Facultad de informática de Barcelona

Cerca i Anàlisi d'Informació Massiva

Session 1: ElasticSearch and Zipf's and Heaps' laws

23/09/2022

Tardor 2022-2023

Marc Camarillas Parés
Marc Nebot Moyano

Marc Camarillas Parés
Marc Nebot Moyano

ÍNDEX

1. Zipf Law 2

2. Heap Law 5

Marc Camarillas Parés
Marc Nebot Moyano

1. Zipf Law

En aquest apartat demostrarem que la ley de Zipf segueix la llei de potencials.

La ley de Zipf és una llei que ens explica que la freqüència d'aparició d'una paraula segueix una distribució inversament proporcional al n-èssima paraula més freqüent, és a dir, el seu ranking. Segueix aquesta fórmula:

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

Primerament, hem calculat la freqüència d'aparició de cada paraula dins dels textos proposats a la assignatura amb un script de python anomenat CountWords.py. Un cop calculades les potències, les hem ordenat per poder saber el ranking de cada paraula i poder aplicar la fórmula observada. Per això, hem necessitat fer una aproximació dels paràmetres *a, b i c.*

Executant el script de python que hem realitzat i aplicant la funció de curve fit hem obtingut els següents paràmetres d'a, b i c sent aquests els millors valors dins d'un intèrval que li hem donat:

	a	b	c
arxive.org	8.63548903e-01	2.96048953e-12	1.00000000e+06
20_news	9.91149852e-01	1.31463725e+00	5.51426863e+05
novels	1.00783850e+00	8.08325608e-01	3.69224200e+05

Per aconseguir aquests valors hem tingut en compte que *a* no podia ser igual a 0 ja que llavors no es tindria en compte el denominador. També sabíem que *a* havia de ser un valor aproximadament 1, per tant, el nostre intèrval de prova era 0.01 per tal d'evitar el 0 i 1.5 per tal de trobar un valor aproximadament 1. Per altra banda, també sabem que b < c per tal de no obtenir freqüències relativament baixes, per tant, c pot ser qualsevol número per això l'interval de c és més gran que el de b. Finalment, cap dels tres valors pot ser negatiu, no tindria sentit obtenir freqüències negatives.

Marc Camarillas Parés
Marc Nebot Moyano

Seguidament, observarem les gràfiques que hem obtingut:

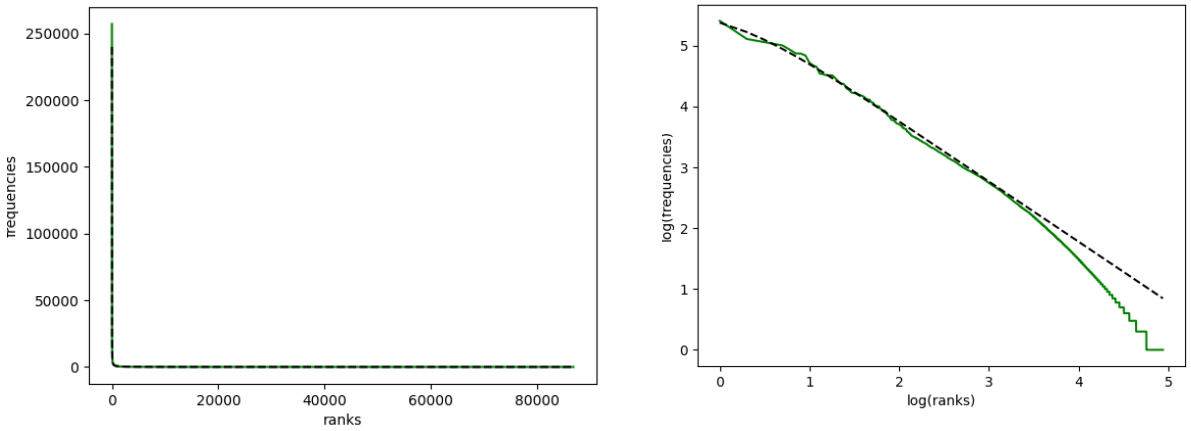

Fig. 1: Representació de la llei de Zipf sense aplicar el log i aplicant-lo a 20_news

