COMP4801 Final year project


Group member:

Kung Tsz Ho, Bob 3035176729
Chan Ka Yeung, Yeung 3035181580
Lor Cheuk Hin, Macro 3035185108
Hui Pak Nam, Jeffery 303518573

Author:

Kung Tsz Ho, 3035176729
Abstract

This paper introduces a new programming learning platform, Code Nova. The scope of this paper will focus on the methodology, final status and future plan of the project. The methodology of this platform contains five core parts: code analyzer, peer programming, exercise suggestion, AI tutoring and data visualization, which form the basis of the platform. Final status of the project will be explained in detail. This part will mainly focus on the implementation and consideration of the project. Future planning of the project will be clearly elaborated. AI and ML techniques used in the platform will be discussed in detail.
Acknowledgement

We would like to thank our professor Dr. Kit for his advice on the project. The project will be hard to process without his suggestions.
Table of Contents

1. Introduction – P. 9
   1.1 Background – P. 9
   1.2 Drawbacks of major learning platforms – P. 9
   1.3 Project aim – P. 10
   1.4 Project scope – P. 11
      1.4.1 Topic research – P. 11
      1.4.2 Website Construction – P. 12
      1.4.3 MI and AI related – P. 12
   1.5 Flow of report – P. 12
2. Methodology – P. 14
   2.1 Basic concept explanation – P. 14
   2.2 Platform set up – P. 14
      2.2.1 Computer hardware properties – P. 14
      2.2.2 Version Control – P. 15
   2.3 System architecture design – P. 15
      2.3.1 Website platform – P. 16
      2.3.2 Personal Code Analyzer – P. 17
      2.3.3 Peer programming – P. 19
      2.3.4 Exercise suggestion – P. 21
      2.3.5 AI tutoring – P. 22
      2.3.6 Data visualization – P. 22
   2.4 User interface design – P. 23
   2.5 Database design – P. 24
      2.5.1 MySQL – P. 24
      2.5.2 SQLite – P. 24
2.5.3 Database structure – P. 24

2.6 User access/security design – P. 25
   2.6.1 Login/Sign up – P. 25
   2.6.2 User level – P. 26

2.7 Test data and test case design – P. 26
   2.7.1 Data collection – P. 26
   2.7.2 Testing data – P. 27

3. Current status – P. 28
   3.1 Overview – P. 28
   3.2 Website construction – P. 28
      3.2.1 Website framework – P. 28
      3.2.2 Database – P. 29
      3.2.3 Login and Authentication – P. 29
      3.2.4 Back End Function – P. 31
      3.2.5 Front End Design – P. 31
      3.2.6 Code editor – P. 31
      3.2.7 Integration of Systems – P. 33
   3.3 Personal Code analyzer – P. 33
      3.3.1 Pre-analyzer step – P. 33
      3.3.2 Analyzing step – P. 34
      3.3.3 After analyzation – P. 34
   3.4 Peer Programming – P. 35
      3.4.1 Distance calculation – P. 35
      3.4.2 Other considerations – P. 35
   3.5 Exercise Suggestion – P. 36
      3.5.1 Learning a new (random) skill – P. 36
3.6 AI tutoring – P. 37
3.7 Data visualization – P. 41

4. Future plan – P. 42
   4.1 Overview – P. 42
   4.2 Platform – P. 42
   4.3 Website – P. 42
   4.4 Personal Code Analyzer – P. 42
   4.5 Peer Programming – P. 43
   4.6 Exercise Suggestion – P. 45
   4.7 AI tutoring – P. 46
   4.8 Website Construction – P. 47

5. Conclusion – P. 48

6. Reference – P. 50

7. Appendices – P. 51
List of figures/tables

1. The trend of demand for CS jobs – P. 8
2. MVC model of Django – P. 14
3. Ability of two students – P. 16
4. A responsive website layout – P. 18
5. A sample MySQL table – P. 19
6. The plan of the final year project – P. 26
Abbreviations

1. ML – Machine Learning
2. AI – Artificial Intelligence
1. Introduction

1.1 Background

Figure 1. The trend of demand for jobs is clear. It is increasing fast from 2011. More students have seen the benefits of receiving computer science education.

Learning programming is a trend in recent years. Coding is becoming an important ability in the world (see figure 1) [1]. Despite the differences in academic background, more students are taking computer courses in universities compared to past decades. Their interests in programming have attracted attention in the education industry. Therefore, many learning platforms have been established over past few years. These platforms’ common intention is clear: to help students learn to code with their own hands. They claim that students could become a professional programmer after finishing tasks assigned by the platform. Learning programming through those platforms seems to be an effective and feasible way.

1.2 Drawbacks of major learning platforms

Although these platforms are highly successful and have thousands of students, they are not perfect and can be improved in many aspects. The project team has found many drawbacks and problems of major learning platforms.
First, tasks given by the platform are all individual programming works. In the real world, though, programming is interactive, with most projects being progressed by a group. Collaboration is a key to programming. From leaving comments in the code to writing code specification for the project, programming is seldom an individual work. The platforms have neglected this need and students are therefore not aware of the importance of collaboration in programming, which should get noticed and be improved.

Second, no personal feedback is given to the user after they finish the tasks. Most of the platform only show the result. If the code passes the test and gives a correct result, it does not mean the code is perfect. How good the codes are, which can be judged from readability, redundancy, the flow of logic and accuracy of the functions, are not covered in the judgment. Attention should be given to the codes and the results rather than the result only.

