Project Title: P8 Analysing the content of README files on GitHub

Author: a1699186 I-Sheng Chris Wang

Part one: Business Case

This is a project about analysing the content of README files on GitHub. Typical README files contain the information about the project whereas the software projects on GitHub often provide high-level documentation in the form of README files. In another form of saying, README files are generated by the computer software distribution, which summarises HTML documents from GitHub. Markdown is often used to format README files and the characteristic between different markdown texts is a key to accomplish this project.

The purpose of this project will be involving the development of the web application that randomly selects several GitHub projects and to analyse different kinds of the content of README files on the GitHub. Specifically, the users can extract section titles and links from the corresponding README files and present the information in the aggregated form. The amount of the quantity and the percentage are the information presented within the aggregated form of the summary. Figure 1 had shown the overall process of the corresponding projects.

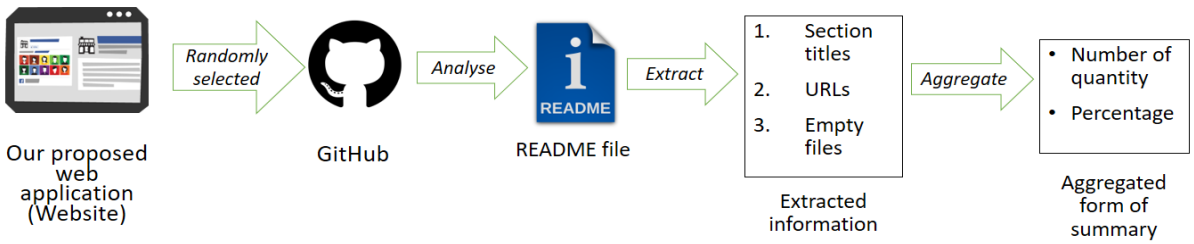
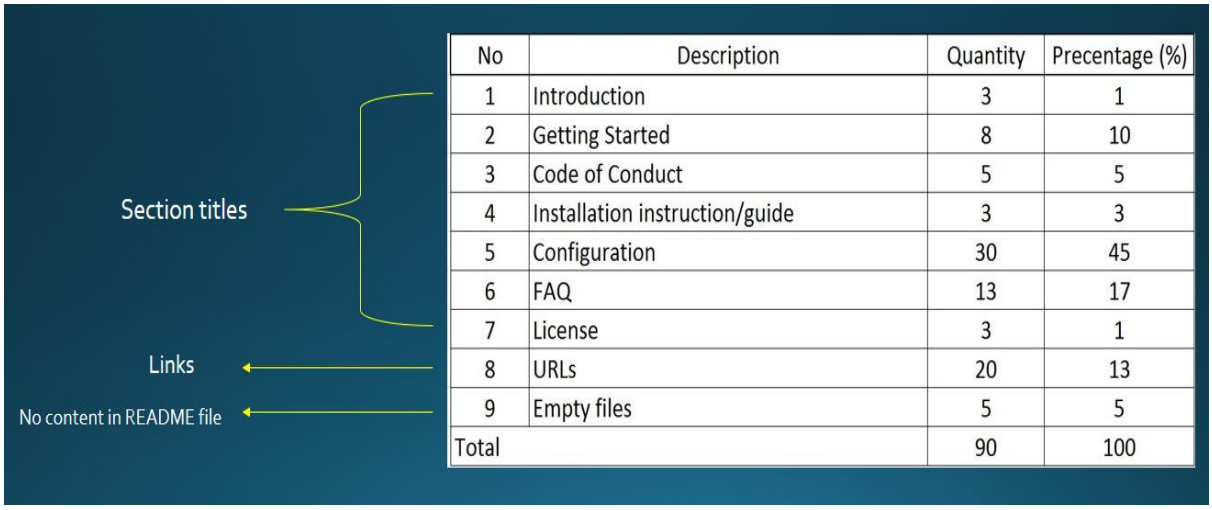


Figure 1: Purpose of the project

The important of this project is to understand the structure of the README file across many GitHub projects. It may also produce a clear-cut overall content of the README files for the users to look after it. If possible, the highest data analysed based on the aggregated form of the information will become README file template on every GitHub project in the future. On the other hand, one of the advantages is it can be a very good reference for those users who intended to write a README file at the first time. With the aggregated form of the summary on the web application, it can act as the template as well as the new users will get to know what is their expectation of creating a new README file. Furthermore, it is superb for the statistical researchers or any other researchers who like to further investigate the user behaviour on the GitHub. They can use this to analyse and compare the information proportion, and the content list made by the application is very clear and easy to see. Therefore, with our web application, it can be a useful tool for the researchers to speed up their requirement gathering as well as the implementations. Most crucially, the results will provide insights into how software developers document their projects and how documentations can be improved.