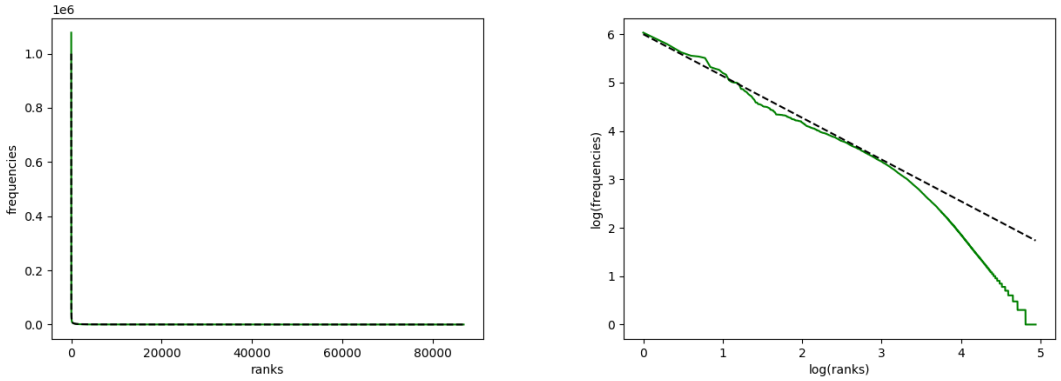

Fig. 2: Representació de la llei de Zipf sense aplicar el log i aplicant-lo a arxive

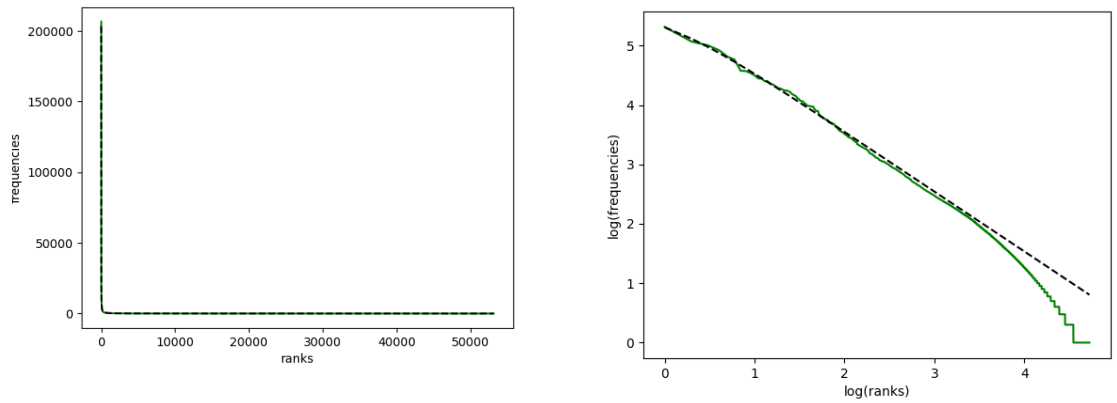

Fig. 3: Representació de la llei de Zipf sense aplicar el log i aplicant-lo a novels

Marc Camarillas Parés
Marc Nebot Moyano

Podem observar que les gràfiques segueixen l'aproximació de Zipf donada i que les freqüències segueixen una distribució inversament proporcional al seu ranking. Aquests resultats són totalment ja que la llei de Zipf s'hauria de complir en la majoria de les llengües existents incloent la llengua anglesa, per tant, compleix com era esperat en aquests textos que hem utilitzat.

Per altre banda, podem observar clarament que on trobem més irregularitats és al conjunt d'arxius que pertanyen a arxive.org, papers científics. Per tant, en trobem moltíssimes paraules no vàlides ja que la majoria eren caràcters numèrics i per tant, al reduir les nostres dades i quedar-nos amb un conjunt tant específic i científic és més probable que trobem irregularitats. Aquestes irregularitats podem veure que no es troben tant en altres conjunts com podrien ser els literaris i col·loquials, ja que estadísticament parlant no utilitzarem un llenguatge tan específic i tècnic com sí ho faríem en articles dedicats a la ciència.