Third, platform learners may not get enough assistance during coding. Most of the platform does not give any advice when the learner is typing the code. If a user encounters any difficulties during the process, the platform does not know the situation. The user can only seek help from others. It may cost hours searching and make users frustrated.

Fourth, platform cannot evolve after collecting students’ solution. Most of the platform did not provide any statistics of the problem. The tags of the problems are all manually added. From the perspective of teaching, it is not using the data collected from students and improve the learning environment.

1.3 Project aim

This project is aimed to establish a new code learning platform, which will overcome the drawbacks and use new techniques to help students to learn to
programme in a friendly environment. The platform is named Code Nova.

The platform will be able to give advice and assist students to make progress in coding tasks. Students can ask the platform to pick an exercise for themselves. In the same time, they can decide to enhance a particular skill of themselves.

Besides, students can learn to programme in many aspects. Not only do the team focus on the cooperation in the learning platform, but the project team provides a system to seek help. Students in the platform can either act as a helper or a person asking for help. Using this platform, students can learn the skills of communication and understand the importance of working together. During the exercise, AI can assist the student to provide hints. Those hints are provided based on the student current process and code.

After completing tasks, the platform will record the students’ skills. The platform will show the profile of students. Students can learn their disadvantage and areas to improve. The platform will keep tracking the user development in multiple aspects, and students can improve their ability in a comprehensive way.

1.4 Project scope

In order to implement the platform, the project is divided into several parts. The first task is topic research. Research of related topics started in August. Previous work and essays on similar topics were studied, which would be useful in platform construction. There are several main focuses of the research.

1.4.1 Topic research

The first one is the extraction of students’ solution. Students’ solution are written in code format. There is no direct way of using code to discover meaningful clues for machine. Large time were spent on investigating the conversion of code into numbers without losing useful information.
The second one is the analysis of the students’ solution. If there are only numbers presented, the result may be hard to understand. The conversion from number to meaningful suggestion is our focus. Multiple testing and investigation are done during the project.

1.4.2 Website Construction

The second part is the website establishment. The website consists of back-end and front-end functions. The website construction has been ongoing since October.

For the front end, the main focus is the data visualization. Attractive web design is one of the concern of the project. Besides, good data visualization will enhance the learning experience for students.

For the back end, the main focus is the concurrency of the text between users. Due to collaborative programming in the platform. The concurrency of editor need to be ensured and stable. It took weeks to solve the problem.

1.4.3 MI and AI related

The third part is the AI and ML related things, which includes four sections. They are personal code analyzer, peer programming, exercise suggestion and AI tutor. All four parts involve AI and ML technologies. Again, studies on related materials are required.

The main focus is the ML part, whether a pattern of code can be found in the students’ solution. Multiple methods are discussed and investigated during the project. The stability and effectiveness of those methods are compared.

1.5 Flow of final report

In this report, the methodology of the work will be explained in developmental approach.

After methodology discussion, this paper will report the final status of project team.
Completeness of the project, difficulties encountered and method consideration will be elaborated in detail.

Then, future planning and possible improvement of the final year project will be shown. The project has lasted for nearly one year. The parts need to be further investigated, functions need to be redefined and methods need to be re-considered are written in details. We list all the possible working directions to have an improvement of the platform.

At last, there will be a conclusion for the final report. A brief feedback of the project is included.
2. Methodology

2.1 Basic concept explanation

Before driven deep into the detail of methodology, the concept of AI and ML are needed to be explained first. In short, AI is an approach of imitating human mind. In this stage of technology, data are collected and stored in the database. Those data are used to train the machine to understand human action. Thus, further action can be taken to help the platform user. For ML, it is a model which it can improve by itself. In other words, the model has the ability to correct itself. The model can be used as error collector in code, which analyses the frequent mistakes appear in the code among platform users.

Although both AI and ML are big topics, only several ideas of those topics are used in this project. The detail will be described one by one. Since the final goal of the project is to establish a programming learning platform, the methodology will be explained in developmental approach.

2.2 Platform set up

2.2.1 Computer hardware properties

For the platform setup, the faculty has provided servers for final year projects. A server running Windows Server 2008 is provided for this project. This server is widely used in the industry and has an Intel Xeon CPU, which is sufficient for the project usage. It is estimated that the ML method will consume the majority of CPU calculation power. This set up is flexible for further changes. If there is any need in future, the change of the set up will not cost much to change in system design.

For any unexpected issues, a backup server is prepared. The server is running on Ubuntu OS instead of Windows. The platform can operate on both OS without
additional management.

2.2.2 Version Control

Besides server implementation, the code are pushed on the GitHub for collaborative programming within the project team [2]. Latest update and source code are uploaded on the master branch. All dependent library are listed in the Readme.md file. The code can be run on local machine after setting the environment properly.

2.3 System architecture design

For the system architecture design, it consists of six major parts. They are website platform, code analyzer, peer programming, exercise suggestion, AI tutoring and data visualization. Six parts have different roles in the platform. Each part will connect with others for the different purpose. To be brief, website platform is responsible for the user to interact with the complete system. In the flow of the system, users choose what he want to practice. Thus, the system will suggest an exercise for the user to test. Next, a typing interface is shown. User can type his code in a text editor. Code typed are first under grammar passed to code analyzer. The code analyzer will analyze the code and judge the programming level of the student. Peer programming will compare his ability with others. The record of that user’s programming skill is kept tracked by the individual skill tree system. After collecting codes and ability records, system will extract meaningful statistics and generate feedback to the student and platform user. The database will store various data of the student during his usage. Each part will be clearly explained in following parts. To be reminded, the flow of collaborative programming will be slightly different from the single user environment. The difference will be clearly elaborated.
Figure 2 The flow of the platform

The flow of the platform is shown as above. Various parts are built by different team members.

