Department of Computer Science
The University of Hong Kong
Final Year Project
Interim Report

The next generation eLearning platform for computing education

Supervisor: Dr. Chui Chun Kit

Group Members: Cai Tung San 3035090470
Chan Wai Lun 3035089835
Cheung Siu Wai 2013588661
Wong Ching Quen 3035098161

Date of Submission: 22nd January, 2017
1. Abstract
Concerning the loose linkage between e-learning platforms and computing education, this interim report aims to provide detailed information on our proposed brand-new e-learning platform which is specially designed for computing students.

Although there are plenty of e-learning platform choices available for schools and lecturers, when talking about computing education, current platforms do not have many aids to assist 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.

This interim report will also include the current status, future plans and schedules as well as difficulties for readers’ knowledge.

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, Dr. Li Yi Wei and Professor N. Mamoulis, which are the instructor of the Coursera machine learning course and HKU machine learning courses and Advanced database Systems respectively. Valuable knowledge of machine learning and data analytics are gained from the courses taught by them.

3. Abbreviations
- The platform: The next generation eLearning platform for computing education
# Table of Content

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>2</td>
</tr>
<tr>
<td>Acknowledgement</td>
<td>2</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>2</td>
</tr>
<tr>
<td>List of Figures</td>
<td>4</td>
</tr>
<tr>
<td>List of Tables</td>
<td>4</td>
</tr>
<tr>
<td>Introduction</td>
<td>5</td>
</tr>
<tr>
<td>Objectives</td>
<td>6</td>
</tr>
<tr>
<td>Prerequisites</td>
<td>6</td>
</tr>
<tr>
<td>Scope</td>
<td>6</td>
</tr>
<tr>
<td>Design</td>
<td>7</td>
</tr>
<tr>
<td>Front end</td>
<td>8</td>
</tr>
<tr>
<td>Back end</td>
<td>8</td>
</tr>
<tr>
<td>Compiler</td>
<td>8</td>
</tr>
<tr>
<td>Collaborative tools</td>
<td>8</td>
</tr>
<tr>
<td>Code editor</td>
<td>9</td>
</tr>
<tr>
<td>Drawing Board</td>
<td>9</td>
</tr>
<tr>
<td>Video Presentation</td>
<td>10</td>
</tr>
<tr>
<td>AI teacher</td>
<td>10</td>
</tr>
<tr>
<td>Data analytics</td>
<td>11</td>
</tr>
<tr>
<td>Learning</td>
<td>14</td>
</tr>
<tr>
<td>Evaluation</td>
<td>15</td>
</tr>
<tr>
<td>Prediction</td>
<td>16</td>
</tr>
<tr>
<td>Current status</td>
<td>17</td>
</tr>
<tr>
<td>Further planning</td>
<td>18</td>
</tr>
<tr>
<td>Difficulties</td>
<td>18</td>
</tr>
<tr>
<td>Conclusion</td>
<td>18</td>
</tr>
<tr>
<td>Reference</td>
<td>19</td>
</tr>
<tr>
<td>Appendix</td>
<td>19</td>
</tr>
</tbody>
</table>
4. **List of Figures**

Figure 1. Architectural diagram of the online eLearning Platform 7

Figure 2. User interface of video presentation 10

Figure 3. General data analytics workflow 12

Figure 4. Classification results of Logistic Regression, Artificial Neural Network and K-means clustering 20-21

Figure 5. The elbow method 16

Figure 6. Computation of silhouette coefficient 22

Figure 7. Progress information of the project 17

5. **List of Tables**

Table 1. Prerequisites of the e-learning platform. 6

Table 2. Comparison between ML candidates 16
6. 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 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 appears 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 problems on the platforms but have 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.

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 would like to produce. Moreover, we state the current status and further planning to illustrate the progress of the project currently. Finally, we mention the difficulties we are encountered and provide
possible solutions that may solve the problems in order to accomplish the final year project as usual.

7. **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.

8. **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</td>
</tr>
<tr>
<td></td>
<td>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-learn, Numpy)</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.*

9. **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.

10. Design
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.
10.1 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.

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.

10.2 Back end
The back end part is setup to store the data created or uploaded by users such as assignments, chat history and forum posts. It also contains auto grader with has compiler inside to grade the code submitted by students. Moreover, as seen in Figure 1, these data will be used to carry out data analytics in the backend. The data analytics process is discussed in Section 8.

