Applying Statistical Language Modeling to Genetic Programming

H. Chase Stevens
s1107496

Supervisor: Dr. I. Stark

University of Edinburgh

M.Sc. Project Proposal

April 2016

1 Introduction

2 Background and related work

2.1 Application of language modeling to formal languages

2.2 Genetic programming

3 Methodology

3.1 Compilation of Python corpus

The creation of a language model for Python will require a large and representative corpus of Python source code. As I am not aware of such a corpus existing at this time, the first objective of this project will be to compile one. To do so, source code will be downloaded from publicly available, open source repositories hosted on GitHub.

Ideally, the Python corpus compiled for this project will consist of idiomatic, useful, and well-written Python source code; for this project, good-quality code is of special importance, as inexecutable code may still be syntactically valid and parsable. In compiling a similar corpus for the Java programming language, Allamanis and Sutton [1] sought to maintain a minimum level of quality in the GitHub repositories used by filtering out those that had not been forked at least once; in this project, for which I am planning to create a smaller corpus, repositories will instead be used if they exceed some minimum number of stars or forks. To find repositories matching these criteria, the GitHub search API will be used, which allows filtering on not only number of stars and forks but also on project language.

3.2 Creation of language model from Python corpus

3.3 Application of language model to genetic programming system

4 Evaluation

5 Timeline

6 Conclusions

References

[1] M. Allamanis and C. Sutton, “Mining source code repositories at massive scale using language modeling,” in *Proceedings of the 10th Working Conference on Mining Software Repositories*. IEEE Press, 2013, pp. 207–216.