CAIM
Informe Primera Práctica:
ley de Zipf - Heap

Yimin Pan (G14)
Eduard Pujol Puig (G14)

Introducción

En esta primera práctica, nuestro objetivo era comprobar las leyes de Zipf y Heap mediante experimentos. En concreto, analizamos documentos de diferentes tipos mediante elasticsearch para ver que efectivamente sucedía lo que dicen estas leyes con unos conjuntos de ficheros proporcionado.

Ley de Zipf

La ley de Zipf es una ley empírica que relaciona una distribución de rango-frecuencia con una ley potencial de este aspecto:

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

Para comprobarlo, primero, indexamos los tres conjuntos de documentos de diferentes contextos: coloquial, literario, científico. De este modo podremos observar si el tipo de texto influye sobre la ley.

Contamos las apariciones de cada palabra en los documentos mediante elasticsearch, con el script llamado CountWords.

En el siguiente paso filtramos los datos. Dado que habían palabras que realmente no eran palabras como: links, números y caracteres…, que podrían afectar el experimento, sólo consideramos palabras aquella que todos sus caracteres pertenecían al abecedario.

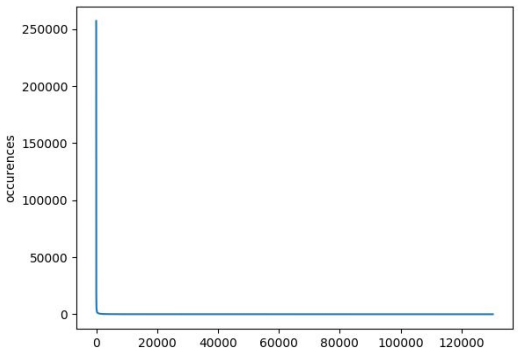

Imagen 1: gráfica original novelas

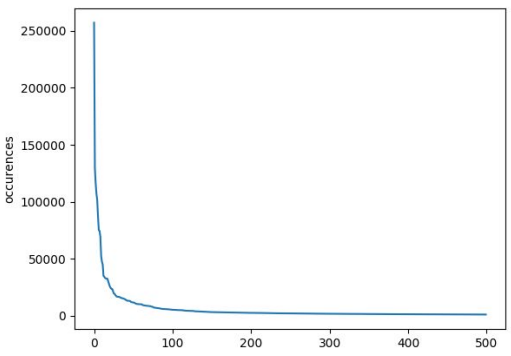

Imagen 2: gráfica con 500 palabras novelas

Al ver que la gŕafica se parecía una L nos extrañó un poco, pero rápidamente nos dimos cuenta que era porque había demasiadas palabras con muy pocas apariciones produciendo un efecto de “zoom out”, que en realidad la gráfica sigue siendo una ley potencial. Dicho esto, decidimos estudiar solo las 500 palabras más frecuentes. Del resultado se observa que la frecuencia de las palabras decae exponencialmente, como se predecía en la ley. (Por eso había tantas palabras con solo una aparición).

En el siguiente paso intentamos aproximar la distribución de rango-frecuencia por la función de Zipf, usando curve_fit de la librería scipy. Sin embargo el primer intento fracasó, el algoritmos no era capaz de resolverlo, no era capaz de encontrar los parámetros a,b y c. Por tal de hacérselo más fácil, le fijamos un intervalo de [0.75, 1.25] al parámetro 'a' (tradicionalmente se usan valores alrededor de 1).

En las siguientes gráficas podemos ver para todos los conjuntos de ficheros proporcionados, la gráfica de la frecuencia respecto el rago y a la derecha la del logaritmo de la frecuencia respecto al logaritmo del rango. La línea roja representa los datos recogidos y la azul representa la función de Zipf que se ajusta mejor a los datos. La gráfica a escala log-log sirve para visualizar mejor la relación entre la función de Zipf encontrada y los datos.

