The next generation eLearning platform for computing education

Supervisor: Dr. Chui Chun Kit

Written by: Cheung Siu Wai 2013588661

Group Members: Cai Tung San 3035090470
Chan Wai Lun 3035089835
Wong Ching Quen 3035098161

Date of Submission: 16 April, 2017
1. Abstract

Concerning lack of linkage between e-learning platforms and computing education, this report aims to provide detailed information on our brand-new e-learning platform which is specially designed for students studying programming courses.

Although there are plenty of e-learning platform choices available for schools and lecturers, when talking about computing education, current platforms do not focus on computing education and there is no assist on students’ computing learning progress. Therefore, the project team would like to set up a new platform which has made improvements in collaborative tools and data analytics. More collaborative tools like drawing boards and code-linked video presentation is provided to construct a CS specialized environment, and data analytics are carried out to reduce the time students needed to finish their assignments. With these two main features of the platform, students could receive automatic but also relevant feedback so that they can be effectively equipped with both technical and practical knowledge on programming.

The team will use C++ as the coding language of the code editor, Laravel as the web framework due to its useful features and good community power. Collaborative tools are provided by either using plugins or developed by ourselves, while data analytics will be done by analytics on the compiled errors and passing of test cases.

2. Acknowledgement

The platform would not be built possibly without the help of many individuals. One of them is our FYP supervisor, Dr. CK Chui. He provided general direction and scope of the platform, and many useful suggestions to problems we encountered in the implementation process. We would like to give our sincere thanks to him.

On the other hand, due to the fact that the platform relies on machine learning techniques for data analytics, relevant knowledge is essential. We would like to thank Dr. Andrew Ng and Dr. Li Yi Wei, which are the instructor of the Coursera machine learning course and HKU machine learning courses respectively. Valuable knowledge of machine learning is gained from the courses taught by them.
# 3. Table of Content

1. Abstract 2  
2. Acknowledgement 2  
3. Table of Content 3  
4. List of Figures 4  
5. List of Tables 4  
6. Abbreviations 4  
7. Introduction 5  
8. Objectives 6  
9. Prerequisites 7  
10. Scope 7  
11. Design and Implementation 8  
   11.1 Front end 9  
   11.2 Web Interface 10  
   11.3 Back end 18  
   11.4 Compiler 18  
12. Collaborative tools 22  
   12.1 Code editor 23  
13. AI teacher 23  
14. Data analytics 24  
   14.2 Learning 27  
   14.3 Evaluation 31  
   14.4 Prediction 34  
15. Future Works 34  
16. Conclusion 34  
17. Reference 35
4. List of Figures

Figure 1 Architectural diagram of the online eLearning Platform ........................................ 8
Figure 2 index page of the platform ......................................................................................... 10
Figure 3 Register page ............................................................................................................ 11
Figure 4 Login page ................................................................................................................ 11
Figure 5 dashboard after login ................................................................................................. 11
Figure 6 Course Interface ......................................................................................................... 11
Figure 7 Forum Interface .......................................................................................................... 12
Figure 8 Course Page .............................................................................................................. 12
Figure 9 Assignment page ....................................................................................................... 13
Figure 10 File uploader in assignment page ............................................................................ 13
Figure 11 coding editor .......................................................................................................... 14
Figure 12 rename function on the platform ............................................................................ 14
Figure 13 Adding user input to the program .......................................................................... 15
Figure 14 error message displayed by the compiler ............................................................... 15
Figure 15 The output of the program ...................................................................................... 16
Figure 16 testing the program with test cases ........................................................................ 16
Figure 17 Hint given in the platform ...................................................................................... 17
Figure 18 Test case page ......................................................................................................... 17
Figure 19 Student Performance page ...................................................................................... 18
Figure 20 General data analytics workflow ............................................................................ 25
Figure 21 Classification results of K-means, DBSCAN and Agglomerative clustering on sample data .................................................................................................................. 31
Figure 22 The elbow method .................................................................................................. 32
Figure 23 Computation of silhouette coefficient .................................................................... 33
Figure 24 Example silhouette plot .......................................................................................... 33

