Compresibilidad de las imágenes

Joan Mata Pàrraga - 1533089

Resum– En el context actual, on l'eficiència en l'emmagatzematge i transmissió de dades és crucial, la compressió d'imatges juga un paper fonamental en la gestió eficaç dels recursos informàtics. El present treball investiga la transformació d'imatges en text com una metodologia innovadora per a la compressió de dades visuals. S'avaluen diferents mètriques de compressibilitat de dades i es comparen els resultats amb les taxes de compressió produïdes per tres mètodes de compressió de text (LZ77, LZ78 i BWT+RLE) aplicats a imatges prèviament transformades en text.

Paraules clau– Compressió de textos i imatges, Compressibilitat de textos, LZ77, LZ78, BWT, RLE, Corba de Morton, Entropia, Compressibilitat d'imatges mitjançant text.

Resumen– En el contexto actual, donde la eficiencia en el almacenamiento y transmisión de datos es crucial, la compresión de imágenes juega un papel fundamental en la gestión eficaz de los recursos informáticos. El presente trabajo investiga la transformación de imágenes en texto como una metodología innovadora para la compresión de datos visuales. Se evalúan diferentes métricas de compresibilidad de datos y se comparan los resultados con las tasas de compresión producidas por tres métodos de compresión de texto (LZ77, LZ78 y BWT+RLE) aplicados a imágenes previamente transformadas en texto.

Palabras clave– Compresión de textos e imágenes, Compresibilidad de textos, LZ77, LZ78, BWT, RLE, Curva de Morton, Entropía, Compresibilidad de imágenes mediante texto.

Abstract– In the current context, where efficiency in data storage and transmission is crucial, image compression plays a fundamental role in the effective management of computing resources. The presente work investigates the transformation of images into text as an innovative methodology for compressing visual data. Various data compressibility metrics are evaluated and the results are compared with the compression ratios produced by three text compression methods (LZ77, LZ78, and BWT+RLE) applied to images previously transformed into text.

Keywords– Text and Image Compression, Text Compressibility, LZ77, LZ78, BWT, RLE, Morton Curve, Entropy, Image Compressibility via Text.

1 Introducción

En los últimos años, el volumen de datos está creciendo exponencialmente, atrayendo a investigadores a desarrollar en un término popular, el Big Data. En el 2018, más de 3,2 mil millones de la población mundial estaba conectada a internet. De cara a 2050, se estima que el 95 % de la población mundial tendrá acceso a internet, con lo cual tendremos una cantidad sin precedentes de datos generados [16]. Este aumento en la producción de datos hace necesaria la implementación de técnicas de compresión de datos, las cuales son cruciales por varias razones.

En primer lugar, la preservación de la gran cantidad de datos producidos consume vastos recursos de almacenamiento en diferentes instituciones. Además, las tendencias históricas de generación de datos son económicamente insostenibles y los recursos de almacenamiento ya están comenzando a limitar los objetivos científicos. La compresión de datos se vuelve primordial para el almacenamiento eficiente y la recuperación de información, así como para la conservación de recursos de almacenamiento. Es igualmente significativo evidenciar que los datos restaurados mantienen la misma información que los datos primigenios. Otra razón de la importancia de la compresión de datos es la eficiencia en la transmisión, ya que los datos comprimidos requieren menos ancho de banda para ser transmitidos, acele-

• E-mail de contacto: Joan.Mata@uab.cat
• Mención realizada: Tecnologías de la Información
• Trabajo tutorizado por: Joan Serra Sagristà (DEIC)
• Curs 2023/24

rando así la transferencia de información a través de redes y reduciendo los costos asociados.

En este estudio, nos enfocaremos en el análisis de la compresibilidad de imágenes utilizando varias métricas como el lenguaje de n-gramas, el tamaño del vocabulario, la longitud promedio de las palabras y la tasa de repetición. También se evaluará la efectividad de la compresión de algunos métodos conocidos como LZ77, LZ78 y la combinación de BWT con RLE [15].

El objetivo principal del estudio es analizar las métricas de compresibilidad para determinar cuán comprimible puede ser un texto o imagen antes de realizar la compresión. Además, se pretende analizar las diferencias de rendimiento de los métodos de compresión anteriormente mencionados cuando se aplican a imágenes transformadas en texto. Es fundamental comprender si estos métodos pueden ofrecer una compresión efectiva para este tipo de datos y cómo se comparan entre sí.

Para llevar a cabo estas evaluaciones, realizaremos una comparación inicial utilizando un pequeño conjunto de datos de texto, donde aplicaremos las métricas y métodos de compresión mencionados para observar sus diferencias y determinar su capacidad de predecir o reducir el tamaño, respectivamente. Posteriormente, extenderemos el estudio a imágenes, aplicando las mismas métricas y métodos a imágenes previamente transformadas en texto.

2 ESTRUCTURA DEL INFORME

Basándose en el contenido y la estructura del documento proporcionado, se ha elaborado una descripción de la organización del artículo, abarcando cada sección con una explicación concisa.

En la **introducción** se expone la relevancia y el contexto del estudio y se comenta el objetivo principal. Además, se explican de manera reducida los algoritmos y métricas utilizada e implementadas en el mismo. Este primer apartado, establece el escenario para la investigación.

En la sección del **desarrollo** se analizan los resultados obtenidos durante la investigación, tanto en compresibilidad como en compresión de textos y de imágenes. Se realizan comparaciones entre los diferentes métodos y métricas. Esta sección se divide en tres subsecciones para una mayor facilidad de comprensión de lo que se ha realizado. Una de ellas se centra en toda la información y procesos relativo a los *textos*. Otra subsección se centra en el trabajo realizado para las *imágenes* que han sido transformadas a texto previamente para hacer el estudio, comparándolas con algunos métodos específicos para compresión de imágenes sin pasar a texto. Por último, una sección dedicada a las *gramáticas unidimensionales* donde se discute las dificultades y limitaciones encontradas respecto a ellas.

La parte final del documento presenta las **conclusiones** donde se dan a conocer los resultados del estudio.

3 ESTADO DEL ARTE

3.1. Métodos de compresión de textos

3.1.1. Lempel-Ziv 1977 (LZ77)

El algoritmo de compresión datos sin pérdidas presentado por Abraham Lempel y Jacob Ziv en 1977 [19] funciona buscando y reemplazando repeticiones de datos en una secuencia.

Se utiliza una *ventana deslizante* para buscar repeticiones de datos. Esta ventana contiene una porción de la secuencia de datos que ya ha sido vista. El algoritmo busca la parte de la secuencia actual que coincide con una parte anterior de la ventana deslizante. Una vez que se encuentra una coincidencia, el algoritmo codifica dos números: el desplazamiento desde la posición actual hacia la coincidencia encontrada en la ventana deslizante y la longitud de la coincidencia. Además de estos dos números, también codifica el siguiente carácter de la secuencia.

El proceso se repite avanzando en la secuencia de datos y ajustando la ventana deslizante para incluir la parte más reciente de la secuencia. El algoritmo continúa buscando y codificando repeticiones hasta que se ha recorrido toda la secuencia de datos. La codificación resultante tiene la siguiente forma:

$$(d_1, l_1, c_1)(d_2, l_2, c_2)...(d_n, l_n, c_n)$$

donde d se refiere a desplazamiento desde la posición actual, l a la longitud de la coincidencia y c el próximo carácter a insertar.

Para convertir en binario la codificación resultante hay que considerar varios aspectos. Primeramente, se ha de encontrar el valor n que hace referencia al número de frases que contiene la codificación. Una frase es cada elemento "(d, l, c)". Los primeros 8 bits de la secuencia binaria es este número n. Seguidamente irían los valores de d, l y c codificados en binario. Los números de d y l se pueden codificar según $\lceil \log_2(n) \rceil$, mientras el carácter c se codifica en 8 bits siguiendo los símbolos de la tabla ASCII.