No	Description	Quantity	Precentage (%)
1	Introduction	3	1
2	Getting Started	8	10
3	Code of Conduct	5	5
4	Installation instruction/guide	3	3
5	Configuration	30	45
6	FAQ	13	17
7	License	3	1
8	URLs	20	13
9	Empty files	5	5
Total		90	100

Figure 2: Expected Output after finished analysing the README files

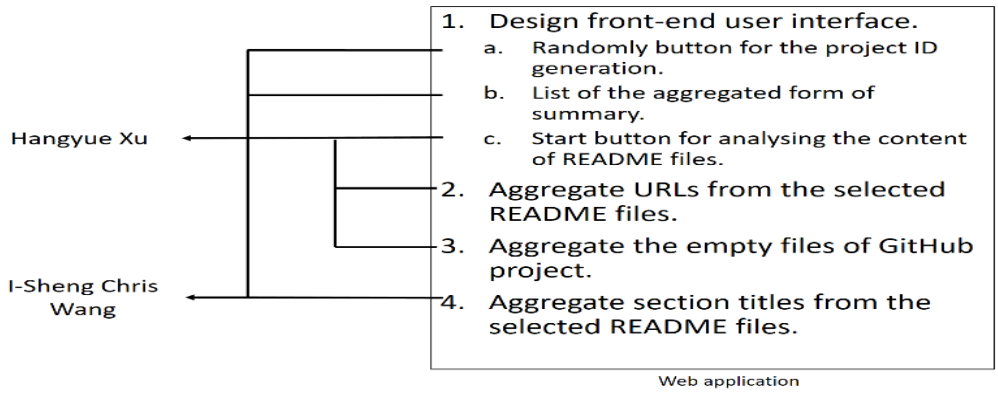
Based on the expected output showed in figure 2, if the user puts a number of 10 on the random number text field, and click the start button to start analysing the content of README files. The application will then aggregate can calculate the information which including the section titles, URLs, and empty files after finished analysing. The number of quantities and the percentage will be presented by the aggregated form of the summary.

If time allowed, more functions will be added that including aggregating the words, characters or even the images of the corresponding README files. On the whole, the goal of this project is developing a web application and systematically analyse the content of GitHub README files. All the features that we mentioned in the earlier stage will be accomplished within this semester.

Part 2: Draft Plan

Team Organisation

Roles & Responsibilities



In this project, Hangyue Xu will be developing the features of aggregating URLs and the empty files from the selected README files on the GitHub. She will be also developing some front-end design such as start button of getting started for analysing the content of README file. Besides that, she may

also responsible of every meeting agenda and meeting minutes after meeting with the supervisors. In the meantime, I-Sheng Chris Wang will be developing the features of aggregating every section titles of README file's content from the randomly selected GitHub project. Besides that, he will design the overall front-end user interface of the web application. He may oversee the entire performance of the project and ensure the milestones are met and delivered on time.

Communication plan

GitHub enterprise is the source code and documentation management platform to grab each other on-going tasks for viewing and amendment. The project URL only can be viewed within the group mates and the corresponding supervisors which was granted access by the University of Adelaide. In the meantime, WeChat is the main communication tool among the group to interact and discuss with group mate. Email is the tool to interact with the supervisors in any kind of discussion and circumstances.

Schedule and Milestones

For the first iteration milestone that will be delivered in week 6, we will be providing the overall user interface of the web application. Anyhow, only partial buttons will have the action performs. One of the text fields will prompt to use as the random button select the number of the project to be analysed. Based on the randomly selected project, the corresponding function should be able to fetch GitHub project ID and download the README files on the local PC. We will restrict up to maximum of 50 projects to be analysed in our first iteration milestone prototyping. However, the first feature we will be delivered is extracting the clickable URLs from the selected GitHub project and calculating how many URLs in total within the README files. The following Gantt chart had shown the project process as well as the first iteration milestones prototyping from week four to week six.