2.3.1 Website platform

The website operation will take control of the duty and give a responsive, instant action to the platform user. The user interface, interactive component, and functions of a website are included in this part.

2.3.1.1 Web framework

In order to satisfy the requirement, Django is chosen to develop the website. By Django, the website establishment is easy. Once installed, a website can be developed within hours. It provides the main structure of the website. Besides, the modification of website is quick, each part of the website is independent. One minor change in part of the website will not cost much to change. It is because Django adopts the Principle of MVC structure.

From the source code, it is clear to see the structure of the website. Django
considers different roles of the webpage as different apps. The development of different apps can be done separately.

2.3.1.2 Text editor

The text editor is used for the online programming. It has to provide the base function and support collaborative programming, which means multiple users can write code and synchronize it instantly.

2.3.1.3 Interface design

![Figure 3 The homepage of the website](image)

The platform will have a home bar in every webpage for redirection. The page design will be different in every webpage. Icon and pictures are added to have a clear identification for platform users.

To conclude, the major role of the website is to receive user command and action. In the flow of usage, the platform user will type his code on the website platform. Then, the code will be passed to the personal code analyzer.

2.3.2 Personal Code Analyzer
Figure 4 The AST tree of the platform

The partial structure of AST tree is shown in the graph. More details is explained in the later part of the report. The root node is the skill tree. The second layer is the main aspect of the graph.

2.3.2.1 AST Formation

It is supposed that user type code on the website platform. After this action, the code analyzer will start tokenizing the code. Tokenizing means chop the lines of code into every single word. Those words are divided into several categories, for instance, variables, string and library functions. In some computer languages, for example, python, there is an official tokenizer library, provided for developers. The official tokenizer is reliable and the result is satisfying. Using official tokenizer, the development efficiency is increased.

Tokenizing the code is the beginning of work in code analyzer. The responsibility of code analyzer is to recognize the code pattern, find the common sequence and identify the possible error in the code. To achieve this function, the tokens, single words generate from the tokenizer, will be processed. From each token, the code analyzer will analyze them and try to form a tree structure of tokens, which means recombination of code in a meaningful order. Thus, further work on this order can be done. The tree formed is called the AST tree. This tree will be used in AI tutoring.
2.3.2.2 Other AST tree usage

Besides, code analyzer will also evaluate the programming skill of the platform user. Each success case of the program will be recorded in the code analyzer. Base on the AST tree, those records contain the ability used in different programming tasks. Those records are necessary for peer programming and individual skill tree. The usage will be mentioned in related parts.

2.3.3 Peer programming

2.3.3.1 Ability evaluation of programmers

The function of peer programming is to find students of similar programming ability and group them together. Code Nova will provide task requiring pair programming and group programming, which needs more than one student work on a task. Finding students with similar programming ability is important.

For instance, there is a task requiring two programmers to work together. System randomly finds two programmers, A and B, pair them together. If A’s ability is much higher than B, it is possible that A finishes the task by himself and B does nothing. If A’s ability is much lower than B, it is possible that A does nothing and B finishes the task by himself. Both situations should be avoided. The expected situation is, both A and B have similar programming ability and work together.
To achieve this situation, peer programming system needs to evaluate the programming ability of each student and find the difference of them. From personal code analyzer, the programming skill of each student is recorded. Peer programming system will transfer this record into numbers. Number is used to represent the level of a user programming skill. The benefits of using a numbering system include easy to understand and compare.

![Ability between two students](image)

Figure 5 The ability of two students. (Numbers are not real. The real part will be much more complicated)

From this figure, it is easy to see the difference of ability of two students. The programming ability is divided into four parts. They are string, loop, object and library. This system is able to measure the level of difference of ability. For simplicity,
Euclidean distance is used. Larger distance represents larger difference in ability. Smaller distance represents similar programming ability.

The evaluation method used is Euclidean method. Based on different methods used, Euclidean distance may provide a suitable value with ease implementation. Other methods like Pearson method, Manhattan distance are considered. But the result may not be as straight as Euclidean method.

Using this system, the platform can search users with similar level of programming skills.

2.3.3.2 Suggestion preference

The ability distance is one of the consideration of the peer pair. They are other things to consider. The one which has completed the same task will have a higher priority than the non-complete one, even the non-complete has a closer value based on Euclidean distance.

Besides, the new version of peer programming will consider the correct solution of other users. The system will judge which method the seeker is using to solve the problem. Then, the system will find a helper which can help seeker to complete his own solution. The detail of implementation will be explained in later part.

2.3.4 Exercise suggestion

Exercise suggestion is based on the user’s skill tree. To explain the exercise suggestion, there is brief introduction of the user skill tree.

2.3.4.1 Individual skill tree

Individual skill tree is simply a profile of each platform user. Students can see their record, previous action and their programming skill in this platform. The platform contains various records and generates a timeline of each user. Students can know the development path of each programming skill.

Individual skill tree works with code analyzer. The AST tree generated from code
analyzer will be converted into skill tree. Partial AST tree items are used to form and update the skill tree. In other words, individual skill tree is a subset of AST tree which contain part of the data only.

2.3.4.2 Three types of suggestions

There are three types provided for suggestions. The first one is learning a new skill. The system will randomly choose a new skill that the user does not know before. Then, a problem that require that new skill will be suggested. The second one is learning a particular skill. The system will ask the user to choose a new skill that the user does not know before. Similar to previous one, a problem that require that new skill to solve will be suggested. The last one is reinforcing skill. The system will choose problems that have all skills learnt by the user.

