Búsqueda y Análisis de Información Masiva

Primer Informe de Laboratorio

Leyes Potenciales, Leyes de Zipf y Heaps

Diego Moreno y Daniel Lagos

Grupo 13 CAIM

Universitat Politècnica de Catalunya

Septiembre 2025

Índice

Orden decreciente

Objetivos

En el primer ejercicio, nuestra tarea es hacer una gráfica de tipo plot que muestre la frecuencia de palabras en el comienzo de la obra *Don Quijote de La Mancha*, en orden decreciente.

También buscamos averiguar si el plot resultante es una ley potencial o si se puede aproximar a una.

Procedimiento

En nuestro archivo de código de Python hemos hemos implementado un bucle for que ordena las palabras en orden decreciente según el número de apariciones de la misma, todo tras haber pasado todo el texto a minúsculas, eliminado los stopwords y comparando solo la raíz de las palabras (**stemming**). Al ejecutar el código, se crea un archivo que nos devuelve el plot que necesitamos para este ejercicio.

Resultados

Como se puede ver en la gráfica generada, el plot del ejercicio es una ley potencial.

Ejes *x* e *y* logarítmicos

Objetivos

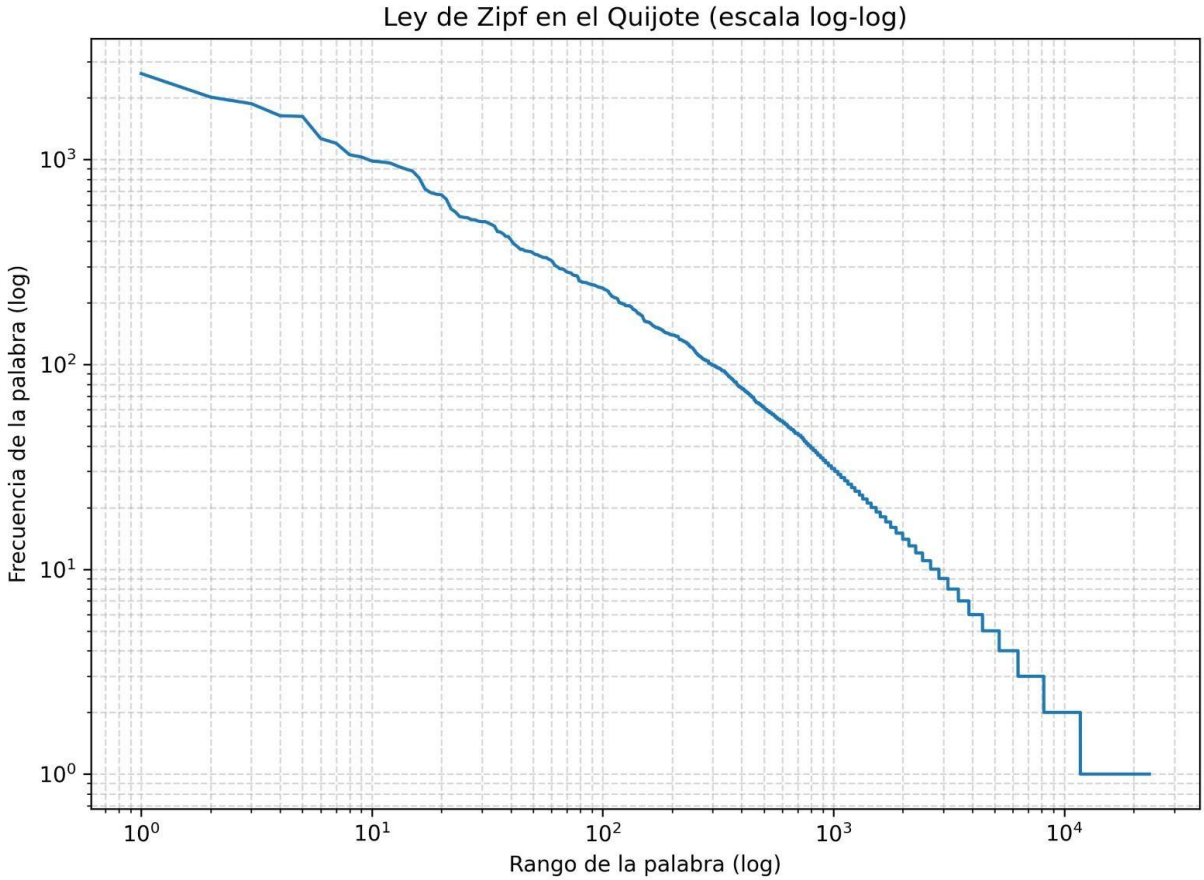
Con el trabajo llevado a cabo para el anterior ejercicio, ahora debemos encontrar una gráfica similar a esta, pero sus valores de los ejes son el resultado del logaritmo de los ejes de la gráfica del ejercicio anterior.