3.1.2. Lempel-Ziv 1978 (LZ78)

El LZ78 es otro algoritmo de compresión datos sin perdidas presentado por Abraham Lempel y Jacob Ziv pero en 1978 [3, 9, 20] que comienza inicializando un diccionario vacío. A medida que se recorre la secuencia de datos, el algoritmo busca subcadenas que ya han aparecido antes en el texto. En lugar de buscar repeticiones como en LZ77, LZ78 busca secuencias que ya han sido vistas y las añade al diccionario. Cuando encuentra una secuencia que no está en el diccionario, la agrega y emite un código que representa la combinación de símbolos ya presentes en el diccionario. Después de emitir un código para una nueva secuencia, el algoritmo actualiza el diccionario agregando esa secuencia como una nueva entrada.

El proceso se repite a lo largo de toda la secuencia de datos, con el diccionario creciendo a medida que se encuentran nuevas secuencias. Cada vez que se encuentra una secuencia no vista antes, se agrega al diccionario y se emite un nuevo código.

Todos los códigos emitidos siguen la siguiente estructura:

$$(i_1, c_1)(i_2, c_2)...(i_n, c_n)$$

donde *i* hace referencia al valor del diccionario y *c* el próximo carácter a insertar.

Para convertir en binario la codificación resultante ha de calcular varios aspectos. Primeramente, se ha de encontrar el valor *n* que hace referencia al número de frases que contiene la codificación. Una frase es cada elemento "(i, c)". Los primeros 8 bits de la secuencia binaria es este número *n*. Seguidamente irían los valores de *i* y *c* codificados en binario. El número de *i* se pueden codificar según $\lceil \log_2(n) \rceil$, mientras el carácter *c* se codifica en 8 bits siguiendo los símbolos de la tabla ASCII.

3.1.3. Transformación de Burrows-Wheeler (BWT)

La Transformada Burrows-Wheeler (BWT) [1] reorganiza una secuencia de datos moviendo las letras de la secuencia de datos de forma cíclica y ordenándolas de acuerdo con las diferentes permutaciones. Después, la BWT toma la última columna de la matriz resultante. Esta columna contiene los caracteres de la secuencia de datos reorganizada en el orden en que aparecen en el texto original.

La BWT no comprime directamente los datos pero la propiedad clave es que agrupa caracteres similares juntos, lo que facilita la compresión.

3.1.4. Run-length encoding (RLE)

La Codificación de Longitud de Ejecución (RLE, Run-length encoding) busca secuencias de datos consecutivos idénticos. Cuando encuentra una secuencia repetida, en lugar de almacenar cada instancia de la secuencia, RLE la reemplaza por un solo carácter seguido de un número que indica cuántas veces se repite esa secuencia.

RLE es efectivo cuando hay repeticiones en los datos. Por lo que reduce la cantidad de datos necesarios para representar la secuencia.

3.1.5. BWT + RLE

La combinacion de BWT con RLE puede producir mejoras significativas en la compresión de datos al aprovechar las capacidades de cada algoritmo para reducir la redundancia y la entropía en la secuencia de datos.

La BWT reorganiza la secuencia de datos para agrupar caracteres similares juntos, lo que facilita la detección de repeticiones por parte de RLE. Al agrupar caracteres similares, la BWT puede hacer que las secuencias repetitivas sean más largas y, por lo tanto, más fácilmente comprimibles por RLE.

Al final de esta combinación, se genera un código de número seguido de carácter y así sucesivamente. Para codificar en binario esta secuencia, se calcula el valor de *n* que es el máximo número de dígitos de todos los números del código RLE. Los primeros 8 bits se destinan para el valor binario de *n*. Seguidamente iría el valor del número que se pueden codificar según $\lceil \log_2(n) \rceil$, mientras el valor del carácter lo codificamos en 8 bits siguiendo los símbolos de la tabla ASCII. Esto se realiza sucesivamente con todos los valores de la secuencia.

3.2. Métodos de compresión de imágenes

Los métodos de la sección 3.1 pueden ser utilizados para la compresión de imágenes, únicamente se han de transformar dichas imágenes a textos antes de la compresión.

3.2.1. Técnicas específicas para imágenes

Ahora se procederá a realizar una breve descripción de algunas técnicas de compresión específicas para imágenes.

Joint Photographic Experts Group (JPEG) [6] es uno de los formatos de compresión de imágenes más populares y ampliamente utilizados. Utiliza un algoritmo de compresión con pérdida que se basa en la transformada discreta del coseno (DCT) para comprimir imágenes fotográficas y otras imágenes con información compleja.

Portable Network Graphics (PNG) [8] es un formato de compresión sin pérdida, a diferencia de JPEG. Utiliza algoritmos de compresión sin pérdida como la compresión DEFLATE (LZ77 + Huffman) para reducir el tamaño del archivo sin perder información [15]. Es ideal para imágenes con áreas de color plano o gráficos con líneas nítidas.

Graphics Interchange Format (GIF) es un formato de imagen que utiliza una compresión sin pérdida, pero está limitado a una paleta de colores de 256 o menos. Es adecuado para imágenes simples con áreas de color plano y animaciones simples.

Tagged Image File Format (TIFF) [7] no es en sí mismo un algoritmo de compresión, es un formato de archivo de imagen que puede contener imágenes sin comprimir o comprimidas con una variedad de algoritmos, incluidos los algoritmos de compresión sin pérdida como LZW (Lempel–Ziv–Welch [15]) o con pérdida como JPEG.

3.3. Métricas de compresibilidad de textos

Las métricas de compresibilidad evalúan la capacidad de un texto de ser reducido de tamaño antes de ser comprimido [5, 11, 12]. Los textos con mayor redundancia y menor entropía suelen ser más compresibles.

3.3.1. Entropía de la Información

La entropía de la información cuantifica la incertidumbre promedio en un conjunto de datos.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2(P(x_i)) \quad (bits) \qquad (1)$$

donde $H(X)$ es la entropía del conjunto de datos X y $P(x_i)$ es la probabilidad de que ocurra el símbolo x_i.

La entropía de la información se utiliza para medir la incertidumbre o la imprevisibilidad en un conjunto de datos. Cuanto menor sea la entropía, más predecible será el conjunto de datos. Esto significa que si un texto tiene una entropía baja, es más probable que sus símbolos estén estructurados de manera predecible, lo que lo hace más compresible.

Sin embargo, la entropía se centra exclusivamente en la probabilidad de ocurrencia de eventos individuales, sin tener en cuenta su disposición o estructura secuencial. Por lo tanto, la entropía no distingue entre conjuntos de datos distribuidos aleatoriamente y aquellos que contienen patrones

o secuencias repetitivas claras. Esto significa que un texto con una alta entropía puede contener patrones repetitivos que podrían ser comprimidos eficientemente, pero la entropía no captaría esta repetitividad [10, 13, 14, 17]. Por lo tanto, aunque es una métrica importante, debe complementarse con otras técnicas de análisis de compresibilidad para obtener una evaluación más completa y precisa.

