Nostril: A nonsense string evaluator written in Python

Michael Hucka[1]

1 Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125, USA

Summary

Nostril (*Nonsense String Evaluator*) is a Python 3 module that can infer whether a given word or text string is likely to be nonsense or meaningful text. A "meaningful" string of characters is one constructed from real or real-looking English words or fragments of real words (even if the words are *runtogetherlikethis*). The main use case is to decide whether short strings returned by source code mining methods are likely to be program identifiers (of classes, functions, variables, etc.), or random or other non-identifier strings.

Discerning real identifiers from nonsense is a surprisingly difficult problem, because program identifiers often consist of words, acronyms and word fragments jammed together (e.g., `ioFlXFndrInfo`). The resulting strings can challenge even humans. Nostril uses a combination of (1) a prefilter that detects simple positive and negative cases using heuristic rules and (2) a custom TF-IDF (Manning, Raghavan, and Schütze 2009) scoring scheme that uses letter 4-grams as features. The software includes a precomputed table of n-gram weights derived by training the system on a large set of strings constructed from concatenated American English words, real text corpora, and other inputs. Parameter values were optimized using the evolutionary algorithm NSGA-II (Deb et al. 2000).

By default, Nostril is tuned to reduce false positives – it is more likely to say something is *not* gibberish when it really might be. This bias is motivated by Nostril's original purpose of filtering source code identifiers for machine-learning applications, where false positives would cause real identifiers to be filtered out and potentially-useful features to be missed. However, the bias and other parameters (and the table of n-grams) can also be retrained if applications require it.

Nostril is reasonably fast: once the package is loaded, a string evaluation takes 30–50 microseconds on average on a 4 Ghz Apple macOS computer. Nostril is accurate: it achieves 99.76% on the the Ludiso identifier oracle (Binkley et al. 2013) and 91.70% on a test set of 1,000,000 machine-generated random strings.

References

Binkley, David, Dawn Lawrie, Lori Pollock, Emily Hill, and K. Vijay-Shanker. 2013. "A Dataset for Evaluating Identifier Splitters." In *Proceedings of the 10th Working Conference on Mining Software Repositories*, 401–4.

Deb, Kalyanmoy, Samir Agrawal, Amrit Pratap, and Tanaka Meyarivan. 2000. "A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II." In *International Conference on Parallel Problem Solving from Nature*, 849–58.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2009. *Introduction to Information Retrieval*. Online edition. Cambridge University Press.