10.3 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. 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.

11.1 Code editor
The platform will embed external code editor to grasp the advantage of code completeness and high degree of functionalities support. Firepad⁴, Ace⁵ and Firebase⁶ 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.

11.2 Drawing Board
The drawing board serves as a graph paper for students to use when they need to mark down outline or brainstorming. The drawing board provides general drawing functions such as colored pen, insertion of shapes and symbols, as well as images, text boxes, etc. In order to provide a CS-specialized environment for students, the drawing board provides drag and drop function of CS graph components. CS graphs such as UML, ER diagram and SSD are outputted by drawing board for the assignment use of students.

---

11.3 Video Presentation

The video presentation serves as the platform of students’ experiences sharing on assignments. In additional to a typical video player, several specialized features are added into the video presentation. There are several time intervals being marked down on the timeline, and each of them owns a short description and refers to a certain part of the code. Figure 2, below illustrates the user interface of the video presentation function on the platform.

```cpp
// foo.h - header file
#include <iostream>
class foo {
public:
  foo();
  foo(const foo&);  
  friend class foo;
private:
  class impl; Code
  std::unique_ptr<impl> impl;
};
// foo.cpp - implementation file
class foo::impl
public:
  void end_internal_work();
private:
  int internal_data = 0;
  std::function<bool([...])> impl_func;
  impl(impl); impl.end_internal_work();
  foo::foo() = default;
  foo::foo(const foo&) = default;
  foo::foo::operator=(foo&) = default;
```

Figure 2. User interface of video presentation

In order to share videos, students may use the video editing system on the platform to add time intervals and description to the video uploaded. The video will be processed by the system and automatically show to other students.

12. 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.

12.1 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.
12.2 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.

12.3 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.

12.4 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?”

13. 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 3. below illustrates the flow of data analytics from data being inputted to auto feedback being produced eventually.
13.1 Preprocessing and getting correlation matrices

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.

13.1.1 Submission Vectors and Delta Vectors

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 \([0 \ 0 \ 1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 0] \). 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 \([1 \ 0 \ 1 \ 0] \).
13.1.2 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 \([1 \ 1 \ 0 \ 0]\), 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 \(0.8^{x-2}\), where \(x\) is the number of “1” entries in the delta vectors, in the case \([1 \ 1 \ 0 \ 0]\), \(x\) is 3, value will be \(0.8^{3-2} = 0.8\), 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.

13.1.2 Penalty function

To model the degree of correlation between test cases more accurately, penalty function is introduced. The penalty function used is \(0.95^{x-2}\), 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.

13.1.3 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 cases are correlated with themselves 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 analyses 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, which sends back the difference between two files. Other data inputted will

---

also undergo normalization or standardization for better comparison and visualization of output.

13.2 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, $O(tkn)$, 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.

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. Figure 4. shows the difference between results of K-means, DBSCAN, Agglomerative clustering.

13.3 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.

13.3.1 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 5, 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.

### 13.3.2 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. After computing silhouette coefficient for every data point as Figure 6, average silhouette value in clusters can be found and plotted as silhouette plot. Similar silhouette distribution is a sign for good clustering result.

### 13.4 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.
14. Current status

Our group consists of four people, so we are separated into two teams each with two people to deal with the two parts of the project. We completed stage one and we are currently at stage two: adding other collaborative tools and using code editor on the platform to generate test data for data analytics.

In the first semester, the platform team learned the framework Laravel and have implemented functions with Laravel. User authentication function was completed by using the middleware group provided from Laravel and classification of the user groups into student, tutor, lecturer and administrator was completed. User belongs to different user groups have different access right to the platform. Also, the main platform was built with basic functions. Lecturers and tutors can add courses and assignment into the platform. Students can view the courses and assignments and use the code editor to code. Drawing boards and video presentation function are building now. In the last month, a request for organizing a meeting with the CS technical staff was made to our supervisor. The ability of hosting our system and reporting to CS technical staff for error reporting from the system is gained.

