Nombre: Julian Camilo Saavedra Rodriguez.

Código: 201411746

Resumen: Esta práctica tiene como objetivo obtener la entropía de un texto, la probabilidad de los caracteres (incluyendo puntos, comas etc..).

I. INTRODUCCION

El objetivo de esta práctica es entender la teoría de la información, el uso de visual Basic en Excel para la lectura de una base de datos, en este caso compuesta por una cadena de caracteres, y aprender el lenguaje de visual para aplicaciones VBA.

II. OBJETIVOS

- Determinar la entropía de una secuencia de caracteres.
- Obtener la cantidad de información de una cadena de caracteres.
- Encontrar las probabilidades de los caracteres.

III. MATERIALES

• Visual Basic

IV. MARCO TEORICO

Teoría de la información Fue desarrollada por Claude E. Shannon para encontrar los límites fundamentales en la compresión almacenamiento confiable y comunicación de datos. Se ha ampliado para encontrar aplicaciones en muchas otras áreas, incluyendo inferencia estadística, procesamiento del lenguaje natural, criptografía, otras redes diferentes a las redes de comunicación como en neurobiología, la evolución y función de códigos moleculares, selección de modelos en ecología, física térmica, computación cuántica v otras formas de análisis de datos.

Una medida clave de la información en la teoría es conocida como entropía, la que usualmente se expresa como el número promedio de bits necesarios para almacenamiento o comunicación. La entropía cuantifica la incertidumbre involucrada al encontrar una variable al azar.

Entropía: mide tanto la falta de información como la información. Estas dos concepciones son complementarias. La entropía también mide la libertad, y esto permite una interpretación coherente de las fórmulas de entropía y de los hechos experimentales. No obstante, asociar la entropía y el desorden implica definir el orden como la ausencia de libertad.

La ecuación que mide la entropía está dada por:

$$S = k * \ln |\Omega|$$

Donde S es la entropía, K es la constante de Boltzmann y Ω es el numero de microestados posibles para el sistema.

Esta ecuación nos indica que la cantidad de entropía de un sistema es proporcional al logaritmo natural del número de microestados posibles.

Fuentes con y sin memoria

Existen diferentes tipos de fuentes de información, veremos 2 las cuales son Con memoria y sin memoria:

Fuentes sin memoria: Los símbolos dentro de un mensaje son independientes entre sí. De esta manera, los símbolos que hayan aparecido hasta el momento no van a condicionar al símbolo presente ni a los posteriores.

Fuentes con memoria: La aparición de los símbolos no es del todo independiente. Es decir, si han aparecido M-1 símbolos, el símbolo M-ésimo esta condicionado por los anteriores.

Aplicaciones de tópicos fundamentales de la teoría de la información incluyen compresión sin pérdida de datos (ej. Archivos ZIP), compresión de datos con pérdida (ej. MP3s), y codificación de canal (ej. Para líneas DSL). El campo está en la intersección de las matemáticas, estadística, ciencias de la computación, física, neurobiología e ingeniería eléctrica.

La cantidad de información para un símbolo según Shannon puede ser hallada como:

$$I(Si) = Log_2\left(\frac{1}{Pi}\right)$$

Donde Pi es la probabilidad para el símbolo Si. La información aportada por un símbolo que es la concatenación de otros dos es la suma de las informaciones de ambos símbolos.

$$I(SiSj) = I(Si) + I(Sj)$$

Así mismo la entropía de un sistema de información esta expresado como:

$$H = \sum_{j=1}^{n} PjLog_2\left(\frac{1}{pj}\right)$$

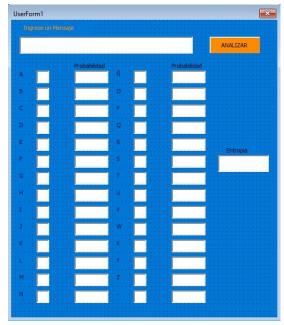
Donde H es la entropía, las Pj son las probabilidades de que aparezcan los diferentes códigos y m el número total de códigos.

V. PROCEDIMIENTO

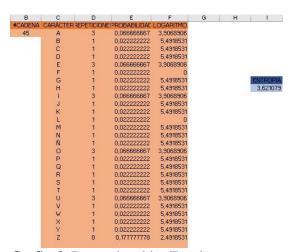
Parte 1

Realizar una interfaz gráfica en visual Basic en Excel, por medio del uso de un botón que de acceso a la misma y que permita ingresar una cadena de caracteres "mensaje" de mínimo 40 caracteres, el cálculo de la probabilidad de los mismos, la cantidad de información "caracteres" y la entropía del mensaje, se debe visualizar en la hoja de Excel en una columna cada carácter y al frente la probabilidad del mismo. Realizar La toma de datos de dicho Excel en Matlab.

La interfaz gráfica obtenida fue la siguiente:



Grafica1. Interfaz Visual Basic.



Grafica2. Datos obtenidos Excel.

