TEORIA DE LA INFORMACION

TRABAJO PRÁCTICO ESPECIAL

PROYECTO FINAL

Integrantes:

- Disteffano, Juan.
  248303
  juandisteffano@yahoo.com.ar
- Leiva, Nicolas.
  247520
  nicoleiva08@gmail.com

RESUMEN

El presente trabajo práctico tuvo como fin investigar e implementar una solución computacional para distintos temas estudiados a lo largo de la cursada de la materia Teoría De La Información, se llevo a cabo en tres etapas cada una sobre un tema distinto pero en todas se trabajo con el mismo conjunto de datos que se obtuvo del procesamiento digital de una imagen dada por la cátedra que tomamos como fuente de emisión de símbolos. En la primer parte del trabajo práctico el tema a tratar fue la codificación de dicha fuente, primeramente se la trato como una fuente sin memoria y luego como una fuente markoviana estacionaria y su extensión de orden 2 para realizar un análisis comparativo sobre la codificación con el método de Huffman entre ambas, también se realizo una comparación de las entropías y las longitudes medias por símbolo obtenidas, en la segunda parte se trato el tema de compresión de datos donde se implementaron dos técnicas de compresión Huffman (semi-estatico) y Run-Length comparando las tasas de compresión obtenidas con cada una y por ultimo en la tercera etapa se trabajo sobre el tema canales de transmisión donde generamos un canal a partir de conjuntos de datos aportados por la cátedra que luego se utilizó para mandar distintos volúmenes de datos y evaluar así las distintas características que posee un canal.

Además el trabajo práctico especial cuenta con el presente informe a modo de resumen general de todo el proyecto, al igual que durante el desarrollo de las entregas anteriores se analizaron dos fuentes de emisión distintas obteniendo las respectivas codificaciones para luego poder comprimirla y comprobar el primer teorema de Shannon.

INTRODUCCION

La teoría de la información está relacionada con las leyes matemáticas que rigen la transmisión y el procesamiento de la información y se ocupa de la medición de la información y de la representación de la misma, así como también de la capacidad de los sistemas de comunicación para transmitir y procesar información.

La representación de información es el primer tema abordado donde se intenta representar distintos símbolos con la mínima cantidad de información basado en una codificación probabilística para luego poder reducir el tamaño de un volumen de datos con diferentes técnicas, también basadas en la probabilidad de aparición, con el objetivo de trasmitir la información de manera más rápida por un canal de comunicación, aunque hoy en día la tecnología ah evolucionado y mandar grandes volúmenes de información no es muy costoso, las velocidades de conexión entre emisor y receptor son muy altas por lo que el tiempo que se tarda es relativamente poco pero en un principio cuando se comenzó a desarrollar esta teoría no era así y reducir el tamaño podía significar un importante ahorro en costo y tiempo, así como también se ahorraba espacio para su posterior almacenamiento.

Shannon definió la entropía de una fuente como medida de la cantidad de información de una fuente, basándose en las probabilidades de emisión de los símbolos. Matemáticamente queda definida:

Para una fuente sin memoria:

$$H(S) = -\sum_{i=1}^{N} p_i log_2(p_i)$$

Para una fuente con memoria:

$$H_1(S) = \sum_{i=1}^{N} p_i^* (-\sum_{j=1}^{N} p_{j\,|i}\, log_2(p_{j|i}))$$

Esta medida es el número promedio mínimo de preguntas binarias para conocer el símbolo emitido.

Una vez obtenida la codificación con la probabilidad de cada símbolo y la longitud de la codificación obtenida se puede calcular la longitud media de la codificación:

$$< L >= \sum_{i=1}^{N} p_i l_i$$

Es evidente que la longitud media del código tendrá una cota inferior mayor que cero ya que los datos se deben codificar con al menos un símbolo, el teorema de Shannon establece la cota mínima como la entropía de la fuente y la cota máxima como la entropía más uno.

Extendiendo la fuente a orden n (el alfabeto queda ahora definido como la concatenación de n símbolos consecutivos).

