DESIGN AND IMPLEMENTATION OF WEB-BASED SURVEY ANALYSIS AND MANAGEMENT SYSTEM

A Project

Presented to

The Faculty of the Graduate School

University of Missouri-Columbia

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

By

Gaofei Zhao

Dr. Yi Shang, Advisor

APRIL 2018

ACKNOWLEDGEMENTS

I would first like to thank my advisor Dr. Yi Shang. He consistently allowed this paper to be my own work but steered me in the right the direction whenever he thought I needed it. His broad scientific knowledge and research experience are my greatest motivation. It is my honor to learn from him.

I would also like to thank Dr. Dong Xu and Dr. Ye Duan for reviewing my report and providing so many insightful and valuable comments and suggestions on this report. I appreciate them taking precious time out of their busy schedule to serve on my report committee.

I would also like to thank everyone from the Computer Science lab, I would also like to show appreciation to Jodie Lenser and Shirley Holdmeier for guiding me through my application process to Mizzou as well as advising me on my courses selection.

Finally, I would like to thank my family and friends for their constant support and encouragement during all these years. Without their understanding, it would be impossible for me to go through all the difficult time.

TABLE OF CONTENTS

[1 Introduction 9](#_Toc510833822)

[1.1 Data Management Applications 9](#_Toc510833823)

[1.2 Data Analysis Applications 10](#_Toc510833824)

[1.3 Problem Description 10](#_Toc510833825)

[1.4 Proposed Solution 12](#_Toc510833826)

[2 Related Work 17](#_Toc510833827)

[3 System Design 19](#_Toc510833828)

[3.1 Client Side Module 19](#_Toc510833829)

[3.1.1 Survey Module 20](#_Toc510833830)

[3.1.2 File Management Module 21](#_Toc510833831)

[3.1.3 Data Demonstration Module 22](#_Toc510833832)

[3.1.4 Rule Management Module 23](#_Toc510833833)

[3.1.5 User Management Module 23](#_Toc510833834)

[3.1.6 Data Selection Module 24](#_Toc510833835)

[3.2 Server Side Module 24](#_Toc510833836)

[3.2.1 Rule-based Notification Module 24](#_Toc510833837)

[3.2.2 Data Preprocessing Module 25](#_Toc510833838)

[3.2.3 Unsupervised Learning Module 25](#_Toc510833839)

[3.3 Technology Stack 26](#_Toc510833840)

[3.3.1 Client Side 26](#_Toc510833841)

[3.3.2 Server Side 27](#_Toc510833842)

[3.3.3 Other utilities 27](#_Toc510833843)

[4 System Implementation 28](#_Toc510833844)

[4.1 Client Side 28](#_Toc510833845)

[4.1.1 Survey Module 28](#_Toc510833846)

[4.1.2 File Management Module 32](#_Toc510833847)

[4.1.3 Data Demonstration Module 34](#_Toc510833848)

[4.1.4 Rule Management Module 36](#_Toc510833849)

[4.1.5 User Management Module 37](#_Toc510833850)

[4.1.6 Data Selection Module 40](#_Toc510833851)

[4.2 Server Side 41](#_Toc510833852)

[4.2.1 Rule-based Notification Module 42](#_Toc510833853)

[4.2.2 Data Preprocessing Module 44](#_Toc510833854)

[4.2.3 Unsupervised Learning Module 45](#_Toc510833855)

[5 Future Work 56](#_Toc510833856)

[6 Summary 57](#_Toc510833857)

LIST OF FIGURES

[Figure 1.1 Survey Example 12](#_Toc510833667)

[Figure 1.2 Google Form Example 13](#_Toc510833668)

[Figure 1.3 System Overview 14](#_Toc510833669)

[Figure 1.4 Analysis Languages Ranking 2015-2017 15](#_Toc510833670)

[Figure 3.1 Overall System Architecture 20](#_Toc510833671)

[Figure 3.2 Survey Implementation Activity Diagram 22](#_Toc510833672)

[Figure 4.1 Client Side Architecture 29](#_Toc510833673)

[Figure 4.2 User Information Data Structure 30](#_Toc510833674)

[Figure 4.3 User Answers Data Structure 31](#_Toc510833675)

[Figure 4.4 Implementation Page Select Survey 32](#_Toc510833676)

[Figure 4.5 (a) single choice questions (b) scale type questions (3) yes or no questions,(4) Text questions (5) time interval questions (6) multiple-choice questions 33](#_Toc510833677)

[Figure 4.6 CSV File Upload 34](#_Toc510833678)

[Figure 4.7 Index/String Switch 35](#_Toc510833679)

[Figure 4.8 CSV File Download 35](#_Toc510833680)

[Figure 4.9 Data Visualization Sample 36](#_Toc510833681)

[Figure 4.10 Rule Management Panel 37](#_Toc510833682)

[Figure 4.11 Rule Panel Question For Select 37](#_Toc510833683)

[Figure 4.12 Select the rule 38](#_Toc510833684)

[Figure 4.13 Login and Register 39](#_Toc510833685)

[Figure 4.14 Oath Protocol 40](#_Toc510833686)

[Figure 4.15 Data Selection 41](#_Toc510833687)

[Figure 4.16 Structure Of The Server-Side 43](#_Toc510833688)

[Figure 4.17 Notification Module Pipeline 43](#_Toc510833689)

[Figure 4.18 Event Structure 44](#_Toc510833690)

[Figure 4.19: A comparison of the clustering algorithms in scikit-learn 46](#_Toc510833691)

[Figure 4.20 K-means Result By Iteration 49](#_Toc510833692)

[Figure 4.21: Clustering examples, with clusters indicated by different symbols (and colors where available). 52](#_Toc510833693)