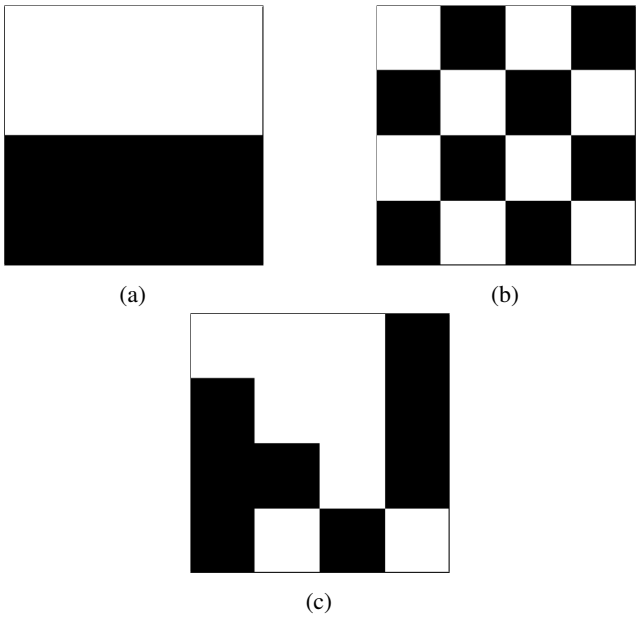


Fig. 1: Imágenes de 4x4 píxeles.

En la figura 1 tenemos tres imágenes de 16 píxeles cada una donde siempre hay 8 píxeles blancos y otros 8 negros, pero cambiamos su distribución en la representación. Podemos decir que la entropía de las tres imágenes es la misma, pero no todas tienen la misma dificultad a la hora de comprimir.

3.3.2. Modelo del lenguaje de n-gramas

El modelo de lenguaje de n-gramas es una técnica utilizada en el procesamiento del lenguaje natural para modelar la probabilidad de una secuencia de palabras o caracteres en un texto, donde un n-grama es una secuencia contigua de "n" elementos, que pueden ser palabras, letras o caracteres.

Este modelo se basa en la suposición de que la probabilidad de ocurrencia de una palabra o carácter depende de las palabras o caracteres previos en la secuencia. Por lo tanto, el modelo de lenguaje de n-gramas estima estas probabilidades observando la frecuencia de ocurrencia de los n-gramas en un corpus de entrenamiento. La estimación de estas probabilidades permite al modelo generar nuevas secuencias de palabras o caracteres que tienen una alta probabilidad de ocurrencia.

Al evaluar la compresibilidad de un texto, el modelo de lenguaje de n-gramas puede proporcionar información sobre la predictibilidad del texto basada en la frecuencia de aparición de secuencias de palabras o caracteres. Un texto con secuencias repetitivas o patrones claros tendrá una estructura más predecible y, por lo tanto, será más compresible. Por otro lado, un texto con una alta variabilidad en las secuencias de n-gramas puede tener una compresibilidad más baja.

Para un texto binario, el modelo de lenguaje de n-gramas se aplica de manera similar, pero en lugar de considerar palabras o caracteres, considera secuencias de bits.

La entropía mide la incertidumbre o imprevisibilidad de los n-gramas en el texto, siguiendo la ecuación 1. Un texto con patrones repetitivos tendrá baja entropía, indicando menor incertidumbre y alta compresibilidad. Por otro lado, un texto con alta variabilidad en los n-gramas tendrá alta entropía, indicando mayor incertidumbre y baja compresibilidad.

Un texto con muchos patrones repetitivos tendrá n-gramas con alta probabilidad, resultando en códigos más cortos y mayor compresibilidad. Un texto con alta variabilidad en los n-gramas tendrá probabilidades más uniformemente distribuidas, resultando en códigos más largos y menor compresibilidad.

La compresibilidad se calcula como la suma ponderada de las longitudes de código. A más cercano sea el valor resultante al valor de "n", mejor compresibilidad.

3.3.3. Tamaño del vocabulario

El tamaño del vocabulario se refiere al número total de palabras únicas en un texto. Es una medida fundamental en el procesamiento del lenguaje natural que indica la diversidad léxica de un texto. Cuanto mayor sea el tamaño del vocabulario, más palabras diferentes se utilizan en el texto, lo que implica una mayor variedad de información y una menor predictibilidad. Se esperaría un valor natural como solución de este método, pero está normalizado al número total de palabras del texto para facilitar la comparación entre diferentes textos.

En el contexto de la compresibilidad, el tamaño del vocabulario puede ser una métrica importante. Un texto con un tamaño de vocabulario más grande tiende a ser menos compresible, ya que hay menos repeticiones de palabras y, por lo tanto, menos oportunidades para la compresión basada en la eliminación de redundancias.

Por otro lado, un texto con un tamaño de vocabulario más pequeño puede ser más compresible, ya que es más probable que contenga repeticiones de palabras y patrones lingüísticos que pueden ser explotados por algoritmos de compresión.

Es importante destacar que el tamaño del vocabulario puede variar según el idioma y el tipo de texto.

3.3.4. Longitud promedio de las palabras

La longitud promedio de las palabras se refiere a la cantidad promedio de caracteres por palabra en un texto. Esta métrica proporciona información sobre la complejidad léxica y la estructura del texto.

En el contexto de la compresibilidad, la longitud promedio de las palabras puede influir en la capacidad de comprimir eficientemente un texto. Algunas consideraciones a tener en cuenta son:

- Mayor predictibilidad: Los textos con palabras más largas tienden a ser más predecibles en términos de longitud de palabra. Esto se debe a que las palabras más largas tienen menos variabilidad en su longitud, lo que facilita la identificación de patrones y repeticiones por parte de los algoritmos de compresión.

II Menos redundancia léxica: Los textos con palabras más largas pueden tener menos repeticiones de palabras debido a su longitud.

III Efecto del idioma y el tipo de texto: La longitud promedio de las palabras puede variar significativamente según el idioma y el tipo de texto.

IV Compresión basada en diccionarios: En algunos casos, una longitud promedio de palabras más larga puede permitir una compresión más efectiva mediante técnicas como la compresión basada en diccionarios, donde las palabras largas y menos comunes se pueden reemplazar por códigos más cortos.

3.3.5. Tasa de repetición de secuencias

La tasa de repetición de secuencias se refiere a la frecuencia con la que aparecen secuencias de caracteres, palabras o patrones en un texto. Esta métrica es importante en la evaluación de la compresibilidad de un texto, ya que las secuencias repetidas ofrecen oportunidades para la compresión eficiente al identificar y codificar estas repeticiones.

Al analizar un texto, se pueden identificar secuencias de caracteres, palabras o patrones que se repiten con frecuencia. Las secuencias repetidas ofrecen oportunidades para la compresión mediante diferentes técnicas.

La longitud de las secuencias repetidas puede variar; cuanto más larga sea la secuencia repetida, mayor será el potencial de compresión; ya que una secuencia larga repetida puede ser codificada más eficientemente que múltiples secuencias cortas.

La tasa de repetición de secuencias puede influir en la selección del algoritmo de compresión más adecuado para un texto dado.

3.3.6. Coherencia semántica y gramatical

La coherencia semántica y gramatical es esencial para la comprensión y compresión eficaz de un texto. Implica tanto la conexión lógica entre las ideas como la corrección en el uso de la gramática dentro del texto.

En términos de coherencia semántica, se evalúa la consistencia y conexión lógica entre las ideas, conceptos y temas en un texto. Un texto semánticamente coherente presenta un tema claramente definido y mantiene su relevancia a lo largo del texto. Esto se refleja en una progresión lógica de ideas y argumentos, con transiciones suaves entre párrafos y secciones. Además, implica el uso apropiado de la sintaxis y la gramática para expresar ideas de manera clara y precisa.

Por otro lado, la coherencia gramatical se centra en la corrección y consistencia en el uso de la gramática dentro del texto. Un texto gramaticalmente coherente está libre de errores gramaticales y presenta una estructura sintáctica clara y coherente. Esto incluye la consistencia en el estilo de escritura, como mantener la misma voz, tono y nivel de formalidad en todo el texto, así como el uso adecuado del vocabulario.

