CSYS 300 PoCS Assignment 5

Thomas Waters

October 5th 2018

Code is located at https://github.com/Evelios/PoCS_Assignment_05

1 Problem 1

This problem is about evaluating a diversity index D, which is easier to intuit about than the entropy number. This diversity index is supposed to represent a different hypothetical text T' which shares the same entropy H as the origional text T. The text T' has all the words appearing in equal frequency $1/D$, so there is $p_i = 1/D$ for $i = 1, ..., D$. This means that we set $H = H'$, substituting the p_i from T with $p'_i = 1/D$ in H' from T'. Also remembering that the sum of the text is from $i = 1$ to D.

Simpson Concentration

$$S = \sum_{i=1}^{n} p_i^2 \tag{1}$$

$$S = S' \tag{2}$$

$$\sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{D} (1/D)^2 \tag{3}$$

$$\sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{D} (1/D)^2 \tag{4}$$

$$\sum_{i=1}^{n} p_i^2 = D(1/D)^2 \tag{5}$$

$$\sum_{i=1}^{n} p_i^2 = (1/D) \tag{6}$$

$$D = \left(\sum_{i=1}^{n} p_i^2 \right)^{-1} = S^{-1} \tag{7}$$

Gini Index

$$G \equiv 1 - S = 1 - \sum_{i=1}^{n} p_i^2 \tag{8}$$

$$G = G' \tag{9}$$

$$1 - \sum_{i=1}^{n} p_i^2 = 1 - \sum_{i=1}^{n} (1/D)^2 \tag{10}$$

$$\sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{D} (1/D)^2 \tag{11}$$

$$\sum_{i=1}^{n} p_i^2 = D(1/D)^2 \tag{12}$$

$$\sum_{i=1}^{n} p_i^2 = (1/D) \tag{13}$$

$$D = \left(\sum_{i=1}^{n} p_i^2 \right)^{-1} = S^{-1} \tag{14}$$

The diversity index for the Simpson Concentration and the Gini Index come out to be the same! The diversity index for the Gini Index can even be written in terms of the Simpson Concentration.

Shannon's Entropy

$$H = -\sum_{i=1}^{n} p_i \ln p_i \tag{15}$$

$$H = H' \tag{16}$$

$$H = -\sum_{i=1}^{D} (1/D) \ln(1/D) \tag{17}$$

$$H = -D(1/D) \ln(1/D) \tag{18}$$

$$H = -\ln(1/D) = \ln D \tag{19}$$

$$D = e^{H} = e^{-\sum_{i=1}^{n} p_i \ln p_i} \tag{20}$$

Renya Entropy, $q \neq 1$

$$H_q^{(R)} = \frac{1}{q-1}\left(-\ln\sum_{i=1}^{n} p_i^q\right) \tag{21}$$

$$H_q^{(R)} = (H_q^{(R)})'$$

$$\frac{1}{q-1}\left(-\ln\sum_{i=1}^{n} p_i^q\right) = \frac{1}{q-1}\left(-\ln\sum_{i=1}^{D}(1/D)^q\right) \tag{22}$$

$$-\ln\sum_{i=1}^{n} p_i^q = -\ln\sum_{i=1}^{D}(1/D)^q \tag{23}$$

$$\sum_{i=1}^{n} p_i^q = \sum_{i=1}^{D}(1/D)^q \tag{24}$$

$$\sum_{i=1}^{n} p_i^q = D(D)^{-q} \tag{25}$$

$$\sum_{i=1}^{n} p_i^q = D^{q+1} \tag{26}$$

$$D = \left(\sum_{i=1}^{n} p_i^q\right)^{\frac{1}{q+1}} \tag{27}$$

Generalized Tsallis Entropy, $q \neq 1$

$$H_q^{(T)} = \frac{1}{q-1}\left(1-\sum_{i=1}^{n} p_i^q\right) \tag{28}$$

$$H_q^{(T)} = (H_q^{(T)})'$$

$$\frac{1}{q-1}\left(1-\sum_{i=1}^{n} p_i^q\right) = \frac{1}{q-1}\left(1-\sum_{i=1}^{D}(1/D)^q\right) \tag{29}$$

