Regex Usage in Two Million Open Source Projects

Carl Chapman and Kathryn T. Stolee
Department of Computer Science
Iowa State University
{carl1978, kstolee}@iastate.edu

***Abstract*—Regular expressions are used frequently in many programming languages for form validation, ad-hoc file searches, and simple parsing. Given their popularity, many researchers have focused on making regular expressions easier to build, understand, and use. Yet, there does not exist a study of regular expression feature usage and diversity. In this paper, we explore how often regular expressions are used, which language features are most common, and how syntactically and semantically similar regular expressions are to one another. To do this, we scraped 2 million open source Python projects from GitHub and explored the regular expressions contained within. Our results indicate that TODO: high level results**

I. Introduction

II. Motivation

III. Related Work

A. Research on Regular Expressions

Visual debugging of regular expressions [1]

B. Research that Depends on Regular Expression Usage

Regular expressions are used as queries in a data mining framework [2]

IV. Study

Some data from the database:
nObserved: 5001.
nSkipped: 3.
nAborted: 5.
nScanned: 9644.
nProjectsWithRegex: 4873.
nPythonFiles: 550532.
nFilesWithRegex: 50369.

REPRESENTING A CORPUS OF REGEXES

1. How can we choose a small list of actual regexes to best represent all regexes within a corpus by usage?
a.) What regexes are most frequently used?
b.) What features are most frequently used?
c.) What regexes are most cloned?
d.) What behavioral clustering can be observed?
e.) What syntactic clustering can be observed?
f.) Various synthesis of a-e.

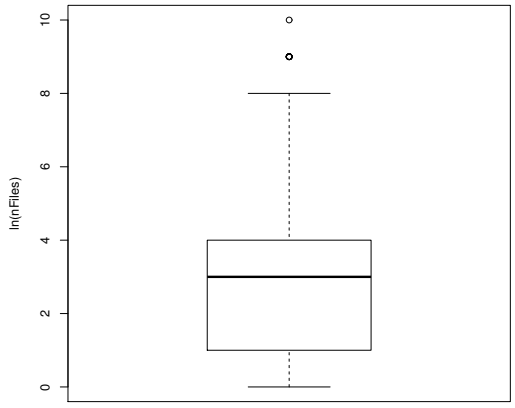

Fig. 1. Example diagram: filesPerProject natural log

CONTEXT AND CORPUS

V. Results

VI. Discussion

VII. Conclusion

Acknowledgment

This work is supported in part by NSF SHF-1218265, NSF SHF-EAGER-1446932, and the Harpole-Pentair endowment at Iowa State University.

References

[1] F. Beck, S. Gulan, B. Biegel, S. Baltes, and D. Weiskopf. Regviz: Visual debugging of regular expressions. In *Companion Proceedings of the 36th International Conference on Software Engineering*, ICSE Companion 2014, pages 504–507, New York, NY, USA, 2014. ACM.

[2] A. Begel, Y. P. Khoo, and T. Zimmermann. Codebook: Discovering and exploiting relationships in software repositories. In *Proceedings of the 32Nd ACM/IEEE International Conference on Software Engineering - Volume 1*, ICSE '10, pages 125–134, New York, NY, USA, 2010. ACM.