Therefore, if a student lacks a particular programming skill, platform will recommend tasks that can train and strength the skill in that particular part.

2.3.5 AI tutoring

AI tutoring is a function that based on the previous record of all students solutions on a particular question.

From various data collected from platform users, it is predicted that patterns can be found in different areas. For instance, system may find common skills or terms needed to solve the problem. Thus, when the user types in the wrong answer, the system will suggest the user with specific sentences. For instance, the system will suggest the user to use a particular library function to solve the problem. All these recommendation are based on the previous data collected from others.

2.3.6 Data visualization

The system will visualization the data using graphs and charts.

For exercise, the user can clearly use the numbers of methods that users used in an exercise. The proportion of solutions of different methods will be shown.
For personal profile, the user can clearly use which skill he has learnt and which skill he does not know.

As mentioned in the previous part, the principle of data visualization is to ensure that user can know about his personal ability. Besides, data visualization will help the user to understand better in a particular problem. Not only for the student, teachers can know more about students from those graphs formed.

2.4 User interface design

User interface design is a website interface. The place where users see, response and interact with the website.

The theme of the website is light, picture and icon based because a great contrast between the background and the text content is preferred in this type of website. Those icons and pictures are easy for users to understand the meaning of the text. It helps establish a user friendly environment.

Besides, an important consideration of the visual art should be mentioned. The website is a responsive one. Responsive means website can adjust its display according to devices’ monitor size. The website is accessible for desktops, laptops, tablets and mobile phones. Each device may have a different size of the monitor, the display of the website may be awkward on some device. Using third-party library can solve this problem in an easier way.

Different from the visual arts, the interactive part of website is implemented by functions. Similar to visual part, HTML and JavaScript file is responsible for the interactive part. HTML contains the structure of a webpage and JavaScript contains the interactive parts of a webpage. Several plugins are used to enhance the website functions. The online text editor is one of them. A Combination of these elements
forms a complete and user-friendly platform.

2.5 Database design

2.5.1 MySQL

For database design, MySQL is chosen for this project. It is able to retrieve data from the database within milliseconds. Compare with traditional text file storage, MySQL database is smaller. The speed of data operation is much faster [2].

The design of the database is simple. A table is created for the platform. Under this table, there is a row for each user. Each column of the user is responsible for different data. From username to the personal skill tree, all these data are stored in the MySQL database. For the task in the platform, a new table is needed to store those data. Each task has its own row. Each column represents different data of the task. All data will be backed up every day. In this stage of development, it is predicted that the usage of the database is not heavy. MySQL will be able to handle any daily usage of the platform. If the database operation exceeds the calculation capability of MySQL in the further, it is easy to migrate the data to other servers. In other words, data is flexible in MySQL.

2.5.2 SQLite

Although MySQL is used for the platform, during development, SQLite can be used for fast testing in local host. SQLite is the default database of Django. If MySQL are not installed and set up, SQLite can be used as a temporary testing database. Because both MySQL and SQLite are SQL-based database, the data between two databases can be transferred without loss.

2.5.3 Database Structure
Figure 6 Table structure of the platform.

The database structure of the platform is shown above. It contains six major tables. Each tables representing different data for different apps in Django. In the figure, PK means the primary key and FK means the foreign key. The arrow is the referencing the key that foreign key is referring.

2.6 User access/security design

2.6.1 Login/Sign up

To use the platform, the user must have an account. No trial version is provided. New user can sign up for the platform. The platform will record the username and password.

For security design, all users need to login in before usage. Login is compulsory. Once logout, users will prompt out to the login page. A webpage cannot be accessed by logout users. For safety issues, login status will be expired after some time and users need to log in again. It is used to avoid account being used by others. Temporary codes are stored in server before logout.
2.6.2 User level

For user access design, three kinds of the user are set. The first type of user is the administrator, which is the developer of the platform. The administrator has the full access and privilege to any data. They can modify the code and change the logic of the platform. No barrier will be set for this kind of user. The second kind of user should be students, the most common type of user of the platform. A student can access his own profile, viewing his skill tree and work on the tasks at his level. The third kind of user should be teachers. Teachers have the same access to the platform as student users. Beyond that, the teacher can work on any tasks on the platform. He can also view the record of students belong to his classes. More functions may be provided to students and teachers in later stages.

The level of user can only be changed be administrator. Any new account signed up are considered as student user.

2.7 Test data and test case design

2.7.1 Data collection

The platform requires test data for ML and AI to train and operate. Therefore, every solutions collected are stored in database for research use of the platform to adjusting the ML and AI functions.

Currently all exercise shown in the platform are collected from other programming websites. Our team are using programming questions and collected solutions from Kaggle. There is a database provided by Code Chef which contains various data of programming.

But for the later development, test data and test case design will be mainly designed by the teachers. Every task should have at least several test cases before releasing to the public. Those test cases should also be designed by the teachers. Students can
design their own test case for each task. Students can submit his own test case to the platform. If the test case passes, it will be adopted as an extra test case and store in the database.

Besides test case of programming tasks, other function of the system will be tested during the implementation. Each part requires individual testing. Fake data will be created during the test. The accuracy and result of each function will be recorded. Those data will be shown in next part. Due to the comparison of different parameters, the team has modified some ML and AI functions for a better result.

2.7.2 Testing data

Collecting from real data may not be enough for the development. Some fake data are created manually for extreme case testing.

