

ABERYSTWYTH UNIVERSITY

PROJECT REPORT FOR CS39440 MAJOR PROJECT

---

# IntelliJ Plugin to Aid With Plagiarism Detection

---

*Author:*

Darren S. WHITE (daw48@aber.ac.uk)

*Supervisor:*

Chris LOFTUS (cwl@aber.ac.uk)

May 3, 2018

Version: 1.0 (Release)

This report was submitted as partial fulfilment of a BSc degree in  
Computer Science (G401)

Department of Computer Science  
Aberystwyth University  
Aberystwyth  
Ceredigion  
SY23 3DB  
Wales, United Kingdom

## **Declaration of originality**

I confirm that:

- This submission is my own work, except where clearly indicated.
- I understand that there are severe penalties for Unacceptable Academic Practice, which can lead to loss of marks or even the withholding of a degree.
- I have read the regulations on Unacceptable Academic Practice from the University's Academic Quality and Records Office (AQRO) and the relevant sections of the current Student Handbook of the Department of Computer Science.
- In submitting this work I understand and agree to abide by the University's regulations governing these issues.

Name: Darren S. White

Date: May 3, 2018

## **Consent to share this work**

By including my name below, I hereby agree to this dissertation being made available to other students and academic staff of the Aberystwyth Computer Science Department.

Name: Darren S. White

Date: May 3, 2018

## **Acknowledgements**

I am grateful to my supervisor Chris Loftus, for the continued support, inspiration, and guidance throughout this project. The organised weekly meetings provided assurance, and were useful to discuss any problems encountered.

I'd like to thank my loving family and close friends, for their support and encouragement each step of the way.

# **Abstract**

Source code plagiarism is an ever-growing issue in academia, primarily in Computer Science. The majority of tools that exist to detect plagiarism only analyse the final piece of code. This paper discusses the research and development of a new tool which detects how the code is written. The tool will be an IntelliJ IDEA plugin and tracks file changes in the editor. The tracked data gives an insight into how plagiarism evolves over the course of development. This method of detection allows direct identification of the specific pieces of code that were plagiarised.

# CONTENTS

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Overview . . . . .	1
1.2	Plagiarism and Unacceptable Academic Practice . . . . .	1
<b>2</b>	<b>Background</b>	<b>3</b>
2.1	Existing Systems . . . . .	3
2.2	IntelliJ Plugin SDK . . . . .	4
2.3	Back-end Server . . . . .	4
2.3.1	Post-processor . . . . .	5
2.4	Analysis . . . . .	5
2.4.1	Continuous back-end server . . . . .	6
2.4.2	Non-continuous back-end server . . . . .	6
2.4.3	Requirements . . . . .	6
2.5	Process . . . . .	7
<b>3</b>	<b>Iteration 0</b>	<b>9</b>
3.1	Planning . . . . .	9
3.2	Implementation . . . . .	9
3.2.1	Initial Investigation of IntelliJ Plugin SDK . . . . .	9
3.2.2	Developing the Initial Data Structure . . . . .	10
3.2.3	Identifying Change Sources . . . . .	11
3.2.4	Squashing the Bugs . . . . .	12
3.3	Retrospective . . . . .	12
<b>4</b>	<b>Iteration 1</b>	<b>13</b>
4.1	Planning . . . . .	13
4.2	Implementation . . . . .	14
4.2.1	Implementing the External Change Detection Algorithm . . . . .	14
4.2.2	Testing the Plugin . . . . .	14
4.2.3	Adding Authentication to the Server . . . . .	15
4.2.4	Deciding on the Server Submission Method . . . . .	15
4.2.5	Parsing and Storing the XML Data . . . . .	15
4.2.6	Using Flask to Deliver Web Content . . . . .	16
4.2.7	Dockerising the Server . . . . .	17
4.3	Retrospective . . . . .	17
<b>5</b>	<b>Iteration 2</b>	<b>18</b>
5.1	Planning . . . . .	18
5.2	Implementation . . . . .	18
5.2.1	Adding the Student Upload Form . . . . .	18
5.2.2	Displaying Previous Student Submissions . . . . .	19
5.3	Retrospective . . . . .	20
<b>6</b>	<b>Iteration 3</b>	<b>21</b>
6.1	Planning . . . . .	21
6.2	Implementation . . . . .	21

6.2.1	Displaying Student Submissions for Staff . . . . .	21
6.2.2	Testing the Server . . . . .	21
6.3	Retrospective . . . . .	22
<b>7</b>	<b>Iteration 4</b>	<b>23</b>
7.1	Planning . . . . .	23
7.2	Implementation . . . . .	23
7.2.1	Squashing More Bugs . . . . .	23
7.2.2	Planning for the Mid-Project Demo . . . . .	23
7.2.3	Adding User Tests for the Server . . . . .	24
7.3	Retrospective . . . . .	24
<b>8</b>	<b>Iteration 5</b>	<b>25</b>
8.1	Planning . . . . .	25
8.2	Implementation . . . . .	25
8.2.1	Starting the Post-Processor . . . . .	25
8.2.2	Implementing the Watch Feature . . . . .	25
8.2.3	Attempting to Detect Automatic Code Generation . . . . .	26
8.3	Retrospective . . . . .	26
<b>9</b>	<b>Iteration 6</b>	<b>27</b>
9.1	Planning . . . . .	27
9.2	Implementation . . . . .	27
9.2.1	Researching Machine Learning Techniques . . . . .	27
9.2.2	Encrypting the XML File . . . . .	28
9.3	Retrospective . . . . .	28
<b>10</b>	<b>Iteration 7</b>	<b>29</b>
10.1	Planning . . . . .	29
10.2	Implementation . . . . .	29
10.2.1	Detecting Rename Refactoring . . . . .	29
10.2.2	Rebuilding Files From the Tracked Changes . . . . .	30
10.2.3	The Post-Processed Result Data Structure . . . . .	30
10.3	Retrospective . . . . .	31
<b>11</b>	<b>Iteration 8</b>	<b>32</b>
11.1	Planning . . . . .	32
11.2	Implementation . . . . .	32
11.2.1	Testing the Post-Processor . . . . .	32
11.2.2	Documenting the Code . . . . .	32
11.2.3	Plotting Charts using Pygal . . . . .	33
11.2.4	The Plagiarism Value Metric . . . . .	35
11.3	Retrospective . . . . .	36
<b>12</b>	<b>Design Architecture</b>	<b>37</b>
12.1	Architecture Diagram . . . . .	38
12.2	Student Plugin Interaction Sequence Diagram . . . . .	39
12.3	Student Server Interaction Sequence Diagram . . . . .	40
12.4	Post-Processor Sequence Diagram . . . . .	41

12.5 Data Structure Class Diagram . . . . .	42
12.6 Front-end Web Application . . . . .	43
12.7 Docker Containers . . . . .	46
<b>13 Testing</b>	<b>48</b>
13.1 Strategy . . . . .	48
13.2 Plugin . . . . .	48
13.3 Server . . . . .	49
13.4 Post-Processor . . . . .	51
13.5 Results . . . . .	52
<b>14 Results Analysis</b>	<b>55</b>
14.1 Example A . . . . .	55
14.2 Example B . . . . .	58
<b>15 Evaluation</b>	<b>60</b>
<b>16 Conclusions</b>	<b>62</b>
16.1 Introduction . . . . .	62
16.2 Goal . . . . .	62
16.3 Future Development . . . . .	63
<b>Appendices</b>	<b>65</b>
<b>A Third-Party Code and Libraries</b>	<b>66</b>
<b>B Ethics Submission</b>	<b>68</b>
<b>C Code Examples</b>	<b>71</b>
3.1 Example MongoDB Student Document . . . . .	71
<b>Annotated Bibliography</b>	<b>73</b>

## LIST OF FIGURES

3.1	Change UML Class Diagram . . . . .	11
3.2	FileTracker UML Class Diagram . . . . .	11
4.1	Web Sign-in Page . . . . .	16
4.2	Web Invalid Sign-in Flash . . . . .	17
5.1	Web New Submission Page . . . . .	19
5.2	Web New Submission Success . . . . .	19
5.3	Web New Submission Page . . . . .	20
11.1	Web Submission Data . . . . .	34
11.2	Web Submission FTS Chart . . . . .	34
11.3	Web Submission Changes . . . . .	35
11.4	Web Staff Dashboard All Submissions . . . . .	36
12.1	Final design architecture diagram . . . . .	38
12.2	Student-plugin sequence diagram . . . . .	39
12.3	Student-server sequence diagram . . . . .	40
12.4	Post-processor sequence diagram . . . . .	41
12.5	Tracked data structure . . . . .	42
12.6	Web Sign-in Page . . . . .	43
12.7	Web Student Dashboard Page . . . . .	43
12.8	Web Staff Dashboard Page . . . . .	43
12.9	Web Submission Data . . . . .	44
12.10	Web Submission FTS Chart . . . . .	44
12.11	Web Submission Changes . . . . .	45
13.1	Plugin tests . . . . .	52
13.2	Post-processor tests . . . . .	52
13.3	Server tests . . . . .	52
13.4	Server Terminal Output 1 . . . . .	53
13.5	Server Terminal Output 2 . . . . .	54
14.1	Submission Results A Data . . . . .	55
14.2	Submission Results A Chart . . . . .	56
14.3	Submission Results B Data . . . . .	58
14.4	Submission Results B Chart . . . . .	58
15.1	Velocity chart . . . . .	61
15.2	Burndown chart . . . . .	61



## LIST OF TABLES

14.1 Submission Results A Changes . . . . .	57
---------------------------------------------	----

# Chapter 1

## Introduction

### 1.1 Overview

Plagiarism is becoming popular among students due to ease of access to the internet. Plagiarism.org states that, “One out of three high school students admitted that they used the Internet to plagiarize an assignment” [1]. Detecting acts of plagiarism is not a simple task, especially when done manually. Humans are simply incapable of analysing multiple pieces of work and finding similarities. At least not efficiently or quickly. Automatic detection systems already exist but they only analyse the final piece of work. These systems are described more in detail in section 2.1. These systems can be improved by advancing the detection algorithms. This may prove difficult over time as these algorithms become increasingly complex.

A system that analyses work from its inception would allow more targeted and sophisticated detection methods to be performed. These methods could directly identify the plagiarised work, and how it may be obfuscated to hide any evidence. This system would not act as a sole detection system, but instead alongside existing tools. This would improve the detection rate and provide more evidence for such cases.

### 1.2 Plagiarism and Unacceptable Academic Practice

Aberystwyth University classifies plagiarism, collusion, and fabrication of evidence of data as acts of UAP (Unacceptable Academic Practice) [2]. The act of plagiarising is to commit fraud by stealing someone’s work and not clearly referencing the owner [3]. Plagiarism has many forms, some of these are described below [2] [3] [4] [5].

- Copying or cloning another’s work without modification (including copying and pasting)
- Paraphrasing or modifying another’s work without due acknowledgement
- Using a quotation with incorrect information about the source
- The majority of the work is made up from other sources
- Copying an original idea from another persons work

- Putting your own name on someone else's work

These forms of plagiarism apply to written work as well as source code. In terms of coding, copying someone's code without modification or acknowledgement is still plagiarism. Automated systems can be put in place to detect plagiarism and UAP.

The main objective of detecting plagiarism, is finding pieces of work which originated from another source. The plagiarised work can often be obfuscated or modified in a way to try and conceal the original work yet keep the same outcome. In source code, this ranges from simple changes to more complex alterations. The program plagiarism spectrum describes the levels of plagiarism that can be done, with level 0 having no modifications and level 6 altering the control flow of the program [6]. The higher levels make it more difficult to compare pieces of work against each other. Tools that only analyse the final piece of work will struggle with identifying the higher levels in the spectrum. Tracking how a piece of work develops over time will show patterns of obfuscation. As an example, evidence of changing control flow will be tracked. This shows us that the potential of this project could be very successful if developed properly.

## Chapter 2

# Background

This project introduces two problems. Tracking the work as it is being written, and identifying when plagiarism has occurred. Tracking code being written will be accomplished by developing a plugin for IntelliJ IDEA. IntelliJ IDEA is a Java IDE (Integrated Development Environment) for software development. IntelliJ has the capability to add and develop plugins. These plugins are developed using the IntelliJ Platform [7]. Identifying plagiarism is the second major task for this project. Due to this system operating differently to existing systems, it will be difficult to determine how to accurately detect plagiarism. Alternatively, Eclipse could have been used to develop the plugin. Eclipse is also a Java IDE for developing Java software. Similar to IntelliJ, Eclipse has the ability to add and develop plugins.

“Plagiarism detection usually is based on comparison of two or more documents” [8]. This is what makes this project stand out. It’s not simply reinventing the wheel, but finding new ways to solve problems which still exist in academia.

### 2.1 Existing Systems

The most popular existing systems are Turnitin, MOSS, JPlag, and YAP3, amongst many others. For this system, MOSS, JPlag, and YAP3 are worth looking into as they all check source code, whereas Turnitin only checks plain text [8]. Although these tools can automatically identify plagiarism, they still require human verification.

MOSS relies on the Winnowing detection algorithm [9]. It works by creating fingerprints or hashes for documents. Creating single hashes for each document allows for exact document comparisons. Single hashes are useful for checking if a document is correct and non-corrupt. Instead, Winnowing, uses multiple k-grams for each document. K-grams allow for partial document comparisons between multiple documents. Comparing between multiple sources does however require a large set of documents beforehand.

JPlag provides an online user interface where documents can be submitted and the results can be viewed. JPlag parses each document into token strings and these tokens are compared in pairs between documents. The percentage of tokens that match is referred to as the similarity factor [10].

Yap3 operates in a similar way to JPlag and MOSS. It uses its own similarity detection algorithm, RKR-GST. The algorithm compares sets of strings in a text much like the other

algorithms [11].

Plagiarism detection tools can use either extrinsic or intrinsic detection algorithms. Extrinsic detection uses external sources to compare against. Using a massive collection of external documents, extrinsic algorithms can be used very effectively although will take a large amount of time to process. Intrinsic detection analyses the document to detect changes in writing style. This allows the post-processing time to be very small in comparison to extrinsic algorithms. All of the existing systems described above use extrinsic detection algorithms.

## 2.2 IntelliJ Plugin SDK

IntelliJ is being used as it is a supervisor requirement. Being unfamiliar with the IntelliJ Plugin SDK, the online tutorials provided by JetBrains [12] are a great place to start. IntelliJ comes bundled with the IntelliJ Platform Plugin SDK so setting up the development environment should be no issue. The IntelliJ Community Edition is open source and contains many plugins [13]. This repository is very useful. Looking at existing plugins is sometimes more useful than reading tutorials.

One aspect of the plugin that's needed is to track keyboard events. This feature is not mentioned in the tutorial, so the Community Edition repository will be used to find any examples of this. `TypedHandlerDelegate` and `TypedHandler` are classes which perform actions upon typing events in the editor of the IDE. Both of these classes could be very useful when starting development.

Saving data to disk is also another feature that would be required by the plugin. This feature is used widely by many plugins and was documented in the tutorial. Persistent state is stored to file using the XML format. The tutorial is understandable but looking at an existing plugin for a real example should provide clarity. The GitHub plugin saves settings to an XML file and so this was helpful to my understanding.

## 2.3 Back-end Server

The following is a list of the possible technologies that could be used for the back-end server. The requirements for the back-end server include being able to submit tracked data and storing it in a suitable database. Authentication will also be required. A front-end web service will be used to display information about the tracked data to authenticated users. Having experience with both Python (Flask) and Ruby on Rails to develop web servers proposes a major decision. Bootstrap and jQuery can both supplement the front-end user experience.

- **Python / Flask** - Micro web framework for Python based on Werkzeug, and Jinja2.
- **Ruby on Rails** - Server-side web framework written in Ruby which uses a MVC architecture and provides default structures. It is very quick to implement a solution.
- **Bootstrap / JQuery** - Bootstrap is a front-end library for HTML, CSS, and JavaScript. JQuery is a JavaScript library.
- **Database** - Either an SQL or NoSQL database will be used to store recorded data for users.

- **Authentication** - Staff (and possibly students) should be able to sign-in to the service

The server could work in various ways. Recorded data could be sent continuously or once. It could receive continuous data from the plugin which would allow the user to seamlessly use the plugin without having to interact with the server in anyway. This would require the user to input some identification in the plugin.

Another way which the server could work is that the user could submit the recorded data in a form using a web application. This would reduce the bandwidth used by the server. But this introduces the issue that the user could modify the saved data before submitting. This would also make it easier to switch between projects and different computers.

A front-end application will also be required for staff and potentially students. Staff should be able to view students' project submissions and the final plagiarism results. Both Ruby on Rails and Python / Flask support this and can be tied into the back-end server module.

### 2.3.1 Post-processor

The post-processor would be a part of the server. Its functionality would be to process the data recorded from the plugin using plagiarism detection algorithms. The resulting data would be stored in the database, which the server can display when requested. There is also the possibility to add machine learning to the post-processor to improve detection results.

## 2.4 Analysis

The project will include the IntelliJ plugin, back-end server, post-processor, and a database. The plugin will be developed using the IntelliJ Plugin SDK using IntelliJ IDEA. The plugin will handle tracking file changes in the editor of the IDE. The tracked changes will be stored in the back-end server database. Each students' project will be stored with identification data. Identification data should include Aberystwyth user id, full name, project title, and the project module.

The back-end server will be developed using Python 3 and Flask. I decided to use Python due to being more confident and experienced with Python and Flask than Ruby on Rails. The back-end server will operate either continuously or non-continuously. This decision is discussed more in detail in Iteration 1 in chapter 4. Both back-end server scenarios are described below in subsection 2.4.1 and subsection 2.4.2. The server should also provide authentication for staff and potentially students. It would be nice to have authentication using Aberystwyth credentials. The post-processor will also be developed using Python 3 for consistency. The job of the post-processor will be to process all the data and store the results in the database. MongoDB will be used for the database. MongoDB stores documents in BSON (Binary JSON) format. Due to the recorded data being stored as XML, it is convenient to convert to a JSON format. Complex queries are not required, only simple queries are needed to store, retrieve, and update student submissions.

### 2.4.1 Continuous back-end server

The student will only need to interact with the plugin in this scenario. The user will be required to enter the identification data upon project creation. This identity data will be sent to the server. All of the tracked data will be sent to the server along with some of the identity data to be stored in the database. To prevent malicious attacks, the connection between the plugin and the server should be encrypted. Once the project is complete the student will click an action which will notify the server of the completion to start post-processing of the data. The front-end application will allow staff to login and view their students' submissions as well as the results from post-processor. This design does require a constant connection to the server, so a fail-safe would be needed for when no network is connected.

### 2.4.2 Non-continuous back-end server

In this scenario, the user will interact with both the plugin and the front-end application. There are no prerequisites to setting up the plugin. Instead the plugin tracks file changes of the editor and saves them directly to an XML file. Once the student has finished their project, they will submit the XML file via the front-end web application. This does present an authenticity issue. Having this data stored in a plain XML file, means that the student may maliciously manipulate the data. This would provide the server with spurious data providing an incorrect plagiarism result. This problem could be solved by encrypting the file. Students should be able view their previously submitted projects. Staff however, should be able to view their students' project submissions and the results from the post-processor. This design does not require a constant network connection.

