
# Laboratori de CAIM

## Pràctica 1

*Walter J.Troiani*

Prof: Ignasi Gómez Sebastià

24/9/2023 - 2023/24 Q1

# 1 Experiment 1: Llei de Zipf

La llei de Zipf és una equació desenvolupada als anys quaranta pel lingüista George Zipf, que proposa una relació inversa entre el rànquing (Posició que ocupa a la llista de paraules quan aquestes són ordenades descendentment per freqüència) i la freqüència d'aparició d'una paraula en concret f(i) d'aquest en un corpus, llenguatge, o conjunt de documents. Té uns coeficients (a, b, c) que depenen del text en concret a avaluar:

$$f(i) = \frac{c}{(i+b)^a} \tag{1}$$

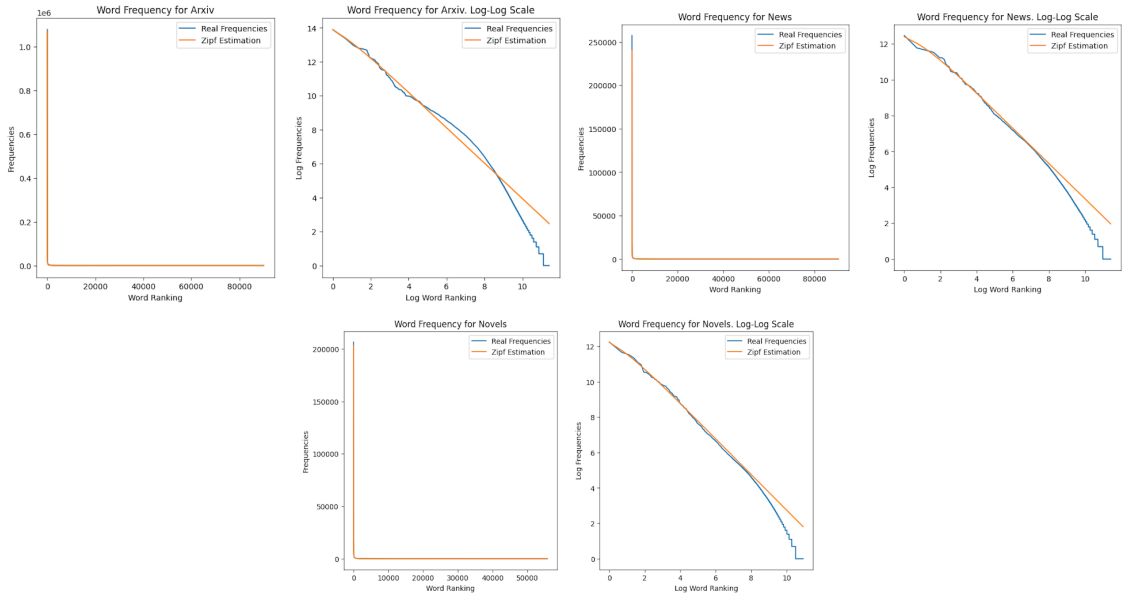
L'objectiu d'aquesta experimentació establir un model (determinant les 3 constants), per a cadascun dels conjunts de documents (índexs) que disposem: novel·les, noticies i arxius científics de arxiv.org, tots en anglès. Els valors que s'han aconseguit usant la llibreria de SciPy amb hipòtesi inicial on:
a = 1, b = 0, c = max f(i), són els següents:

Table 1: Coeficients Zipf pels 3 índexs

| | **a** | **b** | **c** |
|---|---|---|---|
| **Arxiv** | 1.043 | 0.596 | 1745153.519 |
| **Noticies** | 0.984 | 1.282 | 541407.338 |
| **Novel·les** | 1.007 | 0.805 | 368405.245 |

Aquestes constants com bé es pot apreciar, a no s'ha desviat a penes del valor inicial, a diferència de b que ha crescut i c que comparat amb els valors originals ha augmentat força , més d'un 50% (Valors inicials de freqüències màximes eren 1077659, 257240, 206546 respectivament).