After creating fake data for testing, it is predicted to have a fix and improvement on the system. Real data can reveal more accurate result. To make sure that the system functions well, students might be invited to participate, since user feedbacks are vital for a website. Being user-friendly is an important consideration. If users feel hard to use the platform, they may not use this website again. Inconvenient website interface may cause the same effect.
3. Final status

3.1 Overview

The project team works on six parts concurrently. Same as the part mentioned in methodology, they are website construction, code analyzer, peer programming, exercise suggestion, AI tutor and data visualization. Each member is working on different parts and the progress of each part will be explained in this section.

3.2 Website construction

Website construction is fundamental to the project. Without the website construction, functions are not combined together. Team member Marco is working on this part.

A complete website consists of six parts: website framework, database, login and authentication, back-end function, front-end design, code editor and integration of systems. Details will be explained one by one.

3.2.1 Website framework

The website framework of the learning platform is Django, which is mentioned in methodology. The project team has experience of developing a website using Django. The version of Django used in 1.X. The latest version of 1.20 is not used due to compatible problem. To be reminded, the programming language used by Django is Python 3. Functions implemented by others team members can be directly used in Django.

As mentioned in methodology, Django forms the based skeleton of the source code. Each webpage are divided as apps and controls different part of functions. Each apps stores the webpage template.

To be short, Django is easy to use, implement and develop. It suits the project requirement and be reliable in the project.
3.2.2 Database

3.2.2.1 SQL database

As mentioned in methodology, MySQL is used in this project. Each team member is familiar with this database. It is predicted that choosing MySQL is the most confident way to develop among databases. Normal usage of MySQL should not cause trouble in this project.

Also as mentioned in methodology, SQLite is used in this project for local deployment. The data of MySQL and SQLite can be transferred directly. Before hosting on school server, the platform is tested locally using SQLite. It can withstand large data read and write operation.

3.2.2.2 Database buffer

Due to the usage of code editor, the communication between users is required. Codes of different need to be synchronized instantly. To achieve this point, using SQL database is not enough. Read and write of the SQL data took longer time than expected.

As result, the team use Redis as database buffer to temporary store the data before reading and writing to the database. It can faster the speed of synchronization between users.

3.2.3 Login and Authentication

This part was implemented by team member, Marco. Simple testing was started and no bug has been found yet. The user will be redirected to login page when direct accessing other webpage of the platform. If the user has no account yet, user can sign up a new account. The platform will ask the user to type in the new account name, email and password for twice. Then, after registration, the user can use the platform using the new account.

The default setting of every new user is considered as student. To upgrade the usage
privilege of the user, only the administrator can modify it in a totally different panel. The panel is only shown to the administrator. Without administrator account, the panel cannot be shown. Through this panel, administrators can modify the users’ data, including the privilege level.

Figure 7 Login page of the platform

Figure 8 Register page of the platform
3.2.4 Back End Function

Backend function means modification of core functions to suit the online platform. The compatibility of core functions is not considered during construction. The parameters, return items and function names are constant. It is to ensure the stability of the system. The naming and parameters are set as the earliest stage.

3.2.5 Front End Design

As mentioned in user interface design, front-end design is included. The use of HTML, CSS and JavaScript forms the core part of the front end design. Django has provided a simple to develop the webpage. Website template can be used for different page of the platform.

Icons and images are largely used in the website. Mentioned in methodology, it constructs a user friendly environment for every user. For instance, icon reflects the characteristic of the problems.

3.2.6 Code editor

Code editor is the most important part of the website construction. The code editor need to satisfy the following requirements: provide the basic editor functions, support
program functions and able to synchronize the code between users.

3.2.6.1 Basic programming editor

The online code editor can provide basic functions that a programming editor should have. The programming code are highlighted for a better programming experience.

Although the interface of the code editor is hidden, it actually can react to the common shortcuts, for instance control + c, control + v. To achieve these functions, Ace editor is used and modified by the team. It is a stable, reliable and fast-loading editor which suits the platform. The color and UI of the editor is modified for a better looking.

The code editor provide a communication window. When there is only one user, the user can receive message from the platform via the communication window. When there is more than one user, users can speak and treat it as a real-time communication channel. There are other usages of the window. They will be mentioned in following part.

3.2.6.2 Program functions

Due to the use of programming platform, the code editor must provide testing for passing the questions. Therefore, several buttons are implemented below the code editor. User can tape the button and pass the code to run. The codes will need to pass several test case. If the codes has syntax error, the platform will show the error message in the communication window. Besides, for very simple syntax error, the text editor will show the red curved line under the wrong code.

3.2.6.3 Synchronize between users

Synchronization between users is not simple. Mentioned in the previous database part, Redis is used as database buffer to read and write data quickly. It solves the data storage problem. But the communication between users need to establish a stable
network channel first. A library called socket is needed to solve the problem. Socket will automatically arranged the socket used between different users.

Addition to the programming code, users can communicate in the communication window. The text are handled similarly as code, which is transmitted instantly.

3.2.7 Integration of Systems

Systems are integrated in the platform. All functions are tested during the actual usage.

3.3 Code analyzer

The team member Jeffery and Yeung works on this part.

The flow of the code analyzer is:

1. Code analyzer will act as a normal compiler. Any syntax error in the codes will be discovered and shown to the user.

2. If the code is correct in syntax, then test case will be passed to the code. If all test case are passed, go to next step. If not, the error will be shown and the failure test case will be shown. If there is problems running the code, including infinity loop and very large running time, the platform will return a runtime error to the user. The detail will be shown.

3. After passing all test case, the program code can be seen as correct answer. The code will be analyzed. If the code cannot be analyze, an internal error will be record. No error report to the user.