Ambos aspectos de la coherencia, semántica y gramatical, contribuyen a la comprensión y compresión del texto. Un texto coherente en ambos aspectos tiende a ser más predecible y, por lo tanto, más compresible. Esta coherencia puede ser aprovechada por algoritmos de compresión para identificar y comprimir secciones de texto que mantienen una estructura sintáctica y gramatical similar, así como una alta correlación en términos de contenido y significado.

3.4. Compresión gramatical unidimensional

La compresión gramatical unidimensional es un enfoque de compresión de datos que transforma secuencias de texto en una gramática formal diseñada para generar únicamente la secuencia de texto original. Este método se centra en la creación de una gramática a partir del análisis del texto para identificar patrones, repeticiones y estructuras, que son luego representados de manera más compacta mediante reglas de producción. Al construir una gramática que encapsula estas estructuras, se permite la reconstrucción del texto original sin pérdida de información, logrando la compresión mediante la minimización del número y tamaño de estas reglas [2, 4].

Los componentes fundamentales de la gramática formal incluyen variables, también conocidas como símbolos no terminales, que representan patrones o estructuras dentro del texto y pueden ser reemplazados por grupos de otros símbolos según las reglas de producción. Los símbolos terminales constituyen los elementos básicos del texto original que no se descomponen más en la gramática. La variable inicial es el punto de partida para la derivación o reconstrucción del texto original, utilizando las reglas de producción, que son las instrucciones que definen cómo se pueden reemplazar las variables por grupos de símbolos terminales o no terminales.

La Forma Normal de Chomsky (FNC) es una manera estándar de representar gramáticas libres de contexto. Una gramática está en FNC si todas sus reglas de producción siguen el principio de que una variable puede ser substituida como 2 variables o como un símbolo terminal. Un ejemplo de ello es el siguiente,

$$A \to AB|a$$
$$B \to BB|b$$
$$S \to AB|a$$

donde A y B son variables no terminales, a y b son símbolos terminales y S es el símbolo inicial.

Se ha investigado la posibilidad de desarrollar un programa o función capaz de generar una gramática en forma normal de Chomsky a partir de cualquier tipo de texto. Sin embargo, esta tarea excede tanto la capacidad técnica como el tiempo disponible para el proyecto actual.

4 DESARROLLO

Este apartado esta orientado a explicar los datos obtenidos durante el desarrollo de la investigación y las posibles hipótesis o conclusiones que se van obteniendo durante su realización.

- Las *longitudes* de los diferentes textos, tanto el original como los comprimidos, hacen referencia a la cantidad de bits que contiene dicho texto (o imagen) en binario.
- Los *símbolos únicos* son la cantidad de símbolos diferentes que hay en el texto antes de ser comprimido.

En los textos, cada carácter es un símbolo diferente. En cambio, en las imágenes, un símbolo diferente es el conjunto de tres caracteres. En el apartado 4.2.1 se explica con más detalle, pero las imágenes están compuestas de cadenas de texto numéricas, por lo tanto un '000' es un símbolo y '255' es otro símbolos diferente. No existen los símbolos de menos o de más de tres caracteres.

- La *entropía* se refiere a la medida de incertidumbre asociada con el contenido de la información. Si todos los resultados posibles son igual de probables, la entropía es máxima. La entropía máxima es calculada según $\log_2(\textit{número de símbolos únicos})$.
- La *tasa de compresión* es una medida para indicar cuánto se ha reducido el tamaño de un texto después de aplicar un cierto algoritmo de compresión. Se calcula como la relación del tamaño o longitud del texto original y el tamaño del comprimido. Cuánto mayor sea la tasa de compresión más efectiva, lo que significa que se ha eliminado más redundancia o información innecesaria del archivo original.
- La *compresibilidad* es el valor que devuelve cada una de las métricas de compresibilidad estudiadas.

4.1. Textos

Se inicia el desarrollo del proyecto con la creación de los métodos de compresión de textos Lempel-Ziv de 1977 (LZ77), el de 1978 (LZ78) y la Transformación de Burrows-Wheeler (BWT) junto a la Codificación de Longitud de Ejecución (RLE). Seguidamente, se compararán y analizarán los resultados obtenidos con ellos.

Se van a utilizar 4 textos para el estudio. El texto número 1 es la palabra "BANANA". Los textos número 2, 3 y 4 son extraídos del libro *The Silmarillion* de J.R.R. Tolkien [18] de la versión en inglés, pero con los primeros 6.432, 26.394 y 39.004 caracteres respectivamente.

		Texto 1	Texto 2	Texto 3	Texto 4
	Longitud original	48	51.456	211.152	312.032
	Símbolos únicos	3	63	72	75
	Entropía máxima	1,585	5,9773	6,1699	6,2288
	Entropía	1,4591	4,3412	4,3119	4,3237
Compresión	Longitud LZ77	56	99.944	450.224	701.070
	Tasa LZ77	0,85714	0,51485	0,46899	0,44508
	Longitud LZ78	63	35.481	128.255	189.208
	Tasa LZ78	0,7619	1,4502	1,6463	1,6491
	Longitud BWT+RLE	38	54.113	218.849	320.645
	Tasa BWT+RLE	1,2632	0,9509	0,96483	0,97314
Compresibilidad	Nº n-gramas únicos	3	1.601	2.592	2.984
	Entropía n-gramas	1,5	9,6285	9,7012	9,7566
	Entropía máx. n-gramas	1,585	10,6448	11,3399	11,543
	Nº palabras	1	1.127	4.862	7.269
	Nº palabras únicas	1	471	1.134	1.502
	Tamaño de vocabulario	1	0,41792	0,23324	0,20663
	Longitud promedio	6	4,6087	4,3283	4,2585
	Tasa de repetición	0,3	0,20006	0,20002	0,20001

Tabla 1: Comparación texto

4.1.1. Compresión de textos

Se analizaron los resultados obtenidos de cada una de las compresiones ejecutadas como se puede visualizar en la tabla 1 y se puede extraer varias reflexiones.

Primero, se cree importante la relación entre la entropía del texto y la entropía máxima del mismo. Esto nos genera una orientación del nivel de compresibilidad que puede tener el texto, pero no es una medida definitiva. Podría ser interesante para comparar textos con el mismo número de símbolos únicos como son el texto 3 y 4. Se ve como el texto 4, que tiene un mayor número de símbolos, tiene mayor entropía. También es de observar, que la entropía y las tasas de compresión son muy parecidas independientemente que el texto 3 tenga un 32 % menos caracteres que el texto 4.

Seguidamente, hay una diferencia sustancial entre cadenas de texto cortas (como el primer texto) y cadenas más extensas. En el texto corto, el método con mejor tasa de compresión es el BWT+RLE, mientras en los textos con más caracteres la mejor tasa hace referencia al método LZ78. Esto sucede debido a que el método LZ78 puede acabar almacenando palabras que son repetidas en múltiples ocasiones, como la palabra "the", mientras el algoritmo de ordenación BWT no tiene porque conseguir que todas las letras iguales acaben juntas (que seria el caso ideal).

