Linnéuniversitetet
Linnaeus University

Degree Project Proposal

Department of Computer Science

Thesis Topic	**N-Grams as a Measure of Naturalness and Complexity**
Degree level	Master, 15 or 30 credits
Company	Lnu, Computer Science
Description	N-grams are used in Natural Language Processing to classify a text (a sentence) as “odd” or ”ordinary”. This idea has been adapted to Software Engineering where lexical tokens correspond to words, and various source code entities (files, classes, methods) correspond to sentences [On the Naturalness of Software, Hindle et al, Communications of the ACM, 2016]. In this scenario N-grams defines a Statistical Language Model making it possible to associate each source code fragment s (a token sequence) with a probability p(s) that can be interpreted as p(x) > p(y) implies that fragment x is more ”natural” than fragment y. The language model is computed based on a source code corpus C and p(s) expresses a likelihood of finding s in C. Hence, a high value of p(s) indicates that s is ”ordinary” (i.e. a frequently occurring) whereas a low p(s) value identifies s as ”odd” (i.e. seldom used). Hindle et al argue that programmers are creatures of habit and that their code is mostly simple and repetitive, and that code that deviates from this norm is a bit strange and hard to understand. This triggers the question if low p(s) values can be used to identify code that is hard to understand and complex. Or, can we find a correlation between low p(s) values and well-known software complexity metrics like cyclomatic complexity and lines-of-code?
Objectives	15 hp thesis: Implement an N-gram based language model for Java source code and present basic statistics for a few software projects. 30 hp thesis: Investigate if a low p(s) value can be used as complexity metric by trying to correlate it with well-known software complexity metrics.
Requirements	A course in basic statistics (similar to 1MA211), compiler knowledge (similar to 4DV506), and for the 30 credit version, knowledge about software metrics (similar to 4DV607).
Contact Person	Jonas Lundberg (Jonas.Lundberg-at-lnu.se)