$$1-\sum_{i=1}^{n} p_i^q = 1-\sum_{i=1}^{D}(1/D)^q \tag{30}$$

$$\sum_{i=1}^{n} p_i^q = \sum_{i=1}^{D} (1/D)^q \tag{31}$$

$$\sum_{i=1}^{n} p_i^q = \sum_{i=1}^{D} (1/D)^q \tag{32}$$

$$\sum_{i=1}^{n} p_i^q = D(D)^{-q} \tag{33}$$

$$\sum_{i=1}^{n} p_i^q = D^{q+1} \tag{34}$$

$$D = \left(\sum_{i=1}^{n} p_i^q \right)^{\frac{1}{q+1}} \tag{35}$$

We see that the Generalized Tallis Entropy and the Renya Entropy come out to be the same!

Matching the Tallis Diversity as $q \to 1$ to Shannon Diversity

$$D_H = e^H = e^{-\sum_{i=1}^{n} p_i \ln p_i} \tag{36}$$

$$D_T = \left(\sum_{i=1}^{n} p_i^q \right)^{\frac{1}{q+1}} \tag{37}$$

$$\lim_{q \to 1} D_T = \lim_{q \to 1} \left(\sum_{i=1}^{n} p_i^q \right)^{\frac{1}{q+1}} \tag{38}$$

$$\lim_{q \to 1} D_T == \left(\sum_{i=1}^{n} p_i \right)^{1/2} \tag{39}$$

Hmmm, doesn't seem quite right

2 Problem 2

Mandelbrotian derivation of Zipf's law by minimizing the function

$$\Psi(p_1, p_2, ..., p_n) = F(p_1, p_2, ..., p_n) + \lambda G(p_1, p_2, ..., p_n) \tag{40}$$

The 'cost over information' is,

$$F(p_1, p_2, ..., p_n) = \frac{C}{H} = \frac{\sum_{i=1}^{n} p_i ln(i+a)}{-g \sum_{i=1}^{n} p_i \ln p_i} \tag{41}$$

The constrain equation is,

$$G(p_1, p_2, ..., p_n) = \sum_{i=1}^{n} p_i - 1 \;\; (= 0) \tag{42}$$

Minimizing the cost over information...

$$\nabla F(p_1, p_2, ..., p_n) = \nabla \frac{\sum_{i=1}^{n} p_i ln(i+a)}{-g \sum_{i=1}^{n} p_i \ln p_i} \tag{43}$$

We can look at the gradient component wise, because when taking the partial derivative of the summation, all other terms other then the $i = j$ terms cancel out for that particular vector entry.

$$\frac{\partial F}{\partial p_j} = \frac{\partial}{\partial p_j} \frac{p_j ln(i+a)}{-g(p_j \;\, lnp_j)} \tag{44}$$

$$= \frac{\partial}{\partial p_j} \frac{ln(i+a)}{-g} \frac{p_j}{(p_j \;\, lnp_j)} \tag{45}$$

$$= \frac{\partial}{\partial p_j} \left(\frac{ln(i+a)}{-g} \right) \left(\frac{1}{lnp_j} \right) \tag{46}$$

$$= \left(\frac{ln(i+a)}{-g} \right) \left(\frac{1}{p_j (lnp_j)^2} \right) \tag{47}$$

Solving for the gradient of the constrain function...

$$\nabla G(p_1, p_2, ..., p_n) = \nabla \sum_{i=1}^{n} p_i - 1 \tag{48}$$

$$\frac{\partial G}{\partial p_j} = \frac{\partial}{\partial p_j} (p_i - 1) \tag{49}$$

$$\frac{\partial G}{\partial p_j} = 1 \tag{50}$$

Solving for the critical point,

$$\frac{\partial \Psi}{\partial p_j} = \frac{\partial F}{\partial p_j} + \lambda \frac{\partial G}{\partial p_j} = 0 \tag{51}$$

$$0 = \left(\frac{ln(i+a)}{-g}\right)\left(\frac{1}{p_j(lnp_j)^2}\right) + \lambda \tag{52}$$

$$\left(\frac{ln(i+a)}{-g}\right)\left(\frac{1}{p_j(lnp_j)^2}\right) = \lambda \tag{53}$$

$$\frac{ln(i+a)}{-g\lambda} = p_j(lnp_j)^2 \tag{54}$$

I should have reached the following equation, so working on from here.

