Milestone 1 Progress Evaluation

Student Performance Analysis

Shreyas Ugemuge sugemuge2014@my.fit.edu

Yaqeen AlKathiri yalkathiri2013@my.fit.edu

Mohammed AlHabsi malhabsi2013@my.fit.edu

Shiru Hou shou2015@my.fit.edu

Faculty Sponsor: Dr. Phillip Chan pkc@cs.fit.edu

February 21, 2017

Contents

1 Progress of current Milestone

Task	Completion %	Mohammed	Shreyas	Yaqeen	Shiru	To do
1. Select Collaboration Tools	100	20	40	20	20	n/a
2. Investigate packages, languages and tools	100	25	25	25	25	n/a
3. Familiarize with the required technologies	100	25	25	25	25	n/a
4. Requirement Document						n/a
5. Design Document						n/a
6. Test Plan						n/a
7. Progress Evaluation						n/a
8. GIT test push	100	25	25	25	25	n/a
9. Python hello world	100	25	25	25	25	n/a

1.1 Discussion of each task

1.1.1 Select Collaboration Tools

This task included of setting up the following collaboration tools:

1. GIT repository
2. Webpage
3. Slack team communication
4. Slack shared calendar

5. Slack shared and personal to do list
6. Google slides
7. Google docs

1.1.2 Investigate packages, languages and tools

The following data mining tools were explored and pros and cons listed:

Technology	Notes	Pros	Cons
Advanced Miner			
Ghost Miner			
GNOME			
SPAD			
ESTARD			
Statistica			
Tanagra	see 4.1	see 4.1	see 4.1
Rapid Miner	see 4.1	see 4.1	see 4.1

1.1.3 Familiarize with required technologies

This task included of familiarizing with GIT, google docs, Python and 8 of the assigned tools. Hello world like programs were pushed on the repository for python and git.

1.1.4 Requirements, Progress Evaluation and Design Documents, Test Plan

The documents were created according to the guidelines provided on http://cs.fit.edu/ pkc/-classes/seniorProjects/

1.1.5 GIT test push and Python Hello World

The corresponding files were pushed on to the repository.

1.2 Discussion of team member contribution

1.2.1 Shreyas

Created GIT repository and project website and set up Slack team communication. Researched and provided brief report on AdvancedMiner and GhostMiner. Authored to the progress evaluation document and contributed to Requirements, Design and Test Plan document and presentation.

1.2.2 Shiru

Set up calendar and todo list plugin for slack. Researched and provided report on GNOME and SPAD. Contributed to all 4 documents and presentation. Explored new technologies including GIT.

1.2.3 Yaqeen

Provided detailed report on Tanagra, SPAD and Rapid Miner. Set up google slides. Contributed to all 4 documents and presentation.

1.2.4 Mohammed

Provided report on ESTARD and statistica. Set up google docs. Contributed to all 4 documents and presentation.

2 Plan for the next milestone

2.1 Discussion of each task

3 Sponsor feedback on each task for current milestone

4 Appendices

4.1 Yaqeen's tool research report

Finding ways to extract the information that we need from the data is one of the important steps toward completing the goal of this project. The two tools that I have been exploring are Tanagra and SPAD.

SPADis a mature data mining suite that provides powerful exploratory analyses and data mining tools, including PCA, MCA, clustering, interactive decision trees, discriminant analyses, text mining and more, all via user-friendly GUI. SPAD data mining approaches are Classification Discovery, Cluster Discovery, Regression Discovery, Association Discovery, Text Mining, Outlier Discovery, Data Visualization, Discovery Visualization, Social Network Analysis. Although SPAD is a powerful data mining software, it is a commercial licensed software that comes in French. These two points in SPAD will delay the process of extracting information. Also, the only website that we can download SPAD from is www.coheris.com, which is in French, too. In my opinion, SPAD is not the data mining software that we want to be using in this project.

Tanagra, data mining suite, is a free software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. Tanagra is the successor of SIPINA which implements various supervised learning algorithms, especially an interactive and visual construction of decision trees. TANAGRA is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and nonparametric statistics, association rule, feature selection and construction algorithms. Moreover, its an "open source project" as every researcher can access to the source code, and add his/her own algorithms. Also, Tanagras main purpose is to give researchers and students an easy-to-use data mining software, conforming to the present norms of the software development in this domain (especially in the design of its GUI and the way to use it), and allowing to analyze either real or synthetic data. In the other hand, TANAGRA does not include what makes all the strength of the commercial software in this domain; a wide set of data sources, direct access to datawarehouses and databases, data cleansing, interactive utilization ..etc. The last thing, Tanagra originally comes in French as well, but we can find it in English, too. From this concept, I think using Tanagra is better than using SPAD, but I wouldnt recommend it.

While exploring SPAD and Tanagra, I researched rapid miner. Weka, a software that I think is helpful in data mining, which is originally in JAVA. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from the researcher own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Weka is open source software issued under the GNU (General Public License). Also, one of the helpful things about Weka is that you can apply it to big data.

In general, all the data mining softwares above are useful, but in order to choose a software that is most helpful we have to consider the main goal of the project. In my opining, from all the three data mining softwares, Weka is what I think most helpful. SPAD is powerful commercial tool that is not free nor is in English, which makes it hard to work with. Also, Tanagra doesnt include all what a commercial software would offer which makes it less useful in our case. In conclusion, Weka what I see best from what I gathered from my research.

5 Sponsor Evaluation

Sponsor: Please detach this page and return to Dr. Shoaff

Score (0-10) for each member: circle a score (or circle two adjacent scores for .25 or write down a real number between 0 and 10)

Shreyas Ugemuge	0	1	2	3	4	5	6	6.5	7	7.5	8	8.5	9	9.5	10
Yaqeen AlKathiri	0	1	2	3	4	5	6	6.5	7	7.5	8	8.5	9	9.5	10
Mohammed AlHabsi	0	1	2	3	4	5	6	6.5	7	7.5	8	8.5	9	9.5	10
Shiru Hou	0	1	2	3	4	5	6	6.5	7	7.5	8	8.5	9	9.5	10

Faculty Sponsor

_____

Signature

_____

Date