Universidad de La Habana

Facultad: Matemática y Computación

Curso 2024–2025

Probabilidad y Entropía en dos Idiomas

Asignatura: Introducción a la Criptografía

Alumno: Fabio Víctor Alonso Bañobre

Grupo: C-211

Índice

1. Introducción

En este trabajo se analiza la frecuencia de los símbolos del alfabeto en dos idiomas distintos —español e inglés— utilizando textos periodísticos digitalizados de al menos un millón de caracteres cada uno. Posteriormente, se realiza el cálculo de entropía, cantidad de información y análisis de distribución probabilística de los símbolos. Estos conceptos son fundamentales para la criptografía y el tratamiento eficiente de la información.

2. Fundamentos Teóricos

2.1. Probabilidad de los Símbolos

En teoría de la información, la probabilidad $p(x_i)$ de un símbolo x_i representa la frecuencia relativa en la que aparece en un mensaje o en un conjunto de datos.

2.2. Cantidad de Información

La cantidad de información que aporta un símbolo x_i se define como:

$$I(x_i) = -\log_2 p(x_i)$$

2.3. Entropía

La entropía $H(X)$ de una fuente de información es el valor esperado de la cantidad de información:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

2.4. Ley de Zipf

La **ley de Zipf** describe una distribución empírica donde la frecuencia f_r de un símbolo está inversamente relacionada con su rango r en frecuencia:

$$f_r \propto \frac{1}{r^\alpha}$$

donde α es un parámetro cercano a 1 en lenguas naturales. Esta ley refleja que pocas letras son muy frecuentes mientras que muchas otras aparecen esporádicamente. Esta propiedad es común en los lenguajes naturales y refleja una organización no aleatoria.

2.5. Entropía de Dos Símbolos

La **entropía conjunta** para pares de símbolos (bigramas) mide la cantidad total de incertidumbre sobre la aparición de dos símbolos consecutivos:

$$H(X_1, X_2) = -\sum_{i,j} p(x_i, x_j) \log_2 p(x_i, x_j)$$

Cuando $H(X_1, X_2) < 2H(X)$ se evidencia que existe dependencia estadística entre los caracteres.

2.6. Entropía Condicionada

La **entropía condicionada** mide la incertidumbre del símbolo X dado que se conoce el anterior Y:

$$H\big(X|Y\big) = -\sum_{i,j} p(x_i, y_j) \log_2 p(x_i|y_j)$$

También puede calcularse como:

$$H\big(X|Y\big) = H\big(X,Y\big) - H\big(Y\big)$$

cuando se conoce la entropía conjunta. En este trabajo se usa esta segunda forma, asumiendo que $H\big(Y\big) \approx H\big(X\big)$ si las distribuciones son similares.

3. Inciso a: Análisis Probabilístico de los Idiomas

3.1. Metodología

Se seleccionaron dos corpus de texto periodístico:

- Español: artículos del periódico *El País*.
- Inglés: artículos del diario británico *The Guardian*.

Cada corpus tiene más de 1,000,000 de caracteres. Se programó un software en Python para:

- Limpiar los textos (eliminar puntuación, convertir a minúsculas y normalizar acentos: "á"→"a", etc.).
- Calcular la frecuencia de cada símbolo del alfabeto.
- Generar histogramas y calcular las distribuciones.

Se eliminaron signos de puntuación, espacios, números y símbolos no alfabéticos, y se consideró la "ñ" en el alfabeto español. Los datos se normalizaron según el total de letras.

3.2. Distribución Observada

Se observa que ambos idiomas presentan una distribución de tipo Zipf. En español, la letra más frecuente es la "e"; en inglés, también "e", seguida por "t" y "a".

4. Inciso b: Evaluación de la Entropía

4.1. Cantidad de Información Promedio

- Español: $\approx$ 4,18 bits por símbolo.
- Inglés: $\approx$ 4,05 bits por símbolo.

