 7trigranma del texto de entrenamiento

La figura 7 se va ir sustituyendo las palabras por las que tenemos en la figura 8 como ya se comentó.

Figura 8 trigrama del texto cifrado

3. RESULTADOS OBTENIDOS

Después de la implementación, se realizaron pruebas con el texto cifrado, primero realizamos el descifrado de frecuencias solo tomando en cuenta la frecuencia de letras individuales (figura 9) Con esta configuración se obtuvo un 60.2% de acierto comparando letra por letra el texto descifrado que obtuvimos mediante nuestro algoritmo y el texto original no cifrado.

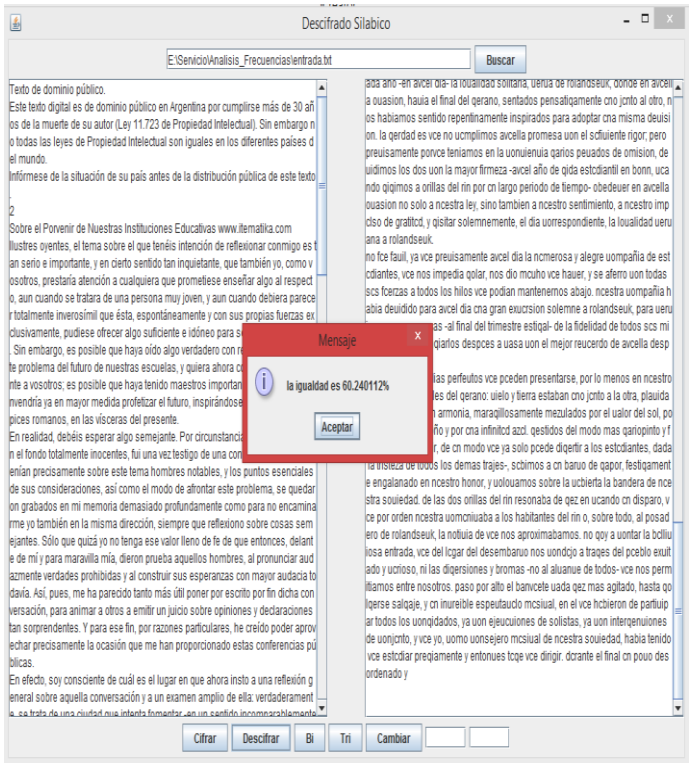

Figura 9 decodificado con una sola letra (unigrama).

Este resultado lo guardamos y aplicamos el bigrama para seguir decodificando el texto cifrado y se obtiene un porcentaje del 61.9.1% de acierto letra por letra Se nota una ligera mejoría en cuando a la comprensión del texto (comparación de palabras) y visualmente el texto obtenido es más entendible, figura 10.

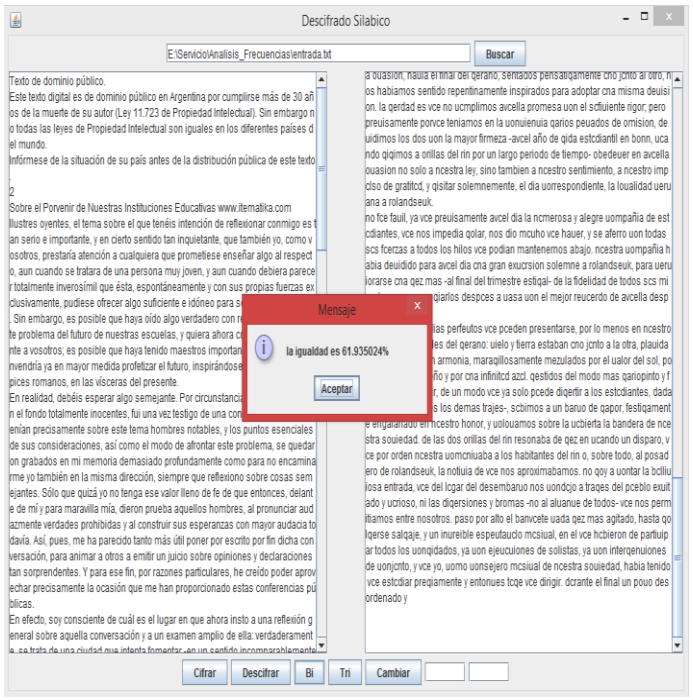

Figura 10.- decodificaciones del texto cifrado utilizando el bigrama

Los resultados obtenidos en la figura 10 se guardan y se le aplica el algoritmo otra vez para los trigramas aquí se obtiene un 65.8%. Que ya es bastante entendible el mensaje, y con nuestro conocimiento en español ya podemos leerlo figura 11.

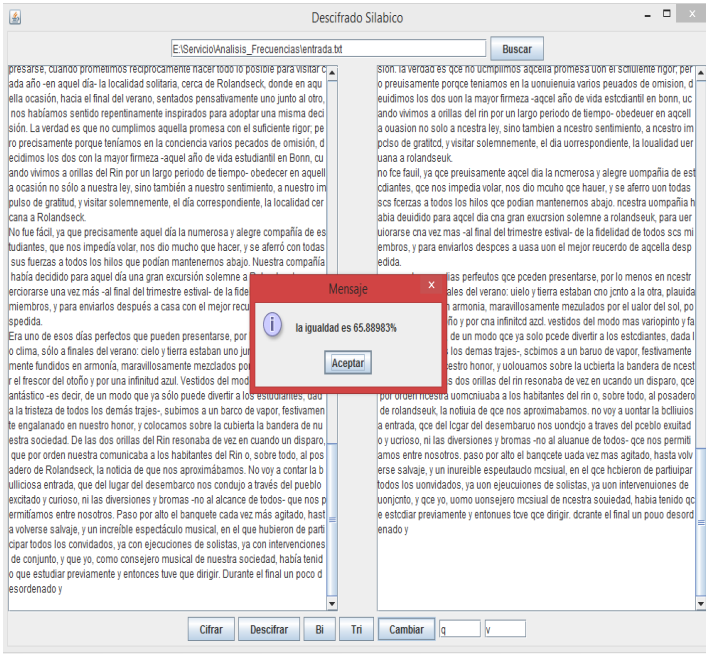

Figura 11.- decodificación del texto cifrado utilizando unigramas, bigramas y trigramas.

4. TRABAJO A FUTURO

Nuestro sistema aún no es muy robusto. Para versiones futuras se trabajará con caracteres especiales, como lo son los acentos o los signos, y las mayúsculas, ya que el texto descifrado es devuelto totalmente en minúsculas.

5. CONCLUSIONES

Con este algoritmo usando análisis de frecuencias para el descifrado de textos, se obtuvieron resultados aceptables, aunque se deberá trabajar más para la mejora de los resultados. Aplicando nuestro algoritmo a un texto cifrado se obtiene visualmente un gran entendimiento del texto obtenido. Pero solo nos sirve para decodificar textos cuyas funciones matemáticas son sobreyecticas [2].

Este tipo de algoritmos de frecuencias para decodificar solo obtiene resultados buenos con cifrados de corrimiento, o aquellos algoritmos que arrojan la misma cantidad de letras que el texto plano y el cifrado.

6. BIBLIOGRAFÍA

[1] Sánchez, B. Bigurra, Diana. *et all*. ***De-Encryption of a text in Spanish using probability and statistics***. 18th International Conference on Electronics, Communications and Computers: isbn 07695 3120 2 march2008.

[2] Sánchez, B. Cruz, S. ***Cesar decryption algorithm, but the method of frequency points in the Spanish language***. International Journal of Engineering and Innovative Technology, vol 3, issui 5 november 2013 issn 22773754: