Universitat Politècnica de Catalunya

Facultad de Informática de Barcelona

Ley de Zipf y Heap

Búsqueda y Análisis de Información Masiva

Autores

Nicolas Llorens

Ricardo Lopez

03/10/2024

Contents

1 Ley de Zipf

La ley de Zipf determina que la frecuencia de aparición de una palabra es inversamente proporcional a la posición que ocupa dicha palabra según su número de apariciones. En este caso la i es la posición de la palabra en la lista y f(i) la frecuencia de la palabra en los textos. Los valores de a,b,c depende del texto.

$$f(i) = \frac{c}{(i+b)^a} \tag{1}$$

1.1 Método y Ajuste de Parámetros

Utilizaremos el dataset **20 newsgroups** para analizar el comportamiento de la ley de Zipf. Para realizar el análisis, hemos indexado los textos en el motor de búsqueda de ElasticSearch.
A continuación realizamos el recuento de las palabras, para obtener su frecuencia.
El siguiente paso fue cribar las palabras que se obtienen. Este cribado se realiza para eliminar aquellas palabras que no existen. En este caso el criterio para aceptar una palabra es aquella que solamente contenga letras, ya sean mayúsculas o minúsculas. Otro cribado que también se puede realizar es el descarte de las palabras llamadas *"StopWords"* . Las palabras *"StopWords"* son aquellas palabras que se consideran inútiles dentro un texto, dado que estas son palabras de un uso muy común en el idioma y aportan una información casi nula del texto. Un ejemplo claro para la mayoría de los idiomas son las preposiciones. En este caso hemos hecho el análisis de la ley tanto con *"StopWords"* como sin *"StopWords"*. Hemos utilizado la librería de nltk de python, para realizar el cribado de las *"StopWords"*.
Finalmente las ordenamos en orden descendente, asignamos el rango a cada frecuencia, siendo rango 1 la palabra con la frecuencia más alta.

El objetivo principal es determinar las 3 constantes(a,b,c) para el dataset escogido. Los valores se obtienen de usando la librería *scipy* de python. Hemos acotado los valores de a b y c entre 0 y $+\infty, dado que la frecuencia de una palabra nunca puede ser negativa.$

1.2 Análisis de los Resultados

1.2.1 Parámetro a

- **Sin palabras vacías:** El valor $a = 0.789$ muestra que la frecuencia de las palabras decrece más lentamente en comparación con el caso con palabras vacías.
- **Con palabras vacías:** El valor $a = 0.487$ refleja que la presencia de palabras vacías suaviza la pendiente de la distribución, ya que estas palabras son extremadamente frecuentes.

1.2.2 Parámetro b

- **Sin palabras vacías:** $b = 9.433$ introduce un desplazamiento notable en el ranking, lo que puede reflejar que el ajuste requiere un desplazamiento adicional para capturar mejor el comportamiento de las palabras más frecuentes.

- **Con palabras vacías:** $b = 0.0$, lo que implica que no se requiere un desplazamiento en este caso para ajustar adecuadamente la curva.

1.2.3 Parámetro c

- **Sin palabras vacías:** $c = 93261.897$ es mucho mayor, ya que las palabras más comunes (como pronombres o artículos) se han excluido.
- **Con palabras vacías:** $c = 151.045$ es significativamente menor, debido a la inclusión de palabras vacías, que suelen ser más frecuentes.

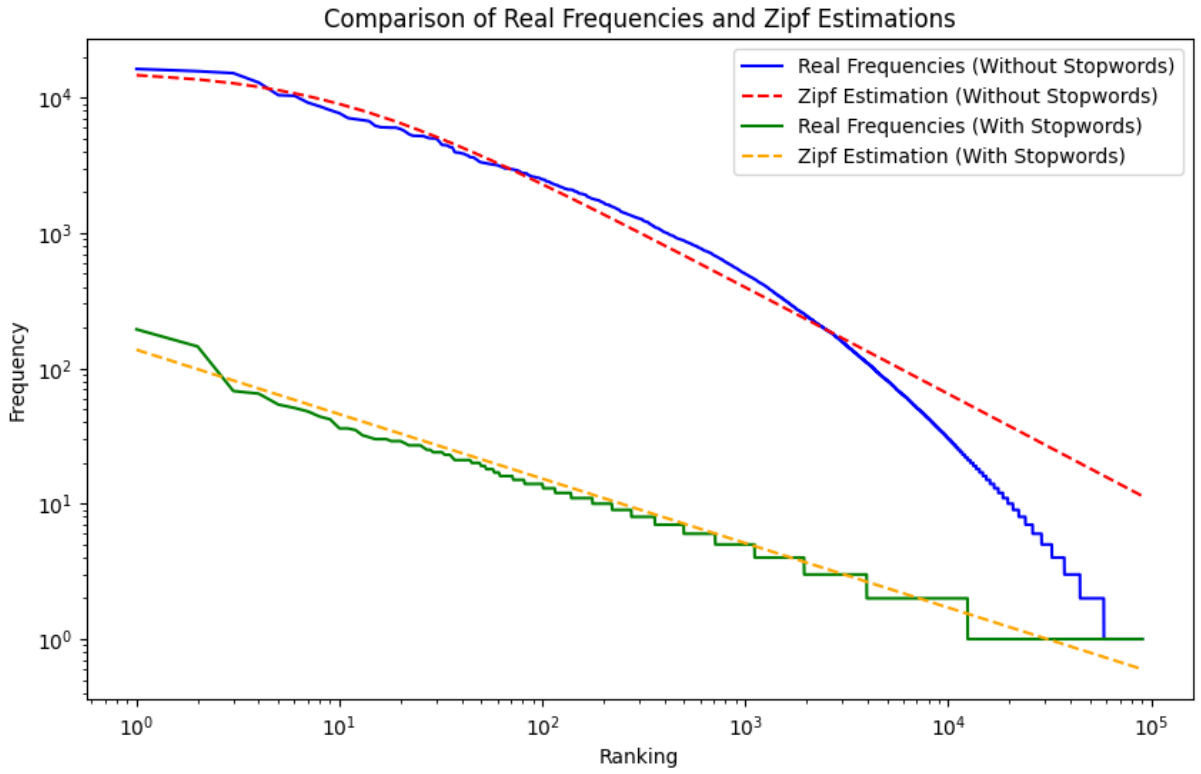

Figure 1: Comparación de frecuencias reales y estimaciones de la Ley de Zipf.

2 Ley de Heaps

La ley de Heap describe una relación entre el tamaño de un corpus de texto y la cantidad de palabras únicas dentro de él. Esta ley, que refleja el crecimiento del vocabulario conforme se añaden más palabras al corpus, se puede representar mediante la fórmula:

$$V(N) = k \cdot N^b$$

Donde:

- $V(N)$ es el número de palabras únicas (vocabulario),
- N es el número total de palabras en el corpus,
- k y b son parámetros específicos de cada corpus.

En este informe analizamos dos tipos de textos: novelas literarias y artículos científicos, con el fin de ver como las características del corpus pueden tener impacto en los parámetros de la Ley de Heap.

2.1 Método Usado y Ajuste de Parámetros

Hemos realizado el ajuste de los valores k y b mediante la función `'curve_fit'`. La función de Heap fue ajustada a los datos de ambos corpus con las siguientes configuraciones iniciales:

- Parámetros iniciales: $params = [1, 0.5]$
- Límites de los parámetros: $k \geq 0$, $b \geq 0$
- Máximo de iteraciones: 10,000

El ajuste se realizó por separado para las novelas y los artículos científicos, obteniendo los siguientes valores:

- **Novelas**: $k = 15.53$, $b = 0.53$
- **Artículos científicos**: $k = 6.87$, $b = 0.62$

En la Figura 2 se muestran las gráficas comparativas entre los datos originales y los valores ajustados con la Ley de Heap.