### 2.4.3 Requirements

A list of key requirements are shown below. Using an agile process, these are the initial objectives identified. These may change in future sprints. Each item will be split into a single or multiple stories depending on the size of the task. More on the agile process used is described in section 2.5

- **Design architecture decisions**

- **Data structure for storing data**

- The data that is recorded must be stored in an adequate and efficient data structure. A Tree or Map implementation may suffice.

- **Data submission method**

- Recorded data could be sent to the server continuously (real-time) or non-continuously.

- **Development**

- **IntelliJ Platform Plugin**

- \* Implementation of source code detection methods:

- Keyboard events
      - Copy/paste
      - Code generation
      - Code auto-completion

- Refactoring
- External changes (using a different editor)
- \* Settings GUI - for student identification information (only required for continuous server)
- \* Storing recorded data in a data structure
- \* Storing recorded data to file (only required for non-continuous server)
- \* Encrypt stored data
- \* Unit tests
- **Back-end server**
  - \* Database storage
  - \* Authentication for staff (and potentially students)
  - \* Docker support for quick deployment
  - \* Mechanism to submit recorded student data
  - \* Unit tests
- **Post-processor**
  - \* Watch for updates in the database - this will allow the post-processor to process new student submissions
  - \* Process recorded data and update the database with the result
  - \* Unit tests
- **Front-end application**
  - \* A web application for staff to use (and possibly students)
  - \* Authentication page for staff (and possibly students)
  - \* Display student project submission data
  - \* Staff should be able to view all students' submission data
  - \* Students should only view their own submissions
  - \* Possible cases of plagiarism should be shown with adequate evidence. Graphs, metrics, and tracked data could be displayed.

## 2.5 Process

The approach chosen to use for this project was an agile methodology, scrum. Scrum was chosen due to having prior experience with this methodology. Scrum usually consists of multiple team members. This project was developed individually and this meant that a modified scrum was used. Scrum uses short iterations, each consisting of planning at the beginning, implementation, review, and then retrospective at the end. Alongside these iterations, short daily stand-up meetings are attended by the team. Daily stand-up is a short 10 to 15 minute meeting where the team discusses what they're working on and of any changes that need to be reviewed. Instead of daily stand-up, weekly meetings on Mondays with my supervisor were organised. These meetings will consist of discussions of the previous and next sprints.

Planning involves discussing and deciding which stories should be worked on during the sprint. A story is a piece of work that needs to be done. The intricate details of each story may not yet be known but they will develop over the course of the iteration. A story will have a time estimate associated with it. The golden ratio is used as a guideline, and the story points are described below.



- **1** - 10 minutes to 1 hour
- **2** - A few hours to half a day
- **3** - A few days
- **5** - A week
- **8** - Over a week, this story should be broken into smaller stories

Implementation and review take up most of the iteration time. This is spent designing, developing, and reviewing code that will end up in the code base. Once the code has been reviewed for a story, it can be marked as done.

Retrospective is a reflective process. It is a discussion of what went well, what didn't go well, and what could change for the next sprint. The retrospective is aimed to improve the scrum process over time.

To track each sprint and its stories, milestones and issues will be used on GitHub [14]. During planning a new milestone will be created, issues assigned to it (creating new issues if necessary), and a goal will be set. The goal should be a general aim for that sprint, which multiple stories would accomplish. During implementation and review, issues can easily be closed by referencing them in a commit message with specific keywords such as `Fixes #IssueNum` [15]. After the sprint is done, the milestone will be closed. Any remaining issues in the milestone would remain in the backlog still marked as open.

The structure of this report will continue with each iteration as a new chapter. The final design will be discussed in detail in chapter 12. The testing approaches used for each of the project components will be discussed in chapter 13. The results produced from the post-processor will be reviewed in chapter 14.

## Chapter 3

# Iteration 0

### 3.1 Planning

This first iteration involved investigating the IntelliJ Plugin SDK and starting the initial development of the plugin. Below is the list of stories and their assigned story points. The list is displayed in completion order. Bugs that were encountered and resolved during this iteration are also included. Bugs do not have story points as they are not planned stories.

- Investigate IntelliJ Platform Plugin SDK component: 2 points
- Research existing plagiarism detection tools: 3 points
- Implement Settings GUI: 2 points
- Investigate data structure for storing data: 1 point
- Detect copy-paste of code: 2 points
- Store recorded data in a data structure: 2 points
- Detect external file changes: 3 points
- Bug: File path is sometimes relative and has no parent
- Bug: Ignore .idea directory when listening for DocumentEvents
- Bug: FileTooBigException when comparing external file changes
- Bug: Check externally changed file exists

### 3.2 Implementation

#### 3.2.1 Initial Investigation of IntelliJ Plugin SDK

The *Getting Started* tutorial from JetBrains was an obvious place to start [12]. Setting up a new project with the IntelliJ Plugin SDK was very simple, as the SDK is bundled with IntelliJ IDEA.

Following the tutorial to add actions to the plugin was straight-forward. `AnAction` is a class which can be subclassed to perform an action when an a menu item or toolbar button is clicked. `AnAction` is registered to a menu item or toolbar button in the `plugin.xml` file.

The `plugin.xml` file is used to register actions and project components. These components can be one of three levels: application, project, or module [16]. An application level component is created and initialised upon the start-up of the IDE. A project level component is created for each project in the IDE. A module level component is created for each module inside of each project in the IDE.

`AnAction` has been added to a new menu. The new menu is specifically for this plugins' actions. The action will display an input dialog when clicked. This will allow the student to enter their Aberystwyth username. When the action is clicked again, an information dialog will display the previously entered username.

`PersistentStateComponent` is a class used to store data to a file. Using this in conjunction with the previously used `AnAction` implementation allows saving the students' username to file and loading it when the project is loaded. This XML file is saved in the `.idea` directory.

The `TypedHandler` and `TypedHandlerDelegate` classes can both be used to listen to typing events in the editor. Implementing the `TypedHandlerDelegate` was a simplistic solution. Although, as with anything simple, there's always a drawback. That drawback was when the auto-complete popup is shown, the typing events were no longer triggered. This means that most of the typing wouldn't be tracked. On the other hand, implementing the `TypedHandler` class to override the default class worked well but also came with a dramatic downside. The auto-complete popup no longer worked at all. There had to be another way to track typing events in the editor. Introducing, `DocumentListener`, this class listens to file changes in the editor. This means that any change to a files' contents triggered an event. This worked perfectly, even with the auto-complete popup. A `DocumentEvent` object contains all of the data that describes what was changed in a file.

### 3.2.2 Developing the Initial Data Structure

The `Change` class is the core of the data structure implementation (see Figure 3.1 below for the UML Class Diagram). Whenever a file is modified a new `Change` object is created to represent the data that was altered. Initially, a `LinkedList` of `Changes` was used to store all recorded data. A `LinkedList` was chosen due to keeping insertion order, but it wasn't fully necessary as each `Change` had an associated timestamp. However, the `LinkedList` was superseded by the `HashMap`. The map key is the file path, and the map value is a `FileTracker` object (see Figure 3.2 below for the UML Class Diagram). The `FileTracker` class now contains the `LinkedList` of `Changes` object. The final data structure design and example XML data is discussed more in chapter 12.

<b>Change</b>
+ offset: int + oldString: String + newString: String + source: Source + timestamp: long
+ Change() + Change(int,String,String,Source,long)

**Figure 3.1:** The UML Class Diagram for the `Change` class. Each of the fields are public and non-final to allow for serialisation which is required for storing state via `PersistentStateComponent`. The default constructor is also needed for serialisation. Initially this class also had a `path:String` field but was removed after moving to a `Map` data structure.

<b>FileTracker</b>
+ path: String + changes: LinkedList<Change> + cache: String
+ FileTracker() + FileTracker(String)

**Figure 3.2:** The UML Class Diagram for the `FileTracker` class. Each of the fields are public and non-final to allow for serialisation which is required for storing state via `PersistentStateComponent`. The default constructor is also needed for serialisation.

### 3.2.3 Identifying Change Sources

Now that file changes can be detected and stored in an adequate data structure, the source of the changes must be identified. Below is a list of identifiable sources. The first source detection that was implemented was clipboard (i.e. copying and pasting). The `CopyPasteManager` class contains methods to retrieve the current clipboard contents. To identify if a change originated from the clipboard, the change `newString` value was compared with the clipboard contents. IntelliJ IDEA supports clipboard history so each content value was checked. If the values match then the change was a copy-paste event. Currently, any unidentified change has a source of `other`. This includes normal typing. The reasoning behind this is to identify all types of source changes and then the remaining unidentified changes should only be from the student typing. But for now, they are classed as `other`.

#### Identifiable Sources:

- Keyboard events
- Copy/paste
- Code generation

- Code auto-completion
- Refactoring
- External changes (using a different editor)

The last story for this sprint was to implement detecting external file changes. External file changes occur when the project is closed. When using an external editor (i.e. any other editor besides IntelliJ IDEA) to edit any of the project files, these are detected when the project is next opened. The `FileTracker` class has a `String` `cache`. This was used to store the last known file contents for that file (encoded as base-64). Each time a new change was added to a `FileTracker` its cache was also updated. This ensures the `cache` is always up-to-date. Upon opening the project, the `cache` of each `FileTracker` was compared against its current file contents. Unfortunately, there wasn't enough time in this sprint to finish implementing an algorithm to determine the differences between the old and new file contents.

### 3.2.4 Squashing the Bugs

Multiple minor bugs were encountered near the end of this iteration. The first was a simple `NullPointerException` when adding a new change for a file. If the file had no parent file, then getting its path would throw the exception. The next issue was less of a bug, and more of an inconvenience. Files in the `.idea` directory were being tracked. This was causing the XML file containing the tracked data to track itself and this caused the a massive increase in the file size very quickly. The last two bugs were associated with the new external file change detection. When reading the new contents of a large file, a `FileTooBigException` was being thrown. This was solved by directly reading from the file `InputStream`. Lastly, externally changed files were not checked if they exist (i.e. they were deleted when the project was closed). Originally, the removed file was removed from the tracked list. But, it was decided to add a change which removed all the content instead.

## 3.3 Retrospective

This first sprint went exceptionally well considering the unfamiliarity with the IntelliJ Plugin SDK. Already having three different methods of detecting file changes in the editor is daunting but the last method definitely looks like the best option. The velocity for this sprint is 15. This is quite high, so the story estimations could be too large. The next sprint will take this into consideration when estimating story points.

## Chapter 4

# Iteration 1

### 4.1 Planning

This iteration revolved around setting up the back-end server and adding basic unit tests to the plugin. Aberystwyth authentication for staff (and potentially students) was a major stepping stone for this sprint. Below is the list of stories and their assigned story points. The list is displayed in completion order.

- Add external change algorithm to find the difference: 2 points
- Investigate testing framework for IntelliJ Plugin: 2 points
- Write a test for copy-paste detection: 1 point
- Write external change detection test: 2 points
- Investigate server authentication: 3 points
- Investigate data submission method: 2 points
- Investigate back-end server software: 3 points
- Docker support for quick deployment: 2 points
- Implement web server: 1 point
- Add database storage: 2 points
- Implement server authentication: 2 points
- Check if user is staff or student: 2 points
- Add dashboard bootstrap template: 1 point

## 4.2 Implementation

### 4.2.1 Implementing the External Change Detection Algorithm

At the end of the last sprint, the external file change detection was implemented in the plugin. This needed an algorithm implemented to check the differences between the old and new file content. This required a complex algorithm. After researching, the *Myers' diff algorithm* was exactly what was needed. Originally, a third-party library implementation was used, but IntelliJ already includes difference comparisons in the IDE. The SDK provides this functionality too. Albeit difficult at first, due to minimal documentation or examples provided. The `ComparisonManagerImpl` class provided methods to compare the characters between two strings and returned a list of `DiffFragments`. After converting these to a list of `Changes`, they were added to the appropriate `FileTracker` instance.

### 4.2.2 Testing the Plugin

The options for adding units tests to the plugin were limited to one choice. Due to the plugin being written in Java, JUnit was chosen. The IntelliJ Plugin SDK also provides an API for testing. The testing infrastructure provided by the SDK allowed running tests in a headless environment. This means that the plugin was tested without the UI while keeping the full functionality of the IDE. Despite the absurdly long class name, `LightPlatformCodeInsightFixtureTestCase`, provided the basics for unit tests with the IntelliJ Plugin SDK. Extending from `LightPlatformCodeInsightFixtureTestCase`, the `BaseTest` class was built to contain all of the useful testing methods for the plugin. This included methods such as `assertChangeListSize` and `assertOneChangeMatches`. The former method tests that the list of changes for a given file path is of a certain size. The latter method tests that the list of changes for a given file path contains a matching `Change` which is validated by the `Predicate`.

Extending from the `BaseTest` class, a test class for copy-paste detection was developed. This class contained one method which tests that when the clipboard was used to paste contents into the editor, it was identified properly. The testing infrastructure allowed a file to be created in the test project, then a string value was added to the clipboard and the editor pasted the clipboard contents. This data was tracked as usual and then the file changes were tested against to verify that it was identified from the clipboard.

The second testing class added was to test the external file change detection. This was more complex than the simple copy and paste detection test. The project needed to be closed and then the file had to be changed. This proved to be quite the challenge. To achieve this, originally, the project was actually closed but this prevented any files to be editable in the project. Instead, the `ProjectDocumentListener` which tracked all the file changes, was notified of the project closing, but the project was not be closed. Upon the project closing, the listener was removed from the project and file changes were no longer tracked. This allowed files to be editable, and in turn, would mimic external file changes. The `ProjectDocumentListener` was then notified that the project is opened and so the external changes would be detected and verified.

### 4.2.3 Adding Authentication to the Server

Prior to the development of the back-end server, the authentication mechanism was to be investigated. A member of IMPACS (Institute of Mathematics, Physics and Computer Science) was contacted about authentication via Aberystwyth University. This allowed staff and students to authenticate with the back-end server using their Aberystwyth credentials. This minimised the amount of development needed for an independent authentication system. It also removed the need for users to register with the new system. Instead, utilising the Aberystwyth LDAP (Lightweight Directory Access Protocol) server, users were authenticated using their Aberystwyth University credentials. After the LDAP3 server details were acquired, this could be tested and verified. The back-end server was written in Python 3. The LDAP3 library provided the necessary API to connect to the Aberystwyth University LDAP server [17]. Following the documentation, a simple authentication method was developed to test the authentication functionality. Using the LDAP3 library, an LDAP server was created with a new connection for each user. Attempting to `bind` the connection would result in either success or failure. If the bind was successful, then the user is authenticated and further user information may be retrieved from the connection.

### 4.2.4 Deciding on the Server Submission Method

Before continuing development on the back-end server, the data submission method was decided. This was the choice of either the continuous or non-continuous back-end server. The Analysis contains details on both of these methods. In the end, the non-continuous server was chosen. This was due to the simplicity of the design. The non-continuous design provides better support for offline development, and students are not be required to interact with the plugin. However, the server will have to provide authentication for both staff and students, as well as separate user interfaces for both. Students will have to submit their tracked data via the user interface but this should be less complex than continuously sending the data through a network protocol.

### 4.2.5 Parsing and Storing the XML Data

Now that the chosen authentication method was implemented and the data submission method has been chosen, the back-end server development can continue. The back-end server needed to parse the tracked XML data file. The database being used to store the tracked data is MongoDB which is a NoSQL database. The parsed XML data needed to be transformed into a JSON-like document. This didn't prove difficult due to the similarities between the structure of JSON and XML.

Using the `xml.etree.ElementTree` module, the XML data was parsed. But due to the format that the `PersistentStateComponent`, the XML parsing needed a little more intelligence. The XML data contained nested lists in a map, so these data structures needed to be correctly parsed. Recursively iterating over each element in the XML tree, each element was processed depending on its data. If the element was a map, then its children would be parsed as key-value pairs. If the element was a list, then its children would be added to a list. Any other element was parsed with a name and a value. All elements were added to a dictionary. Each element is guaranteed to have a name due to the serialisation assigning names to each element for each field.

For MongoDB, the PyMongo library was used to interface with the database connection [18]. Following the documentation, the tutorial was simple enough to follow to setup a connection

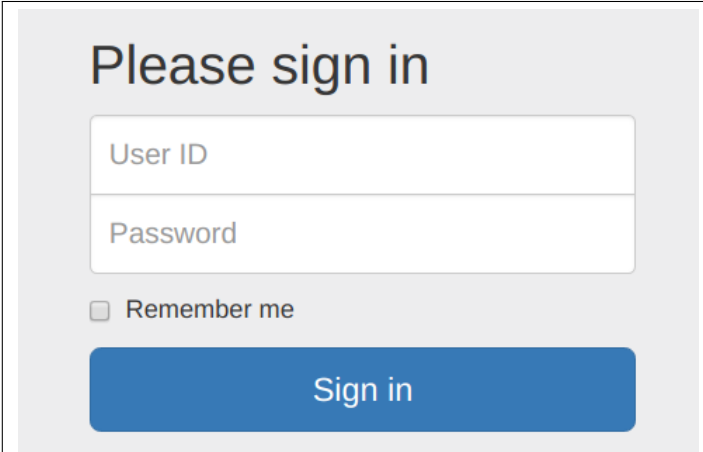


to a local MongoDB instance. The `SubmissionCollection` class was written to provide useful database methods. Primarily, these methods can be used to find, and insert data into the database easily. Example MongoDB Student Document in Code Examples shows the document data structure for a typical user.

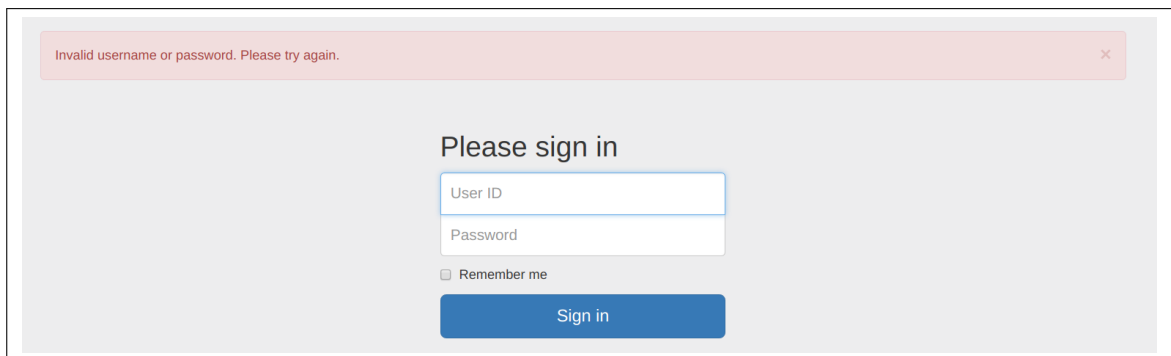
#### 4.2.6 Using Flask to Deliver Web Content

The back-end server had to provide a service to staff and students. Enter, Flask. Flask provides a micro web framework. Flask uses function decorators to specify URL routes. The return value of the function is then displayed when requested. Flask uses a templating language known as Jinja2. Jinja2 allows HTML templates to be dynamically rendered for each request. Pairing the templates with Bootstrap and JQuery provided beautiful dynamic web pages for end users. The front-end web application didn't do much at this point. The end users had to be able to login using their Aberystwyth University credentials. To combine LDAP server authentication with Flask seamlessly, another library was required; Flask-Login.