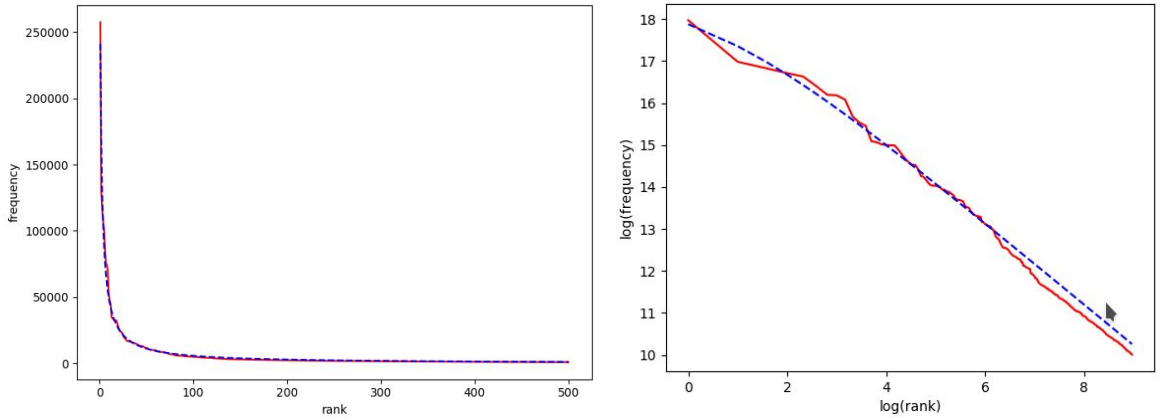

Imagen 3: gráficas con los fitcheros 20_newsgroups (sin escala / log-log)

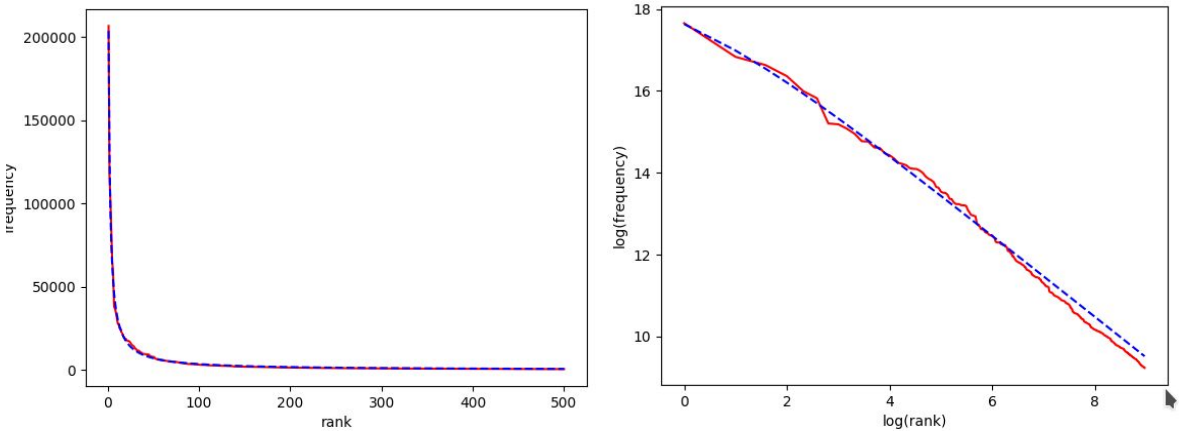

Imagen 4: gráficas con los fitcheros novels (sin escala / log-log)

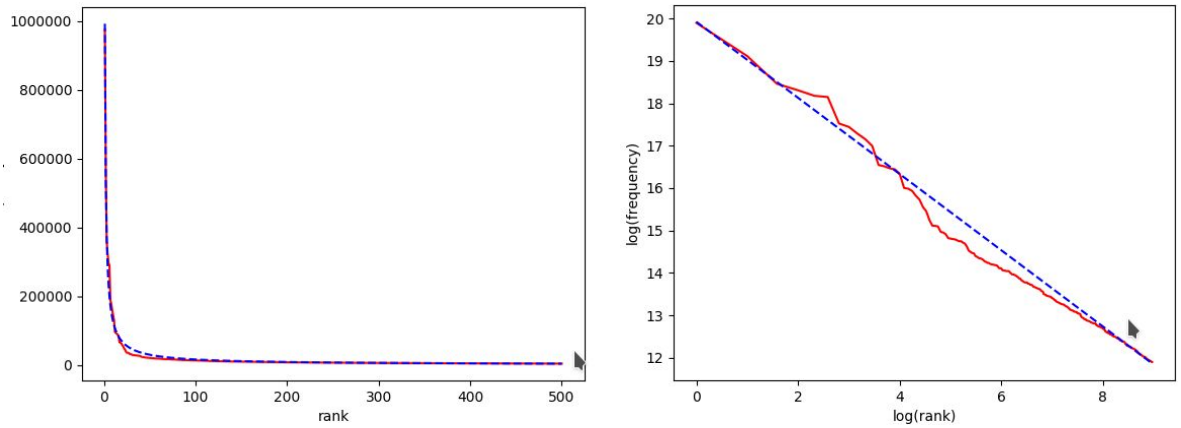

Imagen 5: gráficas con los fitcheros arxiv (sin escala / log-log)

	a	b	c
20_newsgroups	0.974127	1.21529	522973.913
novels	0.9952822	0.757617	356836.039
ArXiv abstracts	0.899612013	0.0082899	1000000

Imagen 7: valores constantes a,b,c de la funcion de Zipf

Conculusión

Al realizar el mismo experimento con los otros dos documentos, observamos dos hechos interesantes.

