Document Classification System

Self-assessment Report

2014/15 Semester A

Delivery Data: 5th Dec 2014

CS3343 LA1 – Acumen

Member Information

Role	**Name**	**SID**
PM	WANG Mingyang	52640166
Assistant PM	PU Jie	52640234
CM	FANG Zhou	52639099
Developer	FENG Xikang	52639689
Developer	WANG Yiji	52639040
Developer	ZHU Renjie	52639014

1. Summary Section

The goal of our project is to classify a given article by categories and improve the correctness of our program. Since we followed both agile development process and test-driven development process, we firstly only built several basic modules such as input, output and counter, and used some JUnit test cases prepared beforehand to ensure the quality of our code. Later on, we consistently and progressively improved, tested and refactored our code through various internal code reviewing.

In our program, we have implemented several well-functioned and encapsulated modules. The first one is ReadContent module. Every file reading is achieved by making use of this class. For document analysis, we are only interested in words consisting of [a-z], [A-Z] or "-", so we need word filtering while reading Since there are a lot of file reading in other classes, it will be more clear to split this function out and build a module just responsible for IO and filtering process.

The second one is CalculateIG module. After summarizing the words in training data on the first stage, we get a word count matrix, which records the useful words and their occurrences in each file. Then, the CalculateIG module will use this matrix to calculate the information gain of each word and assign a weight. By sorting the information gain in descending order, we take the top 10% as feature words, which we will use, in next stage for classification.

The core module is CalculateSimilarity. From stage 2, we get a list of feature words, which are significant for its corresponding category. In this module, the program creates a vector for each training file according to the occurrences of feature words in the document. Also a vector for the input file will be generated. Through calculating the cos value between two vectors, we could find out the similarity between two files. After getting all the similarites, we choose the top 10% similarity values and the category that major files belonging to will be our result. For example, among the top 10% similar files, most of them belongs to games, then the input file is likely to be games too.

The whole development period consist of 3 cycles.

In Cycle 1 (week 1 - 5), we firstly built our team and assigned roles to each member in the team, where WANG Mingyang (Young) served as Project Manager, PU Jie served as Assistant PM, and FANG Zhou (Alex) served as Configuration Manager. Next we brainstormed about our project. After we set down our goal, we started to collect training data and design our core algorithm. Meanwhile, different tasks were assigned to our teammates. (Detailed tasks assignment could be found in our project report.) At the end of this cycle, we implemented a demo program with basic functionalities and a brief project plan.

In Cycle 2 (week 6 - 9), we started to extend and enhance our demo program with adopting both agile development process and test-driven development process. As we found a better algorithm for our project, we re-implemented our code module with the new algorithm. Additionally, we drafted the frameworks for documentations such as object-oriented analysis and design report, bug reports, and test report, and improved our project plan.

In Cycle 3 (week 10 - 14), we refactored our program based on pre-found code smells and added more specific test scripts to fully test our program. Besides, we drew several diagrams to support our report, including domain models and sequence diagrams, and exported our Gatt Graph and working schedule based on our progresses. As well, we polished and integrated our bug reports and team reports. Therefore, all documentations were updated to consummate versions at the end.

2. Self-assessment

Name: WANG Mingyang
SID: 52640166

Severing as Project Manager in this project, I was mainly responsible for planning, coordinating and monitoring the project. In terms of project implementation, I was involved into every part of our project, including analyzing, coding, testing and documenting.

Best practices of software engineering skills
Productive teamwork
Concise and precise documentation

As Project Manager

(1) Lead the team to follow the project management flow: Plan -> Schedule -> Track -> Measure (-> Plan …)
(2) Employ Agile Development Process and Test-Driven Development Process in our project Assist Configuration Manager to set up Git/Github development environment for ever member
(3) Summarize our progress and sketch plans for next phase weekly
(4) Assign tasks to members based on the rotation schedule and their interests
(5) Ensure the quality of our code and documentation with respect to the requirements

As Develop Engineer & Test Engineer

(1) Design the core algorithm with teammates
(2) Integrate modules implemented by teammates
(3) Help to debug and test our code
(4) Help to design the refactoring strategy
(5) Integrate and consummate documentations

Name: PU Jie
SID: 52640234

In this project, my main role is to carry out planning and administration, as well as documentation works during the whole project. However, as we switched roles in the progress, I also did some coding and testing jobs. I have achieved, and gained a lot from the project.

