Ricard Meyerhofer Parra
2016-2017Q1

CAIM - First Deliverable: Introduction to Leucene

Ley de Zipf

El análisis de datos se ha hecho sobre la colección de *corpus novels* debido a que los datos son más limpios. De todo el conjunto de datos obtenidos (full-data-sorted), nos hemos quedado con los que tienen más de 249 apariciones (selected-data) que nos dan un total de 1192 palabras distintas, ha sido una forma rápida y efectiva de limpiar las palabras residuales de la colección generada y quedarnos con las relevantes.

Una vez limpiados los datos, usando excel nos ha salido la siguiente gráfica:

Imagen: Grafico de las ocurrencias respecto al total de palabras, una vez limpiados los datos.

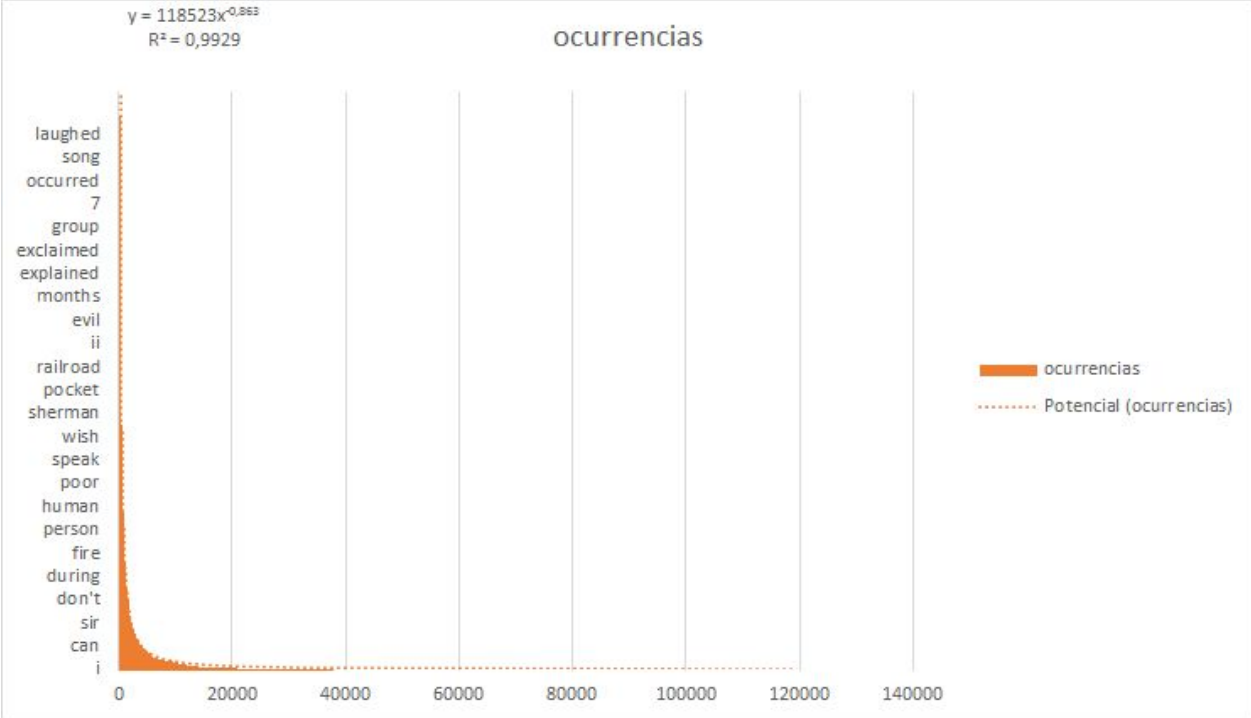
Una vez ya hemos definido los datos que usaremos, procedemos al análisis de estos donde queremos ver que la función es en efecto una ley de potencia (*power law*).

Para ver si es realmente o no una ley de potencia lo que hemos hecho es probar para distintos valores de a, b y c en la siguiente fórmula:

$$f(x) = c * (x + b)^a$$

Para obtener los valores a,b y c lo que hemos hecho es usar excel y decirle que nos diera una línea de tendencia con una función potencial y nos ha salido con una R^2 de 0.9929 que los valores de c, b y a son los siguientes:

$$c = 118523;\ b = 0;\ a = -0.863;$$



Por lo tanto podemos ver que los datos siguen la tendencia de una ley de potencia y que por lo tanto la ley de Zipf tiene la misma tendencia que una función de potencia más concretamente podemos decir mediante nuestros datos que sigue una tendencia similar a

$$f(x) = 118523 * (x + 0)^{-0.863}$$

Ley de Heap

Para demostrar la ley de Heap lo que hemos hecho es crear distintos índices conteniendo cada vez mayor texto. Lo que hemos hecho es coger del total de tamaño de todos los libros (17.4MB) e ir indexando progresivamente cada vez menor "tamaño" de texto (borrando los libros de 2MB en 2MB)

Queremos ver que los datos por lo tanto siguen una función similar a

$$PalabrasDistintas = k * NúmerosPalabras^{b}$$

A continuación el gráfico obtenido de realizar las distintas indexaciones (heap-data)

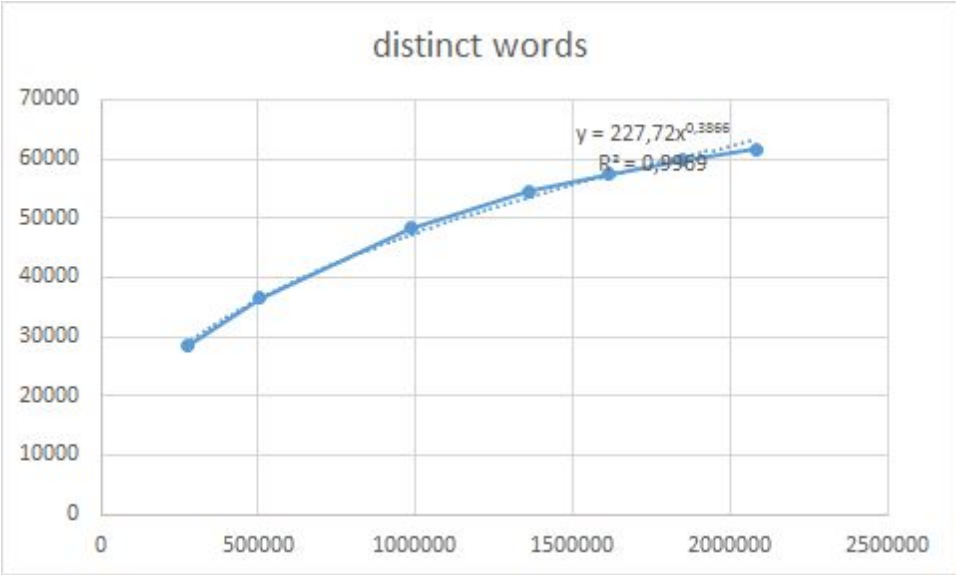


Imagen: Grafico de las ocurrencias respecto a la distintas indexaciones de datos.

Tal y como hemos hecho con la Ley de Zipf, hemos aproximado el gráfico mediante excel donde en este caso nos han salido los valores siguientes:

$$k = 227,72;\ b = 0,3866;$$

Por lo tanto podemos confirmar que la ley de Heap es cierta y que el número de palabras distintas no aumenta de forma proporcional al número total. De hecho, lo que vemos es que el número de palabras distintas sigue aproximadamente la función

$$PalabrasDistintas = 227,72 * NúmerosPalabras^{0.3866}$$