Flask-Login provides user session management for Flask. Using the Bootstrap sign-in template [19] and Flask-Login, the authentication system worked perfectly. A single route was used for the index. The index renders the sign-in template when using the GET request. This is shown in Figure 4.1. However, when a POST request is sent, authentication is performed using the posted form values (username and password). The username and password are used for binding a new LDAP connection. Upon success, the user details are retrieved from the LDAP server and a new user model is created using the `User` class. The `User` class follows the model from Flask-Login providing the required methods. An additional method has been added to test if the user is staff. The user details are stored in the database. If the user already exists in the database, then an update is triggered instead. This will update and out-of-date details. The page is then redirected to the appropriate dashboard. Staff and students have separate dashboards. If authentication is unsuccessful then a flash message is displayed to the user and the sign-in page is displayed. This is shown in Figure 4.2



**Figure 4.1:** The index page of the web-application. This is where users sign-in.



**Figure 4.2:** An error flash message is displayed when a user enters incorrect credentials on the sign-in form.

### 4.2.7 Dockerising the Server

Docker is a software technology providing containers. Docker provides base images or operating systems, such as Ubuntu which can be downloaded and run out of the box. Docker also allows images to be built in layers, meaning that developers can build on top of these base images and distribute their images (which can then be used as base images for others to download). A Dockerfile is a text file which contains commands to build an image. This allows for seamless automation of building Docker images and running Docker containers.

A Dockerfile was created for the back-end server. This provides install instructions for all of the required python packages using pip. The base image used was `python:3.6.4-slim`. This provided the basic Python infrastructure. All of the required files were added to the container upon running the container by mounting the directory to a volume in the container. Another Dockerfile was written for the MongoDB instance. The `docker-compose.yml` file was used by Docker Compose to run multiple-container applications. This allowed deployment of the containers (back-end server, and MongoDB) with one command; `docker-compose up` to start, and `docker-compose down` to stop.

## 4.3 Retrospective

The velocity for this sprint is 25. This is a large increase from the first sprint. Again this sprint had bad estimates for stories. Initially, the sprint ended prematurely, so more stories were added to accommodate this. Alternatively, the sprint could've been closed and a new one started. Aside from the bad estimations, the work completed during this iteration was comprehensive.

## Chapter 5

# Iteration 2

### 5.1 Planning

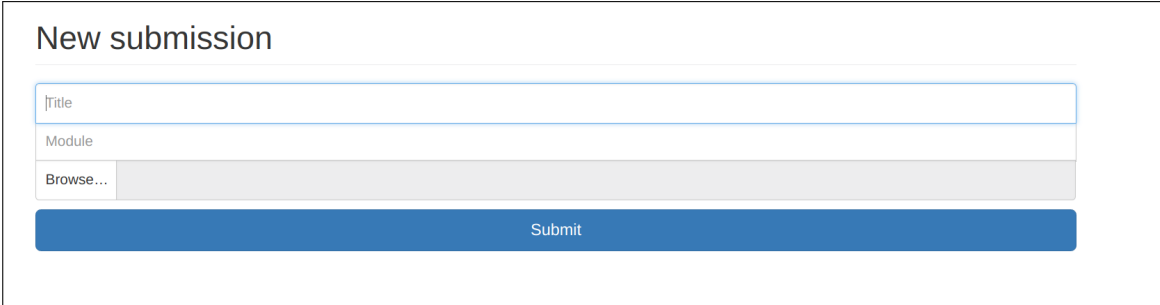
This iteration goal was to allow students to post new submissions and view all previous submissions. Below is the list of stories and their assigned story points. The list is displayed in completion order.

- Add form to upload submission in student dashboard: 2 points
- Display all previous submissions in student dashboard: 2 points

### 5.2 Implementation

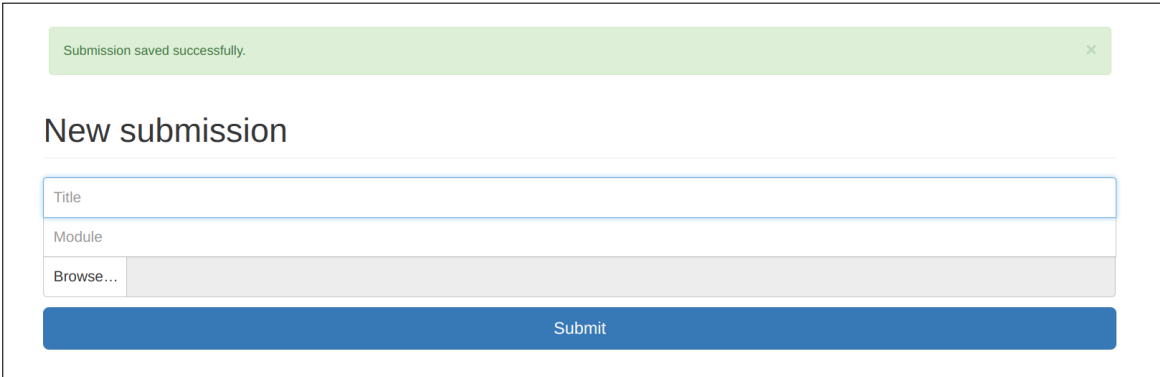
#### 5.2.1 Adding the Student Upload Form

A new route and template were created. This new page contained a form to upload a new submission as shown below in Figure 5.1. Only students can post new submissions, so this page is not accessible by staff. A title, module, and XML file are required. Uploading the XML file was a little more difficult due to the requirements. Following an online tutorial was sufficient enough to overcome this [20]. The file that is submitted must be an XML file. This was checked in the HTML form, and in the Flask route function. The submitted XML file was parsed and the submission is inserted into the student submissions array in the database. A flash message was displayed upon success or failure as shown below in Figure 5.2.



The screenshot shows a web form titled "New submission". It contains three input fields: "Title", "Module", and "Browse..." (which is a file upload button). Below these fields is a large blue "Submit" button.

**Figure 5.1:** The submit page of the web-application. This is where students can submit their tracked data.

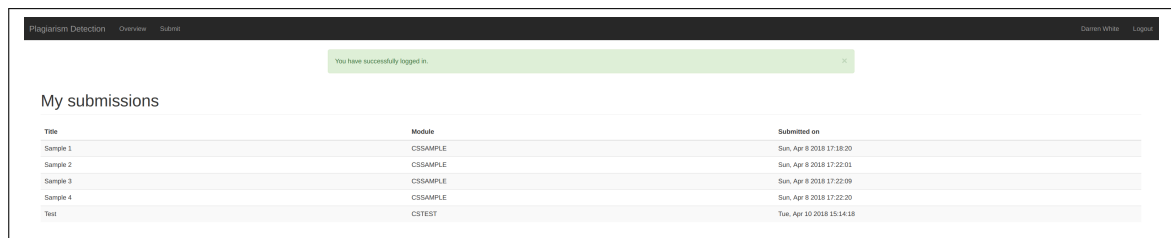


The screenshot shows the same "New submission" form as in Figure 5.1, but with a green flash message at the top that reads "Submission saved successfully." with a close button (X). The form fields and the "Submit" button are identical to the previous figure.

**Figure 5.2:** A success flash message is displayed when the form is successfully submitted and saved.

### 5.2.2 Displaying Previous Student Submissions

The dashboard has been separated into two individual templates. One for staff and one for students. With functionality in place to post new submissions, these were now displayed to the user. Querying the database by filtering `uid`, the current students' submissions were retrieved. The submissions array was passed to the template to render. The template displays the submissions in a table as shown below in Figure 5.3. The submissions were iterated, and each submission was added as a new table row. The submission title and module were displayed in the columns. More data for each submission may be added in future.



**Figure 5.3:** The dashboard page of the web-application for students. This is where students view their previous submissions. The flash message shown is only displayed after successfully signing in. Upon sign-in a redirect is sent to the dashboard.

### 5.3 Retrospective

This sprints' performance is extremely low in comparison to the first two. The sprint velocity is only 4. There were more stories initially for this sprint, but they were moved to the backlog and will be tackled in a future sprint. This was due to unforeseen health issues.

## Chapter 6

# Iteration 3

### 6.1 Planning

This iteration mainly focused on adding unit tests to the back-end Python server. These stories were initially due for the previous sprint, but were not completed. Below is the list of stories and their assigned story points. The list is displayed in completion order.

- Show all submissions on staff dashboard: 1 point
- Add Python nose tests: 3 points

### 6.2 Implementation

#### 6.2.1 Displaying Student Submissions for Staff

Displaying students' submissions on the staff dashboard, was similar to the already existing student dashboard. The only difference was to remove the student uid filter from the database query. Each submission needed to have some form of identification, so each submission was tagged with the user data. The staff dashboard showed the full name of the student in the submissions table.

#### 6.2.2 Testing the Server

Nosetests was used for the unit testing framework in Python 3. Nose extends unittest, which is included in the Python standard library. Both unittest and nose are similar to JUnit. The mock library was also being used for the unit tests. *Mocking* allows replacing parts of the application, to create specific testing scenarios. Function decorators are used to specify which functions will be mocked and which proxy data will be used.

The first unit tests to be developed were the authentication methods. The LDAP3 authentication system was mocked to "bypass" logging in as a user. This allowed the authentication methods to operate as if the LDAP server responded. Instead, the tests will mock

the code to return the required data. If mock was not used, then real credentials would have to be used, and the tests would have to be able to connect to the LDAP server on the Aberystwyth University network. The sign-in test was still incomplete, due to the database methods not being mocked.

MockupDB is a Python 3 package. It is used to create a mock server for testing the MongoDB code. The approach that MockupDB used was different from mocking methods. But the documentation and an online blog was used for reference [21]. The server code was refactored to account for the mocking of the database. Both the test environment and production environment needed to use the same server instances (Flask application and MongoDB client). If not, then the tests would fail. When the tests were set up, the MongoDB client was replaced with the mocked instance.

The sign-in test sends a POST request to the Flask application (in a new thread). Whenever the database receives a query, the test accepts (or denies) the query, then a response was sent with the appropriate data. In this test case, a query was sent to retrieve the user data (the user that was attempting to sign-in). The test acknowledged the query, and then sent back None. None was sent because the user had not signed in previously. If the user had previously signed in, then the user details would be sent back instead - but that's another test case. Another request was received for inserting data into the database. This is received upon successful authentication. The mocked LDAP server was used for this and simply returned True when authentication was tested. The insertion was acknowledged and the data that was to be inserted is validated. This validation either failed or passed the test.

Once this first sign-in test was in place, new test case scenarios would much be easier to write. Some of the other scenarios are, sign-in after previously signing in, and sign-in with incorrect credentials. These new test scenarios and dashboard tests will be developed in a future sprint.

### 6.3 Retrospective

The issues with the previous sprint have continued into this iteration unfortunately. The velocity is 4, which also follows the previous iteration. This was due to unforeseen health issues.

## Chapter 7

# Iteration 4

### 7.1 Planning

The mid-project demonstration date was approaching. Below is the list of stories and their assigned story points. The list is displayed in completion order. Bugs are not assigned story points.

- Bug: Submissions disappear upon sign-in
- Create architecture design: 1 point
- Create user sequence diagram: 2 points
- Start mid-project demonstration plan: 3 point
- Add invalid sign-in test: 2 points
- Add User class tests: 2 points

### 7.2 Implementation

#### 7.2.1 Squashing More Bugs

When a user signed in, a bug occurred which removed all of their previously posted submissions. This was a major bug which required immediate fixing. The code responsible for this was the sign-in database operations. Upon sign-in, the database was updated with the users latest details. But along with that information was the submissions array. A new array was being inserted and overwrote any previous data. This was changed to only occur during the users' first sign-in.

#### 7.2.2 Planning for the Mid-Project Demo

For the mid-project demonstration, diagrams were needed to aid with explaining the system (see chapter 12 for the final diagrams). The architecture diagram was developed to clearly show how the system components interact with one another. A large detailed sequence diagram was



developed to show the interactions between the users and each of the components. The sequence diagram was split into multiple smaller sequence diagrams. An overview or plan for the demo was also needed. This started as a markdown document containing notes, but later was converted into a Google Slides presentation.

### **7.2.3 Adding User Tests for the Server**

More unit tests were added for the back-end server. The user model class is a core part of the authentication and session management system. These unit tests were not very complex and did not require mocking. The tests simply initialise new instances and ensure that the methods worked correctly.

## **7.3 Retrospective**

This sprint was much better in comparison to the past two sprints. The velocity has increased to 10. Although not as high as the first two, it is definitely an improvement.

## Chapter 8

# Iteration 5

### 8.1 Planning

This sprint focused. Below is the list of stories and their assigned story points. The list is displayed in completion order.

- Create post-processor module hello world: 2 points
- Add Docker container to post-processor: 1 point
- Add MongoDB connection to post-processor: 2 points
- Detect automatic code generation: 3 points

### 8.2 Implementation

#### 8.2.1 Starting the Post-Processor

The post-processor was also developed using Python 3. Originally it was to be built-in to the back-end server. But, to allow for modularity, it was developed separately. There was some duplicated code, such as the database code but the post-processor didn't need to directly interact with the server.

The post-processor was based on the server due to the server also being developed in Python 3. The Dockerfile was also reused and modified slightly. The `docker-compose.yml` was updated to add the post-processor service. Once the infrastructure was in place, the major features could be implemented.

#### 8.2.2 Implementing the Watch Feature

The main feature that the post-processor needed was to be able to process new submissions that were inserted into the database. PyMongo provides a `watch` method. This method returns an

iterator which iterates over the changes. It acts as an event notification system. When a student posts a new submission, this is inserted into the database, and the watch iterator will be triggered.

As with most things in the programming world, this did not work at first. There were in fact, multiple problems faced when implementing this. Firstly, MongoDB had to be upgraded to 3.6 to be able to use the `watch` method. This was simple enough to change in the Dockerfile. Next, the database was required to be a replica set for it to work properly. Looking through plenty of online tutorials, there was lots of manual work, let alone only any Docker tutorials (or Docker Compose for that matter). At first it was thought that redundancy had to be added to the database (i.e. 3 nodes instead of 1). This proved difficult with the networking configuration. After the network issues were solved, another issue was encountered. The `replicaSet` was not able to vote for a master node. The two redundancy nodes were removed but the same problem was occurring. It turned out to be that the master/primary node had to be initialized to configure the replica set (only one time). It worked after initialization. But that was manual initialization of the master node. To automate this a bash script was written which would check if its not initialized, and then initialize it. The final issue was that deploying the application a second time would cause the replica state to fail. It turns out the hostname would change each time. The fix was to change the hostname in the `docker-compose.yml` file. This ensured the hostname was the same every time it was deployed.

In summary:

- Upgrade MongoDB to 3.6 to watch collections [22]
- Fix networking issues after upgrading to 3.6 by using `--bind_ip 0.0.0.0`
- Initialise the `replicaSet` with `mongo --eval 'rs.initiate()'`;

### 8.2.3 Attempting to Detect Automatic Code Generation

The last story in the sprint was not completed. Implementing automatic code generation detection was thoroughly investigated. The `ActionManager` class provides a listener which is used to identify the menu action that was performed. This includes the code generation actions. The difficulty comes with identifying the exact editor changes that were executed from that action. The start of the changes was known, but when was the auto generated code finished? Another option was to get the previous/last action performed when detecting editor changes. This had a similar problem, when does the action end? At this point, it was appropriate to ask how much time should be spent trying to find a solution. Should development continue without discrete identifications of source code? On the other hand, while trying to find a solution for automatic code generation, a potential solution for detecting refactoring was found. If renaming files can be detected then that would be enough for the plugin. The primary focus should then shift toward the post-processor.

## 8.3 Retrospective

The velocity for this sprint is 8. Most of the time during this sprint was spent on getting the `watch` method to work with MongoDB. Despite not being able to implement automatic code generation detection, it could be added in the future.

## Chapter 9

# Iteration 6

### 9.1 Planning

This iteration consisted of starting this report, researching machine learning algorithms to detect plagiarism, and encrypting the XML file. Below is the list of stories and their assigned story points. The list is displayed in completion order.

- Begin initial draft of project report: 3 points
- Research machine learning techniques for plagiarism: 5 points
- Encrypt stored data: 2 points

### 9.2 Implementation

#### 9.2.1 Researching Machine Learning Techniques

This iteration had less of a focus on development. Instead, this report was started, and machine learning algorithms to detection plagiarism were researched.

Bandara and Wijayathna state that most plagiarism detection algorithms are based on structured methods [23]. Structured methods are based on the program structure of source code. Another method for plagiarism detection algorithms is attribute counting. This involves extracting various metrics from the source code such as line lengths, and comment frequency. These metrics are used as inputs for the machine learning classifiers. An attribute counting method for detecting plagiarism would work very well with this project. Each change could be used as the metrics. More source code identifications such as automatic code generation, and refactoring, would greatly improve the accuracy. However, the raw data of the changes would be not a viable solution. They would first need to be transformed into a more suitable metric for comparison. In the future, it would be possible to implement these machine learning algorithms in the post-processor. Due to the time restriction on this project, machine learning techniques won't be implemented.

### 9.2.2 Encrypting the XML File

Previously, it was discussed that a student may attempt to generate a fake XML file. But it was recently discovered that the XML file could just as easily be modified before submission. For example, using `sed -i 's/source=CLIPBOARD/source=OTHER/' plagiarism_detection.xml` to replace all clipboard sources with other. This would cause problems with the post-processor identification. The most obvious solution was to encrypt the XML file, or at least to obfuscate the XML data. Previously it was thought that encryption was unnecessary. Due to the `PersistentStateComponent` class handling the XML serialisation and deserialisation internally, it was impossible to implement encryption at this level (i.e. just encrypting the whole XML file).

A selection of methods were attempted to effectively obfuscate the XML data. The first method used was to encrypt the file paths (the map keys in the XML file). This was successful and one way to expand on this would be to manually obfuscate each value in the map. To encrypt the map values, they would first need to be represented as a String. But not only would they need to be encoded as a String, but also decoded back into an Object. Delving into the IntelliJ Plugin SDK, the `XmlSerializer` class handles serialising objects into an XML structure. Utilising this class, the map values were serialised into an XML string. Now each value can be encrypted by performing 128-bit AES encryption on each XML string, then applying base64 encoding. Base64 encoding is used so that the value can be stored as a string value. This string can then be saved as usual and will be used for the persistent state. Upon loading the state, the process is simply reversed. This prevents students from modifying the XML data before submitting the data.

## 9.3 Retrospective

The velocity for this iteration is 10. The investigation of machine learning techniques has proved useful and a similar algorithm could be implemented in the future if possible. However, due to the time limitation, only a basic plagiarism detection algorithm will be implemented. The XML file encryption has solved many issues that were present. The tracking of students file changes is now secured and this data cannot be compromised. It also provides tamper prevention, so that the data will not be altered before submission.

## Chapter 10

# Iteration 7

### 10.1 Planning

This iteration focused on finishing the final touches to the plugin and implementing basic detection methods in the post-processor. Below is the list of stories and their assigned story points. The list is displayed in completion order.

- Detect file name changes and deletions: 2 points
- Gather sample XML files: 3 points
- Reconstruct files from change list: 2 points
- Implement copy-paste detection in post-processor: 3 points
- Implement external file change detection in post-processor: 3 points
- Create post-processed data structure: 2 points
- Update database with processed submission data: 2 points

### 10.2 Implementation

#### 10.2.1 Detecting Rename Refactoring

In the previous iteration, it was mentioned that while investigating detecting automatic code generation, a potential solution was found for detecting refactoring changes. The `RefactoringEventListener` class provides an interface to listen to refactoring events. Specifically, `refactoringStarted`, `refactoringDone`, `conflictsDetected`, `undoRefactoring`. A wrapper class for the refactoring data was developed, `RefactoringData`. This class was used to store the before and after refactoring data. Two methods were developed for applying and undoing the refactoring data. Applying or undoing the refactoring in this class will apply the changes to the appropriate `FileTracker` for the file that was modified. Currently, only renaming classes or files is supported. It was originally planned to also detect file deletions but instead it was decided to keep the file changes (the same as externally deleting a file).