ID	Task Name	Duration	Start	Finish	Predecessors
1	**Requirement Gathering**	**12 days**	**Mon 6/03/17**	**Tue 21/03/17**	
2	Clarified requirement with supervisor	1 day	Mon 6/03/17	Mon 6/03/17	
3	Identify required software tools	5 days	Tue 7/03/17	Mon 13/03/17	2
4	Feasibility studies	11 days	Tue 7/03/17	Tue 21/03/17	2
5	**Development**	**8 days**	**Thu 23/03/17**	**Sat 1/04/17**	**1**
6	Generate random number	1 day	Fri 24/03/17	Fri 24/03/17	
7	Fetch [html_url] from the web	1 day	Mon 27/03/17	Mon 27/03/17	6
8	Look for the first project number of URL	1 day	Tue 28/03/17	Tue 28/03/17	7
9	Fetch the entire web page of the project	1 day	Wed 29/03/17	Wed 29/03/17	8
10	Search and fetch the link of the README file	1 day	Thu 30/03/17	Thu 30/03/17	9
11	Download the selected URL within the project selected	3 days	Tue 28/03/17	Thu 30/03/17	
12	Analysing the content of README files - Aggregate URLs	3 days	Tue 28/03/17	Thu 30/03/17	
13	UI design for randomly button	2 days	Mon 27/03/17	Tue 28/03/17	6
14	UI design for listing	2 days	Fri 31/03/17	Sat 1/04/17	12
15	**Integration**	**2 days**	**Fri 31/03/17**	**Mon 3/04/17**	**5**
16	Integrating the features with user interface	2 days	Fri 31/03/17	Sat 1/04/17	
17	Submission of the first milestone prototyping	1 day	Mon 3/04/17	Mon 3/04/17	

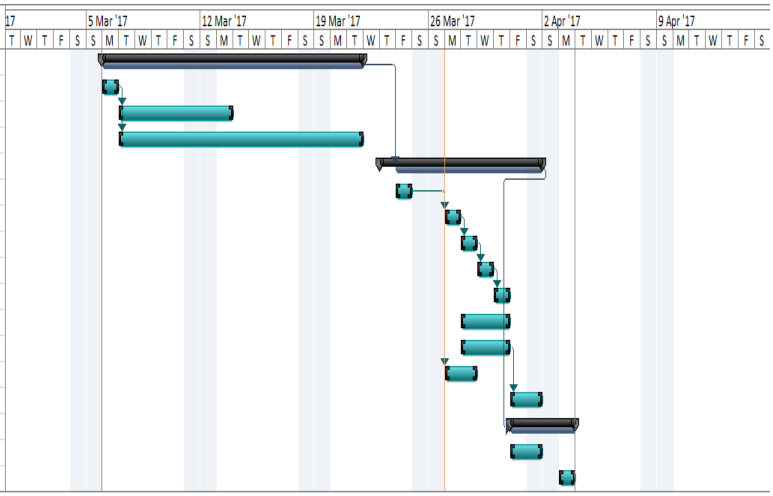


Figure 3: Project process from week 4 to week 6

Baseline Sehedules in detail

No	Week	Action to be done	Assignation	Duration	Timeline/Due date
1	3 - 4	Requirement gathering/Feasibility studies	All	20 hours	06/03/2017 – 12/03/2017
2		Identify the required software tools	All	8 hours	13/03/2017 – 22/03/2017
3		Proposing a front-end interface	Hangyue Xu	2 hours	24/03/2017 – 25/03/2017
4		Generate random numbers	I-Sheng Chris Wang	2 hours	24/03/2017
5		Fetch [html_url] from web	I-Sheng Chris Wang	2 hours	24/03/2017
6		Find the first project number of URL	Hangyue Xu	5 hours	25/03/2017
7		Fetch the entire web page of the project	Hangyue Xu	5 hours	25/03/2017
8		Search and fetch the link of the README file	Hangyue Xu	2 hours	26/03/2017
		Develop business case and draft plan	All	15 hours	25/03/2017 – 27/03/2017
9	5	Business Case and draft plan submission	All	-	28th March 2017
10		Download the selected URL within the project selected	I-Sheng Chris Wang	10 hours	28/03/2017 – 30/03/2017
11		Analyse the URLs from the corresponding README	I-Sheng Chris Wang	3 hours	28/03/2017 – 29/03/2017
12		Extract URLs from the downloaded README files	Hangyue Xu	4 hours	29/03/2017
13		Calculate and present it in the aggregated form	Hangyue Xu	4 hours	30/03/2017
14		Parse the information to front-end	I-Sheng Chris Wang	4.5 hours	31/03/2017
15		UI design: Randomly button for the project ID generation	I-Sheng Chris Wang	2 hours	01/04/2017
16		UI design: List of the aggregated form of summary	I-Sheng Chris Wang	2 hours	01/04/2017
17		Integration		3 hours	02/04/2017
18	6	First milestones prototyping submission	All	-	3rd April 2017
Total hours spend				93.5 hours	