JANUARY 14, 2015

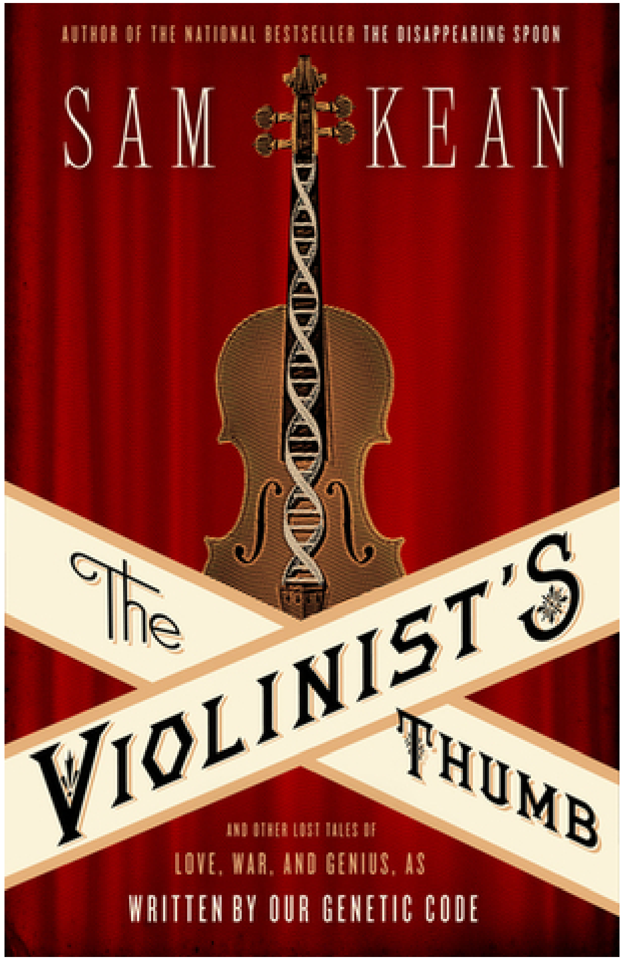
Computer programs and Zipf's law

Notice

> *I have shifted this blog to http://lepisma.github.io/blog. This place is no longer maintained. The link for current article is http://lepisma.github.io/blog/articles/computer-programs-and-zipf-law/*

So, I was reading a book on genomics named **The Violinist's Thumb** when I got the idea behind this post. The basic idea of the book is *how our lives are written by our genetic code*. Along with the **connectome**, I believe the idea defines the whole life of an individual.

26
KUDOS


AUTHOR OF THE NATIONAL BESTSELLER THE DISAPPEARING SPOON
SAM KEAN
The
VIOLINIST'S
THUMB
AND OTHER LOST TALES OF
LOVE, WAR, AND GENIUS, AS
WRITTEN BY OUR GENETIC CODE

Meanwhile in the reading, the book showed me a beautiful empirical law. The **Zipf's law**. When George Kingsley Zipf was playing with classic literary works, he discovered a pattern in the frequency distribution of words.

> - The **second** most used word has **half** the frequency as compared to first most frequent word.

26
KUDOS

> - The **third** most used word has **one third** the frequency as compared to the first most frequent word and so on…

This is Zipf's law. Now the amazing bit. Since then, the Zipfian (the frequency distribution following Zipf's law) distribution has been seen in music, city population, mass extinctions, earthquakes and even DNA.

Originally, the explanation Zipf gave for the phenomenon was the Principle of least effort. According to Zipf, the principle governs the transfer of information via the path of least effort and this explains the relation between *variations* and *repetitions*.