### 10.2.2 Rebuilding Files From the Tracked Changes

Four XML sample files were collected by writing a simple Java application (a bingo caller). Two of the samples did not correctly track the beginning of the project. The `public class ... { }` declaration was not recorded. Using these sample XML files a small algorithm was developed to reconstruct each of the files using the list of changes. Listing 10.1 shows this Python algorithm below. This algorithm worked perfectly, aside from missing the class declaration. The fix for that was rather simple. It works by first checking if the file is not already being tracked. Then it will get the contents (before the change currently being made) and if the contents are not empty it will add the change as an external change.

---

```

1 document = ''
2
3 # Add each change to the document
4 for c in self.changes:
5     # Get change data
6     old_str = c['oldString']
7     new_str = c['newString']
8     offset = int(c['offset'])
9
10    # Get the start and end of the document which shouldn't be
    ↪ modified
11    start = document[:offset] if len(document) > 0 else ''
12    end = document[offset + len(old_str):] if len(document) > 0
    ↪ else ''
13
14    # Insert the new value into the document
15    document = start + new_str + end
16
17 return document

```

---

**Listing 10.1:** Python code to build a string document from a list of changes

### 10.2.3 The Post-Processed Result Data Structure

Example MongoDB Student Document in Code Examples shows the MongoDB document for a student. The `result` value on L26 is the post-processed data. This data structure contains a set of metrics. Each tracked file has its own set of metrics. In the next sprint, these metrics will be used to calculate a plagiarism value on a linear scale.

- **diff\_ratio:** This is the ratio difference between the cached file contents and the reconstructed file contents from the list of changes. This acts as an accuracy modifier.
- **frequency\_total:** The total character count in the file
- **frequency\_clipboard:** The character count from clipboard changes

- **frequency\_external:** The character count from external changes
- **frequency\_other:** The character count from any other changes (keyboard, auto code generation, auto-completion, refactoring, etc.)
- **frequency\_time\_source\_data:** The frequency vs. time data including source. This data will be used to plot a scatter/line graph

Matplotlib was used to easily create graphs. Using the `frequency_time_source_data` values, a line/scatter graph of Code Frequency vs. Time was created. One graph was created per submission. Merging the metrics for each file in a submission allowed for creating a single graph for each submission. This graph shows the character additions/deletions over the time of the project. This allowed easy visualisation for larger chunks of code being added (i.e. large copy/paste). Other metrics which were being generated are; total frequency (character count), clipboard frequency, external frequency, and diff ratios. The frequency values are simply the character for a source. The diff\_ratio is comparing the file cache to the reconstructed document (1.0 is perfect match). Currently these metrics were not being inserted back into the database and therefore not being shown on the dashboard. Updating the submission with the result data was a simple task. The user id and submission id are both used to update the submission with the new data.

### 10.3 Retrospective

The velocity for this sprint is 17. An increase from the past couple of iterations. The stories completed in this iteration developed the fundamentals for the post-processor detection methods. The next step will be to complete these detection methods and display the metrics in the staff dashboard. If the next iteration is as productive as this, then the post-processor and front-end web application should be completed.



## Chapter 11

# Iteration 8

### 11.1 Planning

This main goal for this final iteration was to complete the post-processor and front-end web application. Finishing the detection methods implementations and displaying the metrics and chart in the staff dashboard. Below is the list of stories and their assigned story points. The list is displayed in completion order.

- Add post-processor tests: 3 points
- Add full documentation to each module: 2 points
- Add installation instructions: 2 points
- Add submission graph visuals for staff dashboard: 3 points
- Implement a plagiarism / UAC scale: 1 point

### 11.2 Implementation

#### 11.2.1 Testing the Post-Processor

Due to the post-processor data structure not being well developed before implementation, unit tests were not appropriate at the time. Now that the data structure is mostly complete, it is now reasonable to add these unit tests. Currently, due to time restrictions, only one unit test was implemented. A simple input-process-output test. This test uses sample encrypted XML data, runs it through the post-processor, and a result is returned. The result is compared to the expected data.

#### 11.2.2 Documenting the Code

Over the course of the last few sprints, code documentation has become scarce. A story was put into this sprint to accommodate for the lack of documentation, including installation instructions.

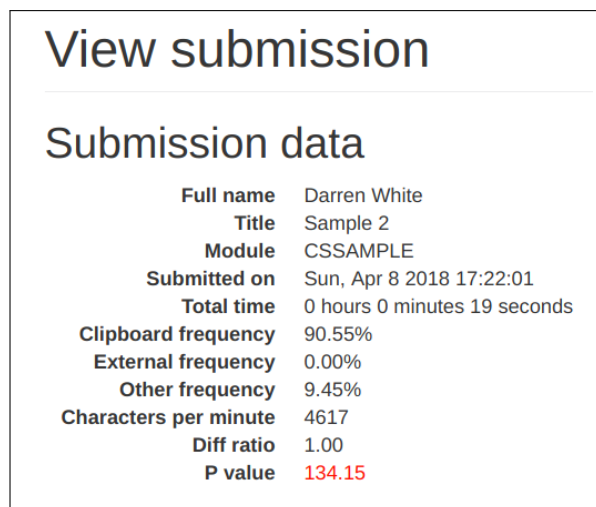
Each module was fully documented with appropriate Javadocs, docstrings, and line comments where necessary. The installation instructions were added to the `README.md` file. These instructions also included system requirements. Separate instructions are documented for each module.

### 11.2.3 Plotting Charts using Pygal

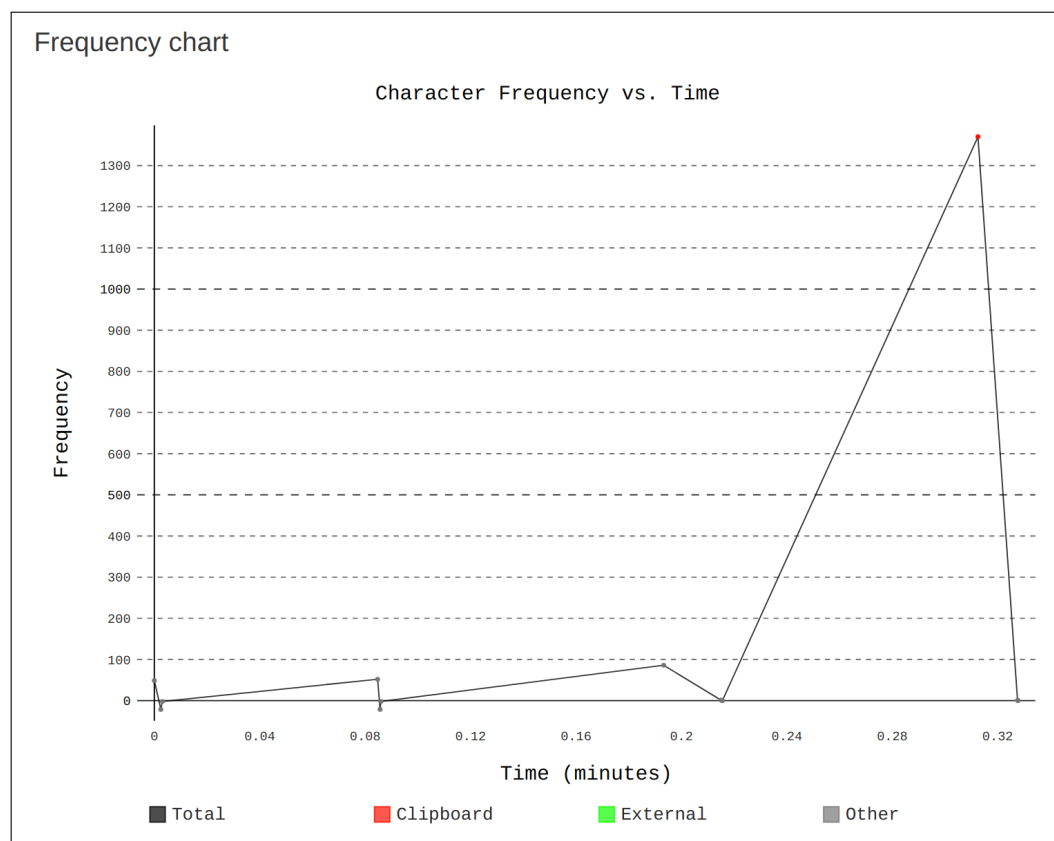
Currently, Matplotlib was used for displaying the frequency vs. time chart. This only worked locally, but was useful for testing the chart data. Pygal will be used to display charts in the Flask application [24]. Matplotlib and Pygal share a very similar method of building charts. This made a smooth transition to Pygal to display the FTS (Frequency Time Source) data. Matplotlib will still be used for local testing of charts.

The Pygal submission charts were added to the staff dashboard. Initially they were embedded directly into the table. This took up a lot of space and was aesthetically unpleasant. It was decided to add a new page which would display all submission details. This would keep the submissions table simple, providing a link to each submission detail page. The submission detail page uses this url format, `/dashboard/submission/<user_uid>/<submission_id>`. The staff dashboard table now has a link to access the submission details view (see Figure 11.4 in subsection 11.2.4).

The submission details page displays the metrics, FTS chart, and a list of *large changes*. The metrics are displayed in a description list as shown below in Figure 11.1. The FTS chart is displayed as an SVG+XML image as shown below in Figure 11.2. This allowed the Pygal chart to provide interactivity. This interactivity includes selecting specific sources and tool tips showing discrete values. The large changes are displayed in a table as shown below in Figure 11.3. A *large change* is a change which meets a specific size requirement. The default size considered for a change to be *large* is 200 characters. This can be configured by specifying a GET parameter in the URL. For example, `/dashboard/submission/abc12/5aca40cca326f6001046e1b8?large_change_size=91`. This url will show a specific submission where large changes are classed as having a size of 91 or greater. The total time metric for submissions was added which led to the creation of the CPM (characters per minute) metric. The CPM is calculated by simply dividing the total frequency by the total time. The average CPM is between 190 and 200 [25]. For professional typists the average CPM is between 325 and 375. However, the CPM when using IntelliJ could be drastically higher. Due to tracking every single change, auto-completion, automatic code generation, and refactoring all count towards the CPM. This will have to be taken into account when evaluating a submission. See chapter 14 for an in-depth review about the submission results.



**Figure 11.1:** An example submission metrics view. This shows the submissions post-processed results.



**Figure 11.2:** An example submission FTS chart. This shows the type of source code changes over the development time of a submission.

Large changes				
Path	Timestamp (minutes)	Offset	Old string	New string
__PROJECT_DIR__\src\BingoCaller.java	0.31251666666666666	0	public class BingoCaller { }	import java.security.SecureRandom; import java.util.*;  public class BingoCaller {  private static final int BINGO_MAX_VALUE = 99; private final SecureRandom rnd = new SecureRandom(); private final List<Integer> history = new ArrayList<>();  public List<Integer> getHistory() { return Collections.unmodifiableList(history); } }

**Figure 11.3:** An example submission large changes. This shows the major changes tracked for a submission.

### 11.2.4 The Plagiarism Value Metric

The final metric to be added was the *p\_value* metric (i.e. the plagiarism value). This was calculated using all the metrics available. It also has an associated colour (low value is green, high value is red). The *p\_value* operates on a linear scale. Currently this scale is 0 to 40. This scale is only a guideline and may change in the future with more testing. A large range of samples would be required for a more accurate scale. The *p\_value* equation is shown below.

$$p\_value = \frac{t}{100} \cdot d \cdot ((c + 1) + (e + 1)^2)$$

where *t* = Total time,

*d* = Diff ratio,

*c* = Clipboard change frequency,

*e* = External file change frequency

The clipboard and external change frequencies are added together. The external frequency is exponential because externally changing a file allows any kind of changes. The diff ratio acts as an accuracy modifier. Usually the diff ratio will be 1.0 so it will not affect the final value. The diff ratio will be less than 1.0 when one or more changes are not tracked. The longer the project was developed for, the smaller the *p\_value*. This means that projects developed extremely quickly will be flagged. The *p\_value* also has a colour associated with it. Depending on the value, the colour will gradually change. Listing 11.1 shows the algorithm below for determining the colour for a *p\_value*. The `__P_VALUE_LIMIT` is defined as 40. But this may change in the future.

---

```

1 mod = 255 / __P_VALUE_LIMIT
2 r = min(255, round(p_value * mod))
3 g = max(0, 255 - round(p_value * mod))
4 b = 0

```

---

**Listing 11.1:** Python algorithm to calculate colour based on the *p\_value*

Plagiarism Detection Overview					
All submissions					
Full name	Title	Module	Submitted on	Details	P value
Darren White	Sample 1	CSSAMPLE	Sun, Apr 8 2018 17:18:20	<a href="#">View</a>	3.77
Darren White	Sample 2	CSSAMPLE	Sun, Apr 8 2018 17:22:01	<a href="#">View</a>	134.15
Darren White	Sample 3	CSSAMPLE	Sun, Apr 8 2018 17:22:09	<a href="#">View</a>	1.36
Darren White	Sample 4	CSSAMPLE	Sun, Apr 8 2018 17:22:20	<a href="#">View</a>	0.75
Darren White	Test	CSTEST	Tue, Apr 10 2018 15:14:18	<a href="#">View</a>	0.66
Darren White	My New Submission	MYMODULE	Sun, Apr 29 2018 14:39:07	<a href="#">View</a>	0.02

**Figure 11.4:** The staff dashboard with a view of all student submissions. The p\_value is shown in the far right column with its associated colour. The staff dashboard now has a link to view the submission details.

## 11.3 Retrospective

The velocity for this final iteration is 11. All of the development was completed during this iteration. The post-processor and front-end web application are now finished. The following chapters in this report will include the design architecture, testing, results analysis, evaluation, and conclusions.

## Chapter 12

# Design Architecture

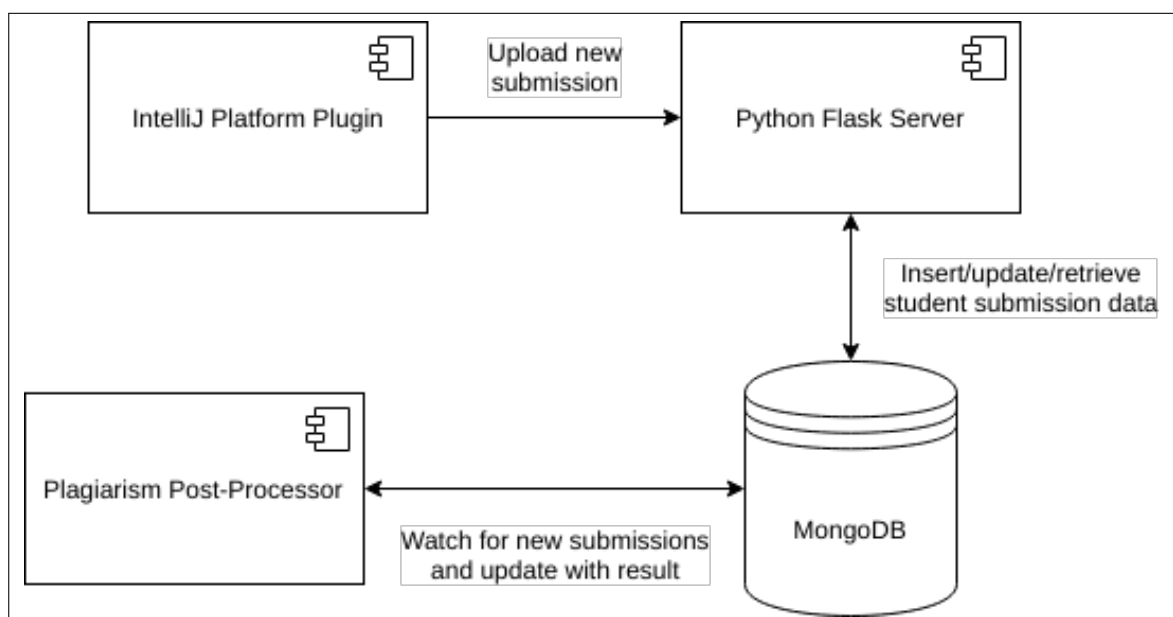
As discussed in Iteration 1 in chapter 4, the non-continuous server design was chosen. The continuous server would be more complex to design and implement than the non-continuous version. This is because a constant connection between the plugin and server was required. If the network connection was interrupted or disconnected then data would not be sent to the server. This means that a fail-safe protocol would have to be implemented, such as writing the unsent data to file. This fail-safe protocol implementation would be required for the non-continuous server anyway.

The server-side of the continuous design would also need to be more sophisticated. The server would have to know when a new project is being started, which project data is being received, and when a project is completed. The non-continuous server only required a method to submit the tracked data file.

The following pages show diagrams to aid with discussing the final design of the system.

## 12.1 Architecture Diagram

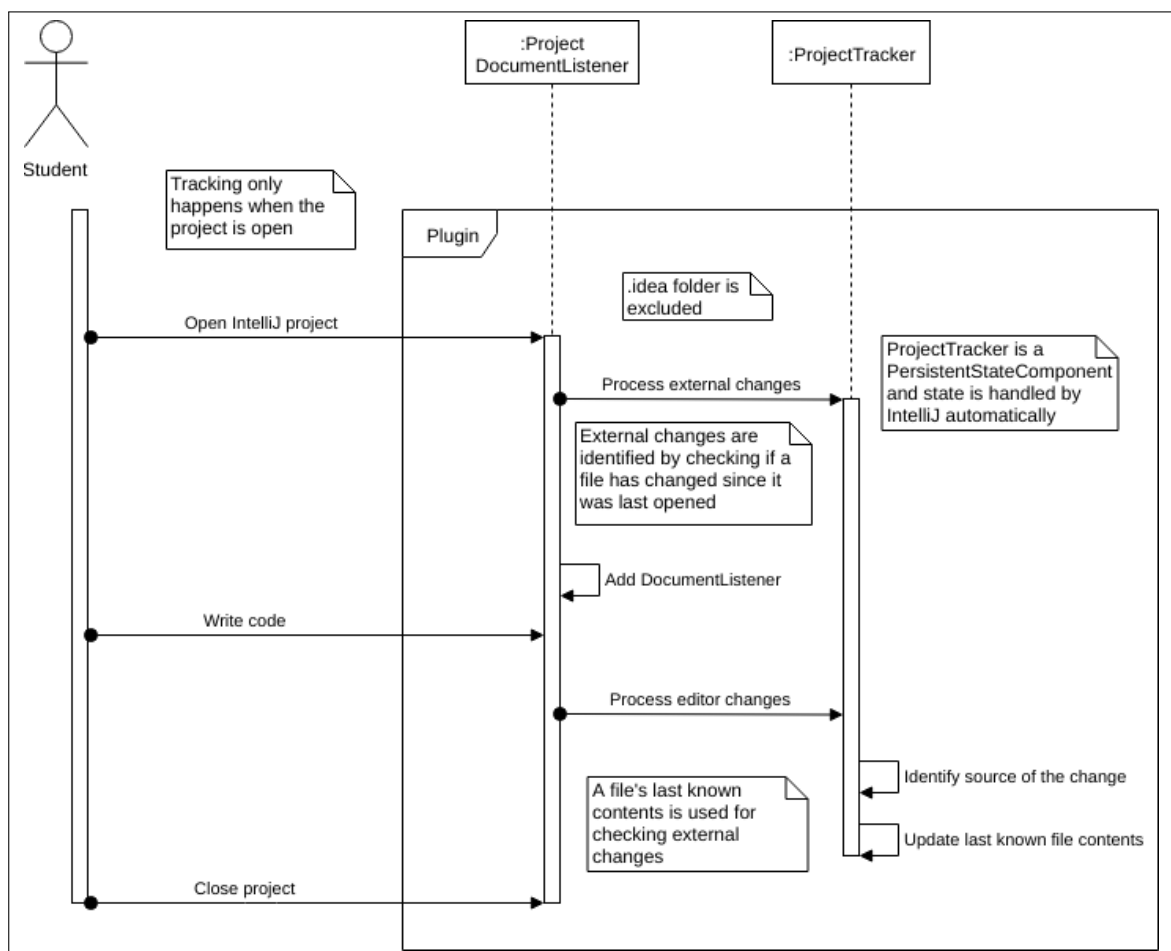
Figure 12.1 shows the interaction between each of the modules. The plugin only needs to track file changes in the editor. This data is saved to an XML file, which the user can submit to the server using the front-end web application. This requires authentication using Aberystwyth credentials. Once the file has been submitted, the back-end server converts the XML data to JSON and stores it in the MongoDB database. The post-processor is continually monitoring the database for new student submissions to process and update the database with the result.



