83GIIN-Compresión y Recuperación de Información Multimedia

Actividad 1.- Teoría de la Información y Entropía

Objetivo

El objetivo de esta actividad es afianzar los conceptos básicos de la Teoría de la Información, en particular lo relacionado con la tasa de compresión y la entropía, y estudiar la importancia de la modelización de una fuente para obtener buenos resultados en la compresión de la misma. Para esta actividad se suministrarán archivos con diferentes características.

Descripción de la actividad a realizar

En esta actividad, se llevará a cabo un análisis de dos tipos de archivos: texto plano y archivos de imagen en formato RAW. El objetivo principal es estudiar los datos contenidos en cada archivo para identificar la diversidad de caracteres o símbolos presentes, analizar su frecuencia de aparición y representar esta distribución mediante un histograma.

Parte 1: Análisis de Archivos de Texto

Trabajaremos con dos archivos de texto en formato plano para examinar la distribución de los caracteres y símbolos que contienen:

- don-quixote.txt
  Contiene el texto completo del libro Don Quijote en inglés, en formato de texto plano.
- ecoli.fa
  Formato FASTA, utilizado para representar secuencias de ácidos nucleicos o proteínas.

Tareas a realizar:

- Identificación de caracteres: Determinar los caracteres o símbolos presentes en cada archivo.
- Frecuencia de aparición: Calcular la frecuencia de aparición de cada carácter o símbolo.
- Histograma: Representar esta distribución en un histograma.
- Cálculo de entropía: Calcular la entropía de los datos de cada archivo. La entropía nos permitirá evaluar la cantidad de información contenida en cada conjunto de datos.

- Análisis de compresión: Explorar la posible tasa de compresión o relación de compresión que se podría alcanzar utilizando un modelo de codificación que maximice la entropía (sin necesidad de implementar un compresor).
- Conclusiones: Comparar los resultados del histograma, la entropía y la cantidad mínima de bits necesarios para representar los símbolos presentes en cada archivo, y sacar conclusiones.

Parte 2: Análisis de Imágenes en Formato RAW

En esta parte de la actividad, analizaremos imágenes en formato RAW. Cada archivo contiene únicamente los valores de los píxeles, representados en una escala de grises (de 0 a 255), y todas las imágenes son de tamaño 64x64 píxeles con un solo canal (escala de grises).

Archivos de imagen a analizar:

- imagen_1.raw
- imagen_2.raw
- imagen_3.raw
- imagen_4.raw
- imagen_5.raw
- imagen_6.raw

Tareas a realizar:

- Análisis de la imagen: Identificar la variedad de valores de píxel presentes en cada imagen.
- Frecuencia de aparición: Calcular la frecuencia de aparición de cada valor de píxel (de 0 a 255).
- Histograma: Representar la distribución de los valores de píxel mediante un histograma.
- Cálculo de entropía: Calcular la entropía de los datos de cada imagen, lo que nos proporcionará una medida de la "incertidumbre" o la cantidad de información contenida en la imagen.
- Análisis de compresión: Explorar la posible tasa de compresión o relación de compresión que se podría alcanzar utilizando un modelo de codificación que maximice la entropía (sin necesidad de implementar un compresor).
- Comparación entre imágenes: Evaluar las diferencias encontradas entre las imágenes en función de la cantidad de símbolos (valores de píxel) distintos presentes y su distribución.

Sugerencias para la implementación

Para esta actividad se puede utilizar el lenguaje de programación que consideres más apropiado (Python, C, Matlab, R/RStudio, etc.) y hacer uso de recursos disponibles en la web siempre y cuando se indique su procedencia.

Detalles sobre la entrega

La entrega debe incluir:

- un **informe** que recoja de forma **bien estructurada** y **organizada** el análisis realizado y las conclusiones a las que lleguéis a partir de dicho análisis. En este informe es importante indicar el lenguaje de programación, librerías o paquetes utilizados para el análisis de los datos.
    - Planteamiento del problema (no es copiar el texto de la actividad)
    - Métodos utilizados para analizar los distintos archivos, es decir cantidad de símbolos presentes en el archivo, frecuencia de los símbolos (histograma), etc.
    - Resultados obtenidos por archivo: la forma ideal de mostrar estos resultados es mediante gráficos. Si colocan un listado de los símbolos con su frecuencia no es posible extraer conclusiones. Calcular la entropía y relacionar el valor con los gráficos.
    - Conclusiones
- **Código fuente**
    - Python: incluir Jupyter notebook con el código para la ejecución de la actividad o su equivalente o código Python.
    - RStudio: incluir el archivo ".rmd" o ".r" con el código que permita su ejecución
    - Matlab: incluir el archivo ".m"
    - Otro lenguaje de programación: incluir código fuente

La entrega se hará a través del Campus VIU en un archivo comprimido .zip

Fecha de entrega sugerida para *feedback*

- Fecha propuesta: 10 de marzo de 2025

Rúbrica para la evaluación

	Niveles de rendimiento			
Criterios	**Excelente 9-10**	**Notable 7-8**	**Aprobado 5-6**	**Suspenso 0-4**
Dominio de los conceptos **Ponderación 40,00%**	**100,00 %** Identifica y explica adecuadamente todos los conceptos relacionados con el tema	**80,00 %** Identifica y explica adecuadamente la mayoría de los conceptos relacionados con el tema	**60,00 %** Identifica y explica adecuadamente alguno de los conceptos relacionados con el tema	**40,00 %** Identifica y explica muy poco o ninguno de los conceptos relacionados con el tema
Relación entre Teoría de Información y compresión **Ponderación 30,00%**	**100,00 %** Identifica y relaciona correctamente los conceptos de la teoría de la información con la compresión de datos, razón de compresión, etc.	**80,00 %** Identifica y relaciona correctamente los conceptos de la teoría de la información con la compresión de datos, razón de compresión, etc.	**60,00 %** Identifica y relaciona algunos de los conceptos de la teoría de la información con la compresión de datos, razón de compresión, etc.	**40,00 %** Identifica y relaciona pocos o ninguno de los conceptos de la teoría de la información con la compresión de datos, razón de compresión, etc.
Estructura, presentación y redacción del informe **Ponderación 10,00%**	**100,00 %** El informe presenta una muy buena estructura con todos los apartados. Aplica correctamente las reglas gramaticales, ortográficas y de sintáxis	**80,00 %** El informe presenta una buena estructura pero falta algún apartado importante. Presenta algunos errores de redacción y gramaticales	**60,00 %** El informe carece de varios apartados importantes. Presenta muchos errores de redacción y ortográficos.	**40,00 %** El trabajo no presenta una estructura de reporte técnico. Se cometen múltiples errores gramaticales
Resultados **Ponderación 20,00%**	**100,00 %** Los resultados obtenidos del análisis son correctos	**80,00 %** Unos pocos resultados del análisis no son correctos	**60,00 %** Una gran parte de los resultados no son correctos	**40,00 %** Los resultados son incorrectos