ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN

INTEGRANTES: SOTO BRUCE – BYRON ORTIZ – JEFFERSON TOAPANTA

FECHA: 30/05/2024

TALLER 03: LEY DE ZIPF Y RECUPERACIÓN DE INFORMACIÓN

Introducción

La ley de Zipf es un fenómeno estadístico que se observa en diversas áreas, incluyendo la lingüística, la economía y las ciencias sociales. Formulada por el lingüista estadounidense George Zipf en la década de 1930, esta ley establece que, en un corpus de texto suficientemente grande, la frecuencia de cualquier palabra es inversamente proporcional a su posición en una lista ordenada de las palabras por frecuencia. Matemáticamente, se puede expresar como:

$$f(r) \approx \frac{C}{r^s}$$

Donde:

f(r) es la frecuencia de la palabra en la posición r.

r es la posición de la palabra cuando todas las palabras están ordenadas por frecuencia.

C es una constante.

s es un parámetro que en muchos casos es aproximadamente 1.

Objetivo

Entender la ley de Zipf y su efecto en la Recuperación de Información mediante la recopilación y análisis manual de datos.

Materiales

- Corpus de texto (disponible en el aula virtual).
- Papel y lápiz, o una hoja de cálculo.
- Calculadora.

Análisis de la Ley de Zipf

- **Grafica en una escala lineal**
  La Ilustración 1 muestra una caída rápida en la frecuencia a medida que aumenta el rango. Esto es esperado y muestra que las palabras más comunes (con rangos bajos) tienen frecuencias mucho más altas que las palabras menos comunes.

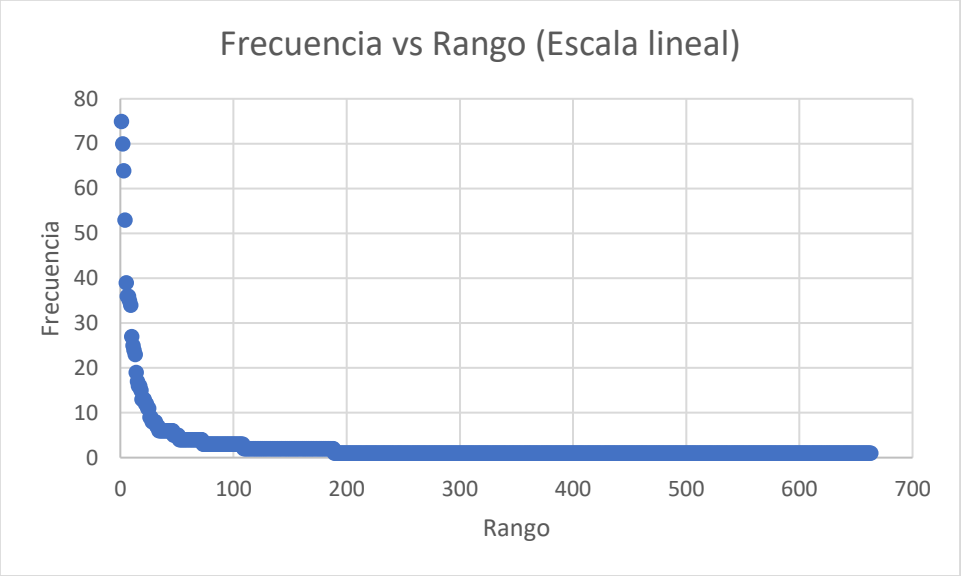

Ilustración 1 Frecuencia vs Rango (Escala lineal)

- **Grafica en una escala logarítmica**
  La Ilustración 2 aunque no es una línea perfectamente recta, la tendencia decreciente aproximadamente lineal en la gráfica logarítmica sugiere que la Ley de Zipf se está cumpliendo en el texto analizado.

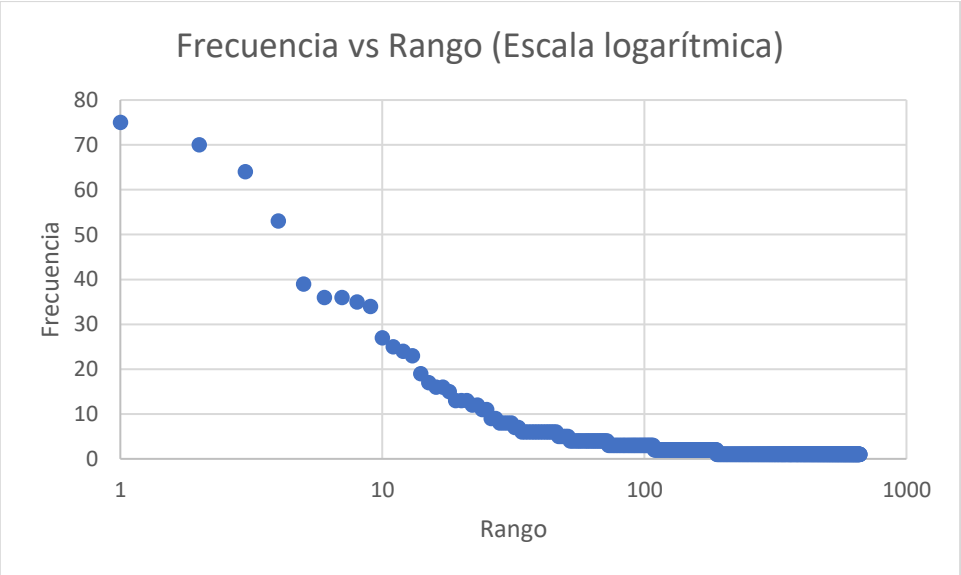

Ilustración 2 Frecuencia vs Rango (Escala logarítmica)

Discusión

- **¿Cómo afecta la alta frecuencia de palabras comunes (como stopwords) en la eficiencia de un sistema de búsqueda?**
  Las palabras comunes (stopwords) tienen una frecuencia muy alta y pueden ocupar una gran parte del índice en un sistema de búsqueda. Esto puede disminuir la eficiencia del sistema, ya que estas palabras no aportan mucho valor semántico. Por lo tanto, es común que los motores de búsqueda filtren estas palabras para mejorar la precisión y eficiencia.

- **¿Qué técnicas se pueden utilizar para manejar palabras extremadamente frecuentes y palabras raras en sistemas de Recuperación de Información?**
  TF-IDF: La ponderación de términos por su frecuencia inversa ayuda a identificar términos relevantes.
  Stemming y Lemmatización: Reducir las palabras a su forma raíz ayuda a agrupar variantes y mejorar la búsqueda.

- **¿Cómo se pueden usar estos hallazgos para mejorar la precisión y el recall de un motor de búsqueda?**
  Precisión: Filtrando stopwords y enfocándose en palabras clave relevantes, se mejora la precisión de los resultados de búsqueda.
  Recall: Utilizando técnicas de ponderación como TF-IDF, se pueden recuperar documentos que contengan términos relevantes aunque no sean los más frecuentes.

  Conclusiones
  Las gráficas obtenidas confirman la Ley de Zipf en el texto analizado. Esta ley tiene importantes implicaciones en la recuperación de información, especialmente en cómo se manejan palabras comunes y raras para optimizar la eficiencia y eficacia de los sistemas de búsqueda.