Se define el primer teorema de Shannon

Para una fuente sin memoria:

$$H(S) \leq \frac{< L_n >}{n} < H(S) + \frac{1}{n}$$

Por lo tanto:

$$\lim_{n \to \infty} \frac{< L_n >}{n} = H(S)$$

Para una fuente con memoria:

$$H_1(S) + \frac{H(S) - H_1(S)}{n} \leq \frac{< L_n >}{n} < H_1(S) + \frac{[H(S) - H_1(S)] + 1}{n}$$

Por lo tanto:

$$\lim_{n \to \infty} \frac{< L_n >}{n} = H_1(S)$$

Lo que indica que la longitud media de código de un símbolo de la fuente no puede ser menor que la entropía y si se incrementa el orden de extensión de la fuente la longitud media de código de un símbolo se puede llegar a alcanzar el valor mínimo de entropía.

DESARROLLO

Para realizar los distintos ítems solicitados se trabajo sobre las siguientes imágenes tomándolas como fuente de emisión de símbolos:

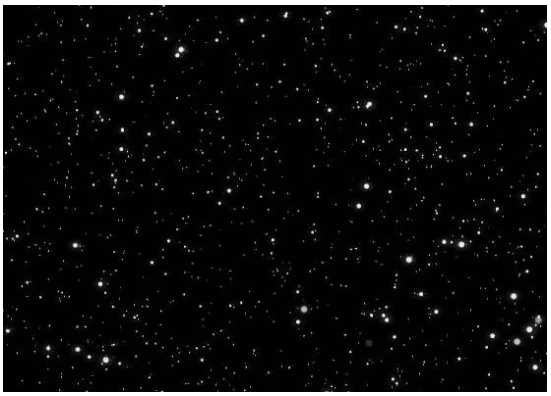


Figura 1. stars_8.bmp

Figura 2. nebula1.bmp

Con la aplicación desarrollada en las entregas anteriores se procedió con análisis de las imágenes primero se genero la codificación de Huffman tomando cada fuente como una fuente sin memoria y se calculo la entropía, a partir de esas codificaciones se obtuvo la longitud media por pixel, comparamos los resultados obtenidos para ambas imágenes.

Para el segundo inciso se realizo la compresión de ambas imágenes con el algoritmo de Huffman semi-estatico desarrollado, teniendo en cuenta el tamaño original del archivo y el tamaño luego de la compresión, se compararon las tasas de compresión obtenidas.

Luego para el tercer inciso, teniendo en cuenta los resultados obtenidos en el primer punto, se hizo la comprobación del primer teorema de Shannon para cada imagen. Esta comprobación se hizo analíticamente a partir de los conceptos teóricos estudiados durante la cursada.

Para el cuarto y último punto de este proyecto final se realizo el análisis analítico y grafico de los resultados obtenidos en el punto 3b) de la entrega del práctico 5. Lo que se resolvió en ese punto fue enviar distintas cantidades de datos por un canal de transmisión y obtener las características de un canal (ruido, pérdida e información mutua) para cada uno de esos volúmenes de datos enviados, generando así una simulación computacional. Luego por cada una de las características se realizo un grafico para analizar la convergencia de cada variable a medida que el volumen de datos enviados aumentaba.

RESULTADOS

1) Luego de generar la codificación Huffman de los pixeles de ambas imágenes obtuvimos los siguientes resultados:

❖ Para la imagen *nebula1.bmp*:

Fuente sin memoria	Fuente con memoria (Extensión orden 2)
Entropía	**Entropía**
H(S) = 2.0238748	H1(S) = 0.7375955
Longitud Media	**Longitud Media (Por símbolo)**
<L> = 2.0480945	<L> = 1.4002008

❖ Para la imagen *star_8.bmp*:

Fuente sin memoria	Fuente con memoria (Extensión orden 2)
Entropía	**Entropía**
H(S) = 0.8719043	H1(S) = 0.7239481
Longitud Media	**Longitud Media (Por símbolo)**
<L> = 1.2672879	<L> = 0.9138184

