Lab1: Llei de Zipf i de Heaps
Q1 2022-23

Integrants:
Pablo Montón Gimeno
Cristian Sánchez Estapé

Introducció

Aquest laboratori consisteix a constatar tant la Llei de Zipf com la Llei de Heaps, tot fent ús del programari ElasticSearch per indexar els documents que ens permetran precisament realitzar aquesta constatació.

Indexació i preprocessament

Per tal d'estudiar ambdues lleis, farem ús de tres tipus de documents diferents: *news*, que corresponen a aquells relacionats amb noticiaris i derivats; *arxivs*, que corresponen a documents acadèmics i que es troben diferenciats en diversos temes (tals com astrofísica, quàntica, etc.); *novels*, que corresponen a aquells estrictament literaris.

A partir de la indexació d'aquests arxius, amb ElasticSearch extraiem informació sobre el nombre de paraules totals, juntament amb la freqüència d'aparició de cada terme. Ara bé, posat que no tots els termes són veritables paraules, s'ha de realitzar un preprocessat per filtrar aquests intrusos. En el nostre cas, el filtratge l'hem fet eliminant les paraules que contenen caràcters especials o dígits, i que no es troben al diccionari britànic (la qual cosa es comprova mitjançant la llibreria *enchant*).

Llei de Zipf

Per comprovar que a les dades que traiem dels documents indexats, la distribució rang-freqüència segueix una funció potencial de la forma $f = \frac{c}{(rank+b)^a}$, hem de provar diferents tripletes de paràmetres (a,b,c) per tal de trobar els valors que segueixin millor la distribució que se'n deriva d'aquestes. Així mateix, aquestes tripletes les hem trobat fent ús de la funció *curve_fit* de la llibreria *scipy*, la qual, de manera iterativa, cerca els valors ideals pels paràmetres d'una funció donada per tal de fer correspondre un conjunt de dades **X** a un conjunt de resultats **Y**. Dit això, les gràfiques, tenen dues representacions: per una banda, les dades tal com s'extreuen i filtren dels índexs (corresponen a *data*) i, d'altra banda, les dades tal com *haurien de ser* segons els paràmetres ideals (a, b, c) (corresponen a *fit*).

Per aquesta llei, segons el codi adjunt, es poden observar les següents figures pels índexs *news*, *arxivs* i *novels*:

