Entropía, Redundancia y Densidad Léxica

La riqueza y complejidad léxica de un texto son medidas importantes que pueden ayudar a los profesionales de la educación a adaptar y personalizar contenidos según las necesidades de los estudiantes. En este artículo, discutiremos las fórmulas utilizadas en nuestra herramienta de análisis de textos para calcular la entropía, redundancia y densidad léxica, así como su interpretación.

En este artículo presentamos la justificación de la **herramienta**:

Análisis de la riqueza y complejidad léxica de los recursos de texto: Calculadora de Entropía, Redundancia y Densidad del Léxico

Entropía (H)

La entropía es un concepto de la teoría de la información, que fue introducido por Claude Shannon en su artículo de 1948, «Una teoría matemática de la comunicación«. La entropía de Shannon es una medida cuantitativa de la incertidumbre o aleatoriedad en un conjunto de datos. En términos simples, describe la cantidad de información promedio que se necesita para identificar un resultado en un conjunto de posibles resultados.

En el contexto de la teoría de la información, la entropía se utiliza ampliamente para analizar y evaluar sistemas de comunicación, compresión de datos y criptografía. La entropía de Shannon se basa en la probabilidad de los diferentes símbolos (por ejemplo, palabras) en un mensaje o conjunto de datos. Cuanto más impredecible es la aparición de un símbolo, mayor es la entropía.

La fórmula para calcular la entropía es:

$$H = -\sum P_i \log_2(P_i)$$

Donde P_i es la probabilidad de cada palabra en el texto. Para calcular P_i, se divide la frecuencia de cada palabra por el número total de palabras en el texto. La unidad de la entropía es el bit.

La entropía, en este contexto, mide la incertidumbre en el uso de palabras, reflejando la **complejidad y riqueza del lenguaje** en un texto. Valores altos de entropía indican mayor diversidad léxica y menor previsibilidad en las palabras, lo que sugiere un contenido más complejo y rico. Valores bajos señalan un contenido más simple y limitado, con mayor repetición y menor variabilidad en el vocabulario.

Redundancia (R)

La redundancia es una medida de la repetitividad en un texto. Cuanto mayor sea la redundancia, más repetitivas y menos informativas serán las palabras. La fórmula para calcular la redundancia es:

$$R = H_{max} - H$$

La entropía máxima (H_{max}) se calcula utilizando la fórmula:

$$H_{max} = \log_2(N)$$

Donde N es el número de palabras únicas en el texto. La unidad de R es el bit.

Densidad Léxica (DL)

La densidad léxica es una medida de la diversidad de palabras en un texto y también recibe el nombre de TTR (*Type Token Ratio*). Se calcula dividiendo el número de palabras distintas por el número total de palabras en el texto:

$$DL = \frac{Número\ de\ palabras\ distintas}{Número\ total\ de\ palabras}$$

No tiene unidades, ya que se trata de una proporción.

Densidad Léxica Estandarizada (DLE)

El método utilizado para calcular la Densidad Léxica Estandarizada proviene de la rarefacción, que es un método estadístico que se origina en la biología, donde se utiliza para evaluar la diversidad de especies en un ecosistema. En el estudio de la

biodiversidad, la rarefacción es un método que permite comparar la riqueza de especies entre diferentes hábitats o comunidades, normalizando el tamaño de las muestras. De esta manera, es posible comparar la diversidad de especies en diferentes entornos sin que los resultados se vean afectados por el tamaño de la muestra.

En el análisis del lenguaje, la Densidad Léxica Estandarizada es una medida que hemos utilizado para **evaluar la riqueza y diversidad léxica en un texto**. Al igual que en la biología, el objetivo es obtener una medida de la diversidad léxica que sea comparable entre diferentes textos, **independientemente de la cantidad de palabras en cada uno de ellos.** Hemos usado la DLE como un método alternativo para hallar la densidad léxica.

La densidad léxica estandarizada es una medida que tiene en cuenta el tamaño del texto y proporciona una estimación de la densidad (diversidad) de palabras en una muestra de tamaño fijo.

