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.

CONTEXT AND CORPUS

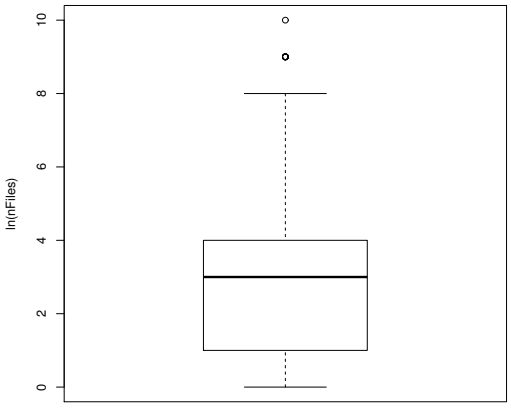

Figure 1: Example diagram: filesPerProject natural log