ElasticSearch and Zipfs and Heaps laws

Pol Renau Larrodé
Felipe Ramis Vásquez

Quadrimestre Tardor 2019 - 2020

1 Ley de Zipf

Nuestro objetivo en este apartado de la practica, es ver si una distribución de rango-frecuencia de las palabras en un gran corpus de datos, sigue la ley de la potencia. Con lo que tendremos que ver si ajustando los parámetros de la siguiente función, encontramos una que describa a los datos, o se asemeje.

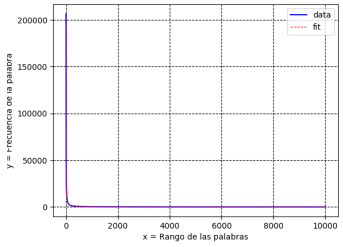
$$f = \frac{c}{(rank + b)^a}$$

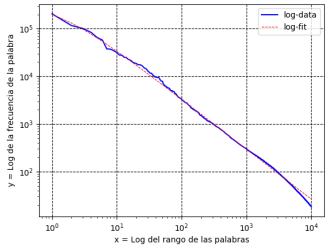
Hemos realizado el anlisis de datos sobre la colección de *"corpus novels"* ya que a nuestro parecer eran los datos mas limpios. Indexamos todo el corpus de textos para extraer todas las palabras y sus respectivas frecuencias, usando el siguiente comando:

$ python IndexFiles.py –index nov –path /path/to/novels

Una vez indexado el texto limpiamos el fichero de salida, eliminando urls, fechas, palabras que contengan números dentro o que tengan símbolos no admitidos en el lenguaje. De esta manera tenemos una entrada limpia con la que podemos empezar a trabajar.

Realizaremos la experimentación con las 10000 palabras con la frecuencia ms alta. Fijamos valores para la *a*, y usamos la función de *curve fit* para encontrar los valores óptimos para b i c. Empezamos con 0.5, no obstante con este valor la grafica de los datos quedaba bastante por debajo que la de "fitting" , así que modificamos el valor de *a* hasta 1.05, en este punto encontramos que la curva de "fitting" es muy parecida a la de datos.





Finalmente vemos que los datos siguen la tendencia de la ley de potencia. Y que entonces la ley de Zipf tiene según nuestros datos una tendencia de:

$$f(x) = \frac{417936}{(x + 1.007)^{1.05}}$$

2 Ley de Heap

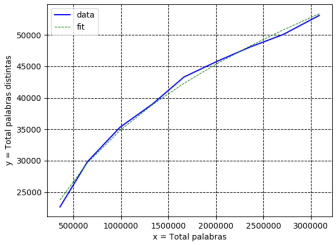
La ley de Heap's es una ley que describe el numero total de palabras diferentes en un Documento:

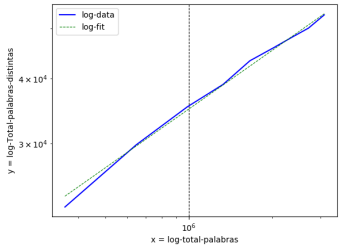
$$d = K * N^{\beta}$$

d representa el numero total de palabras diferentes.
N representa el numero total de palabras en el documento.
K y β són parámetros libres.

Nuestro objetivo es encontrar un valor para K y β para provar que se sigue la ley de Heap's.
Como Tenemos diferentes tamaños de documentos, hemos reducido el tamaño de estos de forma iterativa, reduciendo el corupus aproximadamente para cada iteracin unos 2 MB. Como el tamao inicial del corpus es de 18MB aproximadamente, nos han salido un total de 9 índices diferentes.
Para cada indice, hemos limpiado todas las plabras que no eran válidas url, números, fechas ... hemos contabilizado el total de palabras y el total de palabras distintas.

Finalemente siguiendo la misma metodologia usada en el apartado anterior, hemos usado el metodo *curve fitting* para encontrar los valores que aproximan la ley de Heap's. Estos han sido los valores encontrados: K = 195 β = 0.38





Como podemos ver, hemos encontrado unos valores que se adaptan correctamente a nuestros datos, confirmando así la ley de Heap's. Siguiendo la fórmula siguiente:

$$d = 195 * N^{0.38}$$