PRINCIPLES OF COMPLEX SYSTEMS

ASSIGNMENT 5

January 6, 2023

Maxfield Green
University of Vermont
Comp. Systems and Data Sci.

0.1 Problem 1

Using the framework of the Entropy and Diversity described in the problem statement. Determine the diversity D in terms of the probabilities p_i for the following:

(a) Simpson concentration:

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

After some basic algebra I arrive at: (2)

$$D = \sqrt{\frac{1}{\sum_{i=1}^{n} P_i^2}} \tag{3}$$

(b) Gini Index:

$$G = 1 - S = \sum_{i=1}^{n} P_i^2 \tag{4}$$

$$D = \frac{1}{S} = \frac{1}{P_i^2} \tag{5}$$

(c) Shannon's entropy:

$$H = -\sum_{i=1}^{n} (\frac{1}{D}) ln \frac{1}{D} \tag{6}$$

(7)

Because the sum does not depend on i, the summation can be replaced with multiplication:

$$D = e^{-\sum_{i=1}^{n} p_i ln(p_i)} \tag{8}$$

$$D = e^{H} \tag{9}$$

(d) Renyi Entropy:

$$H_q^R = \frac{1}{q-1}(-ln\sum_{i=1}^{n} p_i^q) \tag{10}$$

$$H_q^R = \frac{1}{q-1}(-(1-q)(-ln(D)) \tag{11}$$

$$H_q^R = (-ln(D)) \tag{12}$$

$$(-ln(D)) = \frac{1}{q-1}(-ln\sum_{i=1}^{n} p_i^q) \tag{13}$$

$$D = \sum_{i=1}^{n} p_i^q \frac{1}{1-q} \tag{14}$$

(e) The generalized Tsallis entropy:

$$H_q^T = \frac{1}{1-q}(1-\sum_{i=1}^{n} p_i^q) \tag{15}$$

$$H_q^T = \frac{1}{1-q}(1-\sum_{i=1}^{n} (\frac{1}{D})^q) \tag{16}$$

$$H_q^T = \frac{1}{1-q}(1-D(\frac{1}{D})^q) \tag{17}$$

$$D = \sum_{i=1}^{n} (p_i^q)^{\frac{1}{1-q}} \tag{18}$$

(f) Show that in the limit $q->\infty$. the diversity for the Tsallis entropy matches up with Shannon's entropy.

In the limit as q-> 1, D_q -> D. The diversity of the Tsallis entropy approaches the Shannon's.

0.2 Problem 2

Complete the Mandelbrotian derivation of Zipf's law by minimizing the function:

$$\Psi(p_1,p_2,...,p_n) = F(p_1,p_2,...,p_n) + \lambda G(p_1,p_2,...,p_n) \tag{19}$$

where the 'cost over information' function is

$$F(p_1,p_2,...,p_n) = \frac{C}{H} = \frac{\sum_{i=1}^{n} p_i ln(i+a)}{-g\sum_{i=1}^{n} p_i ln(pi)} \tag{20}$$

and the constraint function is

$$G(p_1,p_2,...,p_n) = \sum_{i=1}^{n} p_i - 1(=0) \tag{21}$$

to find

$$p_j = e^{-1-H^2/gC}(j+a)^{-H/gC} \tag{22}$$

Then use the constraint equation to show that

$$p_j = (j+a)^{-\alpha} \tag{23}$$

where $\alpha = H/gC$.

We will start this beautiful monstrosity with minimization of Ψ.

$$\Psi = F + \lambda G \tag{24}$$

$$\frac{d\Psi}{dP_j} = 0 \tag{25}$$

$$\frac{dP_i}{dF} = \lambda \frac{dP_i}{dG} \tag{26}$$

$$\frac{Hln(j+a) + Cg(ln(P_j)+1}{H^2} = -\lambda(1) \tag{27}$$

$$e^{Hln(j+a)} + e^{Cg(ln(P_j)+1} = e^{-\lambda H^2} \tag{28}$$

$$(j+a)^H p_j^{Cg} e^{Cg} = e^{-\lambda H^2} \tag{29}$$

$$(j+a)^{\frac{H}{Cg}} P_j e = e^{\frac{-\lambda H^2}{Cg}} \tag{30}$$

$$P_j = e^{\frac{-\lambda H^2}{Cg}} (j+a)^{\frac{-H}{Cg}} e^{-1} \tag{31}$$

$$P_j = e^{\frac{-\lambda H^2}{Cg} - 1} (j+a)^{\frac{-H}{Cg}} \tag{32}$$

Now, we need to deal with λ. We'll solve for λ through H, and plug it back into P_j to arrive at our ultimate minimization of Ψ, showing Zipfs Law from a new angle.