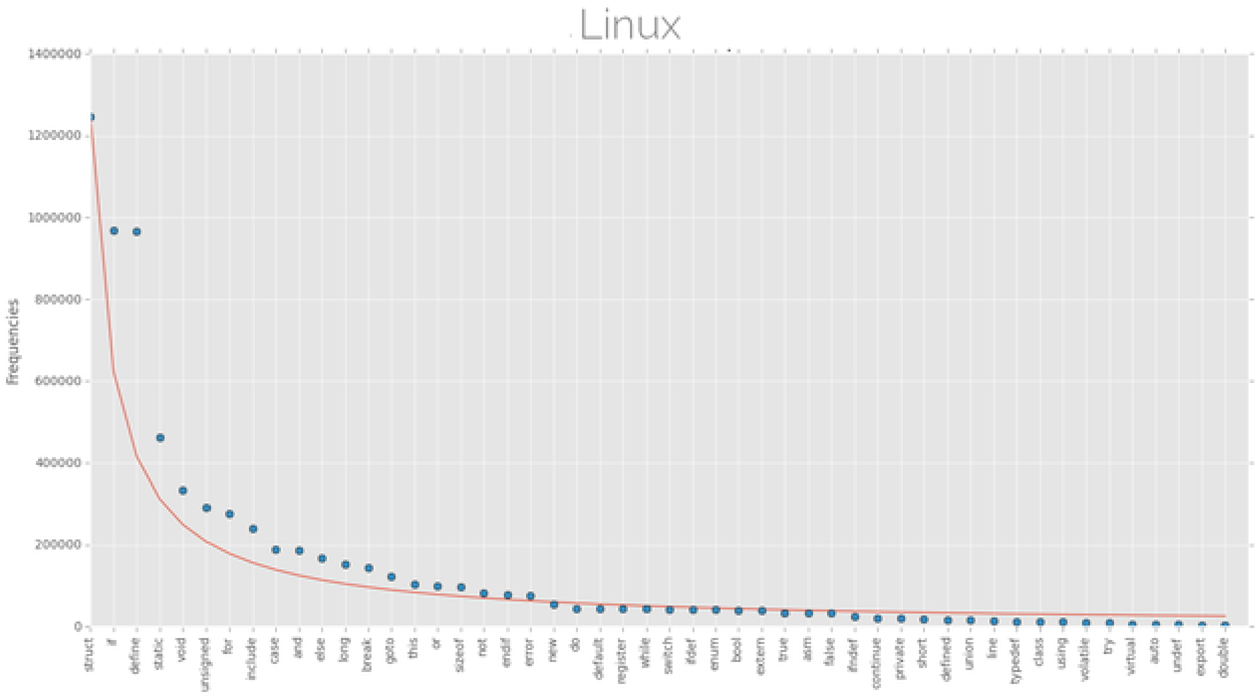
Seeing such a pattern in human languages led me to search if this works for computer languages. Although I didn't expect anything, since the structure of human and computer languages aren't exactly same, but I gave it a try. Similar to words in human language, I took **keywords** in computer languages, since variable names and other things aren't as frequent as keywords. Saying that, I do realize that this is not exactly the way this should be done, but it should work as a crude approximation.

Below are the plots of keyword frequencies (few top keywords) for various software projects along with the approx Zipfian distribution. The projects here are dominated by (almost) one single language and are large projects as far as lines of code are considered, this *might* be better for finding such patterns.