Se espera que el plot de este ejercicio resulte lineal, o una aproximación de ello.

Procedimiento

Aprovechando el código ya existente hemos escalado las variables de abscisas y ordenadas a sus logaritmos. Esto nos permitirá convertir nuestra ley potencial en un plot lineal.

Resultados



Pese a que el plot resultante no es del todo lineal, se aproxima bastante a uno.

Encontrar incógnitas *a*, *b* y *c*

Objetivos

Queremos encontrar los parámetros a,b y c que aplicados a la fórmula de la ley de Zipf ($f = \frac{c}{(rank+b)^a}$) se adaptan mejor a la distribución resultante del análisis de frecuencias de palabras en nuestro test de El Quijote.

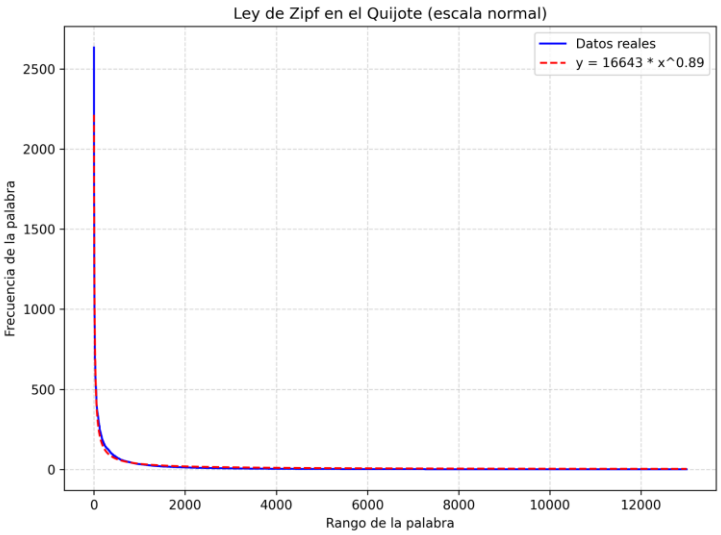
Procedimiento

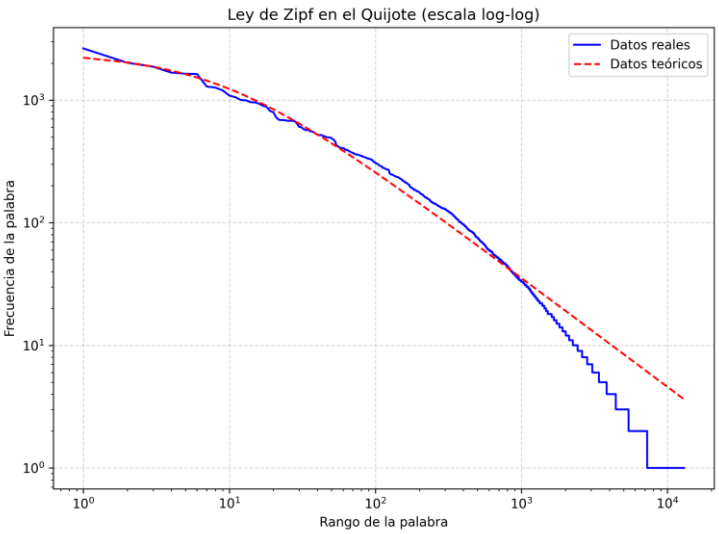
Para encontrar los valores correspondientes de estas variables hemos utilizado el método *curve_fit* proporcionado por la librería de python **scipy** y posteriormente hemos revisado manualmente y comparando valores si los resultados proporcionados se ajustaban a los datos reales.