Cuando más se acerque 'a' a 1, la curva se aproxima mejor a nuestro modelo, que es caso de las novelas. Dicho esto, podemos decir que la ley de Zipf se ajusta mejor a los documentos de tipo literario.

Los textos coloquiales y literarios tienen los valores constantes de la función de Zipf muy parecidos, por tanto siguen una distribución de frecuencia y rango muy parecida. En cambio lo textos científicos tienen unos valores completamente diferentes. Podemos afirmar que los textos literarios y coloquiales son muy diferentes a los científicos a la hora de repetir palabras ya usadas y en la extensión del vocabulario.

En resumen, mediante estos experimentos hemos podido concluir que la distribución rango-frecuencia sigue una ley potencial (ley de Zipf) y que los ficheros de texto cumplen esta ley de una forma muy acurada.

Ley de Heap

La ley de Heap es otra ley empírica que relaciona el número de palabras diferentes con el número total de palabras de un texto. Dicho esto, es asintóticamente equivalente a la ley de Zipf en cuanto a la frecuencia de las palabras.

$$V_R(n) = Kn^{\beta}$$

Vr(n): cantidad de palabras distintas.
n: total de palabras en el texto.
K y B: parámetros de la función.

Por tal de ver si la ley se cumple, coguimos la novela mas grade que encontramos en el conjunto de novelas y la partimos en 16 ficheros diferentes. Cada uno de los ficheros tenía unas cuantas líneas más que el anterior. De este modo obtuvimos diferentes ficheros con tamaños diferentes.

Con la ayuda de elasticsearch calculamos el número de palabras totales (solo palabras que todas sus caracteres eran alfabéticos) y el número de palabras diferentes. En la siguiente gráfica, en la línea roja podemos ver la gráfica con los valores obtenidos y en la azul la función de Heap con los valores que se ajustan más a los datos:

- k = 43.954153
- Beta = 0.47175871

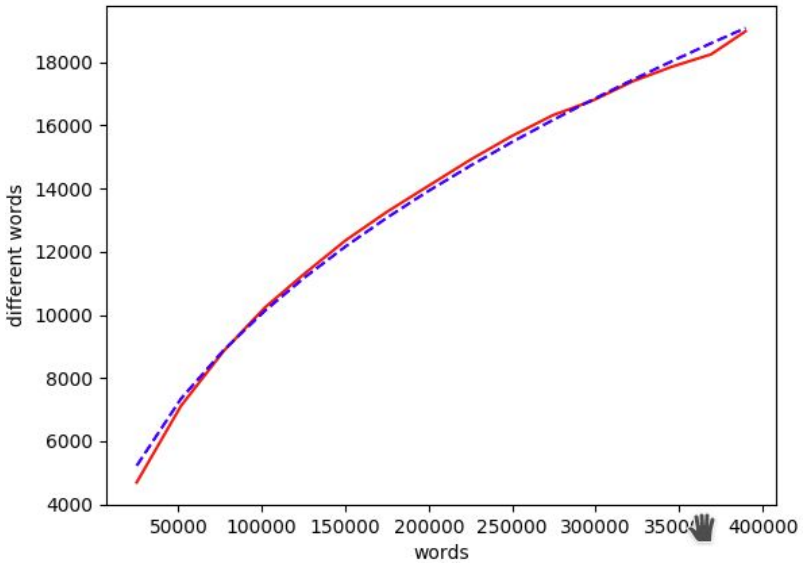

Imagen 8: grafica de la novala más grande del conjunto de novelas

Conculusión

Al realizar el experimento podimos observar dos hechos interesantes.

En las diapositivas se nos explica que la beta de la función de Heap depende del tipo de texto y el lenguaje. Allí también se nos dice que los textos en inglés acostumbran a tener

valores similares a 0.5. En este caso hemos podido comprovar que el texto cojido por nosotros (una novela en inglés) también està muy cerca de ese valor.

Por otra parte hemos visto que la función de Heap representa de forma muy precisa la relación de número palabras con el número de palabras diferentes. Al aumentar el número total de palabras (añadiendo más partición de novela novela original), el crecimiento del número de palabras diferentes decae continuamente. Llegando a un punto en que aunque añadamos más texto, el número de palabras diferentes no va a variar, en la gráfica veríamos que la función tiende a ser paralelo al eje x. Este hecho, si lo pensamos, es bastante lógico, ya que la mayoría de palabras ya lo hemos usado en textos anteriores, por lo que el número de palabras diferentes va a variar poco.