Otra observación es que el método LZ77 no genera una compresión de tamaño, es debido a que este método necesita repeticiones bastante cercanas entre sí. El método LZ77 depende directamente del tamaño de la ventana de trabajo, es decir, si implementamos una venta de 30 caracteres, si en el texto 1 se repite una secuencia del principio al final, no la apreciará y codificará como si fuese nueva. Si se analiza el caso del tamaño máximo de ventana (igual al tamaño del texto) para una cadena muy grande como es el texto 4, y existiera una repetición que implicase moverse 200.000 caracteres hacia atrás. Todos los valores deben ser codificados según el tamaño del número, es este caso con 18 bits por cada número. Esto implica que la codificación aumenta de tamaño de forma sustancial. La idealidad es encontrar un equilibrio ente el tamaño de la ventana y el tamaño del texto, para conseguir incluir el máximo número de repeticiones sin que aumente excesivamente el código comprimido.

Se puede apreciar también que las tasas en el caso de los textos más extensos son relativamente constantes y en el caso particular del método LZ78 va en aumento cuanto más caracteres hay.

4.1.2. Compresibilidad de textos

Se analizaron varias métricas de compresibilidad con el objetivo de evaluar su efectividad para predecir la compresibilidad de un texto antes de aplicar algoritmos de compresión, como se puede ver en la tabla 1.

Los resultados mostraron que el modelo de n-gramas con entropía disminuye significativamente al pasar de textos cortos a otros más largos, indicando que los textos más largos presentan patrones más repetitivos y, por lo tanto, son más predecibles.

En relación a los textos 2, 3 y 4, la compresibilidad del modelo de n-gramas con entropía disminuye ligeramente a medida que aumenta el tamaño del texto, por el mismo principio que se mencionó anteriormente.

Al aumentar el tamaño del texto se aprecia como el tamaño del vocabulario va disminuyendo, indicando que los textos más largos tienden a reutilizar un conjunto más reducido de símbolos, lo cual es un buen indicador de compresibilidad, ya que un vocabulario más pequeño en relación con

la longitud total sugiere más repetición.

La longitud promedio de las secuencias de símbolos disminuyó ligeramente con el aumento del tamaño del texto, sugiriendo que los textos más largos tienen secuencias repetitivas más cortas en promedio. Sin embargo, la tasa de repetición mostró una disminución significativa del texto 1 al resto de textos, pero esta disminución fue mínima entre los otros, indicando que la repetibilidad de las secuencias no varía mucho con el tamaño del texto en este caso.

En resumen, las métricas de entropía y tamaño de vocabulario demostraron ser las más efectivas para predecir la compresibilidad de un texto antes de la compresión. Ambas métricas disminuyen con el aumento del tamaño del texto, indicando mayor repetitividad y previsibilidad, lo cual es crucial para una buena compresibilidad. Por otro lado, la longitud promedio y la tasa de repetición mostraron menos variabilidad con el aumento del tamaño del texto, lo que las hace menos útiles como métricas individuales para predecir la compresibilidad.

4.1.3. Compresión vs compresibilidad de textos

Al comparar las métricas de compresibilidad con los resultados de la compresión, se encontró que la compresibilidad del modelo de n-gramas con entropía más baja se correlaciona con mejores tasas de compresión en los algoritmos LZ77 y BWT+RLE. Pasa algo similar cuanto menor sea el tamaño del vocabulario.

La longitud promedio de las secuencias no mostró una correlación clara con las tasas de compresión, aunque los textos más largos tenían longitudes promedio de secuencias más cortas. La tasa de repetición presentó una correlación moderada con la compresibilidad, siendo mayor en textos muy cortos y disminuyendo para los textos más largos. Sin embargo, esta disminución fue mínima entre los diferentes tamaños de “El Silmarillion” (textos 2, 3 y 4). Por otro lado, LZ78 mostró resultados esperanzadores, con tasas de compresión mejores para los textos más largos, sugiriendo que este algoritmo puede ser más efectivo para textos extensos en comparación con LZ77 y BWT+RLE.

4.2. Imágenes

La idea de este apartado es aplicar los métodos de compresión de textos a imágenes. Para ello, primeramente, se han de convertir las imágenes en textos. Analizamos cada píxel según el sistema RGB, donde acabamos extrayendo tres matrices de datos con el tamaño de la foto. Se ha decidido que cada píxel tenga números de tres caracteres para que sean todos iguales. Como los valores pueden ir desde el 0 hasta el 255 necesitamos 8 bits para codificar cada número.

Para ordenar las matrices en una cadena de texto se han escogido tres métodos diferentes para recorrer la matriz: linealmente por filas (de izquierda a derecha), linealmente por columnas (de arriba a abajo) y según la curva de Morton (o curva Z), como se puede ver en la figura 2. En todos los métodos iniciamos en la esquina superior izquierda y los valores de los píxeles serán sucesivos sin espacios u otros símbolos como separación.

En la figura 3 se pueden ver las imágenes sintéticas escogidas para hacer el análisis. Son parecidas entre ellas por

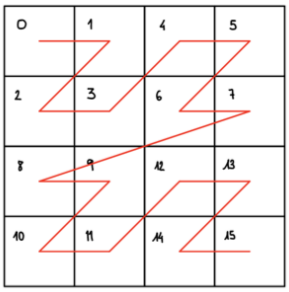
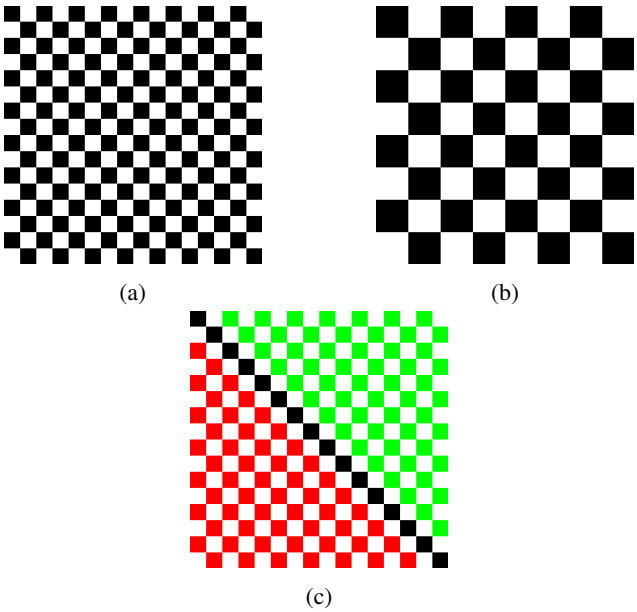


Fig. 2: Escaneo Z o Morton para imágenes 2D.



(a) (b)

(c)

Fig. 3: Imágenes sintéticas de 16x16 píxeles

forma, colores y distribución; con esto se cree que puede ser de ayuda a la hora de comparar resultados y ver cuál de ellas puede tener una compresibilidad mayor. Como se puede observar en la tabla 2 de la página 8, las tres imágenes tienen el mismo número de símbolos únicos, longitud y entropía máxima.

En el caso de la figura 3a, vemos que los primeros 4 píxeles son negro, blanco, negro, blanco. Sabemos que en el sistema RGB el negro es igual a (0, 0, 0) y el blanco es (255, 255, 255), por lo tanto el texto de la matriz R seria “000255000255”. En este caso, las matrices R, G y B son iguales así que el texto completo quedaría como “000255000255···000255000255···000255000255”.