5. List of Tables

Table 1 Prerequisites of the e-learning platform ...................................................................... 7
Table 2 Major differences for three chosen PHP frameworks ............................................... 9
Table 3 Comparison between ML candidates ........................................................................ 29

6. Abbreviations

The platform: The next generation eLearning platform for computing education
7. Introduction

Accessibility, user-friendliness, broad coverage of users and more are the most contributing factors, which make e-learning platforms into an important one in education field during recent decades. It appears to be unavoidable for students nowadays to use e-learning platforms like Moodle and i-Class in their study routine. Students can get access to the platform in anytime and anywhere to retrieve learning materials like lecture slides and readings, which are uploaded by the instructors. E-learning platforms can also serve as a medium for students to view, schedule and submit their assignments. Notice board and announcement system on the platforms can also update students of the latest course information and arrangements in case of emergency. The platform can also help students interact outside the classroom. Students can create new posts on the discussion session on the platform, if they encounter any problem in their study.

In terms of the above basic functions, e-learning platforms seems to work very well in some concentrations. However, when it comes to computer education, the platforms appear to have problem to capture the needs of computing students.

In computing education, students are expected to be equipped with both technical and practical computation knowledge, for instance, the theories and the capability of designing a computational system. Therefore, it is necessary to teach students about different algorithms and data structures in the curriculum. In order to let the students catch up with the learning progress, instructors usually provide programming assignments for students to practice regularly.

However, the linkage between e-learning platforms and programming appears to be weak currently. There appear no tools from the platforms to facilitate the learning processes and to provide learning support to computing students. In present, the platform works with a single focus in terms of assignments. When students work on their programing questions uploaded by their lecturer on the e-learning platform, they may not be able to code the solution on the platforms. Instead, they need to use other code editors or even Integrated Development Environment (IDE) outside the platforms. After they submit the assignment back to the platforms, it takes time for the instructors to review and assess on the students’ work submitted before delivering feedback. The platform serves like a storage for assignments only, which is disconnected with students learning achievements and progress. Also, with different operating system like Window, Mac OS and Linux, lecturer need to spend time on
teaching how to install the related programming software and how to compile the program in different operating system. It certainly wastes valuable time and it can be better utilized on teaching other important course material. Student can also focus on coding once they use this platform.

Spotting the problem, our team aims to provide a platform to fill this position. By providing many collaborative tools and various coding support, students are expected to learn programming in a better way and with aids from peers on the platforms.

The remaining of the report will proceed as follows. First, to make the project direction significantly clear, we provide the objectives of what the project team want to achieve by the platform. The report also lists out the prerequisites and define the scope of the project to indicate the boundaries the project. After that, we describe the design of the project in details to show the approaches that we have adopted and the output we produce. Moreover, we will state the project methodology and carry out trials on the platform to illustrate the results of the projects. Finally, we mention the conclusion of the project and try to recommend some areas that can be improved in the future.

8. Objectives

1. Allow students to learn programming by themselves with minimal efforts from teachers using the coding support and real-time feedback provided by the platform.

2. Provide a collaborative and CS-specialized environment for computing students to communicate and share their ideas using collaborative tools provided by the system.

3. Provide timely coding feedback to students by constructing an AI teacher.
9. Prerequisites

Refer to Table 1 below, it shows the fundamentals of the project, which are critical factors in place.

<table>
<thead>
<tr>
<th>Medium</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server, virtual machine (with high computational power)</td>
<td>1. To compile the codes submitted by students 2. To carry out data analytics</td>
</tr>
<tr>
<td>PHP framework</td>
<td>To act as a base of the e-learning platform</td>
</tr>
<tr>
<td>Data analytics’ tools or libraries (Scikit-learn¹, Numpy²)</td>
<td>To facilitate the process of machine learning</td>
</tr>
<tr>
<td>Students in course COMP2123 (Programming Technologies and Tools)</td>
<td>To provide data to verify and improve our machine learning models</td>
</tr>
</tbody>
</table>

*Table 1 Prerequisites of the e-learning platform.*

10. Scope

The platform will adopt C++ as the coding language to analyze and provide support to students. Students who enroll in COMP2123, a course applying C++ as the teaching language, are the target users of the platform. The platform is assumed to be in line with the course.