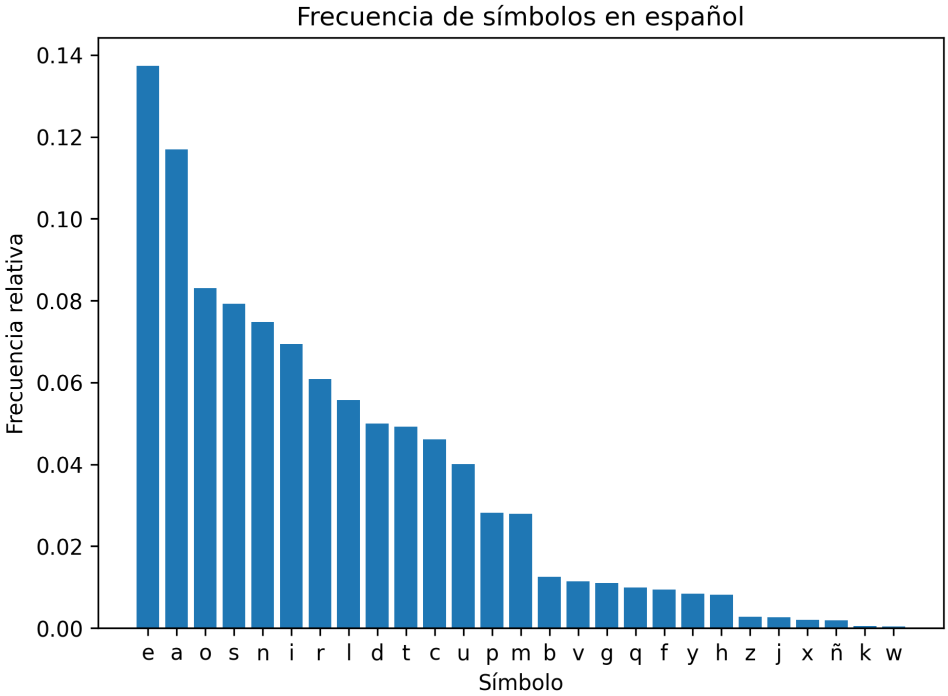

Figura 1: Frecuencia de símbolos en español

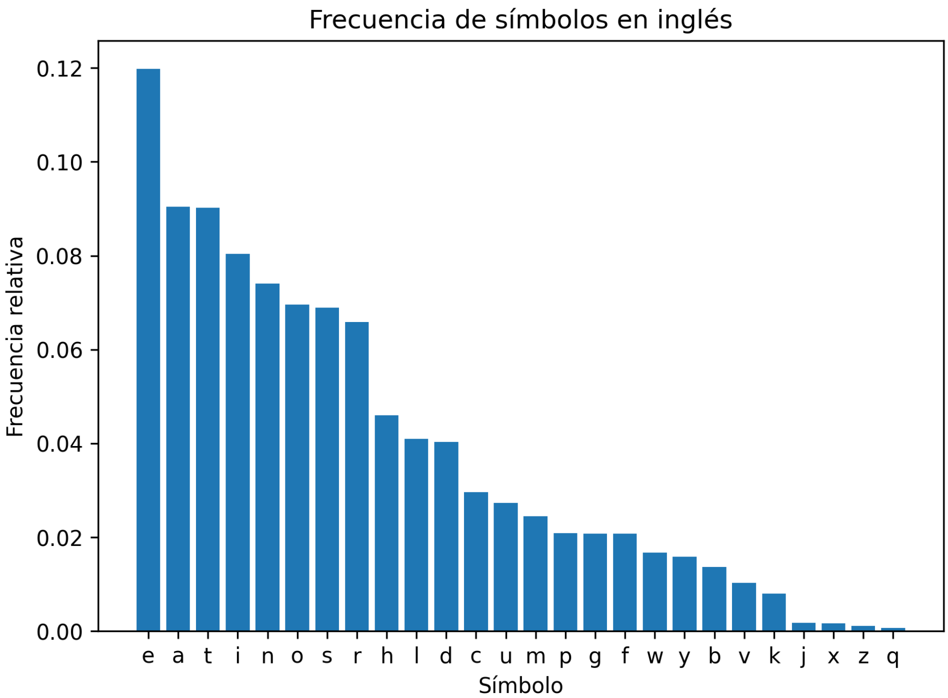

Figura 2: Frecuencia de símbolos en inglés

4.2. Entropía de los Alfabetos

Teórica (Uniforme en Español)

$$H_{\text{esp,u}} = \log_2 27 \approx 4{,}75 \text{ bits}$$

Teórica (Uniforme en Inglés)

$$H_{\mathrm{eng,u}} = \log_2 26 \approx 4{,}70 \text{ bits}$$

Calculada

$$H_{\mathrm{esp}} = 4{,}18 \text{ bits}, \quad H_{\mathrm{eng}} = 4{,}05 \text{ bits}$$

4.3. Entropía de Dos Símbolos

Teórica

$$H^{(2)}_{\mathrm{esp,u}} = 2\log_2 27 \approx 9{,}51 \text{ bits}, \quad H^{(2)}_{\mathrm{eng,u}} = 2\log_2 26 \approx 9{,}40 \text{ bits}$$

Práctica

$$H^{(2)}_{\mathrm{esp}} \approx 7{,}9 \text{ bits}, \quad H^{(2)}_{\mathrm{eng}} \approx 7{,}6 \text{ bits}$$

4.4. Entropía Condicionada

$$H_{\mathrm{cond,esp}} = H^{(2)}_{\mathrm{esp}} - H_{\mathrm{esp}} \approx 3{,}72 \text{ bits}$$
$$H_{\mathrm{cond,eng}} = H^{(2)}_{\mathrm{eng}} - H_{\mathrm{eng}} \approx 3{,}55 \text{ bits}$$

5. Análisis de Resultados

- La entropía real es menor a la teórica, lo que indica redundancia lingüística.
- El inglés presenta una menor entropía promedio, lo que podría deberse a diferencias morfosintácticas o al alfabeto.
- Se observan patrones de bigramas frecuentes: en español “es”, “de”; en inglés “th”, “he”.
- Nuestros resultados coinciden con valores reportados en la literatura: 4.01–4.11 bits para español y 3.9–4.03 bits para inglés.
- Ambos idiomas siguen la ley de Zipf con $\alpha \approx 1$.

6. Conclusión

Los resultados muestran que tanto el español como el inglés presentan una clara redundancia estadística, evidenciada por el hecho de que su entropía real es inferior a la teórica. Esto implica que los datos pueden ser comprimidos eficientemente utilizando métodos como Huffman o codificación aritmética.

La entropía de bigramas y la entropía condicionada revelan dependencia entre caracteres consecutivos. Esta reducción en la incertidumbre es clave en sistemas de predicción, compresión basada en contexto y análisis lingüístico.

Desde la perspectiva criptográfica, esta redundancia representa una vulnerabilidad potencial para sistemas que no consideran estas características. Los algoritmos modernos incorporan mecanismos de difusión para contrarrestar esta predictibilidad.

Bibliografía

Referencias

[1] G. K. Zipf, *Human behavior and the principle of least effort*, Addison-Wesley, 1949.

[2] C. E. Shannon, *A mathematical theory of communication*, Bell System Technical Journal, vol. 27, pp. 379–423, 1948.

[3] Corpus de artículos de El País. https://datos.elpais.com/ (Accedido el 19 de junio de 2025).

[4] Corpus de artículos de The Guardian. https://open-platform.theguardian.com/ (Accedido el 19 de junio de 2025).

[5] T. M. Cover y J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 2006.