4.2.1. Compresión de imágenes

Como se puede apreciar en la tabla 2 (en la página 8), las observaciones podrían ser similares en las imágenes 3a y 3b. Primeramente, el análisis por filas y por columnas es idéntico ya que acaban generando la misma cadena de texto. En ambas imágenes, se observa que el algoritmo de BWT+RLE genera una compresión considerablemente mayor que los otros dos métodos. En la imagen 3a, se puede apreciar que en ambos métodos de Lempel-Ziv, no existe una gran diferencia en cuanto a la forma en que se recorre la figura para almacenar los símbolos. Sin embargo, en la figura 3b, donde se esperaba que, debido a la distribución

		Imagen 3a			Imagen 3b			Imagen 3c		
		Filas	Columnas	Morton	Filas	Columnas	Morton	Filas	Columnas	Morton
	Longitud Original	18.432			18.432			18.432		
	Símbolos Únicos	2			2			2		
	Entropía Máxima	1			1			1		
	Entropía	1			1			0,93773		
Compresión	Longitud LZ77	2252	2.252	2.208	2.252	2.252	3.584	4.352	4.424	3.824
	Tasa LZ77	2,7282	2,7282	2,7826	2,7282	2,7282	1,7142	1,4117	1,3887	1,6066
	Longitud LZ78	2.968	2.968	3.352	3.480	3.480	3.448	3.480	3.464	3.704
	Tasa LZ78	2,0700	2,0700	1,8329	1,7655	1,7655	1,7819	1,7655	1,7726	1,6587
	Longitud BWT+RLE	188	188	116	260	260	260	1.232	1.268	1.062
	Tasa BWT+RLE	32,6808	32,6808	52,9655	23,6307	23,6307	23,6307	4,9870	4,8454	5,7853
Compresibilidad	Nº n-gramas únicos	8	8	8	8	8	8	8	8	8
	Entropía n-gramas	2,6949	2,6949	2,7526	2,7302	2,7302	1,4287	2,9516	2,9516	2,8089
	Entropía máx. n-gramas	3	3	3	3	3	3	3	3	3
	Nº palabras	768	768	768	768	768	768	768	768	768
	Nº palabras únicas	2	2	2	2	2	2	2	2	2
	Tamaño de vocabulario	0,0026	0,0026	0,0026	0,0026	0,0026	0,0026	0,0026	0,0026	0,0026
	Longitud promedio	3	3	3	3	3	3	3	3	3
	Tasa de repetición	0,20013	0,20013	0,20013	0,20013	0,20013	0,20013	0,20013	0,20013	0,20013

Tabla 2: Comparación de imágenes sintéticas de la figura 3[1].

de los píxeles, recorrer la imagen según la curva de Morton resultara en una mayor compresión, esto no ocurre. La tasa de compresión es igual para el LZ78 e incluso inferior para el LZ77.

En la figura 3c se han añadido algunos colores para no tener imágenes simétricas, buscando cuánto implican estas en la compresibilidad de la imagen. La diferencia sustancial que muestra la tabla 2 con la imagen 3c respecto al resto de las imágenes de la figura 3 es que ahora las tasas de compresión se han reducido. También se ve como el recorrer la imagen en curva Z puede generar mejor compresión en algún caso.

(a) (b)

Fig. 4: Imágenes naturales de 64x64 píxeles

Se ha realizado un nuevo estudio cambiando el tipo de imágenes para representar paisajes reales como las de la figura 4. Ambas imágenes tiene el mismo tamaño y aproximadamente la misma entropía máxima y número de símbolos únicos. Cierto es que la figura 4a tiene ligeramente inferior la entropía. Se puede apreciar en la tabla 3 como recorrer la imagen según la curva de Morton obtiene mejores resultados, por lo general, aunque no es tan acentuado como se esperaba. Es posible que se deba a que los píxeles próximos tiene propiedades similares entre ellos.

Otra observación relevante es que ningún método en este caso que logra una tasa de compresión superior a 1.

Por último, se han comprimido las misma imágenes con algunos métodos específicos para imágenes como JPEG, PNG y TIFF y en la tabla 4 se pueden ver los resultados obtenidos. Estos han sido comprados con los resultados del resto de tablas anteriores y se ha llegado a ciertas observaciones.

La primera observación resalta que todas las tasas de compresión mencionadas anteriormente son notablemente inferiores en comparación con las tasas proporcionadas por el algoritmo *Portable Network Graphics (PNG)*. En particular, con la imagen 3b y el método BWT+RLE, podría argumentarse un enfoque optimista al señalar cierta similitud.

Al comparar con el algoritmo *Joint Photographic Experts Group (JPEG)*, se pueden hacer dos observaciones adicionales. Primero, las imágenes naturales de la figura 4 tienen tasas de compresión más bajas que las obtenidas con el algoritmo PNG para las mismas fotos. Por otro lado, para las imágenes sintéticas de la figura 3, las tasas de compresión son más altas, llegando a duplicar las proporcionadas por el PNG. Cabe señalar que el algoritmo JPEG es con pérdidas, a diferencia del resto de algoritmos implementados en este estudio, que son sin pérdidas.

Al comparar las tasas obtenidas con las del algoritmo *Graphics Interchange Format (GIF)*, se evidencia que todos los métodos sin pérdidas tienen peores tasas que la de GIF.

Finalmente, al contrastar los datos con el algoritmo *Tagged Image File Format (TIFF)*, se observa que TIFF tiene las peores tasas de los métodos de compresión sin pérdidas. Específicamente para el método BWT+RLE aplicado a las imágenes sintéticas, las tasas de compresión pueden llegar a ser entre 15 y 35 veces superiores en comparación con el TIFF.

4.2.2. Compresibilidad de imágenes

Se han analizado diversas métricas de compresibilidad para evaluar su eficacia en la predicción de la compresibilidad de textos derivados de imágenes antes de ser sometidos a procesos de compresión.

La entropía de n-gramas mide la aleatoriedad y la complejidad del texto. Una mayor entropía generalmente sugiere menor compresibilidad, ya que indica un mayor grado de aleatoriedad y menor repetición de patrones.

El tamaño del vocabulario es un indicador crucial de la

[1]Para entropía de n-gramas, longitud y tasa de repetición se esperan valores altos, y para tamaño de vocabulario, valores bajos.

		Imagen 4a			Imagen 4b		
		Filas	Columnas	Morton	Filas	Columnas	Morton
	Longitud Original	294.912			294.912		
	Símbolos Únicos	251			256		
	Entropía Máxima	7,9715			8		
	Entropía	7,4445			7,7148		
Compresión	Longitud LZ77	399.788	453.176	392.552	442.088	494.144	431.900
	Tasa LZ77	0,2458	0,2169	0,2504	0,2223	0,1989	0,2276
	Longitud LZ78	148.646	153.350	151.565	156.017	160.343	159.293
	Tasa LZ78	0,6613	0,6410	0,6485	0,6301	0,6130	0,6171
	Longitud BWT+RLE	309.935	347.998	333.667	336.344	380.026	382.525
	Tasa BWT+RLE	0,3171	0,2924	0,2946	0,2933	0,2586	0,2569
Compresibilidad	Nº n-gramas únicos	583	592	592	597	597	597
	Entropía n-gramas	8,8139	8,8254	8,8211	8,9462	8,953	8,9551
	Entropía máx. n-gramas	9,1874	9,2095	9,2095	9,2216	9,2216	9,2216
	Nº palabras	12288	12288	12288	12288	12288	12288
	Nº palabras únicas	251	251	251	256	256	256
	Tamaño de vocabulario	0,0204	0,0204	0,0204	0,0208	0,0208	0,0208
	Longitud promedio	3	3	3	3	3	3
	Tasa de repetición	0,20001	0,20001	0,20001	0,20001	0,20001	0,20001

Tabla 3: Comparación de imágenes naturales de la figura 4.