**Figure 12.1:** Architecture diagram

## 12.2 Student Plugin Interaction Sequence Diagram

Figure 12.2 shows a typical students' interaction with the plugin. Initially, when a student opens a project, the tracked files are checked for external changes. This is done by comparing each files last known contents (which is stored as a base 64 encoded value) with its current file contents. The difference is added to the list of changes for each file. A `DocumentListener` is then added to the project for listening to file changes. Each file change is identified and then added to the list of changes for that file. After every new change that is added, the files last known contents is updated (for checking external changes). The recorded data is saved any time a file is saved. File change tracking stops when the project is closed (or if the plugin is disabled or uninstalled).

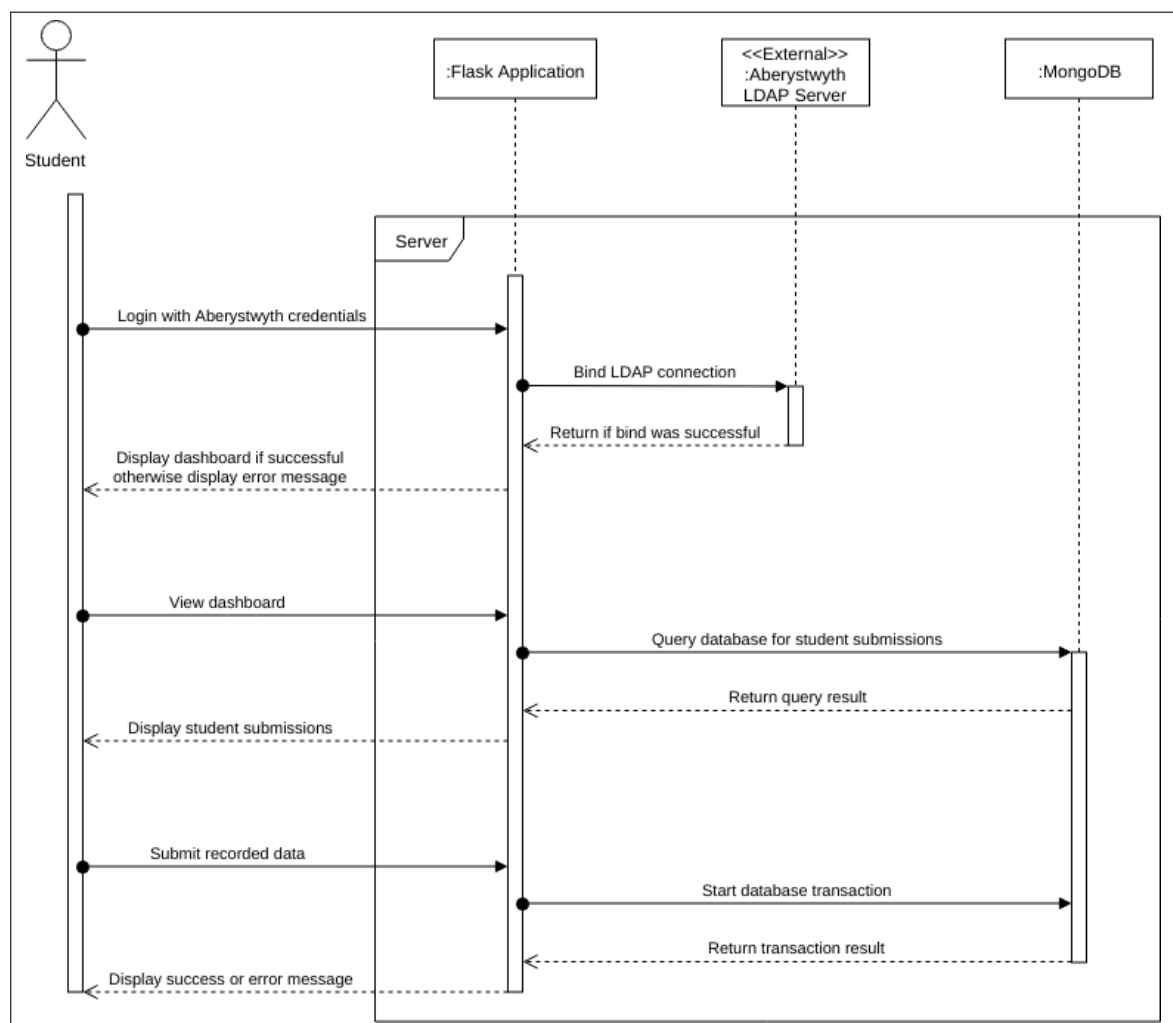


**Figure 12.2:** Sequence diagram showing a student interacting with the IntelliJ plugin



## 12.3 Student Server Interaction Sequence Diagram

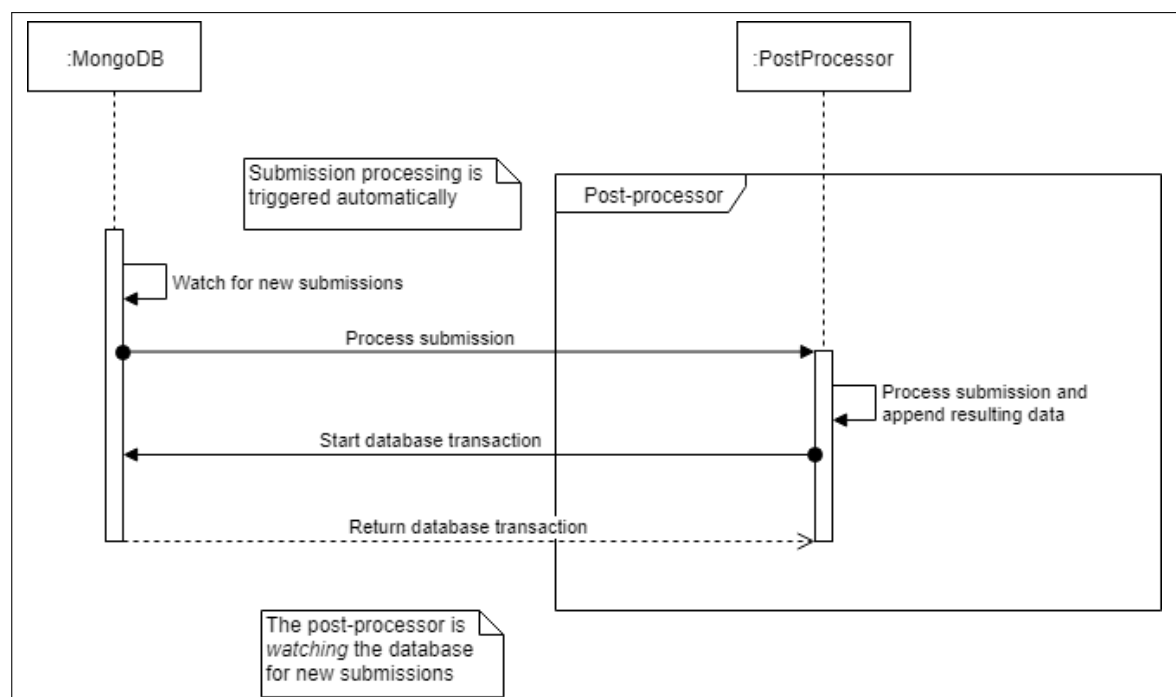
Figure 12.3 shows a students' interaction with the server. Once the project is completed. The student may login to the web application using their Aberystwyth University credentials. This authorisation is achieved using the LDAP server provided by Aberystwyth University. This requires the back-end server to be running on the Aberystwyth University network. The LDAP server returns if authentication was successful or not. If successful, then the student will be redirected to their dashboard. A query is sent to the database to retrieve all of the students' submissions. These submissions are then displayed in a table in the dashboard. Students can post new project submissions by uploading the XML file with additional meta data for identification in a form. The form is submitted via POST. The XML file is decrypted and added to the students submissions in the database. The user will be notified of success or failure via a flash message.



**Figure 12.3:** Sequence diagram showing a student interacting with the server

## 12.4 Post-Processor Sequence Diagram

Figure 12.4 shows when a student posts a new submission via the web application, the post-processor is notified. This is implemented using the `watch` method from PyMongo. The `watch` method notifies the post-processor of all changes in the database in real time. These changes must first be filtered to only identify new submissions. All new submissions are processed and then the resulting data is inserted back into the database using the original submission id and user id. See chapter 14 for an in-depth review about the post-processor results.



**Figure 12.4:** Sequence diagram for the post-processor

## 12.5 Data Structure Class Diagram

The main class that provides all the tracking is the `ProjectDocumentListener`. This class detects file changes, refactoring, and external file changes. When changes are detected they are added to the `ProjectTracker`. The `ProjectTracker` is used to store all of the changes for a single project. All file changes are stored in a Map data structure. The Map keys are the relative file paths as a String. The Map values are `FileTracker` objects. The `ProjectTracker` is also responsible for saving the state of the tracked changes by implementing `PersistentStateComponent`. An inner class, `CipherState` is used to store the encrypted Map data. The Map keys are not encrypted but the values are encrypted with 128-bit AES and base64 encoded. `CipherState` is used to load and save the state. Each tracked file has an associated `FileTracker` object. This object is used to store each files' Change list and cache. The file cache is its last known contents which is used for detecting external changes. The list of Changes can be used to reconstruct the file and compared with its cache.

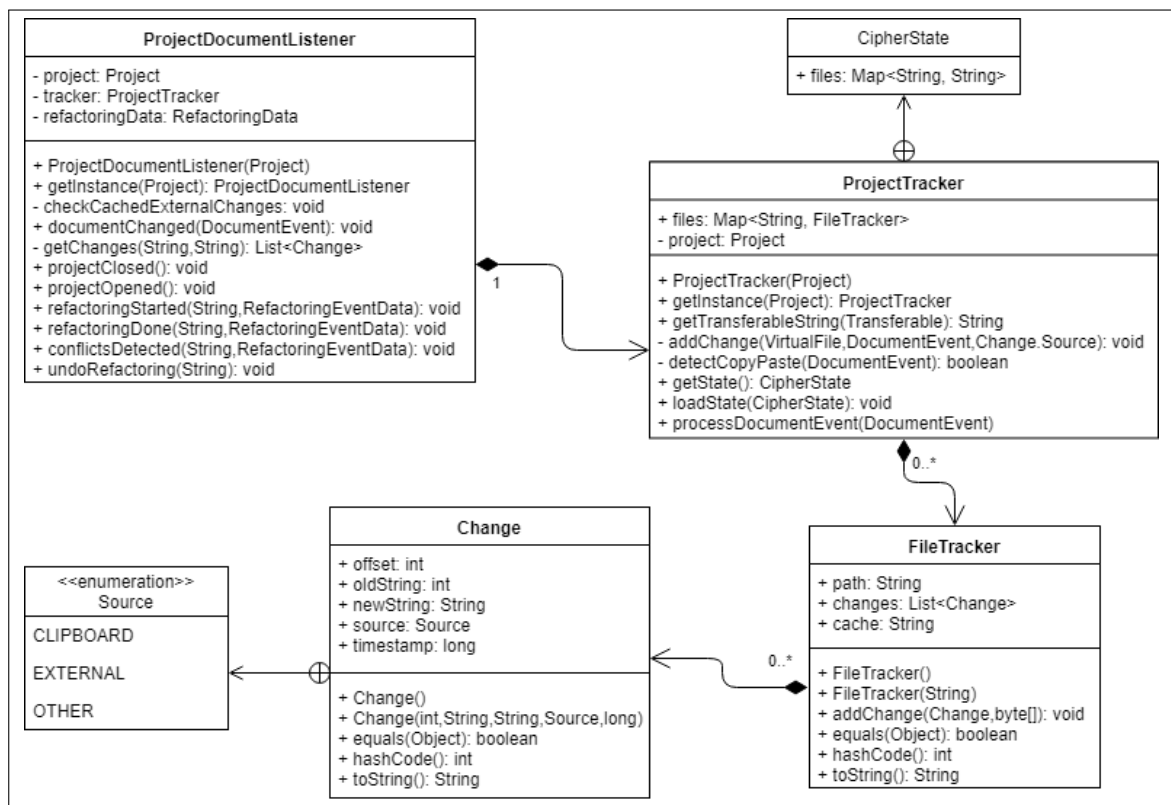


Figure 12.5: Data structure to store tracked data

## 12.6 Front-end Web Application

The approach used for designing the front-end web application was based off of bootstrap templates. The sign-in page is very simple, containing a simple form to enter credentials.

The image shows a simple sign-in form with a light gray background. At the top, the text "Please sign in" is displayed in a large, dark font. Below this, there are two input fields: "User ID" and "Password". Under the "Password" field, there is a checkbox labeled "Remember me". At the bottom of the form is a large blue button with the text "Sign in" in white.

**Figure 12.6:** The index page of the web-application. This is where users sign-in.

After successful authentication, the user is redirected to the appropriate dashboard. Staff and students have separate dashboards. This is to display different data to each. Students will only see their previous submissions and be able to post new submissions as shown in Figure 12.7. Staff will be able to view all students submissions and more in-depth details as shown in Figure 12.8. The staff submissions table contains three extra columns. The student name, details hyperlink, and P value. The student name is to identify the owner of the submission. The details link is to view further details of the submission. The P value is the calculated plagiarism value with its associated colour to ease with identification of possible plagiarism cases.

The image shows a student dashboard. At the top, there is a navigation bar with "Plagiarism Detection", "Overview", and "Submit". A green message box says "You have successfully logged in." Below this, the section "My submissions" contains a table with three columns: "Title", "Module", and "Submitted on".

Title	Module	Submitted on
Sample 1	CSSAMPLE	Sun, Apr 8 2018 17:18:20
Sample 2	CSSAMPLE	Sun, Apr 8 2018 17:22:01
Sample 3	CSSAMPLE	Sun, Apr 8 2018 17:22:09
Sample 4	CSSAMPLE	Sun, Apr 8 2018 17:22:20
Test	CSTEST	Tue, Apr 10 2018 15:14:18

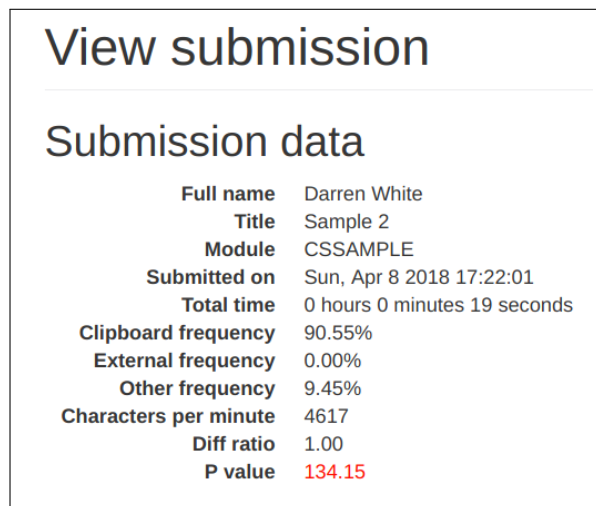
**Figure 12.7:** The dashboard page of the web-application for students.

The image shows a staff dashboard. At the top, there is a navigation bar with "Plagiarism Detection" and "Overview". The section "All submissions" contains a table with six columns: "Full name", "Title", "Module", "Submitted on", "Details", and "P value".

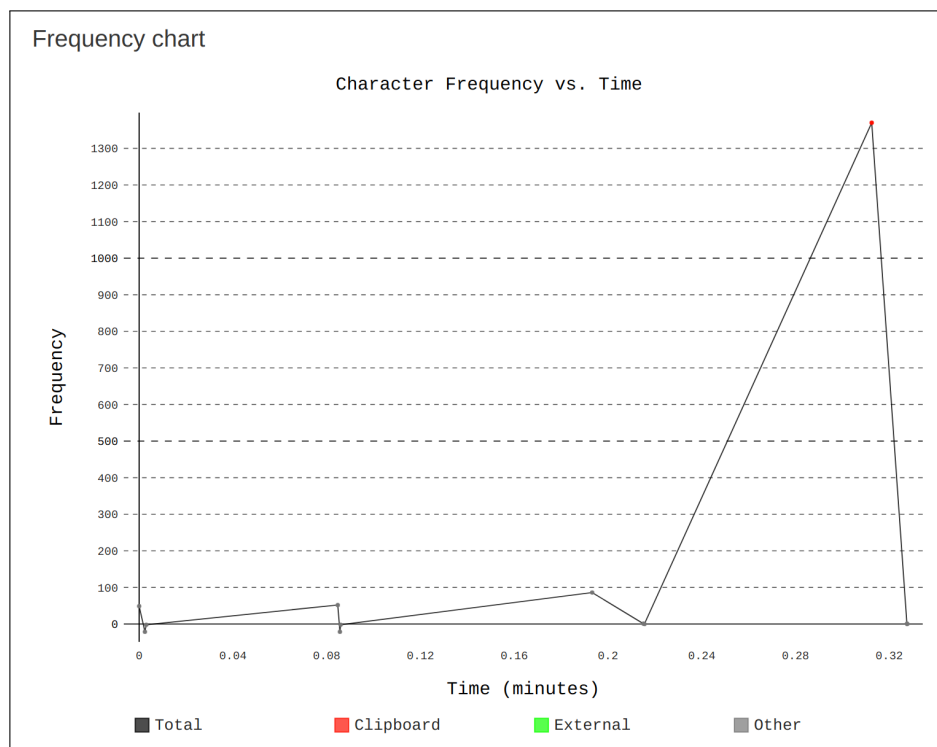
Full name	Title	Module	Submitted on	Details	P value
Darren White	Sample 1	CSSAMPLE	Sun, Apr 8 2018 17:18:20	<a href="#">View</a>	2.77
Darren White	Sample 2	CSSAMPLE	Sun, Apr 8 2018 17:22:01	<a href="#">View</a>	134.15
Darren White	Sample 3	CSSAMPLE	Sun, Apr 8 2018 17:22:09	<a href="#">View</a>	1.36
Darren White	Sample 4	CSSAMPLE	Sun, Apr 8 2018 17:22:20	<a href="#">View</a>	0.75
Darren White	Test	CSTEST	Tue, Apr 10 2018 15:14:18	<a href="#">View</a>	0.66
Darren White	My New Submission	MYMODULE	Sun, Apr 29 2018 14:39:07	<a href="#">View</a>	0.02

**Figure 12.8:** The dashboard page of the web-application for staff.

Staff users are able to view submission details. This includes individual metrics from the post-processor, student submission details, FTS chart, and large changes. Figure 12.9, Figure 12.10, and Figure 12.11 show the submission details information displayed on the page. All of this data can be used to help with identifying plagiarism. This is discussed more in-depth in chapter 14.



