

TEORIA DE LA INFORMACION Y LA CODIFICACION

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 y 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 y 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 número 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.

TEORIA DE LA INFORMACION Y LA CODIFICACION

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:

$$I(S_i) = \log_2 \left(\frac{1}{P_i} \right)$$

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

$$I(S_i S_j) = I(S_i) + I(S_j)$$

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

$$H = \sum_{j=1}^n P_j \log_2 \left(\frac{1}{p_j} \right)$$

Donde H es la entropía, las P_j 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.

#CADENA	CADENA	REPETICIONES	PROBABILIDAD	LOGARITMO				
45	A	3	0.066666667	3.9068906				
	B	1	0.022222222	5.4918531				
	C	1	0.022222222	5.4918531				
	D	1	0.022222222	5.4918531				
	E	3	0.066666667	3.9068906				
	F	1	0.022222222	5.4918531				
	G	1	0.022222222	5.4918531				
	H	1	0.022222222	5.4918531				
	I	3	0.066666667	3.9068906				
	J	1	0.022222222	5.4918531				
	K	1	0.022222222	5.4918531				
	L	1	0.022222222	5.4918531				
	M	1	0.022222222	5.4918531				
	N	1	0.022222222	5.4918531				
	O	1	0.022222222	5.4918531				
	P	1	0.022222222	5.4918531				
	Q	1	0.022222222	5.4918531				
	R	1	0.022222222	5.4918531				
	S	1	0.022222222	5.4918531				
	T	1	0.022222222	5.4918531				
	U	3	0.066666667	3.9068906				
	V	1	0.022222222	5.4918531				
	W	1	0.022222222	5.4918531				
	X	1	0.022222222	5.4918531				
	Y	1	0.022222222	5.4918531				
	Z	8	0.177777778	2.4918531				

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
```

TEORIA DE LA INFORMACION Y LA CODIFICACION

```
TextBox4.Text = Len(cadena)
cadenas = Len(cadena)
Cells([2], [2]) = cadenas

For j = 1 To Len(cadena)

Puntero = Mid(cadena, j, 1)

    If Puntero = "a" Or Puntero = "A"
Then
        contA = contA + 1
    End If

    If Puntero = "b" Or Puntero = "B"
Then
        contB = contB + 1
    End If
.....
.....
Then
    If Puntero = "z" Or Puntero = "Z" Then
        contZ = contZ + 1
    End If
Then
    If Puntero = "-" Then
        contGuion = contGuion + 1
    End If

Next
Range("D2").Select
ActiveCell.FormulaR1C1 = contA
TextBox2.Text = contA
probA = contA / Len(cadena)
Range("E2").Select
ActiveCell.FormulaR1C1 = probA
If probA = 0 Then
    LGA = 0
Else
    LGA = Log(1 / probA) / Log(2)
End If
Range("F2").Select
ActiveCell.FormulaR1C1 = LGA

Range("D3").Select
ActiveCell.FormulaR1C1 = contB
TextBox4.Text = contB
```

```
ProbB = contB / Len(cadena)
Range("E3").Select
ActiveCell.FormulaR1C1 = ProbB
If ProbB = 0 Then
    LGB = 0
Else
    LGB = Log(1 / ProbB) / Log(2)
End If
Range("F3").Select
ActiveCell.FormulaR1C1 = LGB
.....
.....
.....
Range("D28").Select
ActiveCell.FormulaR1C1 = contZ
TextBox54.Text = contZ
probZ = contZ / Len(cadena)
Range("E28").Select
ActiveCell.FormulaR1C1 = probZ
If probZ = 0 Then
    LGZ = 0
Else
    LGZ = Log(1 / probZ) / Log(2)
End If
Range("F28").Select
ActiveCell.FormulaR1C1 = LGZ

Range("D28").Select
ActiveCell.FormulaR1C1 = contGuion
TextBox60.Text = contGuion
probGuion = contGuion / Len(cadena)
Range("E28").Select
ActiveCell.FormulaR1C1 = probGuion
If probGuion = 0 Then
    LGGuion = 0
Else
    LGGuion = Log(1 / probGuion) / Log(2)
End If
Range("F28").Select
ActiveCell.FormulaR1C1 = LGGuion

TextBox3.Text = proba
.....
.....
TextBox59.Text = probGuion
```

TEORIA DE LA INFORMACION Y LA CODIFICACION

```
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 + probÑ * LGÑ + 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--

	Probabilidad		Probabilidad		
A	3	6,6666666	Ñ	1	2,2222222
B	1	2,2222222	O	3	6,6666666
C	1	2,2222222	P	1	2,2222222
D	1	2,2222222	Q	1	2,2222222
E	3	6,6666666	R	1	2,2222222
F	1	2,2222222	S	1	2,2222222
G	1	2,2222222	T	1	2,2222222
H	1	2,2222222	U	3	6,6666666
I	3	6,6666666	V	1	2,2222222
J	1	2,2222222	W	1	2,2222222
K	1	2,2222222	X	1	2,2222222
L	1	2,2222222	Y	1	2,2222222
M	1	2,2222222	Z	1	2,2222222
N	1	2,2222222	-	8	0,1777777

Entropia
3,621079283653

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. CONCLUSIONES

- 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