Hello again! Typically I would start by describing a complicated problem that can be solved using machine or deep learning methods, but today I want to do something different, I want to show you some interesting probabilistic phenomena!

Have you heard of **Zipf’s law**? I hadn’t until recently. Zipf’s law is an empirical law that states that many different datasets found in nature can be described using Zipf’s distribution. Most notably, word frequencies in books, documents and even languages can be described in this way. Simplified, Zipf’s law states that if we take a document, book or any collection of words and then the how many times each word occurs, their frequencies will be very similar to Zipf’s distribution. Let’s say that the number of occurrences of the most frequently occurring word is:

X

Zipf’s law states that the number of occurrences of the second most frequently occurring word will be equal to:

X/2

So basically this word will occur half of the number of times the most frequent word did. The number of occurrences of the third most frequently occurring word would be:

X/3

And so on … So the number of occurrences of the Nth most frequent word would be:

X/N

Most recent studies of this phenomena show that in the case of words, typically there is the same value of 𝞪, and the frequency on Nth word is described as:

X/N𝞪

To check the theory I downloaded a set of the 50,000 most frequent Polish words in subtitles (<https://github.com/hermitdave/FrequencyWords/blob/master/content/2016/pl/pl_50k.txt>) from *OpenSubtitles.org*. Here’s a visualization of real and theoretical frequencies.

To see it more clearly we can use logarithmic scales.

Try it out yourself: a list of example datasets can be found here: <https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists>

You can use this example code to create a similar visualization:

library(ggplot2)

library(dplyr)

library(themes)

library(gganimate)

word\_count <- # Data frame containing words and their frequency

colnames(word\_count) <- c("word", "count")

alpha <- 1 # Change it needed

word\_count <- word\_count %>%

mutate(word = factor(word, levels = word),

rank = row\_number(),

zipfs\_freq = ifelse(rank == 1, count, dplyr::first(count) / rank^alpha))

zipfs\_plot <- ggplot(word\_count, aes(x = rank, y = count)) +

geom\_point(aes(color = "observed")) +

theme\_bw() +

geom\_point(aes(y = zipfs\_freq, color = "theoretical")) +

transition\_reveal(count, rank) +

labs(x = "rank", y = "count", title = "Zipf's law visualization") +

scale\_colour\_manual(name = "Word count", values=c("theoretical" = "red", "observed" = "black")) +

theme(legend.position = "top")

zipfs\_animation <- animate(p)

This experiment is amazing, because language is very complicated: words in text are not random in any sense, and they depend on the previous ones. That’s why it’s so surprising to see such patterns here. We should always remember that the world can astonish us in many different ways! See you next time