**Figure 12.9:** An example submission metrics view. This shows the submissions post-processed results.



**Figure 12.10:** An example submission FTS chart. This shows the type of source code changes over the development time of a submission.

Large changes				
Path	Timestamp (minutes)	Offset	Old string	New string
__PROJECT_DIR__src\src\BingoCafe.java	0.31251666666666666	0	public class BingoCafe { }	import java.security.SecureRandom; import java.util.*;  public class BingoCafe {  private static final int BINGO_MAX_VALUE = 99; private final SecureRandom rnd = new SecureRandom(); private final List<Integer> history = new ArrayList<>();  public List<Integer> getHistory() { return Collections.unmodifiableList(history); } }

**Figure 12.11:** An example submission large changes. This shows the major changes tracked for a submission.

## 12.7 Docker Containers

The back-end server and post-processor were both written with Docker in mind. Docker containers were used to easily deploy each of the server components. These components are the back-end server, post-processor, and database. Each of these have an individual `Dockerfile`. Docker Compose is used to seamlessly deploy all of these applications as separate containers with simple commands. See Listing 12.1 for the `docker-compose.yml` file. The `mongodb` service is built and deployed as a container first. This is because both the post-processor and server services depend on it. The local `data/db` directory is mounted to provide persistence for the database across deployments. This allows the container to stop and the data to remain on disk. The post-processor and server pass through the `PDP_DEBUG` environment variable. This is used by both of the components to enable debug logging when the value is present. The server service mounts the local server directory to the container. This allows Flask to hot-reload when files are changed locally. This removes the need to restart all containers when small changes are made to the server code. Both the `mongodb` and server services open ports for the containers. This is required for `mongodb` so that other containers can connect to it. The server opens the port so users can access the web service.

---

```
1 version: '2'
2 services:
3   mongodb:
4     build: ./mongodb
5     hostname: mongodb
6     ports:
7       - 27017:27107
8     volumes:
9       - ./data/db:/data/db
10  postprocessor:
11    build: ./postprocessor
12    depends_on:
13      - mongodb
14    environment:
15      - PDP_DEBUG=${PDP_DEBUG}
16  server:
17    build: ./server
18    depends_on:
19      - mongodb
20    environment:
21      - PDP_DEBUG=${PDP_DEBUG}
22    ports:
23      - 8000:8000
24    volumes:
25      - ./server:/plagiarism_detection/server
```

---

**Listing 12.1:** The `docker-compose.yml` file. This file is used to deploy multiple containers with Docker Compose.

Listing 12.2 shows the Dockerfile for the back-end server. The Dockerfile uses the Python base image. This provides the basic Python infrastructure, including the Python interpreter and pip (both version 3). The working directory is configured and system requirements are installed (not pip requirements). The entrypoint script is added. The entrypoint script is executed when the container is created as it is defined as the entrypoint on L22. The requirements file is added and pip is used to install the requirements. L17 is used to add the current local directory to the working directory in the container. This will add the server files to the container. The unit tests are executed using nose. If any tests fail then the container will fail to build and will not run.

---

```

1 FROM python:3.6.4-slim
2
3 WORKDIR /plagiarism_detection/server
4
5 RUN apt-get update && \
6     apt-get install -y gcc
7
8 ADD entrypoint.sh ./
9 RUN set -ex && \
10     chmod +x entrypoint.sh
11
12 ADD requirements.txt ./
13
14 RUN set -ex && \
15     pip3 install -r requirements.txt
16
17 ADD ./ ./
18
19 RUN set -ex && \
20     nosetests --with-coverage --cover-erase
21     ↪ --cover-package=server -v
22
23 ENTRYPOINT [ "./entrypoint.sh" ]

```

---

**Listing 12.2:** The Dockerfile for the server

Listing 12.3 shows the entrypoint Bash script for the server as used by the Dockerfile above. This script simply executes the server as a Python module.

---

```

1 #!/usr/bin/env bash
2
3 SCRIPT_DIR="$( cd "$( dirname "${BASH_SOURCE[0]}" )" && pwd )"
4
5 pushd ${SCRIPT_DIR} >/dev/null
6
7 python3 -m server

```

---

**Listing 12.3:** The server Dockerfile entrypoint Bash script



## Chapter 13

# Testing

### 13.1 Strategy

The approach used for testing each of the components was limited to unit testing. Having previous experience with unit testing in both Java and Python lead to this decision. The plan for testing the plugin was simple; test each of the detection methods. The plan for testing the server was to test each of the Flask routes to ensure the correct outcome. The post-processor acts as an I/O tool so the plan was to simply test that I/O process. Due to the lack of experience with setting up CI (continuous integration), this was considered but not set up. This would have eased with ensuring breaking changes are not accepted and merged. However, both the server and post-processor will be built as Docker containers and will run the tests as they are built. Any failing tests will fail the build.

### 13.2 Plugin

JUnit is used for testing the plugin. The IntelliJ Plugin SDK provides a testing infrastructure. This allows testing the plugin in a headless environment. What this means is that the IDE is run without the UI to test all aspects of the plugin. Everything is loaded as usual except the UI. The SDK provides classes and methods to *emulate* user actions such as typing, pasting, clicking menu items, and clicking tool bar buttons. Emulating such actions allows testing of detecting editor changes.

Four classes are included in the tests directory. `BaseClass` provides useful assertion methods. These will check file changes for specific data. `CipherTest` is a simple test case for the 128-bit AES encryption on the tracked data. Encrypted and unencrypted sample data is used to test the both the encryption and decryption methods. `CopyPasteDetectionTest.java` is a core source detection test for detecting copy-paste in the editor. This ensures that all copy-paste actions in the editor are tracked properly. `ExternalDetectionTest` is another core source detection test for detecting external file changes. This test is a unique test because it needs to simulate externally changing a file without using the IDE editor. This works by notifying the `ProjectDocumentListener` that the project has closed, making the changes (and therefore making "external" changes), notify the `ProjectDocumentListener` that the project is opened, which will detect the "external" changes correctly. See the unit test for testing external

change detection in Listing 13.1.

---

```

1  public void testExternalChangeDetection() throws IOException
    ↪ {
2      ProjectDocumentListener listener =
3
4          ↪ ProjectDocumentListener.getInstance(getProject());
5
6      try {
7          // Simulate closing the project
8          // If we actually close the project then we can't
9          ↪ write to the file
10         listener.projectClosed();
11
12         // Append the external content
13         WriteAction.run(() -> VfsUtil.saveText(file,
14             initialContent + externalContent));
15     } finally {
16         // Re-open the project
17         listener.projectOpened();
18     }
19
20     // Check the tracked changes
21     assertEquals(filename, 3);
22     assertEquals(filename, (c) -> c.source ==
23         ↪ Change.Source.EXTERNAL
24         && Objects.equals(c.oldString, ""))
25         && Objects.equals(c.newString, externalContent)
26         && c.offset == 15);
27 }

```

---

**Listing 13.1:** The external change detection unit test. Note that the file is created in the `setUp()` method with the `initialContent`

### 13.3 Server

Nose is used for testing the back-end server. Nose extends unittest to provide extra functionality. Unittest is built-in to Python 3 and works in a similar manner to JUnit. Mock is a library used to replace parts of the system to change functionality. MockupDB is a library used to mock a MongoDB client.

`base.py` provides generic `setUp` and `tearDown` methods, as well as a method to mock or patch the Aberystwyth LDAP connection. This is useful so that the Aberystwyth LDAP server is not directly contacted but instead is replaced with specific values to return. This removes the need to be connected to the Aberystwyth network when running the unit tests, and any

username/password combination may be used for tests. `test_data.py` has all necessary testing data for the unit tests.

`test_auth.py` contains functions which test the LDAP authentication system. The LDAP connection is patched to provide the necessary user information. The scenarios that are tested are: existing user sign-in, first time user sign-in, invalid user credentials sign-in, checking if user is a staff, and checking if user is a student. See Listing 13.2 below for the first time sign-in test.

---

```

1  @BaseTest.patch_connection(gecos=AberUndergrad.GECOS)
2  def test_first_time_signin(self):
3      # Send request in background
4      future = go(self.app.post, '/', data={
5          'uid': AberUndergrad.UID,
6          'password': AberUndergrad.PASSWORD,
7      })
8      # A request is received to check if the user exists
9      request = self.mockdb.receive(
10         Command('find', 'submissions', filter={'uid':
11             ↪ AberUndergrad.UID}))
12     # No user exists
13     request.ok(cursor={'id': 0, 'firstBatch': [None]})
14     # A new User should be inserted
15     request = self.mockdb.receive(
16         Command('insert', 'submissions'))
17     # Request is done
18     request.ok()
19     # Only one user should be inserted
20     documents = request['documents']
21     self.assertEqual(1, len(documents))
22     # Check the inserted user data
23     doc = documents[0]
24     self.assertEqual(AberUndergrad.FULL_NAME,
25         ↪ doc['full_name'])
26     self.assertEqual(AberUndergrad.UID, doc['uid'])
27     self.assertEqual(AberUndergrad.USER_TYPE,
28         ↪ doc['user_type'])
29     self.assertEqual(0, len(doc['submissions']))
30     # Check we are redirected to dashboard
31     response = future()
32     self.assertEqual(response.status_code, 302)

```

---

**Listing 13.2:** The external change detection unit test. Note that the file is created in the `setUp()` method with the `initialContent`. Empty lines have been removed to save space.

`test_dashboard.py` provides functions to test the staff and student dashboard routes. This involves sending various POST and GET requests to sign-in, and view the dashboard. The various database requests are received and the necessary data is sent as a reply to each request. The dashboard request is checked to ensure adequate data in the submissions table.

## 13.4 Post-Processor

Nose is also used for testing the post-processor module due to it also being developed with Python 3. MockupDB is also used for mocking the MongoDB client if necessary. The `base.py` class is very similar to the server `base.py` file but does not have the function for patching the LDAP connection as this is not used in the post-processor. `test_process.py` will test the processing of a submission to ensure the result is as expected. The input is an encrypted XML sample string and the expected output is a dictionary containing the results. The input is processed and the return value is tested. See Listing 13.3 below for the unit test.

---

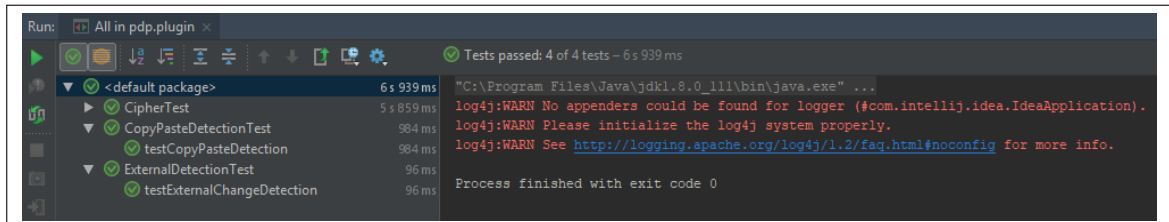
```
1  def test_process(self):
2      # Post process the submission
3      result = self.pp.run(StringIO(SUBMISSION_DATA))
4      # Check the result
5      self.assertEqual(result, SUBMISSION_RESULT)
```

---

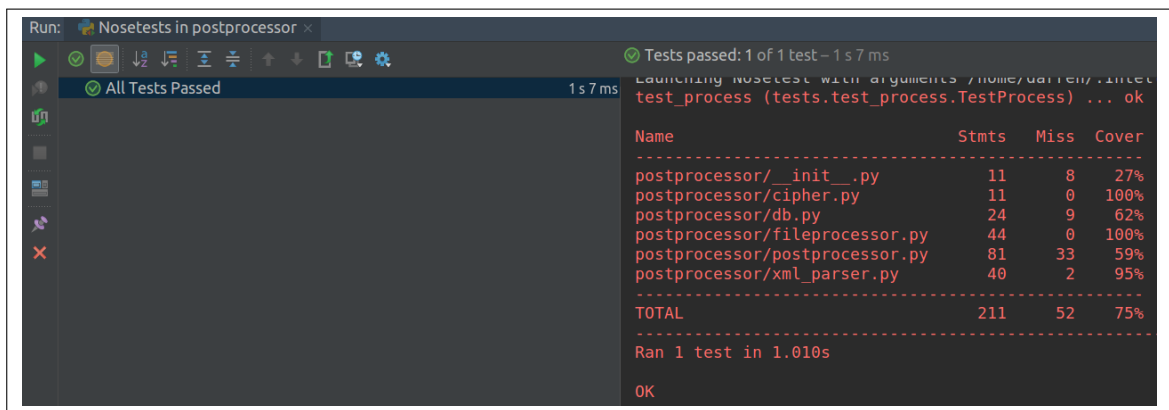
**Listing 13.3:** The Post-Processor I/O process unit test.

## 13.5 Results

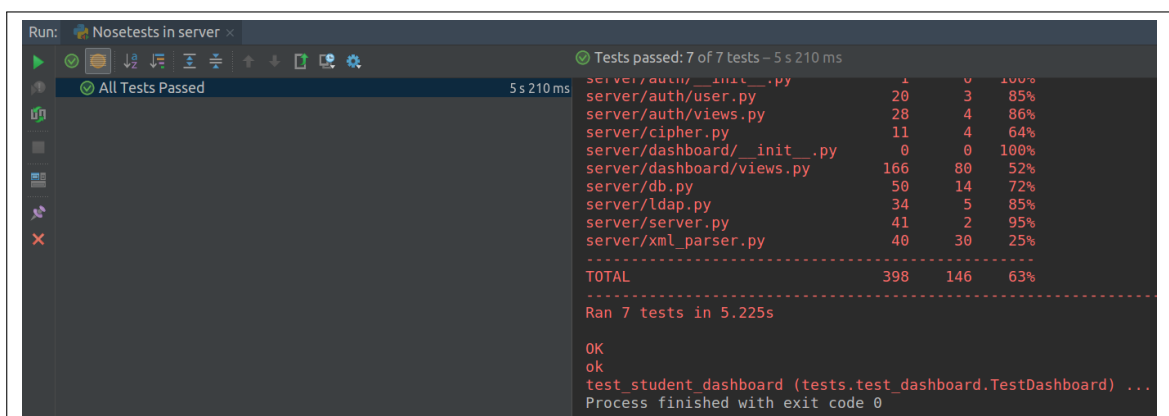
Figure 13.1, Figure 13.2, and Figure 13.3 show that all of the unit tests for each module pass successfully. Figure 13.4 and Figure 13.5 show the output from running Docker Compose to deploy the containers.



**Figure 13.1:** IntelliJ Plugin JUnit tests output



**Figure 13.2:** Post-processor nose unit tests output



**Figure 13.3:** Server nose unit tests output

```

darren@darren-XPS:~/workspace/plagiarism-detection-plugin (master)$ ./run.sh
WARNING: The PDP_DEBUG variable is not set. Defaulting to a blank string.
Removing network plagiarismdetectionplugin_default
WARNING: Network plagiarismdetectionplugin_default not found.
WARNING: The PDP_DEBUG variable is not set. Defaulting to a blank string.
Creating network "plagiarismdetectionplugin_default" with the default driver
Building mongodb
Step 1/5 : FROM mongo:3.6.3
----> 43099507792a
Step 2/5 : COPY entrypoint.sh /tmp/entrypoint.sh
----> Using cache
----> 10f3b2bfe456
Step 3/5 : RUN chmod +x /tmp/entrypoint.sh
----> Using cache
----> c75f20cfbe57
Step 4/5 : EXPOSE 27017
----> Using cache
----> 13ede564b8e1
Step 5/5 : ENTRYPOINT ["/tmp/entrypoint.sh"]
----> Using cache
----> b448246082d2
Successfully built b448246082d2
Successfully tagged plagiarismdetectionplugin_mongodb:latest
Building server
Step 1/10 : FROM python:3.6.4-slim
----> 43431c5410f3
Step 2/10 : WORKDIR /plagiarism_detection/server
----> Using cache
----> 5dafa8ab7688
Step 3/10 : RUN apt-get update && apt-get install -y gcc
----> Using cache
----> 150b6d11f2df
Step 4/10 : ADD entrypoint.sh ./
----> Using cache
----> 7bff999e690a
Step 5/10 : RUN set -ex && chmod +x entrypoint.sh
----> Using cache
----> 3c3a792fb773
Step 6/10 : ADD requirements.txt ./
----> Using cache
----> 8391b81046a4
Step 7/10 : RUN set -ex && pip3 install -r requirements.txt
----> Using cache
----> e0eb88fdabca
Step 8/10 : ADD ./ ./
----> 452ad05b8a6b
Step 9/10 : RUN set -ex && nosetests --with-coverage --cover-erase --cover-package=server -v
----> Running in 8368df227873
+ nosetests --with-coverage --cover-erase --cover-package=server -v
test_existing_signin (tests.test_auth.TestSignin) ... ok
test_first_time_signin (tests.test_auth.TestSignin) ... ok
test_invalid_signin (tests.test_auth.TestSignin) ... ok
test_user_staff (tests.test_auth.TestSignin) ... ok
test_user_student (tests.test_auth.TestSignin) ... ok
test_staff_dashboard (tests.test_dashboard.TestDashboard) ... ok
test_student_dashboard (tests.test_dashboard.TestDashboard) ... ok

Name                               Stmts   Miss  Cover
-----
server/__init__.py                   7        4    43%
server/auth/__init__.py              1        0   100%
server/auth/user.py                 20        3    85%
server/auth/views.py                28        4    86%
server/cipher.py                    11        4    64%
server/dashboard/__init__.py         0        0   100%
server/dashboard/views.py           166      80    52%
server/db.py                         50       14    72%
server/ldap.py                      34        5    85%
server/server.py                    41        2    95%
server/xml_parser.py                 40       30    25%
TOTAL                               398     146    63%

Ran 7 tests in 5.559s

```

**Figure 13.4:** Using Docker Compose to deploy the server. The server and post-processor unit tests are executed in red and all pass successfully.

```

Step 10/10 : ENTRYPOINT ["/entrypoint.sh"]
--> Running in d2b64be00584
Removing intermediate container d2b64be00584
--> 09a9bc9bc8e6
Successfully built 09a9bc9bc8e6
Successfully tagged plagiarismdetectionplugin_server:latest
Building postprocessor
Step 1/10 : FROM python:3.6.4-slim
--> 43431c5410f3
Step 2/10 : WORKDIR /plagiarism_detection/postprocessor
--> Using cache
--> 0616248eb5a5
Step 3/10 : RUN apt-get update && apt-get install -y gcc tk
--> Using cache
--> bd2df473cfd7
Step 4/10 : ADD entrypoint.sh ./
--> Using cache
--> 8a2c49e600ab
Step 5/10 : RUN set -ex && chmod +x entrypoint.sh
--> Using cache
--> 47de0096c9f
Step 6/10 : ADD requirements.txt ./
--> Using cache
--> 6e50681992ac
Step 7/10 : RUN set -ex && pip3 install -r requirements.txt
--> Using cache
--> d71918567833
Step 8/10 : ADD ./ ./
--> Using cache
--> f52136d021e2
Step 9/10 : RUN set -ex && nosetests --with-coverage --cover-erase --cover-package=postprocessor -v
--> Using cache
--> efa800c8e6eb
Step 10/10 : ENTRYPOINT ["/entrypoint.sh"]
--> Using cache
--> 0026ea42c18c
Successfully built 0026ea42c18c
Successfully tagged plagiarismdetectionplugin_postprocessor:latest
Creating plagiarismdetectionplugin_mongodb_1
Creating plagiarismdetectionplugin_server_1
Creating plagiarismdetectionplugin_postprocessor_1
WARNING: Connection pool is full, discarding connection: localhost
WARNING: Connection pool is full, discarding connection: localhost
Attaching to plagiarismdetectionplugin_mongodb_1, plagiarismdetectionplugin_postprocessor_1, plagiarismdetectionplugin_server_1

```

**Figure 13.5:** Using Docker Compose to deploy the server. Each of the containers are built successfully and the containers are deployed.