The data analytics generated by the platform will concentrate on assessing student’s performance by the passing rate of the test cases. All the data in the platform could be used for data analytics. Nevertheless, passing performance of students in test cases will be designated as the first priority in the data analytics. Because it is more likely to produce useful outcomes which is highly related to students’ performance and achievements, instead of analyzing other data in the platform like posts on discussion forums, videos uploaded to the platform.


11. Design and Implementation

Collaborative tools and data analytics are two essential elements in the platform. Former one includes various functions to provide assistance to students from completing the assignments to facilitate the entire learning progress of programming. The later one aims at generating auto feedback to the students in the code editor. Details of the platform are illustrated in Figure 1.

As shown in Figure 1, the platform mainly composed by frontend in the middle of the figure and backend on the right of the diagram. Collaborative tools are included in the frontend and data analytics is handled in the backend. Below are the details of frontend and backend.
11.1 Front end
The front end part is setup as a web system and user interface is provided for users to interact with the platform. For example, in Figure 1, lecturer can use the frontend to input test cases and post assignments; while students can view the assignments and use the collaborative tools on the platform. The collaborative tools will be explained in Section 7.

We considered three of the PHP framework in our early stage of decision. The three PHP frameworks were Laravel, Symfony 2 and Yii. For all three PHP frameworks, the critical library functions such as user access control, security support (i.e. CSRF protection) are available in the frameworks, which can noticeably speed up the development of the it.

All three PHP frameworks use traditional MVC (Model-View-Controller) structure for development. The Model part interacts with and retrieve information from the database. Laravel provides Eloquent ORM which allows us to perform model creation and query to the database easily. Moreover, the View part renders pages to users, all template systems allow us to extend and inherit layouts which make the viewed file clearer to understand. The Controller part handles the interaction between users and the platform. The Middleware and the inheritance allows us to assign action based on user’s interactions. This MVC structure establishes clear separation of duties and makes viewing and modification of codes much easier for other users.

The major difference are summarized in Table 2 below:

<table>
<thead>
<tr>
<th></th>
<th>Laravel</th>
<th>Symfony 2</th>
<th>Yii</th>
</tr>
</thead>
<tbody>
<tr>
<td>Template system</td>
<td>Blade, PHP, custom</td>
<td>Twig</td>
<td>Prado</td>
</tr>
<tr>
<td>Allow PHP usage in template system</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Online resources</td>
<td>Most</td>
<td>Middle</td>
<td>Middle</td>
</tr>
<tr>
<td>Extendability</td>
<td>9000 packages</td>
<td>2830 bundles</td>
<td>2800 extensions</td>
</tr>
<tr>
<td>Logging management</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
</tbody>
</table>

Table 2 Major differences for three chosen PHP frameworks
Laravel\(^3\) is the PHP framework chosen for the platforms’ web system. The major reason of adopting Laravel is the maturity of its online community on offering interactive and timely support for developers. Thanks to the escalating popularity of Laravel, counterparts who encounter similar developing problems as us have heat up the discussion in the Laravel communities. Due to community efforts, solutions to problems are expected to be found easily that can secure the development process. Thus, development time and difficulties could be reduced with the aids of Laravel community.

Another major reason is the extendability of Laravel. Laravel provides more than 9000 packages, fast solutions for desired functions of the platform are expected to be found easily that can speed up the development process.

**11.2 Web Interface**

The web interface mainly use bootstrap as the framework because it is faster and easier to build the interface. Also, we have used several plugin in our platform. For example, the popup window in the coding page is done by the library “jquery-popup-overlay.js”\(^4\). The sorting of assignment and the styles is done by the library “jquery-sortable.js”\(^5\). Also, the download function of the source code is completed by “FileSaver.js”\(^6\).

The following part will show all the web interfaces of the platform.

---


\(^4\) http://dev.vast.com/jquery-popup-overlay/

\(^5\) https://johnny.github.io/jquery-sortable/

\(^6\) https://eligrey.com/blog/saving-generated-files-on-the-client-side/
Figure 2 shows the index page of the platform. If you have not registered an account yet, you can simply go to the register page (i.e. Figure 3) to do the registration. After that you can login to the platform (i.e. Figure 4).