	Imagen 3a	Imagen 3b	Imagen 3c	Imagen 4a	Imagen 4b
Longitud Original	18432	18432	18432	294912	294912
Símbolos Únicos	2	2	2	251	256
Entropía Máxima	1	1	1	7,9715	8
Entropía	1	1	0,93773	7,4445	7,748
Longitud JPEG	4114	4127	4207	3475	4289
Tasa JPEG	4,4803	4,4662	4,3813	84,8668	68,7601
Longitud PNG	104	1091	1138	8777	9834
Tasa PNG	177,2308	16,8946	16,1968	33,6005	29,989
Longitud GIF	93	98	125	1942	4932
Tasa GIF	198,1935	188,0816	147,456	59,6746	59,7956
Longitud TIFF	2402	2402	2402	12644	12644
Tasa TIFF	7,6736	7,6736	7,6736	23,3243	23,3243

Tabla 4: Comparación métodos específicos para imágenes

diversidad de caracteres en el texto. Se obtiene un valor muy pequeño porque vemos que solo hay 2 palabras únicas en las imágenes sintéticas (tabla 2) mientras en las imágenes naturales (tabla 3) el valor aumenta unas 10 veces.

La longitud promedio del texto es otra métrica importante para la compresibilidad, pero en este caso no es útil. Esto es debido a que el valor obtenido siempre es 3 debido a la transformación de imagen a texto.

La tasa de repetición indica la frecuencia de aparición de patrones repetidos en el texto. Una mayor tasa de repetición sugiere una mayor compresibilidad. Sin embargo, en este estudio, la tasa de repetición se mantuvo constante en aproximadamente 0.20013 para todas las imágenes, lo que sugiere que esta métrica, aunque relevante, puede no ser el único factor determinante de la compresibilidad en contextos con estructuras de texto similares.

4.2.3. Compresión vs compresibilidad de imágenes

En el análisis de compresión, los textos con mayor entropía presentaron tasas de compresión menos eficientes. Esto sugiere que la entropía de n-gramas es un buen indicador de compresibilidad.

El tamaño del vocabulario en las imágenes no genera información de interés. Lo mismo pasa con la longitud promedio.

La tasa de repetición constante sugiere que esta métrica sola no es suficiente para predecir la compresibilidad de manera efectiva. No se observaron variaciones significativas en las tasas de compresión relacionadas con esta métrica.

5 Conclusiones

En este estudio, se analizaron diversas técnicas y métricas de compresión de datos aplicadas a textos e imágenes. Las observaciones y resultados obtenidos permiten extraer varias conclusiones importantes.

Primero, la relación entre la entropía del texto y la entropía máxima es un buen indicador de la compresibilidad. Textos con menor entropía respecto a la máxima tienden a ser más predecibles y, por ende, más comprimibles.

Segundo, la eficiencia de los métodos de compresión varía según la longitud del texto. Para textos cortos, la combinación BWT+RLE demostró ser más efectiva, mientras que para textos largos, LZ78 mostró mejor rendimiento debido a su capacidad para almacenar patrones repetitivos.

Tercero, los resultados muestran que, aunque la tasa de compresión de métodos como LZ77 no siempre es favorable, la elección del algoritmo debe considerar la estructura del texto y la longitud de la ventana de trabajo para optimizar los resultados.

Cuarto, al aplicar métodos de compresión de textos a imágenes, se encontró que la curva de Morton puede mejorar la compresibilidad en ciertos casos, aunque los métodos específicos de compresión de imágenes, como PNG y JPEG, ofrecen mejores tasas de compresión debido a su diseño optimizado para datos visuales. Sin dejar de tener en perspectiva que el método JPEG es un método con pérdidas.

Finalmente, las métricas de compresibilidad, como la entropía de n-gramas, el tamaño del vocabulario y la longitud promedio, se correlacionan bien con los resultados de compresión, sugiriendo que son buenos indicadores para predecir la compresibilidad de un texto antes de aplicar algoritmos de compresión. Sin embargo, es crucial utilizar un enfoque combinado que considere múltiples métricas para una evaluación precisa.

Este estudio confirma que la evaluación previa de la compresibilidad puede mejorar significativamente la eficiencia de los procesos de compresión, permitiendo optimizar la gestión de datos textuales derivados de imágenes.

REFERENCIAS

[1] M. Burrows & D. Wheeler. "A block sorting lossless data compression algorithm", *DIGITAL Systems Research Center*, 1994.

[2] N. Chomsky, "Three Models for the Description of Language.", *I.R.E. transactions on information theory 2.3*, 1956, 113–124.

[3] J. Du, L. Cui, J. Zhang, J. Li, & J. Huang, "The method of quantitative trend diagnosis of rolling bearing fault based on protrugram and lempel–ziv", *Shock and Vibration*, vol. 2018, p. 1-8, 2018. https://doi.org/10.1155/2018/4303109

[4] J. E. Hopcroft, J. D. Ullman, "Introduction to Automata Theory, Languages, and Computation" *Addison-Wesley*, 1979.

[5] C. Huyen. "Evaluation Metrics for Language Modeling", *The Gradient*, 10 de octubre de 2019, `https://thegradient.pub/understanding-evaluation-metrics-for-language-models/`

[6] International Organization for Standardization (ISO) & International Electrotechnical Commission(IEC) "ISO/IEC 10918-1:1994 - Information technology – Digital compression and coding of continuous-tone still images – Requirements and guidelines", *International Organization for Standardization*, 10918-1, 1994.

[7] International Organization for Standardization (ISO). "ISO 12234-2:1998 – Photography – Electronic still-picture imaging – Removable memory – Part 2: Image data format". *International Organization for Standardization*, 12234-2, 1998.

[8] International Organization for Standardization (ISO). "ISO 15948:2004 - Graphic technology – Symbols for content management using XML – Elements of image and symbol markup", *International Organization for Standardization*, 15948, 2004.

[9] A. Lempel & J. Ziv, "On the Complexity of Finite Sequences", *IEEE Transactions on Information Theory*, vol. 22, págs. 75-81, enero 1976, ISSN: 15579654. DOI: 10.1109/TIT.1976.1055501.

[10] T. Lin, D. Lee, Y. Hsu, & K. Kuo, "Damage detection of regular civil buildings using modified multi-scale symbolic dynamic entropy", *Entropy*, vol. 24, no. 7, p. 987, 2022. https://doi.org/10.3390/e24070987

[11] S. Madala. "Evaluationg Languuage Models in NLP", *Scaler*, 13 de enero de 2023, `https://www.scaler.com/topics/nlp/language-models-in-nlp/`

[12] K. Nowakowski, M. Ptaszynski & F. Masui, "MiNgMatch—A Fast N-gram Model for Word Segmentation of the Ainu Language," *Information*, vol. 10, no. 317, pp. 1-19, 2019.

[13] P. Pedram, "The minimal length and the shannon entropic uncertainty relation", *Advances in High Energy Physics*, vol. 2016, p. 1-8, 2016. https://doi.org/10.1155/2016/5101389x

[14] Rekha, & V. Kumar, "Study of quantile-based cumulative renyi information measure", *Statistics Optimization & Information Computing*, vol. 9, no. 4, p. 886-909, 2020. https://doi.org/10.19139/soic-2310-5070-1034

[15] D. Salomon, & D. Herrera Pérez (trad.), "Compresión de datos: La referencia completa", *Springer-Verlag London Limited*, 4ta edición, Marzo 2014, ISBN: 978-1-84628-602-5.

[16] M. Sohail Khan, M. Asghar Khan, & M. Mansoor Alam et al. "Impact of Big Data over Telecom Industry", *Pakistan Journal of Engineering, Technology & Science*, 2018.

[17] M. Tabass, G. Borzadaran, & M. Amini, "Renyi entropy in continuous case is not the limit of discrete case", *Mathematical Sciences and Applications E-Notes*, vol. 4, no. 1, p. 113-117, 2016. https://doi.org/10.36753/mathenot.421418

[18] J.R.R. Tolkien. "The Silmarillion", *Houghton Mifflin*, 1977

[19] J. Ziv & A. Lempel. "A Universal Algorithm for Sequential Data Compression", *IEEE Transactions on Information Theory*, vol. 23, no. 3, 1977, pp. 337–43.