El método para hallar la densidad léxica estandarizada se lleva a cabo de la siguiente manera:

1. Se realizan 1000 muestreos, cada uno con 100 palabras seleccionadas al azar. Cada uno de estos 1000 muestreos es independiente de los demás, lo que significa que algunas palabras podrían aparecer en diferentes muestreos. No obstante, dentro de cada muestra individual de 100 palabras, no hay repeticiones.
2. Para calcular la Densidad Léxica Estandarizada, se divide el número de palabras diferentes (únicas) en cada muestra por el total de palabras en esa muestra, que en este caso es 100. Al hacer esto, se obtiene un porcentaje que representa la diversidad léxica en una muestra específica.
3. Después de calcular la DLE para cada una de las 1000 muestras, se hace la media de todas las DLE individuales. Este promedio representa la densidad léxica general en el texto, considerando la variación entre las diferentes muestras. Esta medida proporciona una estimación más precisa de cuán variado y enriquecido es el lenguaje utilizado en el texto completo.

La fórmula utilizada para calcular la densidad léxica estandarizada es la siguiente:

$$DLE = \frac{\sum_{i=1}^{1000} \frac{\text{Palabras Distintas en Muestra}_i}{100}}{1000}$$

Donde: $\text{Palabra Distinta de Muestra}_i$ es el número de palabras distintas en cada muestra que está formada por 100 palabras

Este proceso permite estandarizar la diversidad léxica entre textos de diferentes longitudes y proporciona una **medida más robusta** de la densidad.

Dado que se toman muestras de 100 palabras, cuando el texto introducido tiene 100 o menos palabras, la DLE coincide con la densidad léxica. No tiene unidades, ya que se trata de una proporción.

Interpretación de los resultados

Los rangos seleccionados en la tabla de interpretación de entropía, densidad léxica estandarizada y redundancia están basados en observaciones generales y patrones identificados en diferentes tipos de textos y niveles educativos. Estos rangos tienen como objetivo proporcionar un marco de referencia para la interpretación de los resultados obtenidos al analizar un texto.

Para la entropía y la densidad léxica estandarizada, los rangos se han establecido teniendo en cuenta la complejidad del contenido y el vocabulario utilizado en textos de nivel primaria, secundaria y bachillerato o superiores. Estos rangos pueden variar según el tema y el estilo de escritura de cada autor, pero proporcionan una referencia aproximada para evaluar la complejidad y diversidad del contenido en función del nivel educativo.

En cuanto a la redundancia, los rangos se han determinado para ayudar a identificar el grado en el cual un texto presenta información nueva o conocimientos adicionales. Un porcentaje más bajo de redundancia indica una mayor cantidad de información nueva, mientras que un porcentaje más alto sugiere una mayor repetición de conceptos y menor cantidad de información nueva.

Estos rangos no son absolutos y pueden variar según el tema y el estilo de escritura de cada persona, así como el tipo de texto (por ejemplo, literatura, ensayos

científicos, textos divulgativos, etc.). Además, proporcionan una referencia aproximada para la interpretación de las métricas y no deben ser considerados como límites estrictos.

Interpretación de la entropía y la densidad léxica estandarizada según niveles educativos

Nivel educativo	Entropía (bits/palabra)	Descripción Entropía	Densidad Léxica Estandarizada	Descripción Densidad
Primaria	< 7	Contenido simple y limitado	< 65%	Lenguaje repetitivo o limitado
Secundaria	7 - 8	Contenido moderadamente complejo	65% - 75%	Lenguaje variado y enriquecido
Bachillerato y superiores	> 8	Contenido complejo y diverso	> 75%	Lenguaje rico y sofisticado