Resultados

Usando esta técnica hemos obtenido los siguientes resultados:

- **a = 0.89**
- **b = 8.65**
- **c = 16643**





Podemos observar que con estos parámetros la ley de Zipf claramente se adapta bastante bien a los valores reales en la escala normal mientras que en la escala log-log se separa bastante en la cola. Esto se debe a que la cantidad de palabras que existen en la cola probablemente con frecuencias de 1 o 2 es mayor a la esperada por la ley. Estamos viendo un fenómeno de **heavy-tail** muy grande

Ley de Heaps y escala Log-Log

Objetivos

La ley de Heaps relaciona la cantidad de palabras de un texto respecto a las que són nuevas en el vocabulario creado por el mismo mediante la fórmula: $d = k * N^{\square}$. En este ejercicio utilizaremos como base el texto de El Quijote para comprobar que este fenómeno realmente se cumple y además buscaremos los valores tanto de k como de β para que la gráfica teórica se adapte de la mejor manera a la gráfica real.

Procedimiento

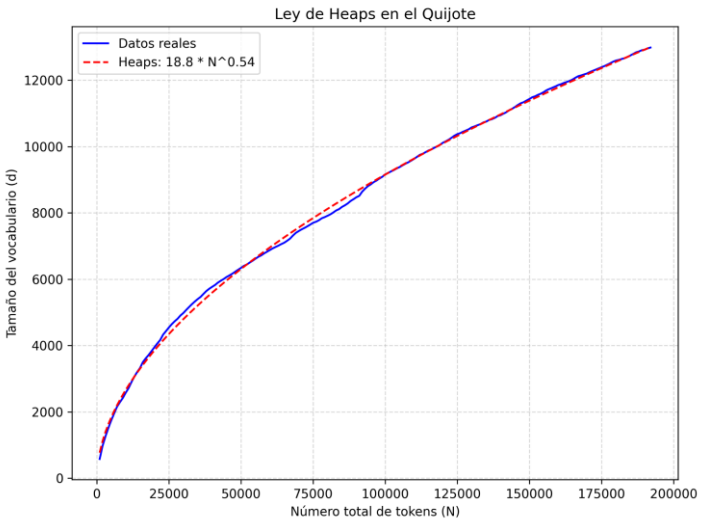
Para llevar a cabo el análisis hemos implementado un código en python que mide el número de palabras nuevas cada cierto número de palabras leídas (step) al cual hemos asignado un valor de 1000 para que el análisis no se haga demasiado complejo temporalmente ya que el

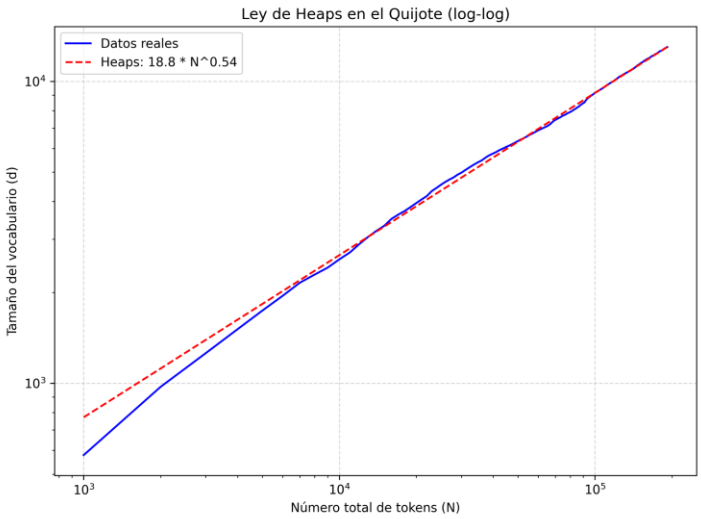
texto completo tiene un total de casi 200.000 palabras en total.

Para medir y calcular los valores de k y β hemos usando el mismo método que con la ley de Zipf, utilizando el método **curve_fit** de la librería de python **scipy**

Resultados

A continuación se muestran las gráficas con los valores reales comparada con la ley de Heaps ajustada a los valores que nos devuelve curve_fit: **k = 18.85** i **β = 0.537.**





Observamos cómo el comportamiento de la gráfica que representa los datos reales se asemeja mucho a la función de Heaps para los parámetros calculados. El comportamiento de la gráfica es el esperado: inicialmente se detectan muchas palabras nuevas del vocabulario y conforme avanza la novela es cada vez menos común encontrar nuevas palabras no detectadas previamente, lo cual provoca la disminución progresiva de la pendiente.