Figure 5 dashboard after login

Figure 6 Course Interface
After login, you will be directed to the interface in Figure 5. If you click the course tab and forum tab on the top navigation bar, you will be directed to the course page and forum page respectively. Course page will list the courses that the student enrolled in it.

When you click the course name in the course interface, it will be directly to the course page as shown in Figure 8. This demonstration will show the course page for teacher and tutor only for simplicity. In the user interface of students, there are some functions which will not be provided including adding assignments and deleting assignment.
After clicking the assignment 1, it will be directed to the corresponding assignment page shown in Figure 9. Teacher and tutor can upload and delete the assignment through this page. The file uploader is shown on Figure 10.

In the bottom of the assignment page in figure 9, there are three buttons with different functions. Students will only have the coding button in which they can code there. Test case and student performance are only available to the lecturer and tutor. In the following demonstration, it is assumed that the question is asking a user to input their HKID number without the check digit and the program will calculate the check digit and output to the console.
After clicking the coding button, it will be directed to the coding editor as shown in Figure 11. Students can code there and there are a lot of function provided at the bottom of the page. Student may rename the cpp file (Figure 12) by clicking the file name test.cpp. Also, they can download the cpp file to the localhost.

If there are any input to the program, students can add it by clicking the input button in order to test the program (Figure 13).
After finishing coding, student can click compile and run to compile the program and check whether there is any syntax error or logic error. If there exist syntax error, the compiler message will be displayed in the console (Figure 14).

If the program can be successfully compiled, it will return the output of the program (Figure 15). In this case, 10 will be returned because the check digit of the HKID Y345678 is A.
If the students think that they can successfully finish the program, they can click the submit and check button. Then the program will be tested by the test cases inputted by the lecturers or tutors. The result is shown in Figure 16. In the console, it will show the testing results and students may try to modify their program according to the test cases that they cannot pass.

Also, there is a button called hint and students may try to amend the program according to the hint (Figure 17).
In test cases page (Figure 18), lecturers and tutors can add and edit test cases of the questions there. They can also add the title of the test cases so that they know which kinds of concept that were testing. For the student performance page (Figure 19), lecturer and tutor can check all of the students’ performance there, including a list of test case that the students pass or fail. Lecturer and tutor can easily know the weakness of each students.
11.3 Back end

The back end part is setup to store the data created or uploaded by users such as assignments and forum posts. It also contains auto grader with has compiler inside to grade the code submitted by students.

11.4 Compiler

The compiler in the platform collects code input and returns results to user, which is an important role in the platform. The project’s supervisor, Dr. Chui, suggests that there is an existing compilation environment in the Moodle system. The compiler is hosted by HKU CS server and use VM (virtual machine) for compilation to prevent any memory issue that commonly arises when user commits a mistake in learning. Considering the convenience of this existing solution, it is proposed that we connect the compiler hosted by HKU CS server with the platform.
11.5 Database implementation

Database is an essential tool in the platform to store data systematically. We use the migration system of Laravel and MySQL to create and synchronize the database. The database schema is listed below.

