Estimating the number of subdomains on a given domain

Stanley Cortes de Sousa

August 1, 2020

Let X be the set of all words $[a-z](k)$, where k is the maximum number of characters. Consider Y the subset of X where every Y_i is a valid host http://Y_i.domain

Let W be another subset of X, where $W_i = f(j)$. Let j be a number on base 26 with at most k digits, $f(j)$ maps j to a word from X with at most k characters. Given the table 0=a ... 25=z, each digit of j is mapped by f on a character, hence:

i_{10}	j_{26}	$W_i = f(j)$
0	0	a
11	B	k
52	20	ba
64	$2C$	bl
178607	$A45D$	jefm
456975	$PPPP$	zzzz

Note that the numerical representation of i and j may discard '0's to the left, hence W would have the word 'a' but not sequences with 'a' to the left, therefore:
$|X| = 26^k$
$|W| = 26^k - 26^{k-1} ... - 26^0$

However, consider n words of X. If $n < |W|$ then we can use n samples of W to represent samples of X since $W \in X$

Finally, let H be the fraction of words in X which are valid subdomains, H is given by:
$H = \frac{|Y|}{|X|}$ (I)

Consider Z a sequence of random variables uniform in $[0, 1]$, and let $g(z) = i$:
$g(z) = int(max_i * z) + 1$
$j = i_{26}$
$W_i = f(j)$

Let $h(W_i)$ be the indicator function which tells us if the word W_i is in Y, that is, $h(W_i)$ checks if http://W_i.domain is a valid host

From (I) we know that:
$E[h(X_i)] = \frac{|Y|}{|X|}$, assuming that $n < |W| \longmapsto E[h(X_i)] = E[h(W_i)]$

We need to estimate $E[h(X_i)]$, the law of the large numbers tells us that:
$M_n \to E[h(X_i)] = \frac{|Y|}{|X|}$, and by definition
$M_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$

Therefore:
$\frac{|Y|}{|X|} \to \frac{1}{n} \sum_{i=1}^{n} h(X_i)$
$|Y| \to \frac{|X|}{n} \sum_{i=1}^{n} h(X_i)$