4. If analyzer runs well, the result will form an AST tree. The AST will the store in the database for later usage.

3.3.1 Pre-analyzer step

Before running the analyzer, the platform need to ensure the code is correct. Therefore, the code are passed into compile first to find out any error in the code.
If there is no problem in the code, test case will be passed into the code. As mentioned, various case will fail the test case including long running time and infinity loop. Accessing unsupported library or doing something not permitted in the code will also return errors. For example, some functions in OS library is banned to avoid damage caused to the server.

3.3.2 Analyzing step

If the code is considered as correct, the code will be passed to analyzer. Analyzer is built based on the AST library of Python3. The library will re-run the code to operate tokenization. Then, at the same time, the team has defined lots of specific rules to identify the tokens. With the assist of library, the tokens are grouped base on their functions. From basic I/O to library calls, all tokens are used to record and the number of appearance are stored to form an AST tree. The team basically defined the aspect of the token into different aspects. They are basic I/O, condition, loop, array, function, class, lambda, comprehension, data structure, py-str, py-builtin, module. Besides, the maximum depth of loop, longest array length and maximum array dimensional are also stored as specific aspects. Below those aspects, there are different detecting items. In total, there are more than 100 items to record in the AST tree. The tree structure is defined in the early stage of the project. After several times of modification, the tree are consistent and not change. It helps the team to work with the tree separately.

3.3.3 After analyzation

The tree formed from the analyzer is the in JSON format. To store in database, it is first converted in string format. Similar after reading data, the string format need to be converted in JSON.

The tree formed will be used to update the personal skill tree. The skill tree is a subset of the AST tree. Only the necessary part is contained in the skill tree. The
personal skill tree is used later in exercise suggestion.

3.4 Peer Programming

Team member Bob is responsible for this part. As mentioned, the number system is decided to represent a user’s programming skill. Then, the system will find the users of similar ability and group them together.

3.4.1 Distance calculation

The comparison methods of programming skill are various. The combination of methods can cause different results. Euclidean, Pearson, Manhattan, those methods are compared. The final choice is the Euclidean method. The distance is the straightest method to see and understand. Pearson method compares the “shape” of all ability instead of compare the difference of each ability. Manhattan is not suitable in the ability comparison. Thus, Euclidean method is used and the result shown is stable.

Although Euclidean method will compare all ability between users, it is improved to take only the necessary part into consideration. To explain, only the skill used in the question will be considered and taken into account. The skill used is evaluated from the previous passed solution submitted by other users. After analyzing, the method of the solutions can be found. Base on the majority method used, the system will only use the skill appeared in the major method. The detail will be explained in the AI tutoring part. It is tested that the improved Euclidean method will return a better and more suitable helper to the seeker.

3.4.2 Other considerations

Difference of ability is not the only consideration. Any user who has completed the question before will have a higher priority in the peer suggestion system. The system treat the user who has completed the question will be able to help the seeker to solve the problem.
It is told that the new Euclidean method will consider the previous passed solutions. If there is no enough solution for reference, the original Euclidean method will be used instead.

Special cases still need to handle with care. For instance, if there is a sudden connection loss, the system find another user to reform a pair or to reconnect with the previous user. Besides, helper may refuse to help others. Each time system will suggest 10 people to help. The first who accept the request will join the room and help. Others will be rejected and request will be canceled.

3.5 Exercise Suggestion

It is mentioned that there is an exercise suggestion in the platform. When a user cannot decide which questions to practice, especially when there are lots of questions provided in the list. The platform will try to suggest an exercise to solve the problem.

There are three types of suggestions provided. They are learning a new (random) skill, learning a particular skill and reinforcing skill.

3.5.1 Learning a new (random) skill

The system will suggest an exercise which contain a skill that the user did not learn
before. It is based on the personal profile of each user. Each user has a profile which record the skill used in the past, which is mentioned in the Analyzer part. At the same time, each question will be marked with tags base on the previous passed code. The exercise suggestion will scan all the exercise in database. Any suitable exercise will be stored in a list. At last, the system will randomly pick a question.

3.5.2 Learning a particular skill

The system will suggest an exercise which contain a particular skill that the user choose. Similar to previous part, personal profile and exercise tag will be considered.

3.5.3 Reinforcing skill

The system will suggest an exercise which the user has all skills required by the exercise. In other words, the skill required by the exercise is a subset of the user learnt skill. It can ensure that user can solve the question.

If there is no question that all skills required is learnt by the user. The system will then suggest the one with most skill learnt. The implementation of comparison of skill is the using set difference.

3.6 AI tutoring

AI tutoring is used for the helping student to solve the problem. Mentioned in methodology, AI tutoring is based on the previous passed solution from other user. When the user type answer that cannot pass the test case, the system will try to lead the student to the correct solution. The flow of AI tutoring is:

1. Student types in wrong answer which has no syntax error but fail in test case.
2. System will analyze the current code.
3. System will compare the current code with other methods used by other students.
4. System will find the most suitable one. Trying to find out the difference.
5. Suggest is given by the system.

3.6.1 Compare the current code with other methods

Before comparing the current code with other methods, the system need to find out the method used in the previous code. The process the platform used is clustering. As passed code form an AST tree, clustering can be done based on the AST tree generated. The flow of clustering is:

1. AST tree is converted from the original passed code.
2. Multiple AST tree is collected to undergo clustering
3. Before clustering, those AST trees, which are JSON data, are converted and undergoing data preprocessing. The data preprocessing used is standardization.
4. After standardization, data cleaning is performed. Only parts of the tree data is remained. The tree data is simplified for a better result.
5. After preprocessing, those data are clustered. The method of clustering is mean-shift clustering.
6. After clustering, the detail of cluster is recorded and stored into the database.