2) A partir de las codificaciones del punto anterior obtuvimos la compresión de ambas imágenes a nivel bit con el método de Huffman Semi-estatico y se obtuvieron los siguientes tamaños de archivo:

❖ Para la imagen *nebula1.bmp*:

- **Tamaño archivo original:** 213 KB
- **Tamaño archivo comprimido:** 108 KB
- **Tasa de compresión:** 1,97 : 1

❖ Para la imagen *star_8.bmp*:

- **Tamaño archivo original:** 215 KB
- **Tamaño archivo comprimido:** 68 KB
- **Tasa de compresión:** 3,16 : 1

3) Con los resultados obtenidos en el punto 1 se comprobó que se cumple el primer teorema de Shannon para los resultados obtenidos de cada imagen:

❖ Para la imagen *nebula1.bmp*:
- Con **n=1** (Símbolos individuales)

$$0.738 + \frac{2.024 - 0.738}{1} \leq \frac{< L_1 >}{1} < 0.738 + \frac{[2.024 - 0.738] + 1}{1}$$

$$2.024 \leq 2.048 < 3.024$$

- Con **n=2** (Pares de símbolos)

$$0.738 + \frac{2.024 - 0.738}{2} \leq \frac{< L_2 >}{2} < 0.738 + \frac{[2.024 - 0.738] + 1}{2}$$

$$1.381 \leq 1.4 < 1.881$$

❖ Para la imagen *star 8.bmp*:
- Con **n=1** (Símbolos individuales)

$$0.724 + \frac{0.872 - 0.724}{1} \leq \frac{< L_1 >}{1} < 0.724 + \frac{[0.872 - 0.724] + 1}{1}$$

$$0.872 \leq 1.267 < 1.872$$

- Con **n=2** (Pares de símbolos)

$$0.724 + \frac{0.872 - 0.724}{2} \leq \frac{< L_2 >}{2} < 0.724 + \frac{[0.872 - 0.724] + 1}{2}$$

$$0.798 \leq 0.914 < 1.298$$

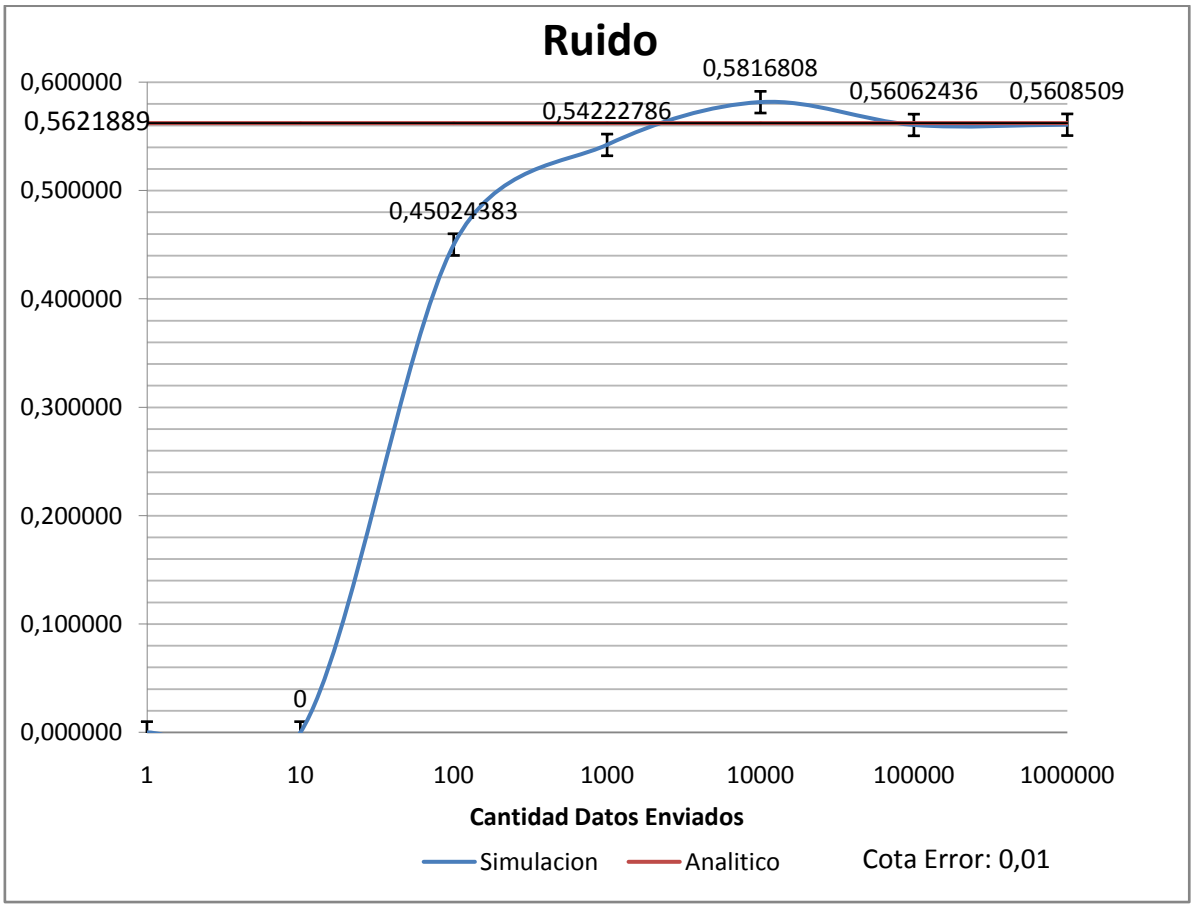
4) Se analizo un canal de transmisión a partir de una imagen que había sido enviada por ese canal y la imagen original, se calcularon analíticamente las características principales obteniendo los siguientes resultados:

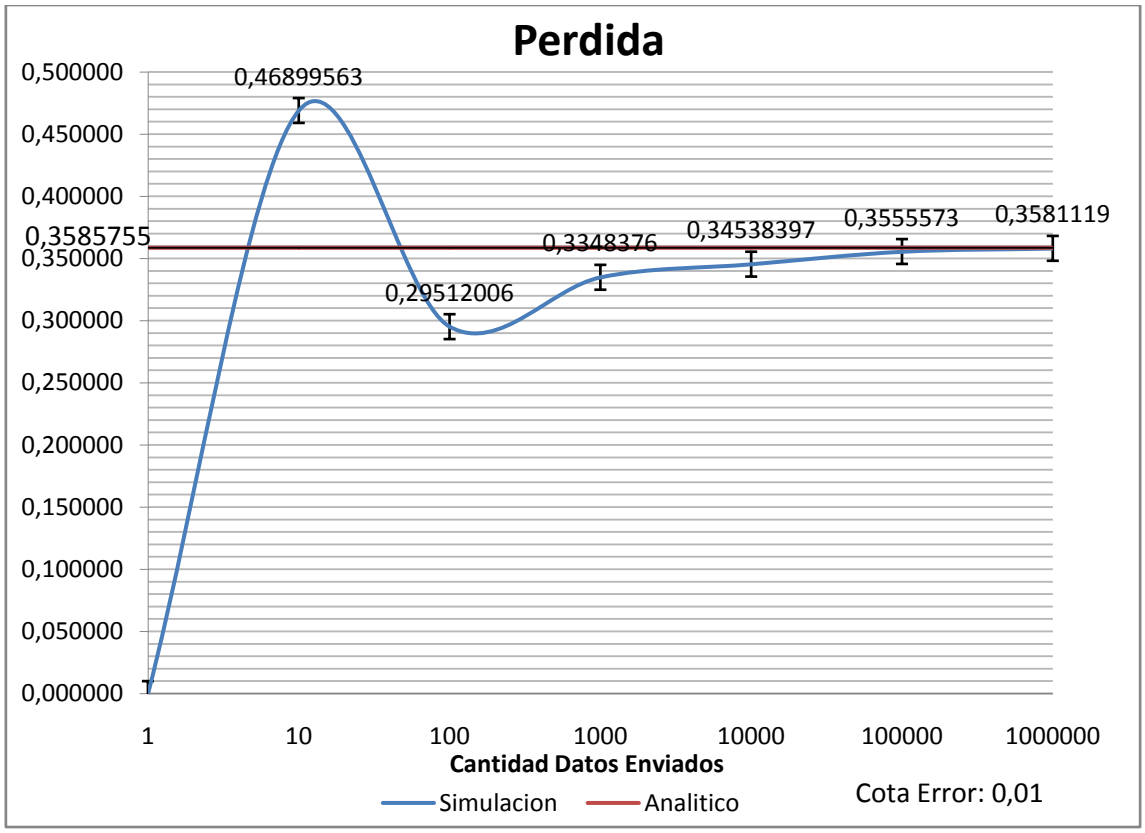
- **Ruido:** 0.5621889
- **Perdida:** 0.35857528
- **Información Mutua:** 0.51347613