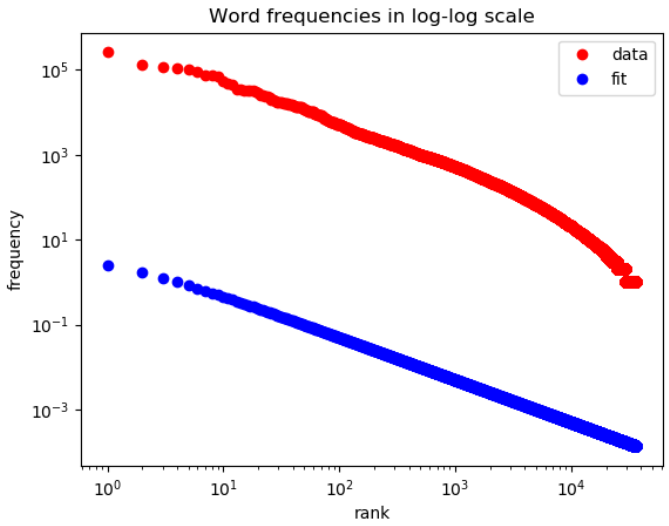

Figura 1: freqüència de les paraules per l'índex news

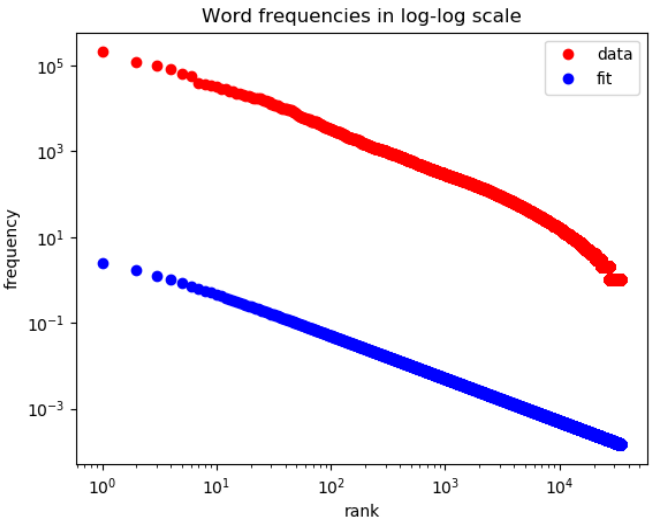

Figura 2: freqüència de les paraules per l'índex novels

	A	B	C	SLOPE (fit)	SLOPE (data)
ARXIV	1.0	1.0	5.0	-0.99868175	-2,17240901
NEWS	1.0	1.0	5.0	-0,99876368	-1,76817569
NOVELS	1.0	1.0	5.0	-0.9987112	-1,65196575

Figura 3: taula amb els valors ideals pels paràmetres de la funció de Zipf i pendent de les rectes associades a les distribucions de dades representades

Segons la funció *curve_fit*, per uns intervals de valors ([1.0, 2.0] [1.0, 2.0] [1.0, 5.0]) per (a b c), els valors ideals són els presentats a la figura 3. És evident que la funció troba els mateixos valors ideals pels paràmetres estudiats, fenomen que atribuïm al funcionament de la funció mateixa: es pot detectar com, tot i que el pendent en les dades originals (la columna *SLOPE (data)*) varia segons el tipus de document estudiat, els paràmetres no canvien, amb la qual cosa resulta evident que el mètode sempre porta a aquesta selecció. Amb això, quelcom destacable a esmentar és el fet que, en indicar els límits en què la funció ha de cercar els valors pels paràmetres mateixos, sempre procura acostar-se tant com pot al mínim establert, i que, en donar llibertat absoluta per fer la tria, els valors oscil·len de tal forma que aquests perden tota possible versemblança (els paràmetres prenen valors o absurdament elevats, o absurdament reduïts, i el pendent del *fit* assoleix un valor igualment anòmal). En aquest cas, posant per límit inferior un valor de 1.0, la funció acota els paràmetres de tal forma que el pendent de la recta associada a la distribució de dades del *fit* és aproximadament -1. Això vol dir que el *fit*, en la present aproximació, no és fiable (o almenys no tant com les dades originals).

D'altra banda, en relació amb la Llei de Zipf, veiem que, mentre que pels documents de tipus *news* i *novels* sembla complir-se, pels *arxivs* la forma de la distribució sembla més una corba que no una recta. A més a més, si es consulta novament la figura 3, es veurà que els pendents detectats en totes 3 distribucions és superior al -1 i, pel cas dels documents de tipus *arxivs*, aquest pren un valor de -2. Amb això podem afirmar que, d'acord amb aquestes mostres, la Llei de Zipf no es compleix: encara que la distribució de les dades s'aproxima a

la forma d'una recta en 2 de 3 figures, els pendents s'allunyen del valor que *idealment* haurien de tenir, i ho fan de manera considerable (i en cap cas negligible).

Llei de Heaps

Per comprovar si els nostres textos compleixen que, per un text amb un nombre total de paraules *N*, el text conté $k \times N^{\beta}$ paraules diferents, buscarem els valors ideals pels paràmetres *k* i *β* per tal de poder acotar la funció de Heaps a les dades extretes dels índexs prèviament esmentats.

