Automatic Identification and Analysis of Commented-out Code

Author:

Blake E Grills

Academic Supervisor:

Michael J Decker

Department of Computer Science

 Bowling Green State University, Ohio

Contents

[**Abstract:** 4](#_Toc26353175)

[**Introduction:** 4](#_Toc26353176)

[**Related Works:** 4](#_Toc26353177)

[**Summary of Reviewed Work:** 4](#_Toc26353178)

[**Analysis and Discussions:** 4](#_Toc26353179)

[**Highlight of Contribution:** 4](#_Toc26353180)

[**Methodology:** 4](#_Toc26353181)

[**Ideology:** 4](#_Toc26353182)

[**Construction of the Data Set:** 4](#_Toc26353183)

[**Python and Value Extraction:** 4](#_Toc26353184)

[**SKlearn:** 4](#_Toc26353185)

[**Results and Analysis:** 4](#_Toc26353186)

[**Proposal for Future Work:** 4](#_Toc26353187)

[**Conclusion:** 4](#_Toc26353188)

[**Acknowledgement:** 4](#_Toc26353189)

[**Reference:** 4](#_Toc26353190)

**Abstract:**

Write this after we finish the Methodology section as we determined that it is the most important section.

**Introduction:**

Mention the study performed in the spring of 2019

**Related Works:**

Or should the study in 2019 be mentioned here?

**Summary of Reviewed Work:**

**Analysis and Discussions:**

**Highlight of Contribution:**

**Methodology:**

**Ideology:**

The construction of the data set was to be the first and most important part of this research, as if it was not done properly then everything learned from the machine learning algorithm, to be explained further later on in the section titled SKlearn, would be without any statistical value. As mentioned in the introduction, this study is in its second iteration, in that previous study, only one language had been chosen for analysis and a primitive system had been designed to simultaneously pull the comments, determine whether or not a comment was code, via a series of simple if statements checking for syntax markers, and then writing a fresh version striped of the commented-out code. the scripts that were scanned were all collected from the projects made in the Software Engineering Capstone Project of the Spring semester in 2019 at BGSU. While these initial results did seem to hold some promise, it was almost immediately noted that using code written by students severely limited the statistical quality of the research, as it can be safely assumed that industry standards and methods have yet to be established in their writing styles.

Due to this concern about the statistical viability of the original study, it was decided that we should be targeting much more skilled and recognized programmers. To this end, it was decided that we would select a series of projects from GitHub that were very well trafficked and recognized, as well as having been worked on for some time in order to increase the likelihood that all of the information collected from the resulting statistics and heuristics will be a better representation of the programming community as a whole.

Furthermore, it was decided that we needed to establish a new method for analyzing the scripts, first because of the sheer size of this new set of scripts, but also to help ensure the quality of the final statistics and heuristics as well as the ability to identify commented-out code via an automated process. In the Construction of the Data Set section, the process of manual verification and the process of pulling the comments from the scripts will be described in detail, as well as the reason the data set was constructed the way it was. In the Python and Value Extraction section, the process developing a numeric representation of the data set will be described in further detail to include why the method was set, and why other methods were rejected. Finally, in the section SKlearn, the process behind the selection of the scikitlearn python module and the way the automation process for identifying commented out code will be explained.

**Construction of the Data Set:**

**Python and Value Extraction:**

**SKlearn:**

**Results and Analysis:**

**Proposal for Future Work:**

**Conclusion:**

**Acknowledgement:**

**Reference:**