[20] J. Ziv & A. Lempel. "Compression of Individual Sequences via Variable-Rate Coding", *IEEE Transactions on Information Theory*, vol. 24, no. 5, 1978, pp. 530–36.

A APÉNDICE

En este apéndice se presentan una serie de ejemplos prácticos para clarificar los métodos y métricas discutidos en el cuerpo principal del documento. A través de estos ejemplos, pretendemos ofrecer una comprensión más profunda y concreta de los conceptos teóricos.

Todos los ejemplos han sido realizados con el texto número 1 del documento que se refiere a la palabra 'BANANA', a excepción de dos métodos (tamaño del vocabulario y longitud promedio) ya que estos requieren de más de una palabra para ser efectivos.

A.1. Ejemplo LZ77

El método de compresión de Lempel-Ziv de 1977 utiliza una codificación resultante como:

$$(d_1, l_1, c_1)(d_2, l_2, c_2)...(d_n, l_n, c_n)$$

donde *d* se refiere a desplazamiento desde la posición actual, *l* a la longitud de la coincidencia y *c* el próximo carácter a insertar.

Al realizar la compresión de la palabra 'BANANA', nos queda una codificación:

$$(0, 0, B)(0, 0, A)(0, 0, N)(2, 2, A)$$

A.2. Ejemplo LZ78

El método de compresión de Lempel-Ziv de 1978 utiliza una codificación resultante como:

$$(i_1, c_1)(i_2, c_2)...(i_n, c_n)$$

donde *i* hace referencia al valor del diccionario y *c* el próximo carácter a insertar.

Al realizar la compresión de la palabra 'BANANA', nos queda una codificación:

$$(0, B)(0, A)(0, N)(2, N)(2,)$$

A.3. Ejemplo BWT + RLE

En este método de compresión, primero hemos de ordenar la palabra según el algoritmo BWT para luego poder aplicar la compresión en RLE.

Permutaciones Cíclicas	Permutaciones Ordenadas	Última Columna
BANANA	ABANAN	N
ANANAB	ANABAN	N
NANABA	ANANAB	B
ANABAN	BANANA	A
NABANA	NABANA	A
ABANAN	NANABA	A

Tabla 5: Pasos de la BWT para "BANANA"

Para la cadena 'BANANA', la BWT reorganiza las letras basándose en todas las permutaciones cíclicas de la cadena y su ordenamiento lexicográfico, resultando en 'NNBAAA'. El proceso se puede visualizar en la tabla 5.

Una vez ordenado el texto, aplicamos RLE al resultado 'NNBAA' obteniendo '2N1B3A'.

A.4. Ejemplo entropía

Se calculará la entropía siguiendo la ecuación 1, para ello hemos de obtener primero las probabilidades de cada carácter:

$$p(A) = 3/6 = 1/2$$
$$p(B) = 1/6$$
$$p(N) = 2/6 = 1/3$$

$$H(X) = -p(A)\log_2(p(A)) - p(B)\log_2(p(B)) -$$
$$-p(C)\log_2(p(C)) =$$
$$= -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{6}\log_2(\frac{1}{6}) - \frac{1}{3}\log_2(\frac{1}{3}) =$$
$$= 1,4591 \quad bits$$

A.5. Ejemplo n-gramas

Para el modelo de lenguaje de n-gramas, escogemos una $n = 3$ como en el estudio.

Principalmente, lo que realiza esta métrica es dividir el texto en variables de longitud *n* y calcula su entropía.

B	**A**	**N**	**A**	**N**	**A**
B	A	N			
	A	N	A		
		N	A	N	
			A	N	A

Para la palabra 'BANANA' se obtienen 4 n-gramas de los cuales solo 3 son únicos, con diferentes probabilidades.

$$[`BAN`] = 1 \rightarrow p(BAN) = 1/4$$
$$[`ANA`] = 2 \rightarrow p(ANA) = 1/2$$
$$[`NAN`] = 1 \rightarrow p(NAN) = 1/4$$

Si calculamos la entropía siguiendo la ecuación 1, obtenemos una entropía de 1,5 bits.

$$H(X) = -p(BAN)\log_2(p(BAN)) -$$
$$-p(ANA)\log_2(p(ANA)) - p(NAN)\log_2(p(NAN)) =$$
$$= -\frac{1}{2}\log_2(\frac{1}{2}) - 2 \cdot \frac{1}{4}\log_2(\frac{1}{4}) = 1,5 \quad bits$$

A.6. Ejemplo tamaño del vocabulario

Al ejemplificar la métrica de tamaño del vocabulario con el texto 'BANANA' no obtenemos resultados validos, ya que la solución es 1.

Hay que tener en cuenta que la operación que realizamos en la división de palabras únicas entre palabras totales.

Si utilizamos otro ejemplo: 'Este documento estudia la compresibilidad', vemos que el tamaño del vocabulario es 1 también, ya que no se repite ninguna palabra. El tamaño del vocabulario es el resultado de las palabras únicas entre las palabras totales.

Al analizar otro ejemplo como: 'El sol es quien calienta el planeta', vemos como en este caso el valor resultante es 0,8571.

A.7. Ejemplo longitud promedio

La longitud promedio de las palabras del texto 'BANANA' es 1, ya que solo hay una palabra. Si calculamos la longitud promedio de otro texto como: '*Este documento estudia la compresibilidad*', la longitud promedio de palabras es 7,4. Esto es debido al numero de caracteres por palabra: *Este (4), documento (9), estudia (7), la (2), compresibilidad (15)*. El resultado se obtiene como la suma total de caracteres (37) entre la longitud de palabras (5).

A.8. Ejemplo tasa de repetición

En la métrica de tasa de repetición, se miran todas las posibles combinaciones con los caracteres que tiene el texto y su longitud máxima. Para ello podemos hacer un proceso similar al del modelo de n-gramas pero comenzando con $n = 1$ y acabando en $n = longitud$.

En el caso del texto 'BANANA', $n = [1, 6]$.

- Si $n = 1$:
  - secuencias = 'b', 'a', 'n', 'a', 'n', 'a'
  - contador = {'b': 1, 'a': 3, 'n': 2}
- Si $n = 2$:
  - secuencias = 'ba', 'an', 'na', 'an', 'na'
  - contador = {'ba': 1, 'an': 2, 'na': 2}
- Si $n = 3$:
  - secuencias = 'ban', 'ana', 'nan', 'ana'
  - contador = {'ban': 1, 'ana': 2, 'nan': 1}
- Si $n = 4$:
  - secuencias = 'bana', 'anan', 'nana'
  - contador = {'bana': 1, 'anan': 1, 'nana': 1}
- Si $n = 5$:
  - secuencias = 'banan', 'anana'
  - contador = {'banan': 1, 'anana': 1}
- Si $n = 6$:
  - secuencias = 'banana'
  - contador = { 'banana': 1}

El contador es acumulativo, por lo cual quedaria como: contador = {'b': 1, 'a': 3, 'n': 2, 'ba': 1, 'an': 2, 'na': 2, 'ban': 1, 'ana': 2, 'nan': 1, 'bana': 1, 'anan': 1, 'nana': 1, 'banan': 1, 'anana': 1, 'banana': 1}.

Para calcular la tasa se calcula como la longitud total de caracteres del texto dividido entre suma de todas las secuencias del diccionario contador, así queda normalizado a 1. En el caso del texto 'BANANA', $6/21 = 0,2857$.