3.6.1.1 Data preprocessing

Standardization is used as data preprocessing method. Other methods such as normalization, are considered and tested. It is considered that normalization may loss the meaning and may be hard for the next step, which is the data cleaning step.

Standardization is used for a better result. It is found that after standardization, the feature of each code can be remained. The clustering result after standardization is much clearer.

After standardization, data cleaning is performed. Only parts of the tree is considered in the cluster method. It is because the AST tree formed is too large. Only part of the tree is useful in the question. Due to the equal weighting of tree item,
larger the AST tree, larger the item it need to consider. Smaller tree can get to a better result. The team has considered several method, including using top-k item and PCA. In this stage, the team has manually cut the tree by throwing some unnecessary part.

3.6.1.2 Clustering

After data preprocessing, clustering is performed. Two clustering method is considered at the stage. The first one is k-mean clustering. The second one is mean-shift clustering. Finally, mean-shift clustering is adopted. The reason of not choosing k-mean is because k-mean clustering requires to indicate the number of cluster first. For single problem, it may not be a problem. But for a smart platform, it is unfeasible. As lots of problems need to be handled, the number of methods cannot be known before clustering. The test-and-modify approach does not suit the situation. But mean-shift clustering is different, this method can decide the number of cluster by itself rather than typing a certain number. The result given by mean-shift clustering is stable and reasonable.

The team has modified some parameters of mean-shift clustering. The bandwidth of mean-shift set in the platform is higher than the default. Lower the bandwidth, more clusters will be formed. The team choose a number will generate a small number of cluster but not too generalized. The number of cluster formed is usually below 10.
3.6.1.3 Recording data cluster

After forming cluster, some data are recorded and used for later usage. The first one is the grouping of the solution. The solution are grouped in clustering method. The group number will be recorded. The cluster point of the cluster will be recorded.

Besides, cluster will be recorded with tags. The tags are formed based on the skill used of solutions under the same cluster. The tags are the characteristic of the cluster. Then, the question tags are formed based on the tags of clustering. They are mainly based on the largest cluster’s tag. (The largest cluster mean the cluster which majority of solution belongs to.)

3.6.2 Finding the most suitable method

The system will compare the AST tree of the user code with other cluster. The cluster with least Euclidean distance will be considered as the most suitable one. By comparing the difference of AST tree, if there is something missing or redundant in the user code. The system will generate comments in the communication windows and try to lead the user to the correct solution. The comments are generated based on
the cluster tags formed after clustering. If there are multiple tags in the cluster, system will return the most significant one. For example, if the user lacks using some skills which is equal to the cluster tag, the system will suggest the user to the skill that the tag means.

3.7 Data visualization

Data visualization means the data collected from various function, like clustering and code analyzer, are presented to the platform user in a much clearer way.

After the clustering, the clustering data are recorded. If the student open the information page of the question, a page will be shown which contains the information of the problem. The problem tag, the method used, the distribution of usage of different method, all those data are shown in graphs and charts. Students can easily know the most used method and the skill tested in the exercise,

For personal skill, the user can see his personal skill tree in the profile page. When the skill is not learnt, the color is in grey or hidden. When the skill is acquired or learnt, the skill is colored.
4. Future plan

4.1 Overview

Due to the time limitation, there are still lots of thing can be improved in the platform. If the project can be extended, improvement of the platform will be the future plan.

4.2 Platform

The platform is set using the school machine. The IP address is not public. Connection to the school machine requires a CS VPN account. It means that the platform is not open to the public. If the platform is opened to the public, a public IP address is required. It will be better to buy a domain name to be accessed easily.

At the current months, the school server connection is not stable. The reason is unknown. It will be better if using other IaaS provider. For example, AWS or Digital Ocean, they can provide a much stable server platform to implement. The SSH management is better than the school server.

4.3 Website

The website can be improved in various way.

4.3.1 Responsive layout

The responsive layout is not stable. Many webpage are not truly responsive. It is recommended to use Bootstrap to be the CSS structure. Then the html should also be modified also for using Bootstrap.

4.3.2 Layout design

The design layout can be improved by adding daily and night mode. More programming website has provided a bottom to switch the webpage from daylight mode to night mode. The user can use the platform in a comfortable way.
4.3.3 User experience

The platform has not asked lots of students to use for a real user testing. If more user can join the testing, the problem or inconvenient part of the website can be indicated. Therefore, the most important thing of improving the website is asking students / volunteers to participate in it. Thus, the platform can be improved in a much user-friendly way.

4.3.4 Exercise provided

At current stage, exercise provided are all collected from other programming website. For later usage, the platform should let users themselves to design problems.

4.3.5 Language

At current stage, only Python3 is supported. For building a better programming learning platform, more languages should be supported.

4.3.6 Code editor

First, code analyzer lacks the modification functions by user. User cannot change the highlight of text. The UI should be changed by the users by their own favorites.

Second, code analyzer should be smarter. Before submitting the answer, the code editor should be able to find out the error by itself.

Third, code analyzer should be able to let user to finish the exercise in the next time. The code analyzer does not store the code temporary in this stage. If the user left the platform, the previous code will be empty. The platform should store the code temporary in cache or in database for user to complete it next time.

4.4 Code Analyzer

Code analyzer can be improved in many ways.

4.4.3 Problem code analyzing

At current stage, the code is analyzed after the code has no syntax error. Error code
are not stored and analyzed. If the code analyzer can record and investigate the error made by the user. Common mistake and the pattern of the mistakes can be found. It will be very meaningful to the user. User can recognize his flaw in the code. The AI tutor can identify the problem in a much more accurate way.

