Analysing the content of README file on GitHub

Hangyue Xu Team08

Executive Summary

This paper will provide the description of the project about analysing the content of README file on GitHub, and clients can use this paper to know the analysing process of this application. Besides that, it will also introduce the specific outcomes, responsibility and schedule. Finally, the first milestone is to create an interface and analyse 50 files at same time.

Part 1: Business Case

The requirement of this project is analysing the content of README file on GitHub. Basically, GitHub is a code storage platform, individuals can use it to store code and share information. README file is a distribution file, which is more like a directory of GitHub. To be more specific, README file summaries HTML files from GitHub and then use Markdown language to write their own documents. The Markdown language is significant for this project since the grammar can distinguish different content. For example, the hash symbol can differentiate titles and normal words. The diagram below shows the process of operation. Firstly, individuals can enter a number from 0 to 70 million on an interface of the web application, and then the application can randomly choose projects from GitHub with this amount. After analysing, section titles, links and empty files can be extracted and listed. But a point needs to be mentioned, due to README files may contain too many links, therefore, these URLs can be categorised by different websites, such as google, government and education. Finally, a proportion table can show the percentage of different content, for example, code is 20% and links take up 10% of the whole content.

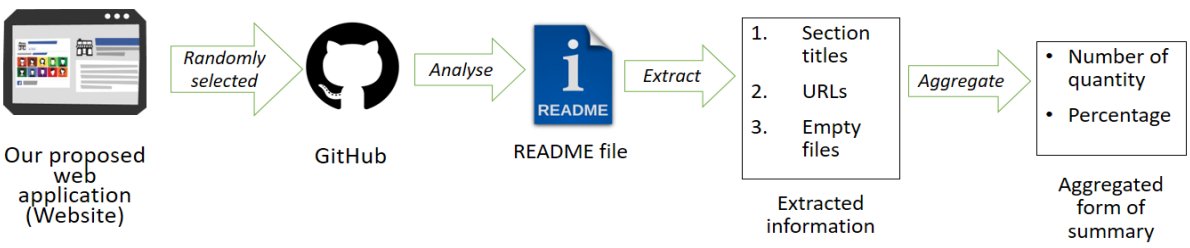

Fig1. Process diagram

For the importance of this project, there is no analogous application at the current stage, therefore, due to this application can analyse plenty of information from random projects, it may become a good example for other applications in the future. Besides that, if researchers want to research about GitHub, this application is a good analysis tool for them to obtain data and references. In addition, since GitHub have many convenient functions, more and more people choose to use it to record codes and share information. Therefore, new users can learn more usages that are not introduced on GitHub. Furthermore, the content with the highest percentage provides a direction, which means more and more individuals like to use this function. According to this, developers will know where they can improve and how they can do.

The final goal of this project is to realise above functions in this semester. The application can successfully analyse the content of README files on GitHub, and then generate a list of titles and links and proportion table. If time allowed, it may have more functions, such as calculate a number of images and characteristics.

Part 2: Draft Plan

Baseline Schedule

The following Gantt chart had shown the project process from week four through week six. The first milestones prototyping will be presented in 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	

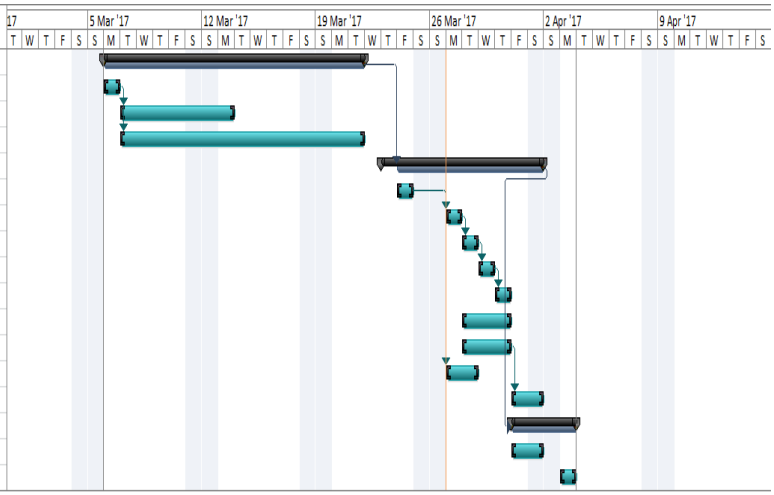

Fig2. Project process from week4 to week6

All background of the project has done before the third week, therefore, the following information will show the specific schedule of the first milestone.