$$\lambda = \frac{Hln(j+a) + Cg(ln(P_j)+1}{H^2} \tag{33}$$

$$\lambda = \frac{Hln(j+a) - Cg(ln(P_j)+1}{H^2} \tag{34}$$

$$\tag{35}$$

Returning to P_j

$$P_j = \frac{Hln(j+a) - Cg(ln(P_j)+1}{HC} (j+a)^{-\alpha} \tag{36}$$

$$P_j = [\frac{-ln(j+a)}{c}][(j+a)^{-\alpha}] \tag{37}$$

$$p_j = \frac{ln(j+a)}{ln(j+a)} (j+a)^{-\alpha} \tag{38}$$

$$P_j = (j+a)^{-\alpha} \tag{39}$$

0.3 Problem 3

Carrying on from the previous problem. We will show how $\alpha \approx 1.73$.

To achieve this, I took a novel approach and showed how the cumulative sum as n –> ∞ differs for different α. We know that the α whose cumulative sum converges to 1 is ≈ 1.73.

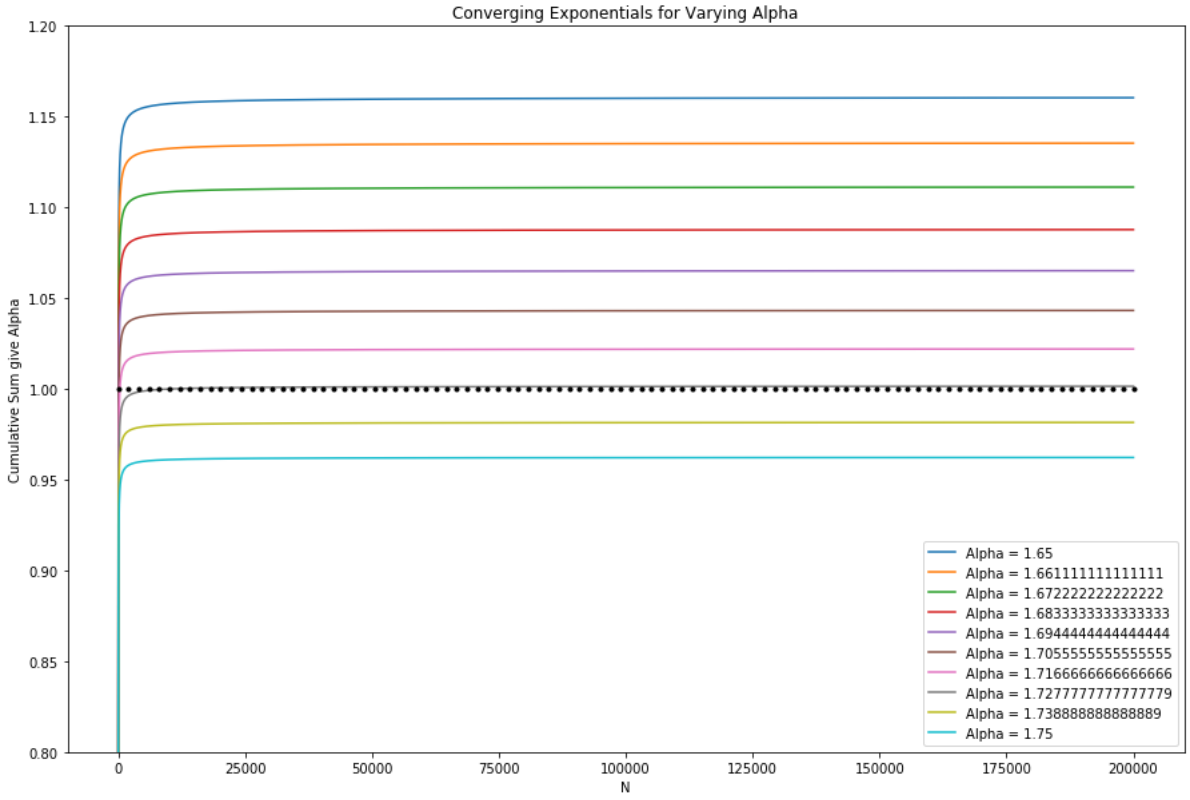

Figure 1: Convergence given select values of Alpha

0.4 Problem 4

Estimate the rare words that are missing from the corpus.

$$N_{\geq k} \approx 3.46X10^8 k^{-0.661} \tag{40}$$

(a) Using the above fit, create a complete hypothetical N_k by expanding N_k back for k = 1 to k = 199, and plot the result in double-log space.