<table>
<thead>
<tr>
<th>Table</th>
<th>Function</th>
</tr>
</thead>
</table>
| Users(id, name, email, password, group, remember_token, created_at, updated_at) | Store the information of the user.  
id is the primary key that uniquely identified a user.  
Name, email and password are as stated.  
Group are used to differentiate different users such as student, lecturer and tutor.  
Created_at and updated_at are the timestamp.  
Remember_token is a field that store the token for “remember me” sessions. |
| Assignments(assignment_id, course_id, created_at, updated_at, order)       | Store the information of assignments.  
Assignment_id is the primary key.  
Course_id is used to refer which course in the table course.  
Created_at and updated_at are the timestamp.  
Order is used to decide the order of the assignment. |
| Courses(course_id, created_at, updated_at, fullName, name, courseCode)     | Store the information of the courses  
Course_id is the primary key.  
Created_at and updated_at are the timestamp.  
fullName is the full name of the course.  
Name is the abbreviation of the course.  
CourseCode is the course code of the course. |
<p>| Files(file_id, assignment_id, created_at, updated_at, filename_original)   | Store the information of the files |</p>
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>filename_storage, size</td>
<td>File_id is the primary key. Assignment_id is used to refer the file belong to which assignment. Created_at and updated_at are the timestamp. Filename_original is the filename of the file when the user uploads it. Filename_storage is the filename of the file when stores on the server. Size is the size of the file.</td>
</tr>
<tr>
<td>Coding(coding_id, course_id, assignment_id, user_id, filename, default_input, created_at, updated_at)</td>
<td><strong>Store the coding information of each user on different course and different assignment.</strong> Coding_id is the primary key. Course_id and assignment_id are used to identify which course and which assignment. User_id is used to refer which user. Filename is the user that name the program file. Default_input is the value of the user saved for compiling the program. Created_at and updated_at are the timestamp.</td>
</tr>
<tr>
<td>Submissions(submission_id, course_id, assignment_id, user_id, created_at, updated_at, content)</td>
<td><strong>Store the submission information of the students.</strong> Submission_id is the primary key. Course_id and assignment_id are used to identify which course and which assignment. User_id is used to refer which user. Created_at and updated_at are the timestamp. Content are the code that the student submitted.</td>
</tr>
<tr>
<td>Table Name</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Testcases(testcase_id, course_id, assignment_id, input, output, created_at, updated_at, group_id)</td>
<td>Store the information of the test cases. Testcase_id is the primary key. Course_id and assignment_id are used to identify which course and which assignment. Input is the test case that uses to test the program. Output is the expected output of the program associated with this test case. Created_at and updated_at are the timestamp. Group_id is the group id of the test case group.</td>
</tr>
<tr>
<td>Testcase_group(id, avg_rate, course_id, assignment_id, title)</td>
<td>Store the information of the test case group. Id is the primary key. Avg_rate is the rate to calculate the score of the submission in this test case group. Course_id and assignment_id are used to identify which course and which assignment. Title is the title of the test case group.</td>
</tr>
<tr>
<td>Test_by(testcase_id, submission_id, user_output, result)</td>
<td>Store the result of the submission when compare to the result of the test case. Testcase_id refers to which test case. Submission_id refers to which submission. User_output stores the output produced by the students with respect to this test case. Result is the Boolean value which store whether the student is correct or not.</td>
</tr>
<tr>
<td>Forum_categories(id, category_id, title, description, weight, enable_threads, thread_count, post_count, privated)</td>
<td>Store the categories of the forum.</td>
</tr>
</tbody>
</table>
12. **Collaborative tools**

One of the objectives of the platform is providing a collaborative and CS specialized environment for computing students.

To achieve this, the platform integrates external libraries or plugins to provide some popular tools like group chat and discussion forum. And customized code editor using external libraries is provided for them to code together and use to submit assignments.

Instead of integrating libraries, the platform also has self-developed tools that are tailor-made for computing students, like drawing board and video presentation.

The details of code editor and self-developed tools will be described below. It is expected that students can share their ideas and ignite inspirations much easier. And students’ learning in programming could also be facilitated by the collaborative tools on the platform.
12.1 Code editor
The platform will embed external code editor to grasp the advantage of code completeness and high degree of functionalities support. Firepad\(^7\), Ace\(^8\) and Firebase\(^9\) are the selected libraries for building our code editor.

Firepad serves as real-time collaborative code editing tool; while Ace serves as a common code editor which provides general features like syntax highlighting, code completion and multiple cursors. Detailed integration tutorial between these two tools is available in their websites that gives developers the greatest fundamentals on functionalities development. Thus, it is believed that the workload and time needed for integrating the tools into the platform has been minimized.

Firebase serves as a native data logging system which records all input of students when they the code editor. The collected data then used in data analytics. It is believed that the use of Firebase minimized the time and workload needed for design and implement of data logging system.

In addition to this embed code editor, the platform will integrate auto feedback session into the system to help students do programming. The generation of auto feedback is done by data analytics process which the data analytics will be explained later in Section 6.5. Test case results and possible compiled errors will be provided in area next to the code editor upon submission of code.