The data analytics team made efforts in learning machine learning algorithms due to the lack of knowledge in machine learning. Machine learning course in Coursera is completed and the machine learning course in this semester is also taken to strengthen their machine learning skills. Some algorithms which may be suitable solutions for the project are identified, which are already mentioned in Section 6.5. After a long period of study in machine learning, the ability of implementing suitable machine learning algorithms was gained by the data analytics team now. A prototype for clustering test cases is finished with given submission vectors.

Figure 7 summarizes the progress information of the project.

---


15. Further planning
After the first presentation, we have received valuable suggestions from our supervisor that we can build a AI teacher to provide useful feedbacks to the students. We now focus on building the AI teacher and its corresponding functions.

The data analytics team will start to test the algorithms with some data generated from the code editor in the platform. Since the algorithm used mainly based on some machine learning libraries like Scikit-learn and Numpy for providing numerical computation, it is expected that the algorithm does not need any large modifications. The data analytics team will also build the bridge to connect to the platform part to provide smooth and automated AI teacher.

16. Difficulties
Owing to the uptight time pressure, we will prioritize our functionalities by its importance in order to achieve our project goal first. Core functions like compiler and AI teacher will be ranked and processed. Effective prioritizing can ensure the timeliness of important functionalities’ delivery.

On account of the variety of selections of programming tools, there are plenty of new applications such as Numpy, Laravel framework for the project team to learn and understand before choosing and applying the best of the best into the platform. The team has to spend time and make continuing effort on the development of the core application. Otherwise, the platform produced will have incomplete functionalities and may have a lot of bugs.

The team also has expected to embrace challenges in making the best decision of machine learning algorithms. This would affect the data analytics part to provide useful feedback to the students. First and foremost, a part of data will be classified to serve as validation set to compare the performance between different algorithms, in order to set up our very first selection criterion. We will also actively seek advice from our supervisor for any support.

For data analytics part, there is difficulty in visualization of high dimensional data. This affects the evaluation of machine learning algorithm since visualization is often needed in evaluation of machine learning system. More research has to be conducted in order to tackle this problem.

17. Conclusion
To create a better online platform for computing education, we have presented a new platform which may greatly facilitate the learning progress of programming.

The platforms may provide different collaborative tools like drawing boards and group chat which let students share ideas effectively. Coding support by customized feedbacks are also provided which would let students learn more efficiently without the help of lecturers.
This report described the architectural diagram of our new e-learning platform and approaches that our team will adopt. This could provide a clear direction for our teams to follow.

The report also stated the libraries and algorithms that are found by our teams which could quicken the development of our final year project.

The future of our work is to implement the full functionalities and select appropriate models for data analytics. This would help our project to achieves the aims letting students learn programming by themselves and to provide a collaborative environment for students to communicate with each other.

18. Reference


19. Appendix

<table>
<thead>
<tr>
<th></th>
<th>K-means</th>
<th>DBSCAN</th>
<th>Agglomerative clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advantage(s)</td>
<td>● Relatively efficient</td>
<td>● Can determine number of clusters</td>
<td>● Manual tagging is not required ● relatively efficient</td>
</tr>
<tr>
<td>Advantage(s)</td>
<td>Disadvantage(s)</td>
<td>Advantage(s)</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>----------------</td>
<td>--------------</td>
<td></td>
</tr>
<tr>
<td>Robust to outliers</td>
<td>• Sensitive to outliers • Need to specify number of clusters manually</td>
<td>• Require manual tuning of parameters • Do not scale up well due to high time complexity</td>
<td></td>
</tr>
</tbody>
</table>

*Table 2. Comparison between ML candidates*
Figure 4. Classification results of K-means, DBSCAN and Agglomerative clustering on sample data
Coherence
\( a(i) \): the average distance (dis-similarity) to all other samples within the same cluster:
\[
a(i) = \bar{d}(x^{(i)}, C) \quad x^{(i)} \in C
\]

Separation
\( b(i) \): the average distance to the nearest cluster that \( x^{(i)} \) does not belong to:
\[
b(i) = \min \bar{d}(x^{(i)}, C) \quad C \in \{C_1, ..., C_k\} \quad x^{(i)} \notin C
\]

Silhouette
\( s(i) \): the silhouette value of \( x^{(i)} \):
\[
s(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))}
\]
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 \( a(i) \ll b(i) \) \to good clustering
- \( s(i) \) is close to \(-1\) \to bad clustering

*Figure 6. Computation of silhouette coefficient*