Instead of plotting the full Nk, I have plotted the complete hypothetical distributions CCDF. It looks smoother and the data line up more closely.

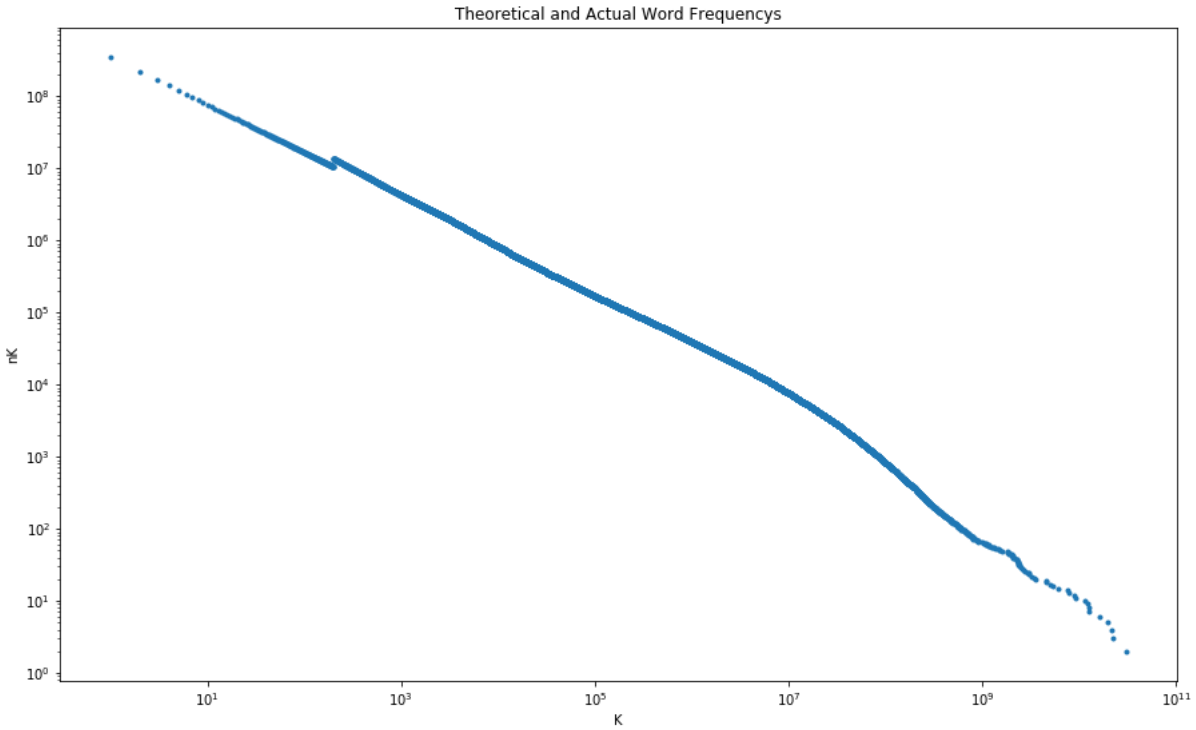

Figure 2

(b) Compute the mean and variance of this reconstructed distribution.
After finding the theoretic distribution of nK's from the theoretic CCDF, finding the mean and variance was easy.

mean = 1462

variance = 86077037994.29

(c) Estimate:

i. the hypothetical fraction of words that appear once out of all words.

We can find this value easily given the CCDF of the reconstructed distribution. The difference between the 1st and 2nd values divided by the 1st value of the distributions CCDF give us the this particular. I'm getting that 36.76%

ii. the hypothetical total number and fraction of unique words in Google's data set.

$$\frac{\sum nK}{\sum [k * nK]} = .1 \tag{41}$$

iii. what fraction of total words are left out of the Google data set by providing only those with counts $k \geq 200$.

$$\frac{\sum nK_{k \leq 200}}{\sum nK_{reconstructed}} = 0.9698 \tag{42}$$