Con la implementación del siguiente código:

Private Sub CommandButton1_Click()
Dim cadena As String
Dim contA As Integer
Dim cadenas As Double

cadena = TextBox1.Text

If Len(cadena) < 40 Then
MsgBox ("DEBE INGRESAR MAS DE
40 CARACTERES")

Else

		ProbB = contB / Len(cadena)
Tex	xtBox4.Text = Len(cadena)	Range("E3").Select
cadenas = Len(cadena)		ActiveCell.FormulaR1C1 = ProbB
Cel	ls([2], [2]) = cadenas	If $ProbB = 0$ Then
		LGB = 0
For $j = 1$ To Len(cadena)		Else
		LGB = Log(1 / ProbB) / Log(2)
Puntero = $Mid(cadena, j, 1)$		End If
Then		Range("F3").Select
	If Puntero = "a" Or Puntero = "A"	ActiveCell.FormulaR1C1 = LGB
	contA = contA + 1	
	End If	Panga("D29") Salaat
Then	If Puntero = "b" Or Puntero = "B"	Range("D28").Select ActiveCell.FormulaR1C1 = contZ
	ii runteio – b Oi runteio – b	TextBox54.Text = contZ
	contB = contB + 1	probZ = contZ / Len(cadena)
	End If	Range("E28").Select
		ActiveCell.FormulaR1C1 = probZ
		If $probZ = 0$ Then
Then		LGZ = 0
Then	If Puntero = "z" Or Puntero = "Z" Then	Else
	contZ = contZ + 1	LGZ = Log(1 / probZ) / Log(2)
	End If	End If
		Range("F28").Select
	If Puntero = "-" Then	ActiveCell.FormulaR1C1 = LGZ
	contGuion = contGuion + 1	
	End If	D (IID 20 II) G 1
N.		Range("D28").Select
Next		ActiveCell.FormulaR1C1 = contGuion
Range("D2").Select		TextBox60.Text = contGuion
ActiveCell.FormulaR1C1 = contA TextBox2.Text = contA		probGuion = contGuion / Len(cadena) Range("E28").Select
probA = contA / Len(cadena)		ActiveCell.FormulaR1C1 = probGuion
Range("E2").Select		If probGuion = 0 Then
ActiveCell.FormulaR1C1 = probA		LGGuion = 0
If $probA = 0$ Then		Else
LGA = 0		LGGuion = Log(1 / probGuion) / Log(2)
Else		End If
LGA = Log(1 / probA) / Log(2)		Range("F28").Select
End If		ActiveCell.FormulaR1C1 = LGGuion
Range("F2").Select		
ActiveCell.FormulaR1C1 = LGA		
D (IID 0II) G 1		TextBox3.Text = proba
Range("D3").Select		
ActiveCell.FormulaR1C1 = contB		
TextBox4.Text = contB		TextBox59.Text = probGuion

Entropia = probA * LGA + Prob * LGB + probC * LGC + probD * LGD + probE * LGE + probF * LGF + probG * LGG + probH * LGH + probI * LGI + probJ * LGJ + probK * LGK + probL * LGL + probM * LGM + probN * LGN + probN * LGN + probO * LGO + probP * LGP + probQ * LGQ + probR * LGR + probS * LGS + probT * LGT + probU * LGU + probV * LGV + probW * LGW + probX * LGX + probY * LGY + probZ * LGZ Range("I9").Select

ActiveCell.FormulaR1C1 = Entropia

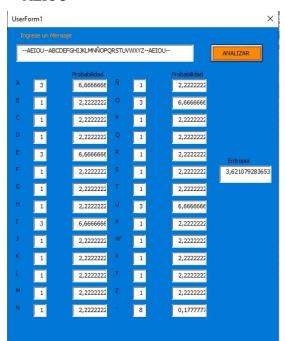
TextBox58.Text = Entropia

End If

End Sub

Para probar nuestra interfaz usamos el siguiente mensaje:

--AEIOU— ABCDEFGHIJKLMNÑOPQRSTUVWXYZ —AEIOU—



Nuestra palabra cumple nuestra condición inicial que es <40 caracteres y dar click en el botón analizar nuestro programa empezara a analizar el mensaje para poder encontrar las veces que se repite cada carácter, la

probabilidad y al final la entropía total que se encuentra con ayuda de la formula donde tenemos en cuenta las probabilidades de ocurrencia de cada carácter.

VI. CONCLUISIONES

- Al realizar la práctica podemos decir que el concepto de incertidumbre surge frecuentemente, ya que, si el diseño del código no es adecuado, crecerá dicha incertidumbre en el sistema que se desee analizar.
- Al enfatizar más con el fenómeno de entropía lo podemos asociar como ruido o desorden en una señal, para el caso de un texto, concluimos que es la cantidad de veces y la probabilidad de que un símbolo llamado se repita. Pero para un sistema más dinámico es conveniente que los símbolos que se repitan más, contribuyan con información más que los símbolos que se repiten menos.
- Se puede afirmar que según la teoría de la información un mensaje tendrá datos relevantes y máxima entropía cuando todos los símbolos son igualmente probables.

BIBLIOGRAFIA

- [1] Abramson N. (1963), Teoría de la Codificación. McGraw Hill.
- [2] Jair N. Bernal. TEORIA DE LA INFORMACION Y CODIFICACION
- [3] Sistemas de Comunicación Digitales y Analógicos Leon W. Couch 7ed