Source Code Volatility (SCV) to Spot Dead Code

Yegor Bugayenko

April 14, 2020

1 Introduction

Volatility of source code is an experimental metric that shows how big is the difference between actively and rarely changed (possibly dead) code. It is assumed that a big percentage of dead code is an indicator of maintainability problems in the project.

2 Details

First, by looking at Git history, it is observed how many times every source code file i out of N was touched during the lifetime of the repository:

$$T = [t_1, t_2, \dots, t_N] \tag{1}$$

Then, t that relate to the files already absent in the repository are deleted and the array T is "normalized" to keep all values within the $[0, 1]$ range:

$$X = [x_1, x_2, \dots, x_M], \quad \text{where } x_i = \frac{t_i - \hat{T}}{\check{T} - \hat{T}} \tag{2}$$

Then, the mean μ is calculated as:

$$\mu = \frac{1}{M} \sum x_i \tag{3}$$

Finally, the variance is calculated as:

$$Var(x) = \frac{1}{M} \sum |x_i - \mu|^2 \tag{4}$$

The variance $Var(x)$ is the volatility of the source code. The smaller the volatility the more cohesive is the repository and the smaller the amount of the dead code inside it.