## Chapter 14

# Results Analysis

In the following sections, the results produced from the post-processor will be discussed. The results include the submission data, FTS chart (Frequency vs. Time including Source), and large changes. Two different submissions will be used as examples.

### 14.1 Example A

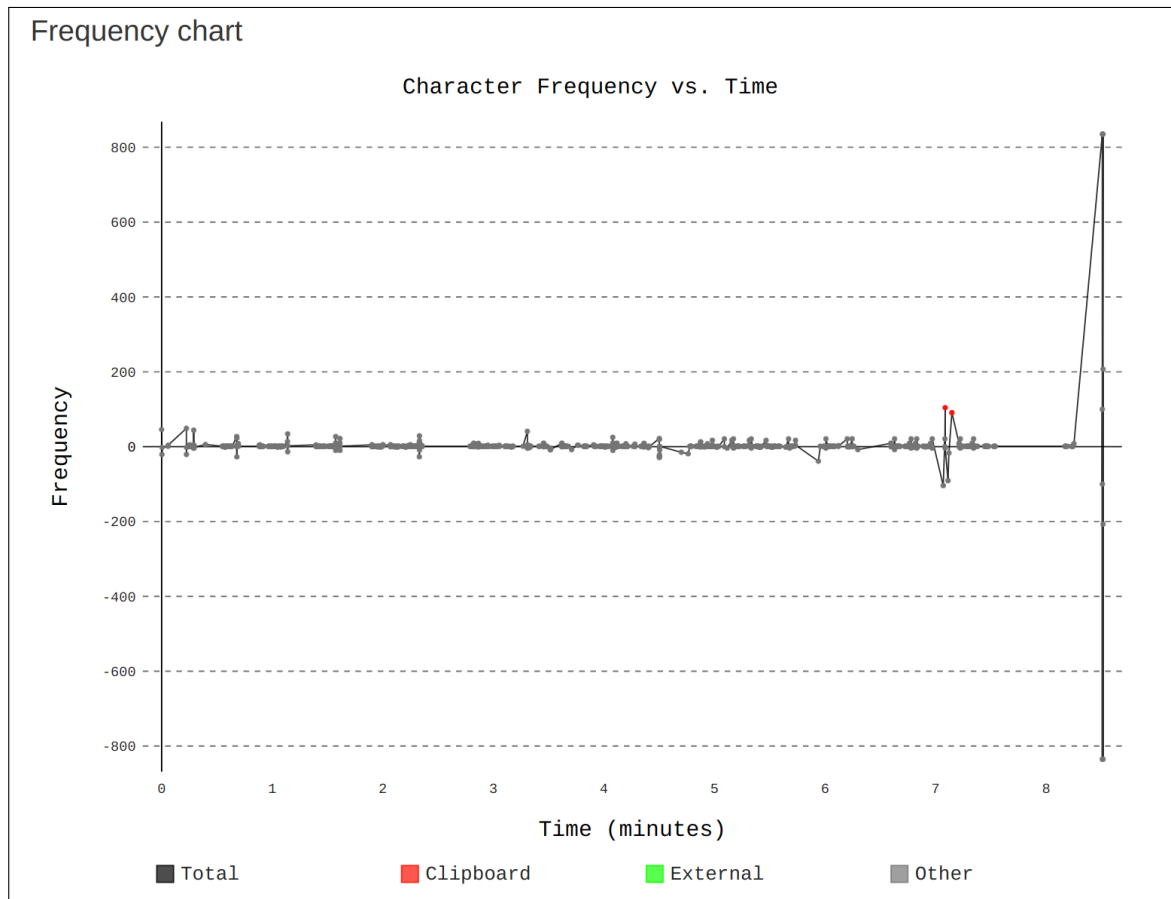
The submission data is shown below in Figure 14.1. This shows several metrics which can be used to quickly identify potential submissions which contain plagiarism or UAC. The **P value** is calculated from the other six values using the algorithm described in subsection 11.2.4 of chapter 11. A linear scale is used for the **P value** and has an associated colour to help with quick identifications. The colour ranges from green to red, with red being a higher chance of plagiarism. This example has a value of 3.77 which is quite low. However, it does contain some copy/paste code, as indicated by the **Clipboard frequency** which is at 12.89%.

Submission data	
Full name	Darren White
Title	Sample 1
Module	CSSAMPLE
Submitted on	Sun, Apr 8 2018 17:18:20
Total time	0 hours 8 minutes 30 seconds
Clipboard frequency	12.89%
External frequency	0.00%
Other frequency	87.11%
Characters per minute	178
Diff ratio	1.00
P value	3.77

**Figure 14.1:** Example A submission data



Figure 14.2 shows the frequency of each source change over the time of development. There are several substantial changes which are noticeable immediately. Two clipboard changes occur just after 7 minutes, but notice that there were deletions of similar sizes just before these. This could potentially just be using copy/paste to move a piece of code. Similarly with the last two large changes, these could be part of a large refactoring method or reformatting. These changes can be individually looked at in detail in the *large changes* below the chart.



**Figure 14.2:** Example A submission FTS chart. Useful for identifying large changes.

Table 14.1 shows the large changes for this example submission. The size of a large change was defined as 91 in this case (the default is 200). This can be changed by setting the `large_change_size` URL GET parameter. The content in the **OldString** and **NewString** columns have been replaced with letters. The content is not particularly interesting in this case. Note that each letter used denotes a unique piece of code. For example, all cells with *A* represent the same piece of code.

Let's start with the first two rows. Row 0 shows us that *A* was removed and in row 1 it was pasted in a different location. As the **NewString** cell of row 0 is empty we can identify that it was removed. We know *A* was then pasted in a different location due to the row 1 **Source** being CLIPBOARD and the row 1 **Offset** being different to that of row 0. Also note that the time difference between rows 0 and 1 is around 1 second. The same scenario occurs for rows 2 and 3.

The next four pairs of rows, all share a similar scenario too. Firstly, the time difference between each pair is approximately 1ms. This indicates that this is most likely from an automatic refactoring or reformatting process. We can also see that the **NewString** content is immediately removed from another location showing that it was most likely moved.

After analysing the metrics, FTS chart, and large changes, this submission seems to have no identifiable acts of plagiarism.

<b>Id</b>	<b>Path</b>	<b>Timestamp</b>	<b>Source</b>	<b>Size</b>	<b>Offset</b>	<b>OldString</b>	<b>NewString</b>
0	src/BingoCaller.java	7.069433	OTHER	104	825	<i>A</i>	
1	src/BingoCaller.java	7.087116	CLIPBOARD	104	717		<i>A</i>
2	src/BingoCaller.java	7.112766	OTHER	91	605	<i>B</i>	
3	src/BingoCaller.java	7.147666	CLIPBOARD	91	838		<i>B</i>
4	src/BingoCaller.java	8.508266	OTHER	835	1090		<i>C</i>
5	src/BingoCaller.java	8.508283	OTHER	835	254	<i>C</i>	
6	src/BingoCaller.java	8.508283	OTHER	100	254		<i>D</i>
7	src/BingoCaller.java	8.508300	OTHER	100	1190	<i>D</i>	
8	src/BingoCaller.java	8.514716	OTHER	835	1191		<i>E</i>
9	src/BingoCaller.java	8.514733	OTHER	835	355	<i>E</i>	
10	src/BingoCaller.java	8.514733	OTHER	207	355		<i>F</i>
11	src/BingoCaller.java	8.514750	OTHER	207	1398	<i>F</i>	

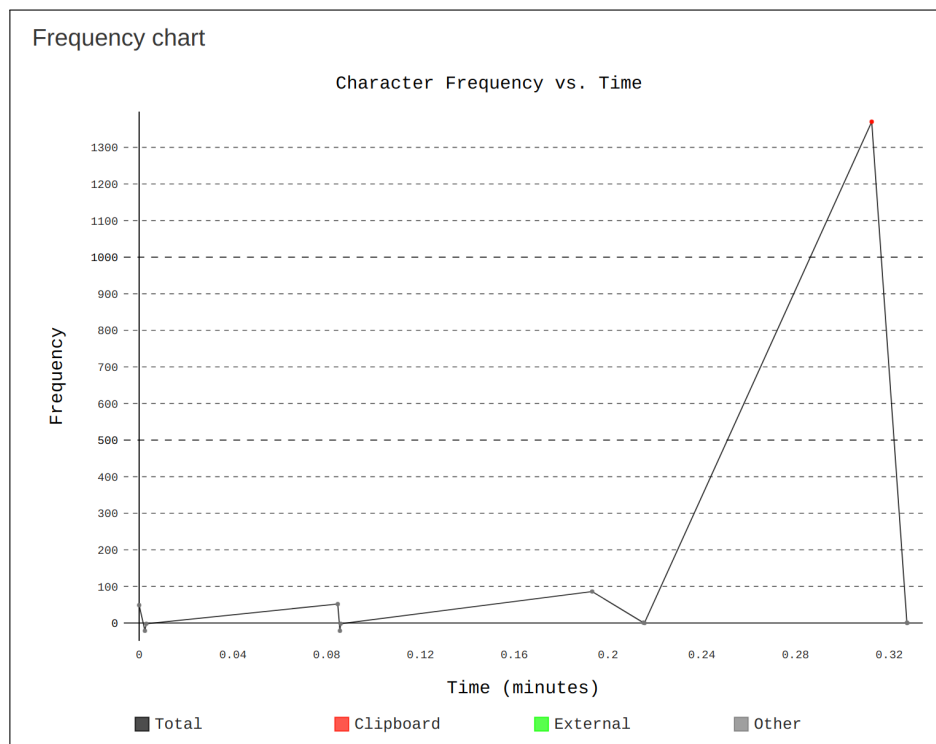
**Table 14.1:** Example A submission large changes. Useful for identifying possible cases of plagiarism. *Note: The code samples in the **Old String** and **New String** columns have been replaced with symbols. The content in these is not of particular use.*

## 14.2 Example B

Similar to Example A, we will look at the submission data, FTS chart, and large changes. Figure 14.3 already shows a major red flag with the **P value** being 134.15. Looking more closely at the other values, it is easy to see why the **P value** is so high. 90.55% of the code was copy/paste and it was completed in 19 seconds. The FTS chart shown in Figure 14.4 also backs this up. The chart shows a massive spike where code was pasted. The chart also shows that no code was deleted from the project at all.

View submission	
Submission data	
Full name	Darren White
Title	Sample 2
Module	CSSAMPLE
Submitted on	Sun, Apr 8 2018 17:22:01
Total time	0 hours 0 minutes 19 seconds
Clipboard frequency	90.55%
External frequency	0.00%
Other frequency	9.45%
Characters per minute	4617
Diff ratio	1.00
P value	134.15

**Figure 14.3:** Example B submission metrics



**Figure 14.4:** Example B submission FTS chart. Useful for identifying large changes.

Below the **Old String** and **New String** values are shown for the CLIPBOARD change. Listing 14.1 shows the original file with an empty public class statement. In Listing 14.2 the pasted contents are displayed. The bulk of the pasted contents are not displayed as it is not valuable to this analysis. However, the pasted contents accounts for 90% of the project and completes the file. This provides clear evidence of plagiarism.

---

```
1 public class BingoCaller {  
2 }
```

---

**Listing 14.1:** Example B Change Old String

---

```
1 import java.security.SecureRandom;  
2 import java.util.*;  
3  
4 public class BingoCaller {  
5  
6     // Lots of copy and pasted code here  
7     // Removed it as it is unnecessary  
8  
9 }
```

---

**Listing 14.2:** Example B Change New String

## Chapter 15

# Evaluation

The major requirements for this project were an IntelliJ Platform Plugin, back-end server, post-processor, front-end application, and a database. Each of these components were developed up to the point where they operate as intended. The settings GUI for the plugin was not implemented due to the using the continuous server architecture. Code generation, and code auto-completion of the source code detection types were not implemented due to the complexity of the identification process.

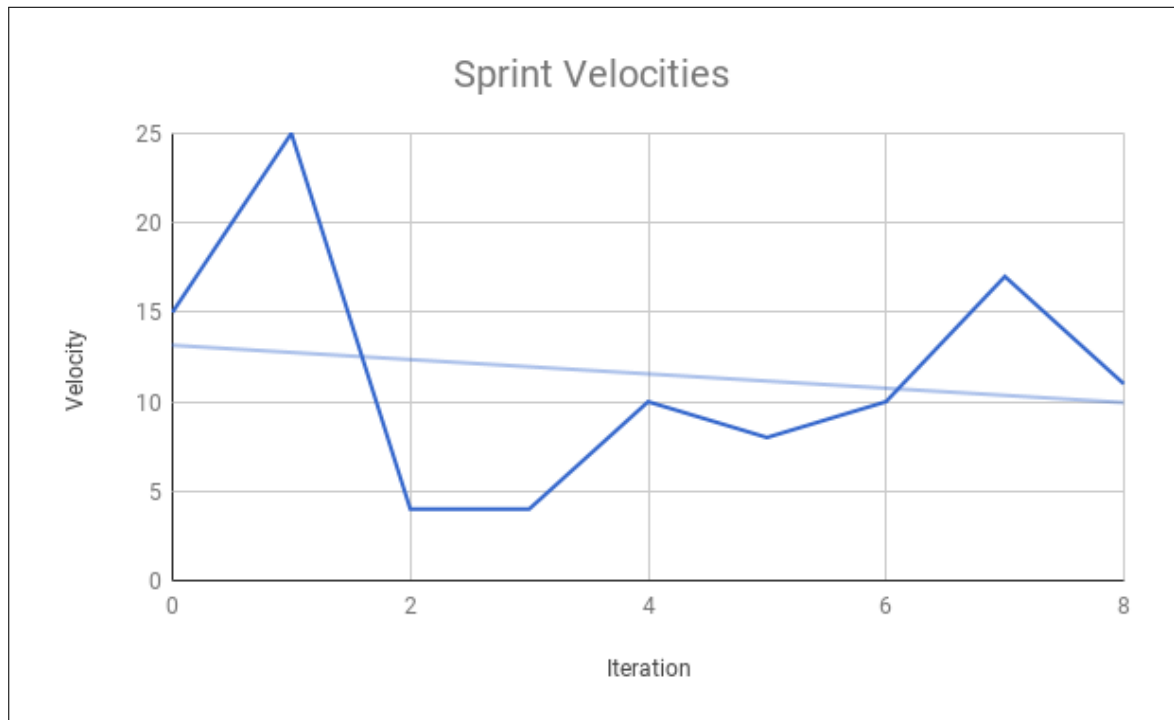
The primary design decision between using a continuous or non-continuous server architecture was a huge decision during this project. It impacted the method in which students would interact with the plugin and submit their tracked data. This in turn affected the back-end server internals. This decision was required before fully developing the server. If the continuous design was used then it would have most definitely increased the workload. Testing the direct connection between the plugin and server would be significantly more difficult than being able to individually test each component.

Using MongoDB was a very clear and strong design decision. If a RDBMS was to be used then the tracked data would need to be completely different from its current form. The use of Python has a clear advantage for using MongoDB. The simplicity of using dictionaries lead to simple interactions with the database.

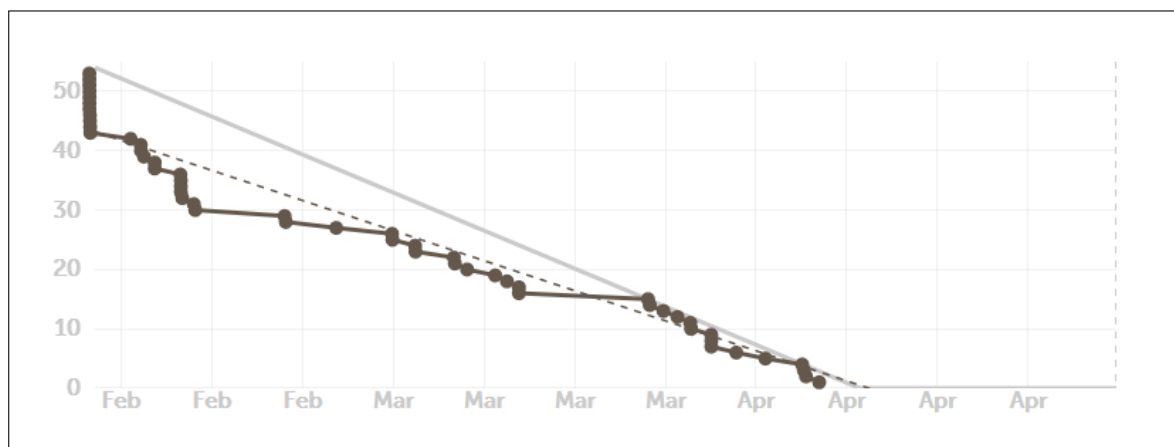
If the project was to be written from scratch again, it would be interesting to have a greater focus on the post-processor. Concentrating more on the identification of possible cases of plagiarism would provide a much better result. The result would be better due to the simplicity of the current identification algorithm being used. Integrating complex algorithms in the post-processor would allow for a more concrete result for each submission.

IntelliJ IDEA offers many ways to interact with the editor. Many were explored during the project including typing, copy and pasting, auto-completion, automatic code generation, and refactoring. Other methods that were not explored were using IdeaVim, and applying Git operations. IdeaVim emulates Vim in the editor. How would the use of this affect the detection methods? Would yank and paste still register as copy and paste? Using Git to stash changes, change branches, or even simply pulling from upstream can all modify files. Would these changes be tracked? If so, would this result in conflicts in the XML file? The XML file would need to be staged in each branch and committed before changing branch. Both IdeaVim, and Git could have been investigated.

Figure 15.1 shows the velocity for each sprint. The average sprint velocity was 11. Figure 15.2 shows the burndown for all iterations. Both of these charts show that development started well, slowed in the middle, and gained speed again towards the end. Using the scrum methodology proved successful for developing this project. It allowed, at maximum, a week to be wasted. If a waterfall approach was to be used, a lot more time could have been wasted if problems arose.



**Figure 15.1:** Velocity for each iteration



**Figure 15.2:** Burndown chart for all iterations

## Chapter 16

# Conclusions

### 16.1 Introduction

As stated in chapter 1, the main objective of detecting plagiarism, is finding pieces of work which originated from another source. This higher level objective has been kept in mind throughout this whole project. This chapter will conclude if that objective has been met, and to what level it has been achieved.