A continuació es mostren els resultats estimats per aquests 3 models basats en la llei de zipf, en comparació als valors de les freqüències reals:




# 2 Experiment 2: Llei de Heap

L'altra troballa estadística important en el món de l'estadística dels textos és la llei d'Heap. Aquesta llei relaciona el creixement del nombre de paraules d'un text, amb el nombre de paraules diferents que també va creixent, a un ritme sublineal. L'explicació darrere d'això es força complexa, però la llei és simple:

$$d(n) = k * n^{\beta} \tag{2}$$

Per obtenir les dades necessàries (Parelles de N,D , és a dir, parelles de valors de nombre total de paraules d'un índex i nombre total de paraules úniques) hem seguit el procediment següent. De l'índex de novel·les original, hem anat fent subconjunts forçosament més grans, successivament, per així establir valors creixents de N i així observar com varia el valor de D.

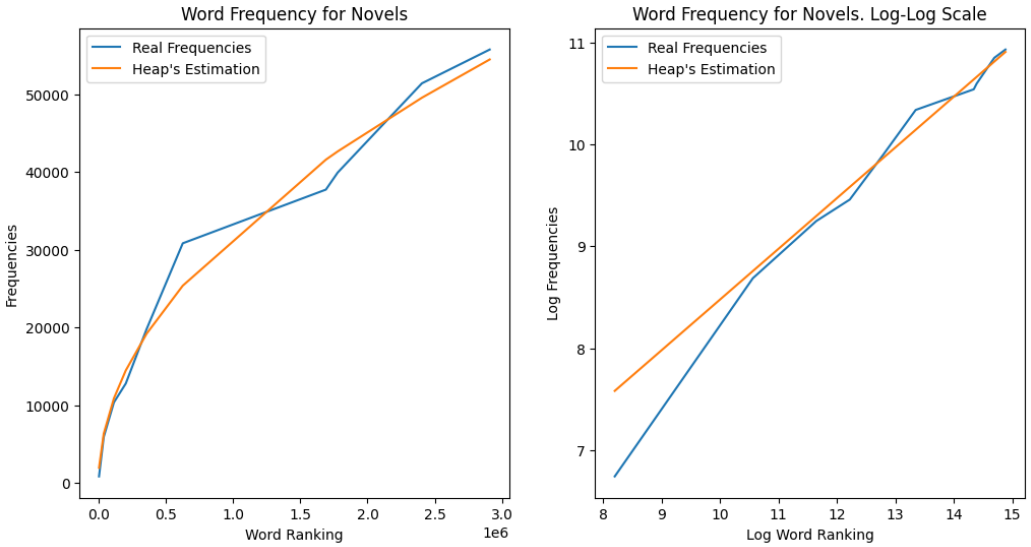
$$|subindex_1| < |subindex_2| < |subindex_3| ... < |subindex_k| \quad (3)$$

La manera de fer aquests subconjunts de novel·les $subindex_j$ ha sigut ordenar-ho en ordre ascendent i anar fent subconjunts de mides de documents ascendents, és a dir , primer 1 , després 2, 3 , 5 .... Fins a 33 documents. El cardinal d'aquest subconjunt, sigui $|subindex_j|$ ha estat escollit a mode de prova i error. Mètodes més elegants podrien haver sigut executats, però per manca de temps en l'experimentació acollirem aquesta solució que serà prou bona.

Seguirem la mateixa metodologia que abans per a calcular els valors òptims per ambdós constants i veure si realment s'ajusten bé i, per consegüent, la llei d'heap es confirma un cop més. Estudis anteriors (Cercar : Wikipedia Heap's Law) han demostrat que el rang típic d'aquestes constants és $0 \leq k \leq 100$ , $0.4 \leq \beta \leq 0.6$, per a documents anglosaxons. Escollirem com a valors inicials $k = 50$ i $\beta = 0.5$:

Table 2: Coeficients Heap

| **k** | $\beta$ |
|---|---|
| 33.214 | 0.497 |



# 3 Experiment 3: Preprocessat i ús de "Stopwords"

Cal comentar que en tot moment, s'ha usat un petit preprocessat mitjançant una expressió regular per així no tenir en compte paraules d'altres idiomes, amb caràcters especials o números. L'expressió, que ha sigut sobre simplificada per simplicitat i evitar tractar amb un nombre enorme d'excepcions i regles lingüístiques, és la següent:

$$\{a, b, \ldots, z, A, B, \ldots, Z, -, '\}^*$$

Conté guions (Que poden ser múltiples, en paraules com un-re-elected politician) i apòstrofs (Que també poden ser múltiples, com shouldn't've o fo'c'sle), però evidentment permet paraules d'altres idiomes com "hola" o bé paraules invàlides com "-'-gasda'd–dasd", però com que els articles científics, noticies i novel·les, típicament no els haurà escrit un gat caminant pel teclat sinó científics, redactors i escriptors, la probabilitat de trobar una paraula de l'estil és ínfima, llavors l'expressió anterior es suficient bona per al nostre propòsit.