$$p_j = e^{-1-\lambda H^2/gC}(j+a)^{-H/gC} \tag{55}$$

Using the constrain equation,

$$\sum_{j=1}^{n} p_j = \sum_{j=1}^{n} e^{-1-\lambda H^2/gC}(j+a)^{-H/gC} = 1 \tag{56}$$

$$\left(e^{-1-\lambda H^2/gC}\right)\sum_{j=1}^{n}(j+a)^{-H/gC} = 1 \tag{57}$$

$$\sum_{j=1}^{n}(j+a)^{-H/gC} = e^{1+\lambda H^2/gC} \tag{58}$$

Expected Output

$$\alpha = H/gC \tag{59}$$

$$p_j = (j+a)^{\alpha} \tag{60}$$

Solving for λ. I am using the equations for p_j, H, and C below,

$$C = \sum_{j=1}^{n} p_i ln(j+a) \tag{61}$$

$$H = -g\sum_{j=1}^{n} p_j \ln p_j \tag{62}$$

$$p_j = e^{-1-\lambda H^2/gC}(j+a)^{-H/gC} \tag{63}$$

$$ln\, p_j = (-1-\lambda H^2/gC)(-H/gC)ln(j+a) \tag{64}$$

The rest follows by starting with the equation for H

$$H = -g\sum_{j=1}^{n} p_j \ln p_j \tag{65}$$

$$H = -g\sum_{j=1}^{n} p_j(-1-\lambda H^2/gC)(-H/gC)ln(j+a) \tag{66}$$

$$H = -g(-1-\lambda H^2/gC)(-H/gC)\sum_{j=1}^{n} p_i ln(j+a) \tag{67}$$

$$H = -g(1+\lambda H^2/gC)(H/gC)(C) \tag{68}$$

$$1 = -(1+\lambda H^2/gC) \tag{69}$$

$$\lambda H^2/gC = -2 \tag{70}$$

$$\lambda = \frac{-2gC}{H^2} \tag{71}$$

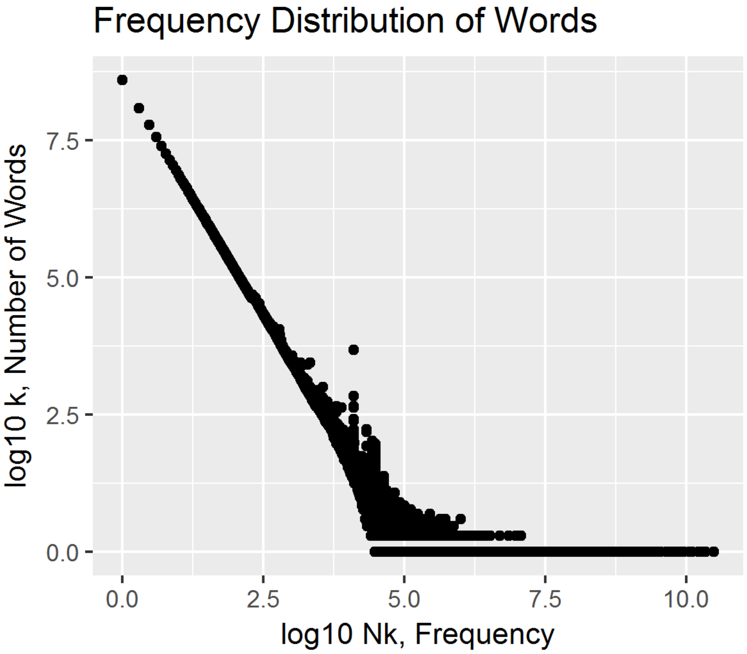
3 Problem 3

4 Problem 4

$$N_{\leq 200} \approx 3.46 \times 10^8 K^- 0.661 \tag{72}$$

From my graphs from homework 2, I got a value of γ and trying to retrofit the values for $1 \geq k \geq 199$ I needed to change the slope to $\gamma = 1.73$ as per the value of my home work. I don't know how I could get a different value here, but the graph seems to line up closely with this slope adjustment. Unfortuantely this throws off all the other calculations.

From my graph, I got a mean $\mu = 3,355$, standard deviation $\sigma = 870,661$, and a varaince $\sigma^2 = 758,050,650,139$.


Frequency Distribution of Words
log10 k, Number of Words
log10 Nk, Frequency