13. AI teacher
In order for to assist students on their coding as well as helping teachers to see students’ progress, an AI teacher is provided to provide coding support and feedbacks to students and calculating results for teachers. The AI teacher has four features provided and are listed below. The method of providing these features are described in Section 13.

13.1 Test cases classification
The AI teacher can classify different test cases into different clusters, which corresponds to different programming concepts. When new test cases are input by teachers, the AI teacher tries to classify them to determine whether this is a test case.

---


to test new concepts in the code or just another test case which test the concepts like some other test cases already stored.

Having test cases testing new concepts allow teachers to test students’ codes much comprehensively and ensure students' codes have less bugs.

13.2 Ranking Test cases

Test cases are divided into different clusters in Section 12.1, the AI teacher can also rank the clusters in different difficulties. Scoring are done by the passing of the number of clusters instead of the number of test cases that are passed. Higher weighting can also be added to cluster with higher difficulties.

13.3 Suggesting Test cases group

Classifying test cases into different groups and ranking different groups allows the AI teacher to know which test cases groups the students have not passed and the AI teacher could also know the level the students are currently in. Therefore, the AI teacher can suggest test cases groups that are in the same level with students, so that students can deal with test cases that are easier for them to deal with in their current levels.

13.4 Show code changes or Provide code change directions

Students may still struggle on how they can pass the test cases groups, the AI teacher can also use other students’ code as a reference and show to the students about the code changes that leads to the passing of those test cases groups. The AI teacher may also provide some basic directions for the students to improve theirs code such as “Can adding a while loop solve the problems?”

14. Data analytics

The data analytics is carried out to provide the features of the AI teacher mentioned in Section 12. The passing of test cases by students are used for the data.

Machine learning is used to group the test cases into different categories. Figure 4. below illustrates the flow of data analytics from data being inputted to auto feedback being produced eventually.
The students result is graded by the percentage of test cases they pass for that assignments. A test case contains both input and output. When students submit a program code, input is used to test the code and try to generate output using the code. The student’s output are then compared with the test case output, if they are the same, then that test case is passed. Usually the students cannot pass all the test cases by the first time they hand in a submission, they will try to modify their code and then submit again. Therefore, one students may have many submissions regarding just one assignment.

These submissions leading to the changes in passing of test cases are used to calculate correlation matrices and for data analysis.

These data will be split into training data, validation data and test data in about 50-25-25 proportion. The training data is responsible for setting parameters in the classification algorithm. The validation data is responsible for fine-tuning to prevent underfit or overfit of models as well as comparing performance between algorithm candidates. The test data is responsible for verification and evaluating the accuracy of the algorithm.

### 14.1 Preprocessing and getting correlation matrices

The submissions of students are modelled as submission vectors. For example, if student passed test cases #3, #5, #6 out of the 10 provided test cases, the
submission vector will be \([0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0]\). The submission of students will then be combined as submission matrix by concatenating submission vectors. The difference between submission vectors, delta-vectors, are computed using XOR operator. To illustrate, for submission #1 \([0 \ 1 \ 1 \ 0]\) and submission #2 \([1 \ 1 \ 0 \ 0]\), the delta vector will be \([1 \ 0 \ 1 \ 0]\).

### 14.1.2 Correlation tuples

The delta vectors calculated are used to generate correlation tuples. Correlation tuples are generated between “1” and “1” entry in each delta vectors. For examples, for delta vector \([1 \ 1 \ 1 \ 0 \ 0]\), test cases #1, #2, #3 are the 1s in the delta vector, correlation tuples are the combination between these 3: (#1, #2), (#1, #3), (#2, #3).

Correlation tuples are used to update the correlation matrices by a function of \(0.8 \times x^{-2}\), where \(x\) is the number of “1” entries in the delta vectors, in the case \([1 \ 1 \ 1 \ 0 \ 0]\), \(x\) is 3, value will be \(0.8 \times 3^{-2} = 0.8\), the value is used to update the correlation matrices by the correlation tuples. (#1, #2) represents update row 1 and column 2 in the correlation matrix.