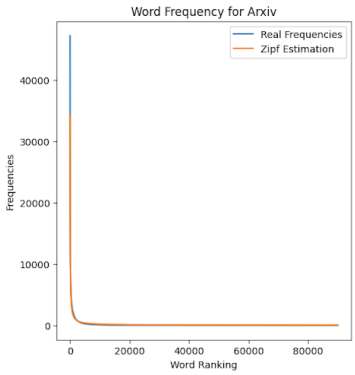
Seguidament, es mostra un experiment que ha estat eliminant les "StopWords" dels textos, és a dir, les paraules més freqüents que a pesar que uneixen les paraules i la llengua per donar coherència i sentit al text, manquen de sentit per relacionar texts similars i impliquen un gran soroll estadístic. Això ha estat aconseguit mitjançant la llibreria de tractament del llenguatge natural de python (NLTK), que conté preposicions, pronoms, verbs auxiliars, conjuncions, articles... típiques de l'anglès.

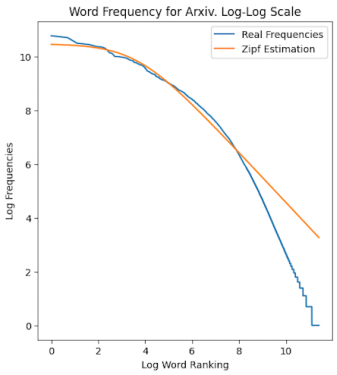
(a) Coeficients Zipf Sense Stopwords

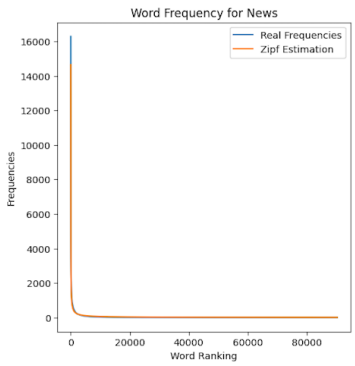
| | **a** | **b** | **c** |
|---|---|---|---|
| **Arxiv** | 0.930 | 38.931 | 1060294.323 |
| **Noticies** | 0.789 | 9.434 | 93262.934 |
| **Novel·les** | 0.791 | 9.458 | 55944.711 |

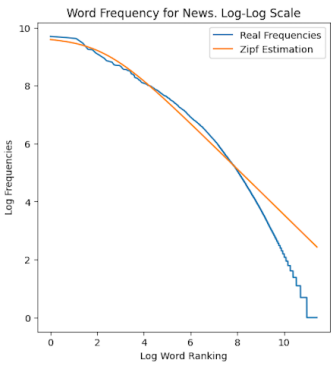
(b) Coeficients Heap Sense StopWords

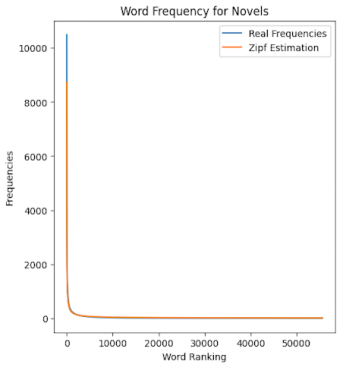
| **k** | $\beta$ |
|---|---|
| 45.718 | 0.497 |

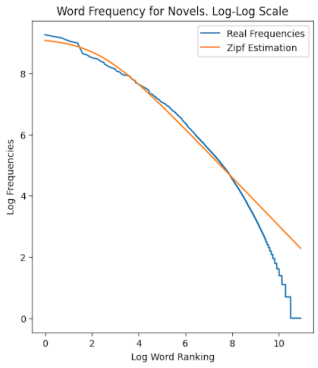

Word Frequency for Arxiv
Real Frequencies
Zipf Estimation
Frequencies
Word Ranking

