

Automated Evaluation of Programming Assignments *Data Collection*

Rishabh Manoj

August 25, 2017

1 Dataset Collection

CodeForces is a platform dedicated to competitive programming. It started nearly 8 years ago which gives us a lot of programming problems and submissions to choose. The problems are divided into 5 category *A, B, C, D* and *E* with *A* being the simplest and *E* being the hardest. CodeForces has recently released a set of *API's* with which we can get access to their data. Using these *API's* I have written a script to download all solutions (correct and incorrect) of a particular programming languages (eg: C, C++) for a problem. [\[script.py\]](#)

I have downloaded **C** and **C++** submissions that have been accepted by the online judge and submissions that failed some test-cases for problems in category *A*. I decided to keep the data-set simple (i.e) made sure that the solutions could be compiled and executed without any problems so I have not included submissions which had the verdict's *Time Limit Exceeded*, *Runtime Error* or *Compilation Error*. [\[Data-set\]](#)

Summary:

- Downloaded submissions for 10 problems in category *A*.
- Total no. of submissions is 3300.
- An average of 300 submissions with 100 correct solutions and 200 incorrect solutions per problem.

Note: Most of the submissions are in C++ with a few minority in C. In future if more data is required, it can be easily obtained with the use of the script mentioned above.

Problem Links: [\[Theatre Square\]](#), [\[Winner\]](#), [\[Shortest path of the king\]](#), [\[Watermelon\]](#), [\[Chat Server's Outgoing Traffic\]](#), [\[Triangle\]](#), [\[Kalevitch and Chess\]](#), [\[Train and Peter\]](#), [\[Die Roll\]](#), [\[Power Consumption Calculation\]](#)