Per fer-ho, a partir dels 20 textos de *novels*, hem fet subconjunts de diferent mida (1, 5, 10, 12 i 15 textos diferents) per diversificar els resultats. Després de calcular les dades de paraules totals i diferents, utilitzant *curve_fit* (com amb la Llei de Zipf), hem trobat els següents valors de *k* i *β* per cada subconjunt:

nº de novels	**1**	**5**	**10**	**12**	**15**
k	228.3153012	233.7936815	224.2061209	185.2177408	219.4926253
β	0.76293992	0.84762358	0.9017101	0.92947968	0.92839855

Figura 4: taula amb els resultats dels valors ideals pels paràmetres k i β

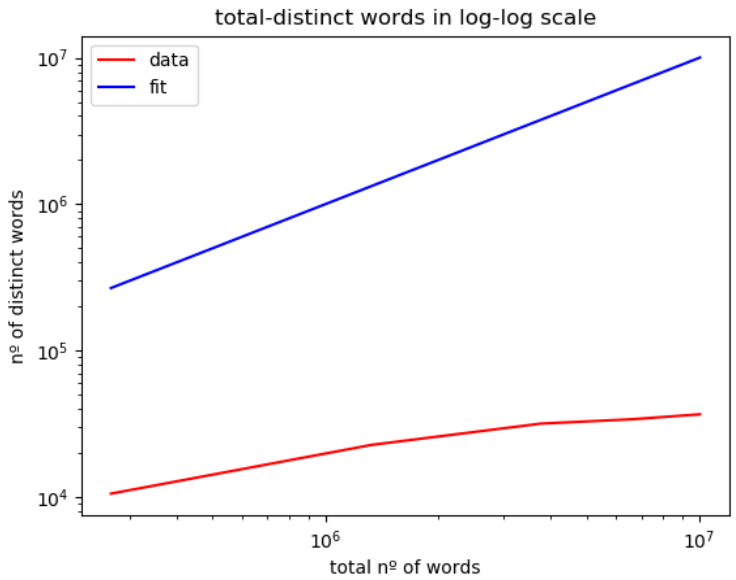

Figura 5: nombre de paraules diferents respecte el nombre total de paraules

	SLOPE
DATA	0.34615613
FIT	1.0

Figura 6: taula amb els pendents associats a la representació gràfica

Observant la figura 4, podem veure que els valors ideals pels paràmetres *k* i β, a diferència de la Llei de Zipf, resulten més versemblants. Donant-li a la funció *curve_fit,* per la tupla de paràmetres (k β), uns intervals de valors màxims entre ([0.0, 500.0] [0.0 1.0]), aquest fa oscil·lar la *k* entre 185 i 233, mentre que β oscil·la entre el 0.76 i el 0.93, fet que, pels llindars que se li proporciona a la funció, fa que els resultats obtinguts, a diferència del que succeïa amb la Llei de Zipf, juguin més amb l'espai de possibilitats que se'ls hi dona, amb la qual cosa porten a pensar que aquests hauran de ser més fiables. Ara bé, sabent que el *Quixot* té una β aproximada de 0.806, i tenint en compte que el castellà és una llengua generalment més rica que no pas l'anglès, sembla que els valors per la *β* resulten relativament erronis, la qual cosa, en representar les dades gràficament (tal com la figura 5 mostra), veiem que sembla ser certa: tenint en compte els pendents de les rectes acotades a les dades representades (vegeu la figura 6), les dades originals tenen un pendent de 0.34 quan se sap que, en escala log-log, el pendent hauria de trobar-se al voltant de valors com 1.0, tal com es veu a partir del pendent de la recta associada a les dades del *fit* (a banda del fet que la representació gràfica de les dades originals té la forma d'una recta deformada). Amb aquesta sèrie d'indicis, podem afirmar amb relativa certesa que les dades no semblen complir tampoc la Llei de Heaps. Tot i això, val la pena afegir que, tot i que les dades del *fit* semblen presentar certes anomalies, en ser representades presenten la forma d'una recta amb pendent 1, amb la qual cosa valdria la pena no descartar completament la hipòtesi que aquests documents la compleixin.