26
KUDOS

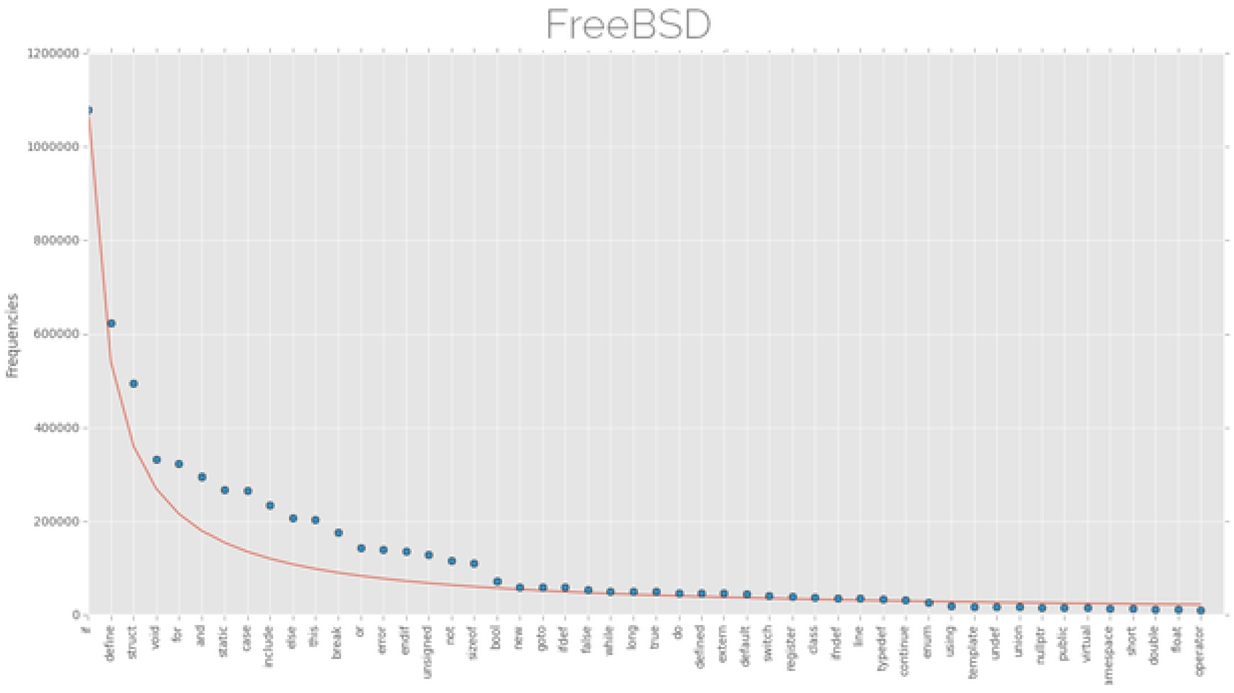
Linux kernel

The language of the project is predominantly C. I ran the frequency analysis for C/C++ language keywords.



FreeBSD

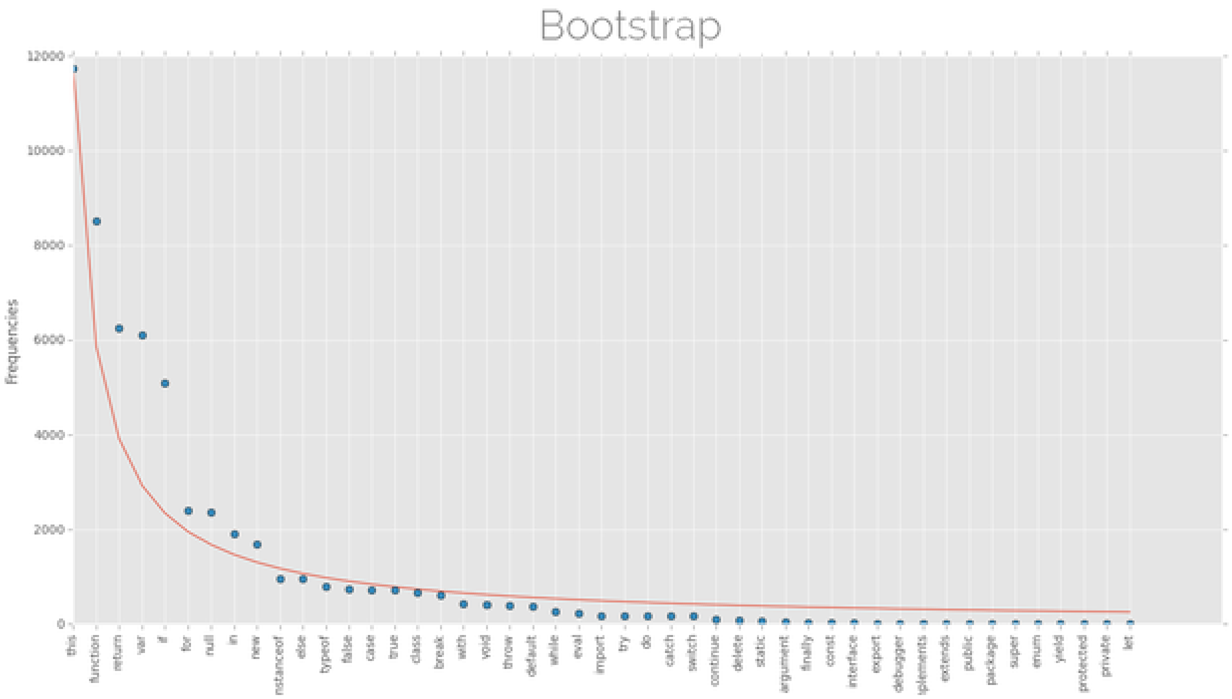
The language again is mostly C/C++, did the same analysis here.