### 14.1.3 Penalty function

To model the degree of correlation between test cases more accurately, penalty function is introduced. The penalty function used is \(0.95 \times x^{-2}\), where \(x\) is the number of non-0 entries in the delta vector. Consider a simple example consists of only 2 submissions, submission #1 \([0 \ 1 \ 1 \ 0 \ 0]\) and submission #2 \([1 \ 1 \ 0 \ 0 \ 1]\). The delta vector obtained according to above section is \([1 \ 0 \ 1 \ 0 \ 1]\). In the delta vector, there are three non-zero entries. The output of the penalty function would be \(0.95^{3^{-2}} = 0.95\), which is named penalty coefficient. All of the entries of the delta vector is multiplied with the penalty coefficient. Following the example, the final delta vector would be \([0.95 \ 0 \ 0.95 \ 0 \ 0.95]\). The penalty function can be seen as a measure for regulating correlation. By imposing higher penalty, the correlation between test cases are reduced.

### 14.1.4 Correlation Matrices

After obtaining delta vectors from submissions, correlation matrix can be calculated. The entries in the matrix range from 0 to 1, implying low degree of
correlation and high degree of correlation correspondingly. Each row in the correlation matrix is initialized to identity matrix with dimension equal to number of test cases. The choice of identity matrix is based on the assumption that each test case is correlated with itself and different test cases are not related at first. The test cases will have correlation only after being analysed by the system. The update of the correlation matrix will be based on the clustering result.

Data first passing in the process will undergo preprocessing before being analysed to provide much relevant data for analysis. For example, code history has to undergo DIFF function to get code changes. The function works similar to the “diff” command in Bash\(^\text{10}\), which sends back the difference between two files. Other data inputted will also undergo normalization or standardization for better comparison and visualization of output.

### 14.2 Learning

After receiving data for input, learning algorithms has to be decided for classification. Currently, we consider clustering algorithms only since they are unsupervised learning method, which suits our unlabelled data. There are multiple possible algorithm candidates, they include K-means, DBSCAN, Agglomerative clustering, which will be explained below.

**(a) K-means**

Among the clustering algorithms, K-means has a relatively efficient time complexity, \(O(tk_n)\), where \(t\) is number of iterations; \(k\) is number of clusters; and \(n\) is number of data points. However, K-means algorithm requires the parameter \(k\) (number of clusters) to be specified in advance. Extra experiments are required to be conducted to find optimal set of parameters. The algorithm also has the disadvantage of being unable to handle outliers. Dataset with outliers can have classification result with low accuracy since the algorithm attempts to put outliers in clusters.

**(b) DBSCAN**

DBSCAN stands for Density-based spatial clustering of applications with noise. DBSCAN is associated with two parameters, $\varepsilon$ and $\text{minPts}$. For a unvisited point $p$, if there are at least $\text{minPts}$ points with distance not greater than $\varepsilon$, $p$ is concluded as core point. In other words, $p$ will become a core point if there are $\text{minPts}$ points inside the circle area with radius $\varepsilon$ and centered at $p$. After all points have been processed and all core points have been found, the core points and the points within their radii will form groups. If two groups’ core points are within the radius of their counterparts, the two groups will be merged. Hence, Clusters are formed and outliers are separated since outliers will not be able to form cluster. The advantage of DBSCAN is the number of clusters is not required as parameters and robust to outliers. It can also form arbitrary shape of cluster, which may be favoured depends on situation. However, it required large amount of experiments to find out suitable values for the two parameters for good clustering result.

(c) Agglomerative clustering
Agglomerative clustering assigns each data point to its own cluster first. By using a evaluation metrics like single link, clusters are merged together. The process continues until termination condition is reached. The evaluation metrics determines whether clusters should be merged together. Take single link as example, if there is two data points in two clusters which has distance smaller than a certain threshold, the two clusters are merged. Eventually all clusters will be merged to a single cluster which contains all data points. For desired clustering result, termination condition is introduced. Common termination condition includes number of clusters in current result smaller than a threshold. A variant to this clustering algorithm is divisive clustering, which reverse the process. Instead of merging clusters, a single big cluster is divided until termination condition is reached. This clustering algorithm can determine number of clusters with the help of termination condition. However, the algorithm requires $O(n^2)$ time and hence does not scales up well.