2.2 Análisis de los Resultados

2.2.1 Novelas

El valor de $k = 15.53$ y $b = 0.53$ para las novelas refleja una tasa de crecimiento del vocabulario más alta en términos absolutos en comparación con los artículos científicos. Esto se explica con la naturaleza del corpus ya que hay mas diversidad de vocabulario en novelas que en textos científicos

El exponente $b = 0.53$, menor que el de los artículos científicos, sugiere que, aunque la diversidad léxica es alta en términos absolutos (debido al alto k), la tasa de incremento

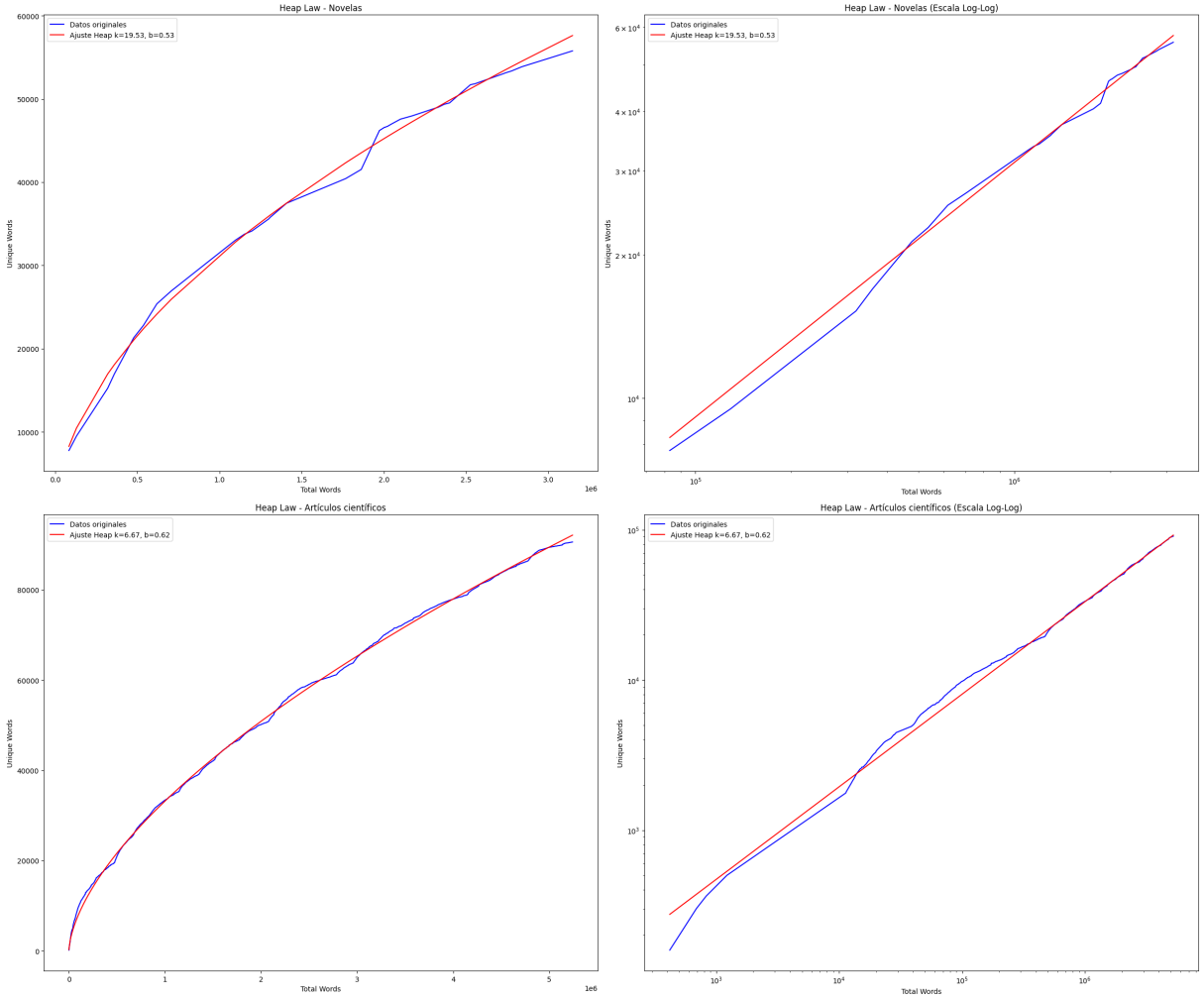

Figure 2: Comparación entre los datos originales y el ajuste de la Ley de Heap para novelas y artículos científicos.

en la diversidad tiende a estabilizarse más rápidamente. Esto podria explicarse porque, aunque las novelas emplean un vocabulario diverso, la repetición de términos comunes y la estructura narrativa acaban imponiendo un límite al crecimiento del vocabulario único conforme avanza el texto.

2.2.2 Artículos Científicos

En contraste, el ajuste para los artículos científicos da como resultado $k = 6.87$ y $b = 0.62$. Aunque el valor de k es notablemente menor, lo que implica que los artículos científicos tienen un vocabulario menos diverso en términos absolutos, el valor mayor de b indica que la tasa de crecimiento del vocabulario único es más sostenida a medida que el corpus crece.

Esto es esperable debido a la naturaleza del lenguaje científico, que tiende a ser más técnico y repetitivo en comparación con el lenguaje literario. Sin embargo, a medida que se exploran más artículos y se introducen conceptos nuevos y específicos, el vocabulario sigue creciendo de manera más sostenida, lo que explica el mayor valor de b.

2.2.3 Método Utilizado

El ajuste de los parámetros se realizó utilizando la función `'curve_fit'`, que es una función de optimización para encontrar los valores de k y b que mejor se ajustan a los datos observados. Utilizamos una estimación inicial estándar de $k = 1$ y $b = 0.5$, asumiendo un crecimiento moderado del vocabulario.