Marc Camarillas Parés
Marc Nebot Moyano

2. Heaps Law

La llei de Heaps es una llei del la lingüística, que descriu el número de paraules diferents en un text en funció al nombre de paraules totals del document. Sent la formula:

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

Per demostrar la llei de Heaps, hem fet servir el conjunt de textos corresponents a les novel·les, amb els quals hem fet 6 grups, cadascun d'aquests amb menor número de novel·les. Un cop fet això obtenim sis valors d'N amb els corresponents valors de paraules diferents que apareixen al conjunt de textos, on N correspon al nombre total de paraules.

A continuació representen els valors obtinguts en un gràfic on l'eix X representa els valors d'N i l'eix Y ens indica el nombre de paraules diferents. Això ens donarà una corba que passa per als 6 punts obtinguts.

Aleshores, per tal de comprovar que la llei es compleix, hem de trobar un parell de paràmetres k i β, que s'ajustin a la corba obtinguda de manera experimental. Això ho aconseguim amb la funció curve_fit().

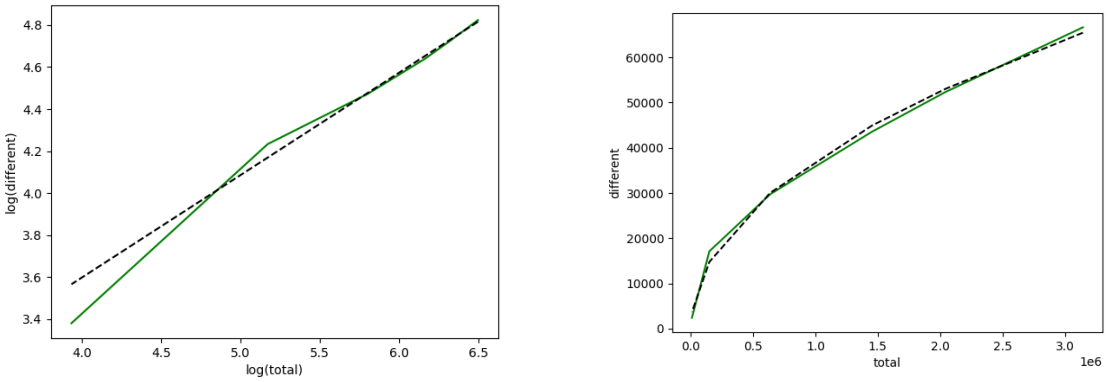

Fig. 4: Representació de la llei de Heaps, aplicant-li el log i sense

Marc Camarillas Parés
Marc Nebot Moyano

Com podem observar a la figura 4, la línia verda representa la corba obtinguda mitjançant l'aproximació dels paràmetres k i β i la de color negre ratllada correspon a la obtinguda de manera experimental. El valors dels paràmetres obtinguts és de 44.45 i 0.48 respectivament, sent k i β un valor raonable en l'àmbit de la llengua anglesa, que ronda el rang de [10,100] per la k i 0.5 per la β.

Mirant les gràfiques podem concloure que la llei de heap es compleix en el conjunt de dades obtingut, ja que hem aconseguit trobar una corba que segueix la funció de la llei de heap, s'ajusta a les dades i té un valor coherent de les variables k i β en la llengua anglesa. No obstant això, en la gràfica logarítmica de la figura 4, veiem que a l'inici no s'ajusta del tot. Això pot ser degut a els dos primers conjunt de novel·les que hem agafat, la relació entre el el nombre de paraules total i el nombre de novel·les es menor que en els altres casos, fent apareguin més paraules diferents de les que haurien d'aparèixer en un text normal.

A més també es pot donar el cas de que no s'han filtrat bé algunes no-paraules i viceversa, fet que pot afectar una mica i que la corba obtinguda de manera experimental no sigui perfecte.