Below are the details.
When working on the planning of the project, we had several meetings, and by cooperating with other teammates, we soon came out with a preliminary idea of the project, how it looks like, and how to achieve that. I drafted and maintained the project plan, which is an important part in the project, and also worked out the milestones, project schedule, as well as the gantt chart. A concise and precise documentation work is achieved. During the process of coding and testing, I used several different methods learnt in the lectures. I wrote several test cases to accomplish full coverage of the classes.

In the progress of developing the project, I have a concise learning of the test-driven project developing methodology and whole image of the IT project developing process. The most important thing learnt is the significance of teamwork. All teammates need to cooperate with each other well to succeed in finishing a large project.

Name: FANG Zhou
SID: 52639099

Responsibility:

(1) Configuration Manager

- Manage the github projects and check the modification of codes between each pull request.
- Add comments to the change for future tracking.
- Help teammates update their local projects and file hierarchy

(2) Data collection

(3) Resolve the bug 0001~ 0005, provides the code change for team code review.

(4) Discuss with a small group about the algorithm to classify documents.

(5) Finish the calculating information gain and similarity code part. Participate in the code refactoring process.

(6) Generate javaDoc and user manual.

(7) Build executable jar file.

Knowledge obtained from project:

(1) Test-debug-refactor development. Using Junit to test the correctness of programs.

(2) The collaboration between teammates when working on a big project. Each team member should be responsible for a certain part and the importance to keep project consistency for intergration.

(3) The format of useful bug report, which can help developer find the bug easily.

(4) Code Review and refactoring . It is a good practice to review the code change with team mates and find the potential influences to other parts.

(5) The coverage of Junit testing. It is very hard to ensure 100% code coverage, however some testing strategy like bottom up and top down are being using in the testing process to test the correctness of program.

Name: FENG XiKang
SID: 52639689

Responsibility:

(1) Data collection (Computer part)
(2) Coding the IO class to read one file or all files in one folder for version 1
(3) Finding the core algorithm of our project
(4) Draft object oriented analysis and design report
(5) Coding all test cases of our project and refactor it to achieve its coverage more than 90%
(6) Participating in every meeting

What learnt from this project:

(1) Better understanding of Testing-Debugging-Refactoring Cycle
(2) Learn to use Git to cooperate with our teammates.

For example,

1). Pull other teammates' files from their branch and fix confliction between our version
2). Push my work to our master branch
3). Reset the repositories to older version if something goes wrong

(3) Bug fixing skills
(4) Better understanding of test comprehensiveness and test organization by coding Junit test for our project
(5) Know the requirements which a good bug report should achieve

Name: WANG Yiji
SID: 52639040

Responsibility:

(1) Recording meeting minutes
(2) Data collection (games part)
(3) Testing for version 1 and 2
(4) The initialization part of the programming (coding for classification, matrix building and information gain calculating)
(5) Object oriented analysis and design report (Introduction, user story, combined user story, use case, domain modeling diagram, domain entity, integration of the report)
(6) Participating in every meeting

What learnt from this project:

(1) Learn testing methods
(2) Correct data collection skill
(3) Bug fixing skills
(4) Importance of version control
(5) Review user story, use case and domain modeling
(6) Meeting minutes taking skills

Name: ZHU Renjie
SID: 52639014

Responsibility:

(1) Collected available clustering algorithms at the beginning
(2) Performed data collection for office software
(3) Drafted the introduction for project plan.
(4) Implemented the function of dictionary generation for each category, and useless word selection
(5) Conducted bug fixing, composed bug reports and tracked the status for each bug.
(6) Drafted the use case diagram, sequence diagram and summary for the Object oriented analysis and design report
(7) Participated in every group discussing and proposed my ideas for program design.

What I learned from this project:

Firstly, having cooperated with other teammates on this project for over three months, I earned the precious experience on the complete software development circle, and how on earth a piece of software is developed in IT industry.

Secondly, while composing project plan and Object oriented analysis and design report, I become more familiar with software design skills like composing use case, and drawing sequences.

Thirdly, I learned some version control skills while developing the program. Fourthly, I also got the chance to know the good characteristics of a bug report.