Es importante definir limites a k y b para asegurar que estos valores fueran positivos evitando ajustes erróneos. Y el alto numero de repeticiones era para asegurar que `'curve_fit'` convergiese.

3 Conclusiones

La gráfica muestra que los datos siguen aproximadamente el comportamiento esperado por la Ley de Zipf. En el caso **sin palabras vacías**, la curva de frecuencia (en azul) muestra una disminución suave en las primeras posiciones, seguida de una caída más pronunciada a medida que aumenta el ranking. La estimación de la Ley de Zipf (línea roja discontinua) sigue esta tendencia bastante bien, aunque sobreestima ligeramente las frecuencias en los rangos intermedios.

En el caso **con palabras vacías**, la curva de frecuencia (en verde) y la estimación de Zipf (línea naranja discontinua) también siguen la misma tendencia general. Sin embargo, el ajuste con palabras vacías presenta un ajuste más lineal, lo que sugiere que las palabras vacías siguen mejor una distribución de Zipf con una pendiente más suave.

En general, los resultados son consistentes con la Ley de Zipf: las palabras más frecuentes son extremadamente comunes, y las frecuencias decrecen de manera inversamente proporcional al ranking. El ajuste de la ley se comporta mejor para los datos **con palabras vacías**, donde la curva sigue un patrón más regular y el ajuste es más cercano a los datos reales.

Los resultados destacan la influencia del contenido de los textos en los parámetros de la Ley de Heap, que era lo que queríamos comprobar inicialmente. El hecho de que los artículos científicos tengan una tasa de crecimiento más alta sugiere que, en corpus con un enfoque más técnico o especializado, el vocabulario sigue expandiéndose a medida que se cubren más temas y áreas de conocimiento.