Redundancia

Descripción	Redundancia en %	Implicaciones para el análisis del texto
Baja redundancia	< 15%	**Gran cantidad de información nueva y conocimientos adicionales.** El texto presenta una baja redundancia, lo que indica una gran cantidad de información nueva y conocimientos adicionales. Esto puede sugerir que el contenido aborda el tema de manera amplia y diversa, lo que puede resultar en un enfoque más profundo y detallado. Un texto con baja redundancia puede ser más atractivo para el lector y proporcionar una mayor cantidad de información útil.
Redundancia moderada	15% - 35%	**Equilibrio entre información nueva y repetición de conceptos**. El texto presenta un nivel moderado de redundancia, lo que indica un equilibrio entre la introducción de nueva información y la consolidación de ideas clave. Un texto con redundancia moderada puede ser apropiado en situaciones donde se busca enfatizar ciertos conceptos o facilitar la comprensión del lector. Puede ser especialmente útil en textos educativos o de divulgación.
Alta redundancia	> 35%	**Mayor repetición de conceptos y menor cantidad de información nueva.** El texto presenta una alta redundancia, lo que indica una menor cantidad de información nueva y conocimientos adicionales en el contenido. Un texto con alta redundancia puede centrarse en aspectos fundamentales de un tema y reforzar conceptos clave. Sin embargo, la alta redundancia puede hacer que el texto sea menos atractivo para el lector y puede no ser adecuado para el análisis profundo de un tema.

Interpretación de las métricas

Concepto	Definición	Valor bajo	Valor alto
Entropía	Cantidad de información que aporta cada palabra por término medio. Se expresa en bits/palabra	Contenido más simple y homogéneo	Mayor diversidad y complejidad en el contenido
Porcentaje de redundancia	Proporción del texto que es redundante y se repite, por lo que no aporta información nueva	Mayor cantidad de información nueva	Menor cantidad de información nueva
Densidad léxica estandarizada	Proporción de palabras únicas en relación con el total de palabras del texto, expresada en forma de porcentaje, ajustada mediante un método estadístico de muestreos repetidos del texto para que su longitud no influya en el resultado.	Lenguaje repetitivo o limitado	Lenguaje rico y sofisticado
Entropía máxima	Máxima cantidad de información que puede contener el texto considerando todas las posibles combinaciones de palabras	-	-
Densidad del léxico	Proporción de palabras únicas en relación con el total de palabras del texto, expresada en forma de porcentaje	Lenguaje más repetitivo o limitado	Lenguaje más variado y rico

Ejemplos con diferentes fuentes

Libro	Entropía (bits/palabra)	Redundancia en %	Densidad léxica estandarizada
El ingenioso hidalgo don Quijote de la Mancha (Miguel de Cervantes)	9.62	33.75%	74.7%
La Mare Balena (cuento por Caterina Albert / Víctor Català)	9.41	23.96%	76.1%
On the Origin of Species (Ch. Darwin)	9.16	28.62%	76.2%
Los cómics como fuente histórica (trabajo de investigación de bachillerato de J. Zhan)	8.84	17.75%	74.7%
Los tres cerditos (cuento popular)	7.00	10.84%	63.76%
A un olmo seco (Antonio Machado)	6.37	7.19%	56.6%

Fuentes consultadas

- Capsada Blanch, R., & Torruella Casañas, J. (2017). Métodos para medir la riqueza léxica de los textos. Revisión y propuesta. *Verba: Anuario galego de filoloxia*, *44*.
- Kraker-Castañeda, C., & Cóbar-Carranza, A. J. (2011). Uso de rarefacción para la comparación de la riqueza de especies: el caso de las aves de sotobosque en la zona de influencia del Parque Nacional Laguna Lachuá, Guatemala. *Naturaleza y Desarrollo*, *9*(1), 60-70.
- Morales, H. L. (2011). Los índices de riqueza léxica y la enseñanza de lenguas. In *Del texto a la lengua: La aplicación de los textos a la enseñanza-aprendizaje del español L2-LE* (pp. 15-28). Asociación para la Enseñanza del Español como Lengua Extranjera-ASELE.
- OpenAI. (2023). *ChatGPT–4* (versión 23 de marzo).
- Riffo, K. F., Osuna, S. H., & Lagos, P. S. (2019). Descripción de la diversidad y densidad léxicas en noticias escritas por estudiantes de periodismo. *Revista Brasileira de Linguística Aplicada*, *19*, 499-528.