Exercise 1 – IRRS

- CAROL AZPARRENT ESTELA
- MOHANNA FATHOLLAHI

Exercise 4

We have a document collection with a total of 10^6 term occurrences. Supposing that terms are distributed in the texts following a power law of the form

$$f_i \cong \frac{c}{(i+10)^2}$$

give estimates of (1) the number of occurrences of the most frequent term; (2) the number of occurrences of the 100-th most frequent term; (3) the number of words occurring more than 2 times. *Hint:* $\sum_{i=11}^{\infty} \frac{1}{i^2} \cong 0.095$.

N

$N = 10^6$ { total # of tokens

#1) $\underbrace{\text{the number of occurrence}}_{\text{\# of tokens}}$ of $\underbrace{\text{the most frequent term}}_{\text{rank} = 1}$

rank=1

$f(1) = \text{token}_1$

rank=2

$f(2) = \text{token}_2$

$\vdots$

rank=V

$f(V) = \text{token}_V$

$$\underbrace{\sum_{1}^{V} \text{token}}_{\sum_{r=1}^{V} f(r) = 10^6} = N = 10^6$$

$$f_i \simeq \frac{c}{(i+10)^2} \Rightarrow \sum_{r=1}^{V} \frac{c}{(r+10)^2} = 10^6$$

$$c * \sum_{r=1}^{V} \frac{1}{(r+10)^2} = 10^6$$

$$\boxed{\sum_{r=1}^{\infty} \frac{1}{(r+10)^2} = \sum_{r=11}^{\infty} \frac{1}{r^2} \simeq 0.095}$$

$$\Rightarrow c * 0.095 = 10^6$$

$$c = \frac{10^6}{0.095}$$

$$c = 1.05 * 10^7$$

Hint: $\sum_{i=11}^{\infty} \frac{1}{i^2} \cong 0.095$

$\hookrightarrow \quad f_i = \dfrac{1.05 * 10^7}{(i + 10)^2}$

rank = 1 $\longrightarrow$ $f(1) = \dfrac{1.05 * 10^7}{(1 + 10)^2} = 86842\underline{3}$

#2) $\underbrace{\text{the number of occurrences}}_{\#\text{ of tokens}}$ of $\underbrace{\text{100-th most frequent term}}_{\text{rank} = 100}$

$* \; f(100) = \text{token}_{100} \; *$

rank = 100 $\longrightarrow$ $f(100) = \dfrac{1.05 * 10^7}{(100 + 10)^2} = 8684$

#3) the number of words occurring more than 2 times.

$= \quad f(r) > 2$

$$f(r) > 2 \Rightarrow \frac{1.05 * 10^7}{(r + 10)^2} > 2$$

$$1.05 * 10^7 > 2 * (r+10)^2$$

$$\frac{1.05 * 10^7}{2} > (r + 10)^2$$

$$(r+10)^2 < \frac{1.05 * 10^7}{2}$$

$$r^2 + 20r - \frac{1.05 * 10^7}{2} < 0$$

$$r^2 + 20r - 5.25 * 10^6 < 0$$

$\Rightarrow r_1 = -2301.3$

$$\boxed{r_2 = 2281}$$

$\therefore$ the number of word occurring more than 2 times is 2281.

Exercise 5

We are given a random sample of 10,000 documents from a collection containing 1,000,000 documents. We count the different words in this sample, and we find 5,000. Supposing that the collection satisfies Heaps' law with exponent 0.5, give a reasoned estimate of the number of different words you expect to find in the whole collection.

$d = 5000$

$\beta = 0.5$

$$5000 = K * N_1^{0.5}$$

$$K = \frac{5000}{N_1^{0.5}}$$

total amount of documents : $d = \frac{5000}{N_1^{0.5}} * N_2^{0.5}$

* Assuming : $N_2 = 100 * N_1$

$$d = \frac{5000}{N_1^{0.5}} * (100 * N_1)^{0.5}$$

$$d = \frac{5000}{\cancel{N_1^{0.5}}} * 100^{0.5} * \cancel{N_1^{0.5}}$$

$$d = 5000 * \sqrt{100}$$

$$d = 50000$$

∴ there are 50000 amount of words in the whole collection approximately

Exercise 6

Let us deduce Heaps' law from Zipf's law.

- Let a collection have N word occurrences, with the frequence f_i of the i-th most common word proportional to $i^{-\alpha}$, $\alpha > 1$.
- Figure out (from previous exercises) the proportionality constant.
- Estimate the rank i such that f_i is likely to be less than 1.
- Explain why this should roughly be the number of distinct words we expect to see in the collection.
- Deduce that this number is $k \cdot N^{\beta}$. Tell the values of k and β as a function of α.

[Note: The given formulation of Zipf's law cannot, for obvious reasons, be taken too literally: If for some large i we have $c \cdot i^{-\alpha} = 0.03$, it makes no sense to say that the ith word appears 0.03 times in the collection. More abstractly, one could imagine texts generated by some random process which assigns probability $P(w)$ to the event that a random position in the text contains the word w. Then the word with rank 1 is the w with highest $P(w)$, etc. Zipf's law is a statement about the form of the probability distribution P. One can then compute rigorously the expected number of distinct words in a text of length N according to this probabilistic model. Let us just say that we this way we obtain the same β but a different k.]
[Note 2: It is also possible but a bit more involved to deduce a power law (generalizing Zipf's law) from Heap's law]

* $\alpha > 1 \rightarrow f(i) = \frac{c}{i^\alpha}$

* V: total # of distinct words

$$\sum_{i=1}^{V} f(i) = N$$

$$\sum_{i=1}^{V} \frac{c}{i^\alpha} = N$$

$$c * \sum_{i=1}^{V} \frac{1}{i^\alpha} = N$$

$$c = \frac{N}{\sum_{i=1}^{V} \frac{1}{i^\alpha}} \quad \zeta(\alpha)$$

$$c = \frac{N}{\zeta(\alpha)}$$

Zeta Riemann

$$\zeta(\alpha) \approx \sum_{i=1}^{V} \frac{1}{i^\alpha}$$

* $f(i) < 1$

$$\frac{N}{\zeta(\alpha)} * \frac{1}{i^\alpha} < 1$$

$$i^\alpha > \frac{N}{\zeta(\alpha)}$$

$$i > \left(\frac{N}{\zeta(\alpha)}\right)^{\frac{1}{\alpha}}$$

the words that have a rank "i" higher than $\left(\frac{N}{\zeta(\alpha)}\right)^{\frac{1}{\alpha}}$ have frequency $f(i)$ greater than 1.

$\Downarrow$

these words appear more than 1 time at the collection

↳ i : the number of distinct word of the collection. and in Heap's law is equal to d.

$$d \approx \left(\frac{N}{\zeta(\alpha)}\right)^{\frac{1}{\alpha}}$$

$$d \approx \frac{N^{1/\alpha}}{\zeta(\alpha)^{1/\alpha}}$$

$K = \dfrac{1}{\zeta(\alpha)^{1/\alpha}}$

$\beta = \dfrac{1}{\alpha}$

$d = K * N^\beta$ } Heaps' law