### 16.2 Goal

Two problems were identified in chapter 2; tracking the work as it's being written, and identifying plagiarised work. Both of these problems have been researched, implemented, and solved.

Tracking the work has been achieved by developing an IntelliJ IDEA Plugin. This plugin tracks all file changes in the background of the IDE editor and saves these changes to an 128-bit AES encrypted XML file.

Identifying the plagiarism involved developing two components; the back-end server, and post-processor. The server provides a front-end web interface for both students and staff to sign-in to. Students may submit their tracked data and view their previous submissions. Staff are able to view all students' submissions and the resulting plagiarism outcome. The post-processor's job is to process student submissions and identify plagiarism from the tracked data. Although the algorithms used to identify plagiarism are simple, they are effective, and provide a new path for detecting plagiarism. These algorithms can be dramatically improved in the future. The plugin tracking system works perfectly, but it too has room for improvement. The future improvements for both the tracking, and detection algorithms are described in section 16.3.

In comparison to the other existing systems, this project seems to be one of a kind. All existing systems previously discussed in chapter 2 do not actively track the development of work. This system achieved that goal. However, the existing systems do have complex algorithms which far surpass the expectations of the algorithms used in this system.

Overall, the main objective of finding pieces of work which originated from another source has been achieved. The complexity of detection may not be as high as initially expected, but this can be improved.

## 16.3 Future Development

All of the components can be improved dramatically. The plugin tracking can include more source classifications. The only sources identified during this project were, external file changes, clipboard, file renaming, and other (everything else, including keyboard typing, is classed as other). More classifications will improve the detection algorithms to be more accurate and will lead to pattern recognition. A much larger improvement would be to implement the continuous server design as described in chapter 2. This would remove the need for students to interact with the web application, improving the overall user experience.

The back-end server has plenty of room for improvement. Adding a load balancer with multiple servers would reduce load and add redundancy preventing any problems if the server was the fail. This can be done with the added support of Docker and Docker Compose. Other container orchestration tools could also be added to improve system scalability and stability, such as Docker Swarm or Kubernetes. The visual and functional aspects of the web application can be further refined. Multiple filtering options for submissions could be added. These filtering options may include, module, name, date, and plagiarism value range. A useful automated feature would be notifications for potential plagiarised submissions. This would notify staff when a new submission has been detected to contain plagiarism or UAC.

The post-processor is the brains of the system. Implementing more intelligent plagiarism detection algorithms would vastly improve the detection of plagiarism in submissions. A great starting point would be to automate the analysis performed in chapter 14. Being able to automatically identify large changes and detect if they cause plagiarism or not would be a big stepping stone and would remove some of the manual analysis required. More sophisticated algorithms would include machine learning. A similar approach to that of Bandara and Wijayathna would be a good place to start [23]. Being able to learn how each students develop their code, then analysing their work for outliers would allow for direct identification of plagiarism code.

Several stories in the backlog were not completed during development. These stories were not a high priority, but may be completed in the future. These stories were:

- **Detect automatic code generation** - To improve the detection algorithms, the tracking classification of sources can be improved to be more specific. Detecting automatic code generation such as, *insert constructor*, and *generate toString* will be a major improvement.
- **Detect code auto-complete** - As with detecting automatic code generation, detecting code auto-completion will help improve plagiarism detection.
- **Detect code refactoring** - Class renaming has already been implemented, but there are many more forms of refactoring which can be detected and will only improve the system further.
- **Allow resubmissions for previous submissions in student dashboard** - Allow resubmissions for students will improve the user experience of the web application. This will reduce duplicate submissions.
- **Add a module filter to staff dashboard** - Currently the staff dashboard will display all student submissions. Being able to filter the submissions would greatly improve the user experience.



- **Add pagination to submissions table** - Another user experience improvement. This would also reduce bandwidth usage.
- **Improve submission post to use transactions if possible** - Using transactions are a key part of databases when being used by many users concurrently.
- **Investigate how to gather training and testing data** - Gathering training and testing data would be the start of implementing machine learning techniques.
- **Improve CPM calculation** - Currently the CPM calculation includes the time when the project is closed which dramatically reduces the CPM. Improving this will improve the detection methods.

# **Appendices**

## Appendix A

# Third-Party Code and Libraries

- **IntelliJ IDEA 2018.1 and IntelliJ Plugin SDK Build 173.0** - The plugin was developed for the JetBrains IDE, IntelliJ IDEA. This used the open source IntelliJ Plugin SDK. This SDK is released under the Apache License. This library was used without modification.
- **JDK 8** - IntelliJ IDEA is built using Java and so the JDK was also used to develop the plugin. The JDK is licensed under the GNU General Public License. This library was used without modification.
- **JUnit 4** - JUnit was used to write tests for the plugin. JUnit is released under the Eclipse Public License. This library was used without modification.
- **Python 3** - Python 3 was used to write and develop the back-end server and the post-processor. Python 3 is released under the Python Software Foundation License. This library was used without modification.
- **Python 3 Packages** - Multiple packages were also used with Python 3:
  - **Coverage v4.5.1** - Coverage was used for unit tests coverage. Coverage is released under the Apache License. This library was used without modification.
  - **Flask v0.12.2** - Flask was used to develop the web service for the back-end server. Flask is licensed under a three clause BSD License. This library was used without modification.
  - **Flask-Login v0.4.1** - Flask-Login was used to provide authentication services for the back-end server. Flask-Login is licensed under the MIT License. This library was used without modification.
  - **ldap3 v2.4.1** - ldap3 was used to provide a connection the an LDAP server for authentication for the back-end server. ldap3 is licensed under the LGPL v3 license. This library was used without modification.
  - **Matplotlib v2.2.2** - Matplotlib was used by the post-processor for displaying the FTS charts. Matplotlib uses a license based on the Python Software Foundation License. This library was used without modification.
  - **Mock v2.0.0** Mock was used by the post-processor and server to mock parts of the system for unit tests. Mock is licensed under the BSD License. This library was used without modification.

- **MockupDB v1.3.0** - MockupDB was used to mock the MongoDB client for the server and post-processor. MockupDB is licensed under the Apache License. This library was used without modification.
  - **Nose v1.3.7** - Nose was used for unit testing the server and post-processor. Nose is licensed under the GNU Lesser General Public License. This library was used without modification.
  - **PyCrypto v2.6.1** - PyCrypto was used by the server and post-processor to encrypt and decrypt the XML data. PyCrypto is licensed under Public Domain License. This library was used without modification.
  - **Pygal v2.4.0** - Pygal was used to display the FTS chart on the web application. Pygal is licensed under the GNU Lesser General Public License v3. This library was used without modification.
  - **PyMongo v3.6.0** - PyMongo was used by the post-processor and server for connecting to the MongoDB server. PyMongo is licensed under the Apache License. This library was used without modification.
- 
- **Docker v18.03.0-ce** - Docker was used to build and deploy containers for the server, post-processor, and database. Docker is licensed under the Apache License. This library was used without modification.
  - **Docker Compose v1.8.0** - Docker Compose was used to deploy the server, post-processor, and database containers together. Docker Compose is licensed under the Apache License. This library was used without modification.
  - **MongoDB v3.6** - MongoDB was used to store the submitted XML data. MongoDB is licensed under GNU AGPL v3.0. This library was used without modification.

## **Appendix B**

# **Ethics Submission**

The ethics submission is shown on the following pages. The Ethics Application Number is: 9702.

**AU Status**

Undergraduate or PG Taught

**Your aber.ac.uk email address**

daw48@aber.ac.uk

**Full Name**

Darren White

**Please enter the name of the person responsible for reviewing your assessment.**

Chris Loftus

**Please enter the aber.ac.uk email address of the person responsible for reviewing your assessment**

cwl@aber.ac.uk

**Supervisor or Institute Director of Research Department**

cs

**Module code (Only enter if you have been asked to do so)**

CS39440

**Proposed Study Title**

MMP: IntelliJ Plugin to Aid With Plagiarism Detection. Tracks file changes in the editor of IntelliJ IDEA to help identify UAC or plagiarism. Not used with students during the course of the project. Could be used with students post-completion.

**Proposed Start Date**

01/02/2018

**Proposed Completion Date**

04/05/2018

**Are you conducting a quantitative or qualitative research project?**

Mixed Methods

**Does your research require external ethical approval under the Health Research Authority?**

No

**Does your research involve animals?**

No

**Are you completing this form for your own research?**

Yes

**Does your research involve human participants?**

No

**Institute**

IMPACS

**Please provide a brief summary of your project (150 word max)**

IntelliJ IDEA plugin to track file changes in the editor to help with detecting plagiarism or UAC. During the research, student participants were not involved. However, post-completion the plugin could be used in academia involving students.

**Where appropriate, do you have consent for the publication, reproduction or use of any unpublished material?**

Not applicable

**Will appropriate measures be put in place for the secure and confidential storage of data?**

Yes

**Does the research pose more than minimal and predictable risk to the researcher?**

Not applicable

**Will you be travelling, as a foreign national, in to any areas that the UK Foreign and Commonwealth Office advise against travel to?**

No

**Please include any further relevant information for this section here:**

**If you are to be working alone with vulnerable people or children, you may need a DBS (CRB) check. Tick to confirm that you will ensure you comply with this requirement should you identify that you require one.**

Yes

**Declaration: Please tick to confirm that you have completed this form to the best of your knowledge and that you will inform your department should the proposal significantly change.**

Yes

**Please include any further relevant information for this section here:**

## Appendix C

# Code Examples

### 3.1 Example MongoDB Student Document

Below is an example MongoDB student document. The document contains the user information retrieved from the Aberystwyth University LDAP server. The submissions array contains all of the students tracked project data. Each submission entry has a one-to-one relation with the XML file. The submission `_id` is unique for each submission. The `_id` also contains the date which is useful to know when submission was posted. The submission contains input from the student, such as the title and module. The XML file that is submitted is processed and transformed into the correct formatting (JSON) which contains all of the tracked files. Each file has its path, changes, and cache. After post-processing, the result is added to the submission.

---

```

1  {
2    "_id" : ObjectId("5aca40bda326f6001046e1b7"),
3    "uid" : "abc12",
4    "full_name" : "John Smith",
5    "user_type" : "ABUG",
6    "submissions" : [
7      {
8        "_id": ObjectId("5accc6ba7041650010aa3a37"),
9        "title": "A Submission",
10       "module": "CS12345",
11       "files": {
12         "src/Test__java": {
13           "path": "src/Test.java",
14           "changes": [
15             {
16               "offset": "0",
17               "oldString": "",
18               "newString": "package PACKAGE_NAME;\n\npublic
↵ class Test {\n}\n",
19               "source": "OTHER",
20               "timestamp": "1522599829159"

```



```
21         }
22     ],
23     "cache": "cHVibGljIGNsYXNzIFRlc3QgewoKfQ=="
24 }
25 },
26 "result": {
27     "src/Test__java": {
28         "diff_ratio": 1,
29         "frequency_total": 45,
30         "frequency_clipboard": 0,
31         "frequency_external": 0,
32         "frequency_other": 45,
33         "frequency_time_source_data": [
34             {
35                 "f": 45,
36                 "s": "OTHER",
37                 "t": NumberLong("1522599829159")
38             }
39         ]
40     }
41 }
42 }
43 ]
44 }
```

**Listing C.1:** Example student JSON MongoDB document

# Annotated Bibliography

- [1] Plagiarism.org. (2018) Plagiarism: Facts & stats - plagiarism.org. Last accessed: 2018-04-06. [Online]. Available: <http://plagiarism.org/article/plagiarism-facts-and-stats>

*This web page contains results from surveys performed in schools on plagiarism. The outcomes from these surveys show how many students commit plagiarism.*

- [2] A. University. (2018) Aberystwyth university - regulation on unacceptable academic practice. Last accessed: 2018-04-06. [Online]. Available: <https://www.aber.ac.uk/en/aqro/handbook/regulations/uap/>

*This shows Aberystwyth University's regulation on UAP and plagiarism. It describes the definition of UAP and the multiple forms it can take.*

- [3] Plagiarism.org. (2018) What is plagiarism? - plagiarism.org. Last accessed: 2018-04-06. [Online]. Available: <http://plagiarism.org/article/what-is-plagiarism>

*This web page has detailed information on what exactly plagiarism is. It states what the definition of plagiarism is and example forms of plagiarism.*

- [4] Turnitin. (2018) Turnitin - the plagiarism spectrum. Last accessed: 2018-04-06. [Online]. Available: [http://turnitin.com/assets/en\\_us/media/plagiarism-spectrum/](http://turnitin.com/assets/en_us/media/plagiarism-spectrum/)

*This web page shows a list of 10 different ways to plagiarise with example texts.*

- [5] P. Clough and D. O. I. Studies, "Old and new challenges in automatic plagiarism detection," in *National Plagiarism Advisory Service, 2003*; <http://ir.shef.ac.uk/cloughie/index.html>, 2003, pp. 391–407.

*In-depth details on lexical and structural changes in program code plagiarism.*

- [6] A. Parker and J. Hamblen, "Computer algorithms for plagiarism detection," *IEEE Transactions on Education*, vol. 32, no. 2, pp. 94–99, may 1989. [Online]. Available: <https://doi.org/10.1109/13.28038>

*Details on the plagiarism spectrum. The different levels of plagiarism. Level 0 has no changes and level 6 has the control logic changed.*

- [7] JetBrains. (2018) What is the IntelliJ platform? Last accessed: 2018-04-07. [Online]. Available: <http://www.jetbrains.org/intellij/sdk/docs/intro/intellij-platform.html>

*This page describes what the IntelliJ Platform is.*

- [8] R. Lukashenko, V. Gaudina, and J. Grundspenkis, “Computer-based plagiarism detection methods and tools,” in *Proceedings of the 2007 international conference on Computer systems and technologies - CompSysTech '07*. ACM Press, 2007. [Online]. Available: <https://doi.org/10.1145/1330598.1330642>

*Useful attribute table containing information on tools for detecting plagiarism such as Turnitin, and MOSS.*

- [9] S. Schleimer, D. S. Wilkerson, and A. Aiken, “Winnowing: local algorithms for document fingerprinting,” in *Proceedings of the 2003 ACM SIGMOD international conference on on Management of data - SIGMOD '03*. ACM Press, 2003. [Online]. Available: <https://doi.org/10.1145/872757.872770>

*This went into detail on the ideas behind MOSS, such as document fingerprints and k-grams.*

- [10] L. Prechelt and G. Malpohl, “Finding plagiarisms among a set of programs with jplag,” vol. 8, 03 2003.

*Detailed information on how JPlag detects plagiarism using tokens.*

- [11] M. Wise, “Yap3: Improved detection of similarities in computer program and other texts,” vol. 28, 04 1996.

*In-depth detail on the RKR-GST algorithm used by YAP3 and the comparison between the YAP versions.*

- [12] JetBrains. (2018) Creating your first plugin. Last accessed: 2018-04-09. [Online]. Available: [https://www.jetbrains.org/intellij/sdk/docs/basics/getting\\_started.html](https://www.jetbrains.org/intellij/sdk/docs/basics/getting_started.html)

*The JetBrains tutorial for developing an IntelliJ Plugin.*

- [13] —. (2018) JetBrains/intellij-community: IntelliJ idea community edition. Last accessed: 2018-01-31. [Online]. Available: <https://github.com/jetBrains/intellij-community>

*The IntelliJ IDEA Community Edition GitHub repository provides code for plugins. This helped with understanding the IntelliJ Platform Plugin SDK.*

- [14] GitHub. (2018) About milestones - user documentation. Last accessed: 2018-04-06. [Online]. Available: <https://help.github.com/articles/about-milestones/>

*Describes using issues and milestones on GitHub.*

- [15] —. (2018) Closing issues using keywords - user documentation. Last accessed: 2018-04-06. [Online]. Available: <https://help.github.com/articles/closing-issues-using-keywords/>

*Explains how to close issues using commit messages by including specific keywords.*

- [16] JetBrains. (2018) Plugin components / intellij platform sdk devguide. Last accessed: 2018-04-17. [Online]. Available: [https://www.jetbrains.org/intellij/sdk/docs/basics/plugin\\_structure/plugin\\_components.html](https://www.jetbrains.org/intellij/sdk/docs/basics/plugin_structure/plugin_components.html)

*Information on project components and the differences between application, project, and module level components.*

- [17] G. Cannata. (2018) Welcome to ldap3's documentation - ldap3 2.5 documentation. Last accessed: 2018-04-19. [Online]. Available: <http://ldap3.readthedocs.io/>

*The LDAP3 library documentation for Python 3.*

- [18] B. Hackett. (2018) Pymongo 3.6.1 documentation - pymongo 3.6.1 documentation. Last accessed: 2018-04-19. [Online]. Available: <https://api.mongodb.com/python/current/>

*The PyMongo library documentation for Python 3 and MongoDB.*

- [19] Bootstrap. (2018) Signin template for bootstrap. Last accessed: 2018-04-20. [Online]. Available: <https://getbootstrap.com/docs/3.3/examples/signin/>

*A signin template using Bootstrap. This uses a form to submit a username and password in a form.*

- [20] Flask. (2018) Uploading files - flask documentation (0.12). Last accessed: 2018-02-27. [Online]. Available: <http://flask.pocoo.org/docs/0.12/patterns/fileuploads/>

*Flask documentation containing a detailed tutorial on uploading files with a HTML form.*

- [21] A. J. J. Davis. (2013) Test mongodb failure scenarios with mockupdb. Last accessed: 2018-03-04. [Online]. Available: <https://emptysqua.re/blog/test-mongodb-failures-mockupdb/>

*A blog post on using MockupDB with MongoDB in Python 3. The online documentation was not straight-forward and this cleared up many questions and any confusion. Seeing this real example helped massively when developing the back-end server unit tests.*

- [22] MongoDB. (2018) Change streams - mongodb manual 3.6. Last accessed: 2018-04-22. [Online]. Available: <https://docs.mongodb.com/manual/changeStreams/>

*MongoDB documentation for change streams which are required for using the watch method on a collection. These are only supported in 3.6.*

- [23] U. Bandara and G. Wijayrathna, "Detection of source code plagiarism using machine learning approach," *International Journal of Computer Theory and Engineering*, pp. 674–678, 2012. [Online]. Available: <https://doi.org/10.7763/ijcte.2012.v4.555>

*This article used three machine learning techniques to predicate the other of source code files. Source code metrics such as line length, and comments frequency were used as inputs to the machine learning classifiers.*

- [24] M. Driscoll. (2018) Using pygal graphs in flask — the mouse vs. the python. Last accessed: 2018-04-23. [Online]. Available: <https://www.blog.pythonlibrary.org/2015/04/16/using-pygal-graphs-in-flask/>

*A tutorial using pyGal for displaying charts in a Flask application.*

- [25] LiveChat. (2018) Free typing speed test - check your typing skills. Last accessed: 2018-04-24. [Online]. Available: <https://www.livechatinc.com/typing-speed-test/>

*This website provides a typing speed test. It also includes the average WPM (words per minute and CPM (characters per minute).*

- [26] P. S, R. R, and S. B. B, "A survey on plagiarism detection," *International Journal of Computer Applications*, vol. 86, no. 19, pp. 21–23, jan 2014. [Online]. Available: <https://doi.org/10.5120/15104-3428>

*This article presents a survey on plagiarism detections tools and methods.*