Taller 03: Ley de Zipf y Recuperación de Información

Prof. Iván Carrera

30 de mayo de 2024

1. 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.

En este ejercicio, exploraremos la ley de Zipf y su impacto en la Recuperación de Información a través de la recopilación y análisis manual de datos.

2. Objetivo

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

3. Materiales

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

4. Pasos a Seguir

4.1. Recopilación de Datos

- Cada grupo debe leer el texto y contar la frecuencia de cada palabra.
- Crear una lista con todas las palabras y sus respectivas frecuencias.

4.2. Ordenación de las Palabras

- Ordenar las palabras por frecuencia, de mayor a menor.
- Asignar un rango a cada palabra basado en su posición en la lista ordenada (1 para la palabra más frecuente, 2 para la siguiente, etc.).

4.3. Análisis de la Ley de Zipf

- Graficar las frecuencias (eje Y) contra los rangos (eje X) en una escala lineal.
- Graficar las frecuencias (eje Y) contra los rangos (eje X) en una escala logarítmica.
- Analizar si las gráficas siguen una línea recta, lo cual confirmaría la ley de Zipf.

4.4. Discusión

- Reflexionar sobre cómo la frecuencia de las palabras afecta la indexación y la búsqueda de información.
- Preguntas:
  - ¿Cómo afecta la alta frecuencia de palabras comunes (como *stopwords*) en la eficiencia de un sistema de búsqueda?
  - ¿Qué técnicas se pueden utilizar para manejar palabras extremadamente frecuentes y palabras raras en sistemas de Recuperación de Información?
  - ¿Cómo se pueden usar estos hallazgos para mejorar la precisión y el recall de un motor de búsqueda?

5. Entrega

- Cada grupo presenta los resultados del análisis de la Ley de Zipf.
- Resumir las implicaciones de la ley de Zipf en la Recuperación de Información y cómo esta ley puede influir en el diseño de sistemas de búsqueda y procesamiento de textos.