1. First Iteration Milestones Prototyping

   After pitch presentation, each student has to prepare a business case as well as a draft plan of the first milestones prototyping, this writing process may take up 15 hours from 25/03 to 27/03 for every team member.

   Due date: Week 5 (28th March 2017)

   The first iteration milestones prototyping will be conveyed as shown as following:

   - Hangyue Xu used 19 hours to generate a general framework for the user interface, according to the project number to find the webpage, and then fetch it with README files (24/03-26/03). Hangyue Xu also needs 4 hours to take the responsibility to extract links from downloaded README files (29/03). Besides that, calculate these links may take up to 4 hours on March 30th.
   - I-Sheng Chris Wang used 8 hours to design random number function and obtain html-url from the webpage (24/03). In addition, he will use 13 hours to download and analyse links from README files (28/03-30/03). I-Sheng Chris Wang also needs 7.5 hours to design user interface (31/03-01/04). To be more specific, after 4.5 hours (31/03) analysing the information from interface, he will use 4 hours to realise the button click function and 4 hours to generate the summary form (01/04).

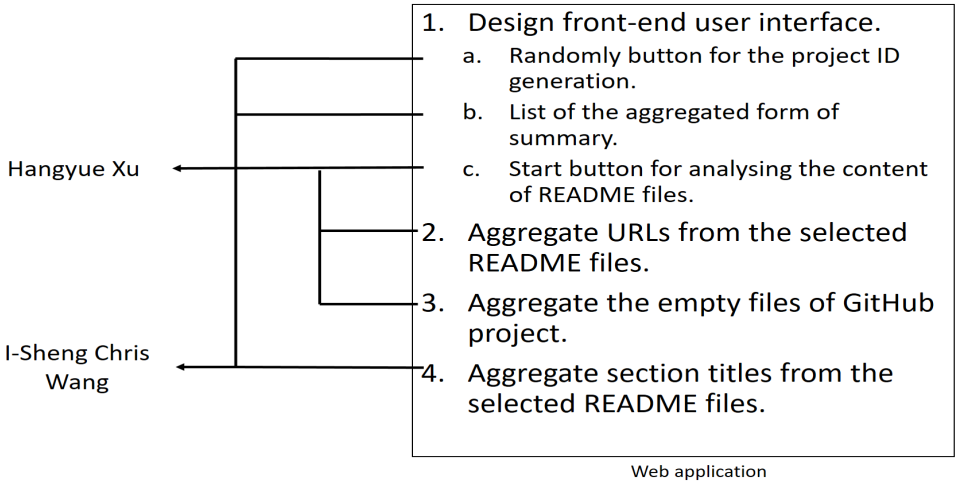

Fig 3. Project process from week4 to week6

In week six, the first milestone may submit on the Monday in the tutorial class.

Milestones

We will propose the overall user interface of the web application and one of the functions stated in the requirement given by the supervisor. The buttons of the random number generation will be used to generate a random project ID number and link to the GitHub project API with the corresponding project ID selected. Yet, for the first iteration milestone, only maximum up to 50 projects can be analysed at the same time. Furthermore, after successfully analysed the content of the README files, another first milestone will be delivered is extracting the URLs from the downloaded README files and present it into the aggregated form of the summary.

Team Organisation

Roles & Responsibilities

Hangyue Xu will be handling the functions of aggregate the empty files and the URLs from the selected README file as well as the start button to get started of analysing the content of README file. She will do every meeting agenda and meeting minutes. On the other hand, I-Sheng Chris Wang will be in charge of other related front-end user interface of the web application and aggregate the section titles from the selected README files. He will make sure the all the milestones delivering on time and manage well of the overall project performance. Overall, all team members are designers and each one will be the tester for others' code part.

Communication plan

GitHub enterprise will be our main communication tool for documentation and program transferring for the entire semester. This is a platform to allow us to view or edit each other work and keep up-to-date from time to time. The University of Adelaide can only grant the intended audiences access to our project. On the whole, WeChat will be our informal interaction tool to have our discussion, while email is considered to be used as a formal tool to communicate with the clients and supervisors.