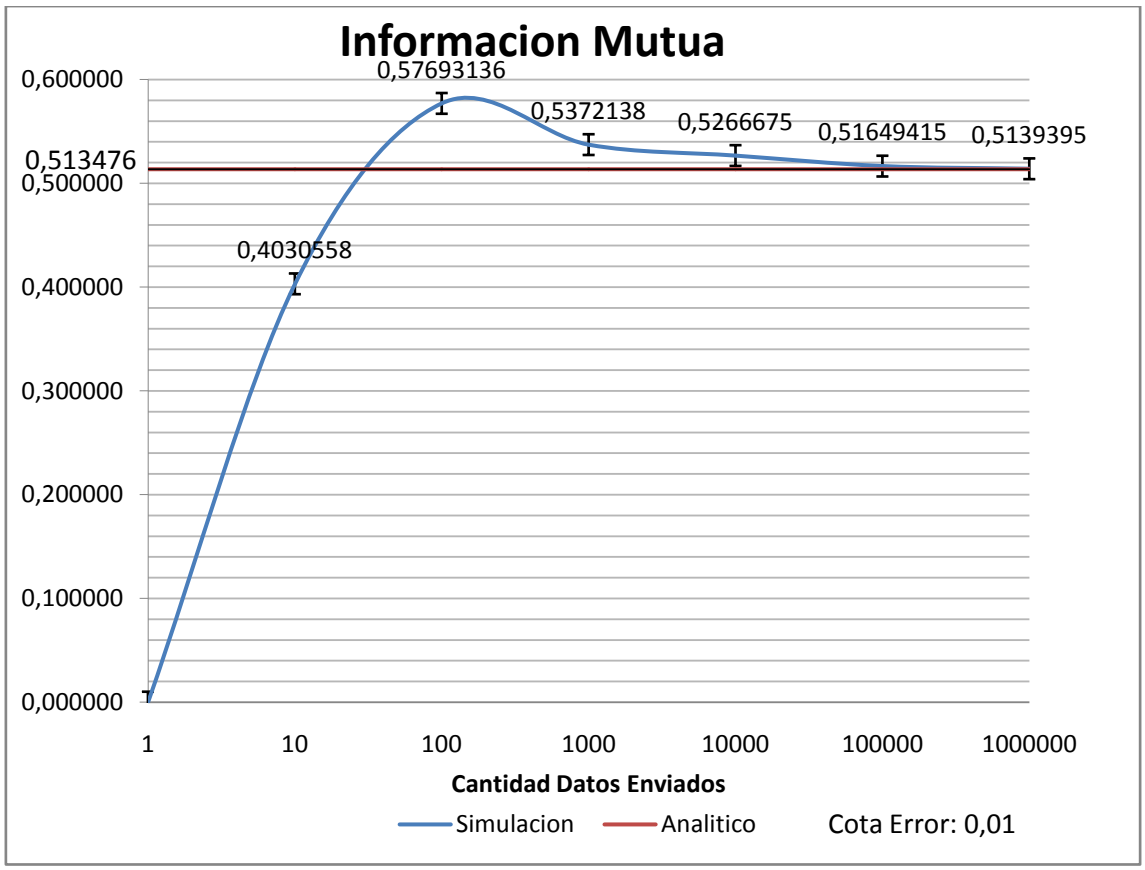
Luego se enviaron por ese canal distintos volúmenes de datos y se volvieron a calcular dichas características para cada envío:

Cantidad Datos Enviados	Ruido	Perdida	Información Mutua
10	0.0	0.46899563	0.4030558
100	0.45024383	0.29512006	0.57693136
1000	0.54222786	0.3348376	0.5372138
10000	0.5816808	0.34538397	0.5266675
100000	0.56062436	0.3555573	0.51649415
1000000	0.5608509	0.3581119	0.5139395

Los gráficos resultantes fueron:




Perdida
0,500000
0,450000
0,400000
0,350000
0,300000
0,250000
0,200000
0,150000
0,100000
0,050000
0,000000
0,46899563
0,3585755
0,29512006
0,3348376
0,34538397
0,3555573
0,3581119
1
10
100
1000
10000
100000
1000000
Cantidad Datos Enviados
Simulacion
Analitico
Cota Error: 0,01


Informacion Mutua
0,600000
0,500000
0,400000
0,300000
0,200000
0,100000
0,000000
0,57693136
0,513476
0,4030558
0,5372138
0,5266675
0,51649415
0,5139395
1
10
100
1000
10000
100000
1000000
Cantidad Datos Enviados
Simulacion
Analitico
Cota Error: 0,01

CONCLUSIONES

Una vez generada la codificación correspondiente a cada fuentes la imagen *star_8.bmp* obtuvo menor valor de longitud media por pixel para los dos tipos de fuentes estudiados (sin memoria y Markoviana), esto se debe a que los símbolos emitidos por esta fuente generan mayor información por lo que se necesitan en promedio menor cantidad de bits para representarlos individualmente en la codificación con el método Huffman, podemos concluir que la longitud media depende directamente de la probabilidad de emisión de cada símbolo, también podemos observar que la longitud media no depende de la entropía sino del método de codificación elegido ya que tomando las fuentes como Markovianas de orden 2 resultan entropías muy similares y si embargo la longitud media por símbolo difiere en mas 60% en relación a los valores obtenidos.

Luego de realizar la compresión de ambas imágenes se obtuvo una tasa de compresión de 3,16 : 1 para la imagen *star_8.bmp* contra 1,97 : 1 de *nebula1.bmp*, como se estudio en la materia obtener una buena compresión depende de dos factores, la entropía y la longitud media obtenida luego de la codificación, cuanto menor sea la entropía y mas se le acerque el valor de la longitud media mejor será la tasa de compresión obtenida, esto se ve claramente reflejado en nuestros resultados donde la imagen *star_8.bmp* tiene una longitud media por pixel menor que *nebula1.bmp*, cabe destacar que para realizar la compresión se utilizo la codificación de los símbolos individuales de la fuente con el método Huffman tomada como una fuente sin memoria.

A partir de los resultados obtenidos al comprobar el primer Teorema de Shannon para ambas imágenes analizadas se puede afirmar que a medida que aumentamos el orden de extensión de la fuente disminuye la longitud media por pixel acercándose a la entropía, esto significa que a medida que aumentamos el orden se logran codificaciones más eficientes y este es el objetivo para luego poder reducir un volumen de datos.

Se puede observar en los gráficos como a medida que se incrementa la cantidad de datos enviados se estabilizan los resultados tendiendo a los valores exactos del canal analizado. Una simulación con pocas muestras puede arrojar un resultado erróneo como es el caso del ruido para 10 datos enviados, lo ideal sería establecer una cota de error para las tres variables y no finalizar la simulación hasta que dos resultados sucesivos difieran en un número menor a ese valor puesto como cota. Definimos la cota de error en 0,01 y la diferencia entre valores sucesivos en la simulación realizada se alcanza entre los 100000 y 1000000 de datos enviados.