Detecting Python Syntax Errors with Machine Learning

Monica Bui, Joshua Campbell, Andrea McIntosh, Dr. Abram Hindle,
University of Alberta, Dept. of Computing Science, Software Engineering Lab

Purpose

- Find out the types of common programming mistakes that people make
- Automate the error finding process using machine learning

Overview

- Finding syntax errors manually is often a time consuming task because the program cannot run without fixing these first
- We use a machine learner called Weka and it decides if the contents in the GitHub commits contain an error
- The goal is for Weka to get the best accuracy in classifying mistakes

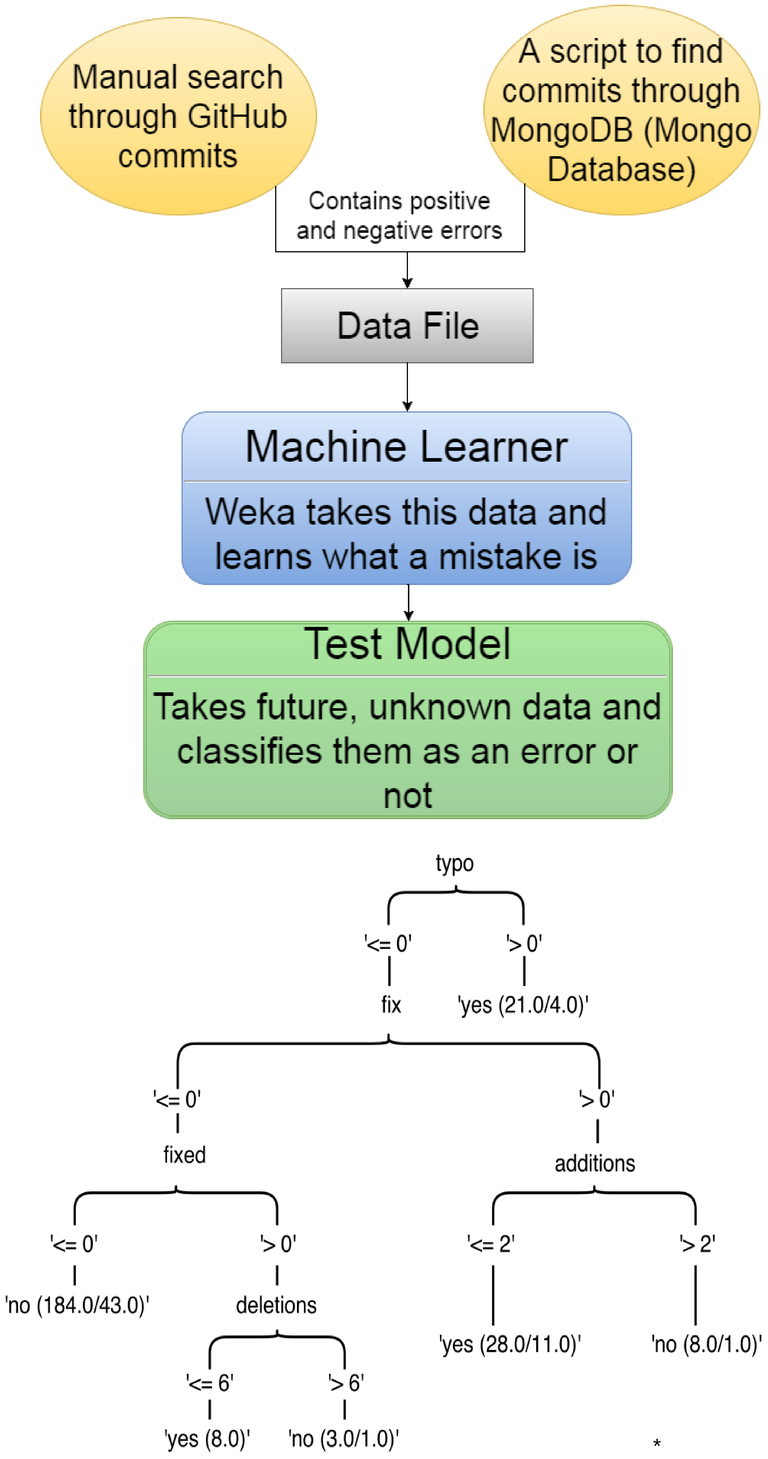
```
Python 2.7.11+ (default, Apr 17 2016, 14:00:29)
[GCC 5.3.1 20160413] on linux2
Type "help", "copyright", "credits" or "license" for more information
>>> for i in list
  File "<stdin>", line 1
    for i in list
               ^
SyntaxError: invalid syntax
>>>
```

Method

- Created a Python script to get data based on additions/deletions in commit, and keywords like “Typo” in the commit message
- Used Weka algorithm classifiers to find accuracy results

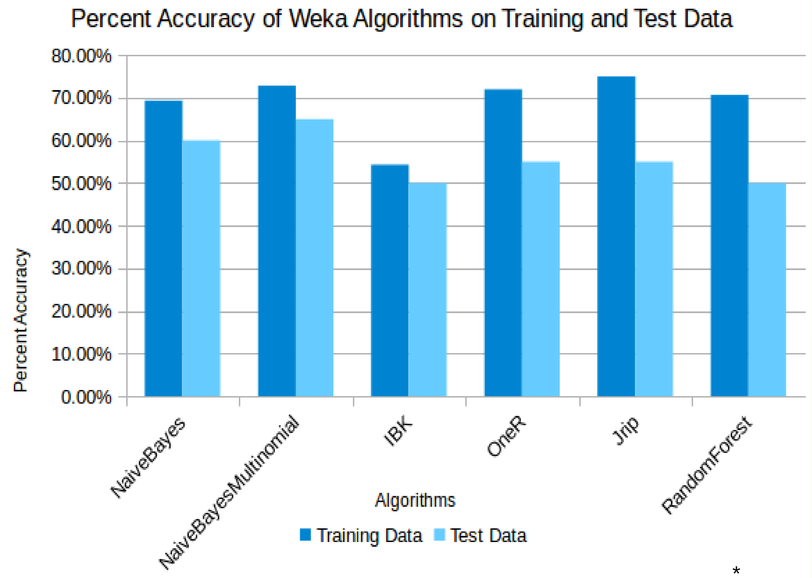
Our classified errors	Computer Predictions
1:yes	2:no
1:yes	1:yes
1:yes	2:no
1:yes	1:yes
1:yes	1:yes
1:yes	1:yes
1:yes	1:yes
1:yes	2:no
1:yes	1:yes
1:yes	2:no

*



*

Results



*

Conclusion

- Commit messages are not useful for detecting errors
- NaiveBayes algorithms gave the best accuracy (~70%) for testing
- Most common error is indentation
- Learning about different types of mistakes will allow for future improvement in teaching programming to others

Acknowledgements

- I would like to thank my WISEST student lab partners: Eisha Ahmed and Hannah Stormer, lab supervisors: Joshua Campbell and Andrea McIntosh, and my supervisor Dr. Abram Hindle for their support throughout my internship
- Thank you to the Software Engineering lab and the High School Internship Program (HIP) for presenting this opportunity

*Some visuals in this poster have been borrowed and modified from my lab partners