These three algorithms both have their merits and demerits. Table 3. summarizes the features of the ML candidates and compares them.
<table>
<thead>
<tr>
<th>Advantage(s)</th>
<th>K-means</th>
<th>DBSCAN</th>
<th>Agglomerative clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Relatively efficient</td>
<td>• Can determine number of clusters&lt;br&gt;• Robust to outliers</td>
<td>• Manual tagging is not required&lt;br&gt;• relatively efficient</td>
</tr>
<tr>
<td>Disadvantage(s)</td>
<td>• Sensitive to outliers&lt;br&gt;• Need to specify number of clusters manually</td>
<td>• Require manual tuning of parameters</td>
<td>• Do not scale up well due to high time complexity</td>
</tr>
</tbody>
</table>

Table 3 Comparison between ML candidates

The project team will try to implement them and the algorithm with higher prediction accuracy will be the major selection criterion. To understand the difference between classification result of different classifiers, sample data set was used to plot graphs and the characteristics of the resulting graphs were studied. The sample data used is Iris flower data set, which is included in the Scikit-learn library.\textsuperscript{11} Figure 5. shows the difference between results of K-means, DBSCAN, Agglomerative clustering.

14.3 Evaluation
The number of clusters is a parameter required in several clustering algorithms. It plays a prominent role in affecting the clustering result. To obtain accurate clustering result, several methods are used to determine the number of clusters.

14.3.1 Elbow method
The elbow method plot a graph of error against number of clusters. The optimal number of clusters will be the elbow point.
From Figure 22, we can observe that the elbow point is $n=4$. This implies the optimal number of clusters is 4. This can be verified by computing second derivative for every point and find the maximum absolute value.

### 14.3.2 Silhouette analysis

The silhouette analysis mainly measure 2 quantities, coherence in between cluster and separation between clusters. High coherence and high separation implies good clustering. The silhouette coefficient can be computed as follow:
Coherence

\[ a(i) = \bar{d}(x^{(i)}, C) \quad x^{(i)} \in C \]

Separation

\[ b(i) = \min_{C \in \{C_1, \ldots, C_k\}} \{ \bar{d}(x^{(i)}, C) \} \quad x^{(i)} \notin C \]

Silhouette

\[ s(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))} \]

Which can be spelled out as:

\[ s(i) = \begin{cases} 
1 - \frac{a(i)}{b(i)} & a(i) < b(i) \\
0 & a(i) = b(i) \\
\frac{b(i)}{a(i)} - 1 & a(i) > b(i)
\end{cases} \]

So \(-1 \leq s(i) \leq 1\)

- \(s(i)\) is close to 1 if we have \(a(i) \ll b(i)\) → good clustering
- \(s(i)\) is close to \(-1\) → bad clustering

\(\text{Figure 23 Computation of silhouette coefficient}\)

After computing silhouette coefficient for every data point as Figure 7, average silhouette value in clusters can be found and plotted as silhouette plot. An example of silhouette plot is in Figure 8.
The red line shows the average silhouette coefficient. From the plot, it can be observed that the 3 clusters share similar silhouette distribution, which is another sign for good clustering result.

14.4 Prediction

The final model after evaluated could then be used for prediction. New data including test cases’ rates comes from students could be inputted to the model. The analysis will attempt to classify the test cases and offering feedback back to the code editor. These data can also be used for further tuning of the model and algorithms in the evaluation part.

15. Future Works

Due to the limitation of time, some of the function that planned have not been implemented. The collaborative tools like the video presentation and collaborative editing can be included into the platform in the future.

Moreover, we may improve the storing method of code because we store every submission of code into the database now. However, some students may modify their code with a little bit portion. It will waste a lot of space in the database. We can try to find a method to compress code in order to reduce the space that the code used.

16. Conclusion

This report described the architectural diagram of our new e-learning platform and approaches that our team have adopted. We also stated the libraries and algorithms that are found and used by our teams which is essential to build our platform.

To create a better online platform for computing education, we have created a new platform which may greatly facilitate the learning progress of students in programming and reduce the workload of lecturers and tutors in marking assignments done by the students. We really hope that this platform can benefit all of the stakeholders, including lecturers, tutors and students.
17. Reference