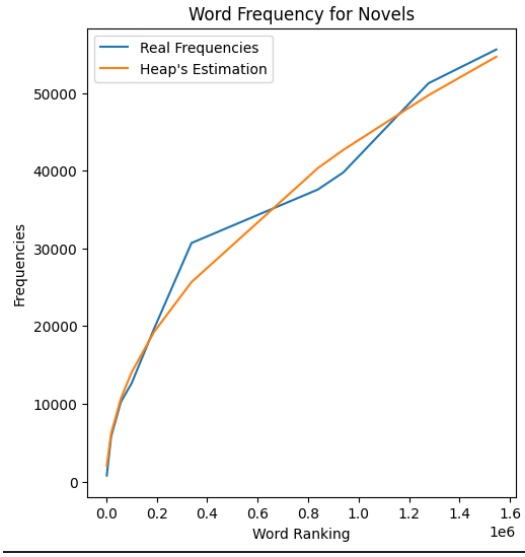

Word Frequency for Arxiv. Log-Log Scale
Real Frequencies
Zipf Estimation
Log Frequencies
Log Word Ranking

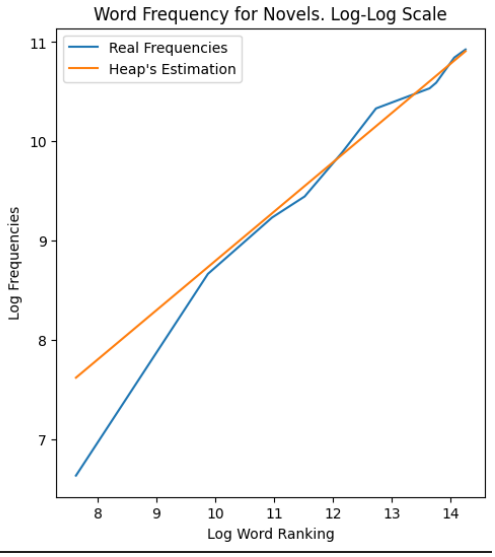

Word Frequency for News
Real Frequencies
Zipf Estimation
Frequencies
Word Ranking


Word Frequency for News. Log-Log Scale
Real Frequencies
Zipf Estimation
Log Frequencies
Log Word Ranking


Word Frequency for Novels
Real Frequencies
Zipf Estimation
Frequencies
Word Ranking


Word Frequency for Novels. Log-Log Scale
Real Frequencies
Zipf Estimation
Log Frequencies
Log Word Ranking


Word Frequency for Novels
Real Frequencies
Heap's Estimation
Frequencies
Word Ranking
1e6


Word Frequency for Novels. Log-Log Scale
Real Frequencies
Heap's Estimation
Log Frequencies
Log Word Ranking

# 4 Breus conclusions i Dificultats tècniques

Al llarg d'aquesta experimentació , diferents dificultats tècniques i decisions de disseny crítiques han anat apareixent, com l'elecció de com filtrar paraules mitjançant expressions regulars , que per limitació de temps i personal s'ha hagut de sobre simplificar. També ha sigut tot un repte que s'ha descartat per aquest experiment, l'avaluació estadística de si els models basats en llei de Zipf/Heap aconseguits mitjançant ajustos de corbes, són prou bons, quant és l'error. Primerament, es va intentar usar la mètrica $R^2$, després d'una bona reflexió em vaig adonar que aquesta mesura era inútil, ja que no es tracta d'un problema de regressió lineal. Finalment, vaig provar de fer proves d'hipòtesi, per veure si havien diferéncies significatives, amb la distribució $chi^2$, però vaig haver-la de descartar, ja que la distribució inicial em vaig adonar que no complia la premissa base i llavors era invàlida. Vaig arribar a la conclusío que mètriques més complexes han de ser emprades.

Un altre punt important, és adonar-se de com varien els diferents coeficients en el cas de Zipf un cop tretes les StopWords. Si mirem les gràfiques logarítmiques, semblen ajustar-se pitjor els models que no pas en el cas **amb** StopWords. Llavors potser aquestes paraules no són tant soroll estadístic com arribàvem a pensar inicialment. També cal adonar-se que pels 3 índexs les constants varien força, a causa de la naturalesa dels textos. Arxiv sempre té una (a) superior als altres, el que pot indicar una distribució més esbiaixada ("skewed") que les altres, probablement degut a factors com un nombre major de paraules amb freqüència baixíssima (Ja que hi ha paraules noves, tecnicismes o llatinismes poc habituals).

També cal esmentar que observant els gràfics d'Heap, es nota una irregularitat en el nombre de paraules que pot ser provocat per la poca quantitat de punts (x,y), o per la possible tria esbiaixada de subíndexs. Seria un bon punt a millorar si es repetís l'experiment.

Finalment, concloem que efectivament ambdues lleis són bons models empírics per descriure els comportaments dels textos, donat al bon ajust que hem observat, a pesar de que no s'hagi pogut quantificar.