Bootstrap

Javascript

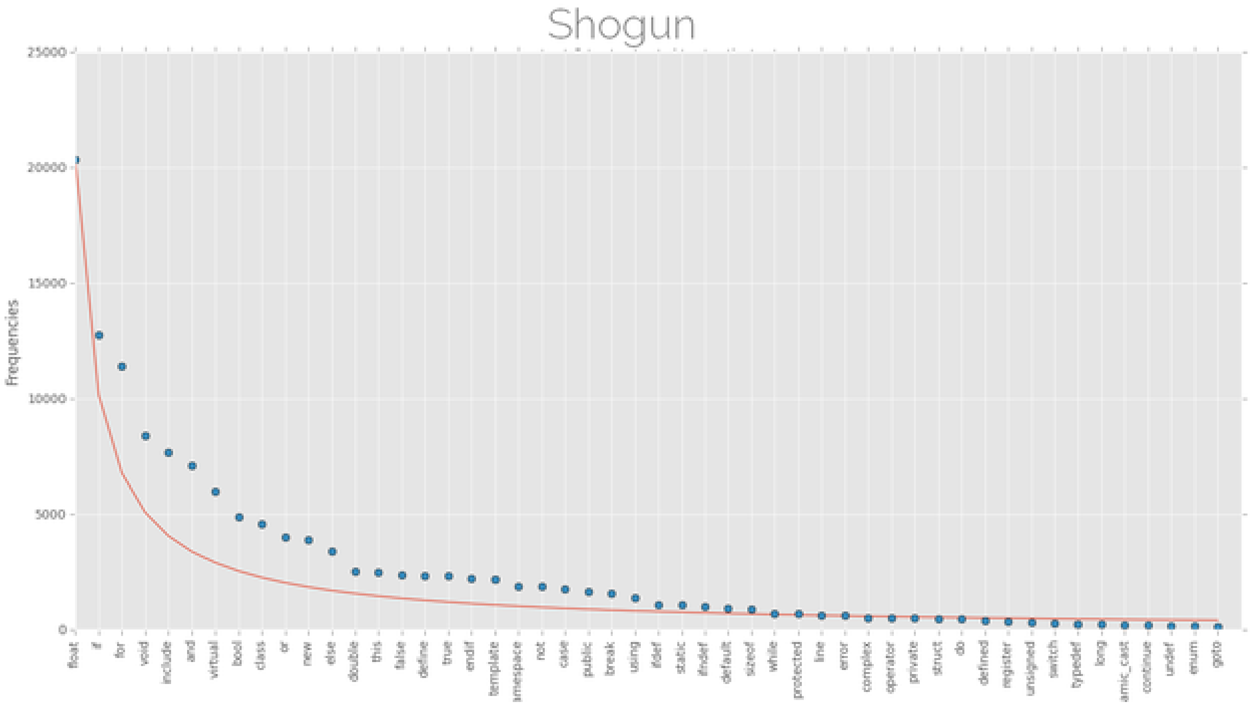
26
KUDOS



Shogun

C++

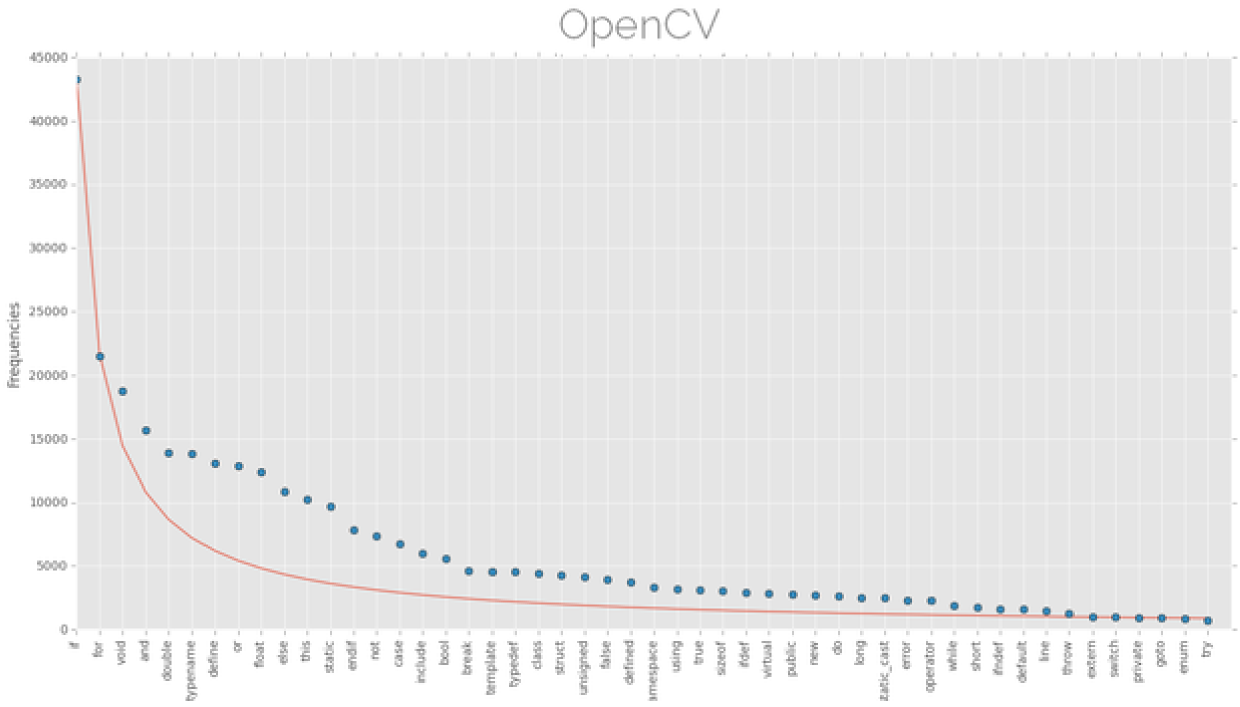
26
KUDOS



OpenCV

C++

26
KUDOS



Ah! I don’t know if I can declare this, but the distributions look nice. The law of least effort or anything else. Although there is high probability that these are not representatives of Zipf’s law, but nevertheless, its fun to see something like this.

26
KUDOS

A note on the script I used

The script is here and is just a "kind-of-works" thing.
A more rigorous analysis can be done, since the script neglected many things.

There are many power laws in nature that tend to line up observations in a specific order. Individually, things don't show up, but on a **crowdish** scale, the patterns begin to emerge, consider population, species, social networks, crowd behavior, star clusters, sub atomic particles … everywhere.

This feels like toy mathematics, like knot theory, but there are serious implications, and there maybe more. Here is an excerpt from the Wikipedia article on long tail

> *In his Wired article, Chris Anderson cites earlier research [17] by Erik Brynjolfsson, Yu (Jeffrey) Hu, and Michael D. Smith, who first used a log-linear curve on an XY graph to describe the relationship between Amazon.com sales and sales ranking. They found that a large proportion of Amazon.com's book sales come from obscure books that were not available in brick-and-mortar stores.*
> *They then quantified the potential value of the long tail to consumers. In*

26
KUDOS

> *an article published in 2003, these authors showed that, while most of the discussion about the value of the Internet to consumers has revolved around lower prices, consumer benefit (a.k.a. consumer surplus) from access to increased product variety in online book stores is ten times larger than their benefit from access to lower prices online. Thus, the primary value of the internet to consumers comes from releasing new sources of value by providing access to products in the long tail.*

Anyways, the exciting thing about living in current time is - We have data. Much more than what Zipf had.

◉ **26** KUDOS

NOW READ THIS

Why I stopped recommending GNOME ?

Notice I have shifted this blog to http://lepisma.github.io/blog. This place is no longer maintained. The link for current article is http://lepisma.github.io/blog/articles/why-i-stopped-recommending-gnome/ “In the beginning was GNOME”... Continue →

◉ **26** KUDOS

26
KUDOS