4.4.4 Multiple language supporting

Only Python 3 is supported in this stage. For supporting multiple language, a multi-language supporting code analyzer is needed. The first step may be implement a Python 2 supported analyzer.

4.4.5 Path recognizing

Students may achieve the correct answer by trying and submitting the solution for many times. The code analyzer should analyze the path of the code, which means the code change after submission. Then, after recording it and finding the path, the system can know whether the user is getting helping, following the AI tutor or falling into another new problems.

4.5 Peer Programming

Peer programming can be improved in various way.

4.5.1 Number of helpers

At current stage, the system cannot know which the correct number of helper is. There is no limitation on the number of helpers in this stage. More helpers are asked to help the seeker does not mean the increase in the helping efficiency. The system should find a way to identity the suitable number of helpers of each question.

4.5.2 Way of help

In the earliest stage, the team design that the helpers may have limitations during helping. Instead of directly typing code to complete the exercise, the help should give suggestion to assist the seeker to finish the problem by himself. The idea is not
implemented due to various limitation. If difficulties are encountered, roles of helpers should be defined. Some helpers may only help the seeker by typing comments in communication windows.

4.5.6 Finding helpers

The current method of find helpers is based on the helper’s ability, method used in the problem and whether the helper has completed the solution before or not. The algorithm used is Euclidean. The improvement can be combining various distance, considering the algorithm has different weighting and evaluating the result in a totally different way.

Besides, helper’s may have a record after helping the seeker. It will be better if the helping ability of helping is considered. It requires modification of the skill tree, which treats helping as one of the new skill and recorded in database.

4.6 Exercise suggestion

Peer programming can be improved in various way.

4.6.1 Difficulties ranking

Exercise suggestion is based on the problem tag and the ability learnt by the user. But the difficulties of the problem is not calculated and considered. There should be a difficulties ranking system. For a new programmer, only easy and medium level problems are suggested. For an experienced programmer, hard and even more challenging problems should be suggested.

4.6.2 Distance Evaluation

The problem should include the distance evaluation between the users’ ability and the problem difficulties. Continue from the previous part, the distance should be calculated for a better suggestion.

In previous version of exercise suggestion function, Euclidean method is used. It
should be combined with current method and improve the function.

4.7 AI tutor

AI tutor is the one of the most important part in the platform. It can be improved in various way.

4.7.1 Clustering

Many clustering method are researched and tested during implementation. Mean-shift is the suitable way in the current stage. But to achieve a better result, the platform should combine the result of various algorithms and give the weighting of each result. The result can be better.

Besides, the AST tree collected are flatten before clustering. The hierarchy structure of the AST tree are not used. Hierarchy clustering should be tried and tested. It is believed with considering the hierarchy of AST tree, a much better result can be returned.

4.7.2 Data Cutting

Because there are too many items in the AST tree, which lower the weighting of each item, the importance of the code cannot be shown. Therefore, data cutting method are necessary. Currently the AST tree is manually cut, which the team only pick the important thing for every problem. In later stage, the system should be able to do it by the system itself. PCA and various algorithm should be considered. Then, the code analyzer can perform better in different questions.

4.7.3 Suggestion made

The suggestion made by the platform is based on the analyzing result. The sentence formed by the platform is not easy to understand enough. The tone of the AI tutor is still cold and not user-friendly. An improvement can be made by changing the AI tutor tone and forming a human-like sentence.
4.7.4 Speed of clustering

The speed of clustering is not fast. The library used in the cluster is scikit-learn. Scikit-learn library support CPU only in this stage. GPU is needed to calculate fast for clustering. Therefore, ML part may need to merge to Tensorflow or Pytorch to enable GPU calculation.

4.8 Data visualization

The data visualization should be done by adding more fancy way to present the data. Same as the website UI, this part can be improved by asking the actual user review. Their suggestion will be helpful for the team to improve.
5. Conclusion

The platform design is elaborated in several aspects. The platform is composite of several parts. They are website platform, code analyzer, peer programming, exercise suggestion, AI tutor and data visualization. Each part has its own role and has functionality in different ways. These parts are elementary and form the basis of the website. The implementation, methodology, final status and future plan of each part are described in detail in this paper.

During the project, it is observed that there is a trend in using AI and ML technique in the education industry. More researchers are using AI and ML to assist students to learn new things. Learning platforms are trying to adopt more AI and ML features to improve effectiveness. It is believed that the AI and ML usage in education industry will expand in coming future.

Besides, the platform of the project is not easy to implement as expected at the beginning of project. The feasibility of the function is the largest difficulty of the project. Many aspect are not tried by others before. The study on this field is still empty.

The AI and ML techniques used in the platform need further discussion. Although AI and ML can achieve a lot, it still has limitations. It needs a huge amount of data for analysis. In the beginning of platform, the effectiveness of AI and ML will be decreased due to lack of data.

Further studies are required to get a clear understanding of the model. There are several approaches in ML and AI. Combining the several ML and AI methods are required to have a better result. The code specification has not been written to unify the coding standard. The project team needs to ask volunteers to use the platform to get the real data. It is necessary to invite students to use this platform, which will
enable us to get feedback from students to improve the platform. As a result, the ease of learning is increased. To be truly user-friendly, many works still needs to be done. Finally, it is expected that the project can really help user to learn programming easier with the help of ML and AI.
6. References


7. Appendices

All figures in the report are made by the team. All results are captured from the real website and real test cases.