[Figure 4.22: The example of K-means and Spectral clustering 54](#_Toc510833694)

[Figure 4.23: Using Ward’s method to form a hierarchical clustering 55](#_Toc510833695)

LIST OF TABLES

[Table 1: A comparison of the clustering algorithms 46](#_Toc510831300)

ABSTRACT

In recent years, data-driven research in the psychology field is moving from traditional paper-based questionnaires and basic statistical models to smart mobile device based ambulatory survey and sensor data collection and data analytics using advanced machine learning and data mining techniques. User-friendly Web-based user interface for real-time data monitoring, data visualization, and data mining becomes very important for efficient and effective data collection and analysis. In this project,

a Web-based Survey Analysis and Management System is developed for researchers that integrates real-time data monitoring, data visualization, and application of some commonly used data-mining and machine learning techniques on the collected data. In particular, the system consists of two parts: 1) a Web dashboard developed using AngularJS, which collects, organizes and displays data for researchers to manage user accounts and data, set up data monitoring rules, and visualize data and data processing result; 2) a server developed using Flask and NodeJS, which monitors users’ responses, processes data, and applies unsupervised machine learning methods, including K-Means and Spectral clustering, etc., to survey data. This system has been applied to the data collected in an alcohol drinking study to demonstrate its usability.

1. Introduction

A self-report is any test, measure, or [survey](https://www.verywellmind.com/what-is-a-survey-2795787) that relies on the individual's own report of their symptoms, behaviors, beliefs, or attitudes.[1] Self-reports are commonly used in psychological studies for the reason that a person’s report contains substantial personal and diagnostic information, which is valuable for researchers or clinicians.

The Internet is a powerful and increasingly commonplace platform that can be used to conduct survey research. Instead of using traditional paper-based questionnaires and basic statistic models, this report introduces a web-based application for researchers to collect substantial amounts of survey data, which provides a convenient and efficient tool for real-time data monitoring, collecting and applying machine learning algorithms to analyzing.

* 1. Data Management Applications

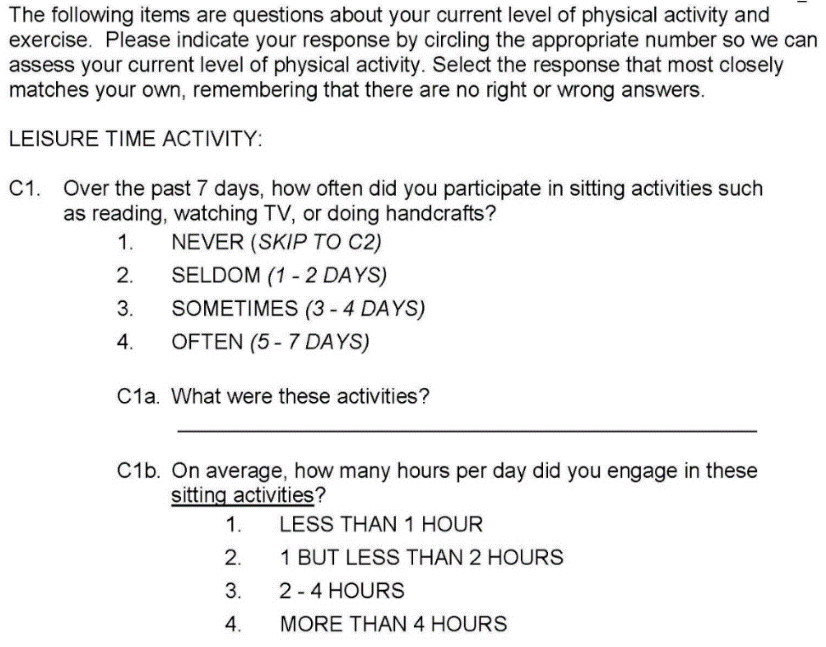
Collecting data is one of the most significant thing in the whole process of research for it’s importance of helping researchers understanding their particular subject. So the application should provide a convenient and effective environment for both researchers and users to conduct or complete the surveys. For most of the time, raw data is meaningless before definition or modeling, which accelerate the demand for the data management system. This application also provides data visualization such as data table and charts, which make it clear for users to review and get the profound grasp into the data. In some particular cases, the administrators will be notified by email immediately when a user's answers match the preset rules, eg. severe depression.

* 1. Data Analysis Applications

Unsupervised learning method is a precisely and scientific way to analyze data. Users can change models and parameters when using machine learning clustering methods to analyze the data. It provides an automatic tool for researchers to utilize machine learning algorithms on data analyzing and study.

* 1. Problem Description

The questionnaire survey is an effective method of assessment.[2][3] The traditional questionnaire survey is based on the distribution of a paper version of the questionnaire to different testers. After the results have been collected, statistical analysis is performed. However, this method is inefficiently, and it is very likely to cause information leakage. Even for current web-based solution, eg. google form, is hard to analyze the data online and have some customized functions, like rule-based email notification module.



[2]Figure 1.1 Survey Example

In psychology, the response represents patients’ psychological state. Sometimes the patient's response is pretty critical, especially when the patients show an extreme mental state, which should pay a lot of attention at once. The doctor will be notified immediately and respond based on the answer. This feature is crucial for real-time data surveys, for which the administrators can quickly respond to emergencies and prevent the problems become worse.

Also, when the questionnaire is completed, the researcher would like to make a preliminary statistical analysis based on the results of the survey, and the results obtained to serve as an overview of the answer.[4] Machine Learning is a powerful tool that intersects statistical, probabilistic, computer science and algorithmic aspects for learning iteratively from data and finding hidden insights.

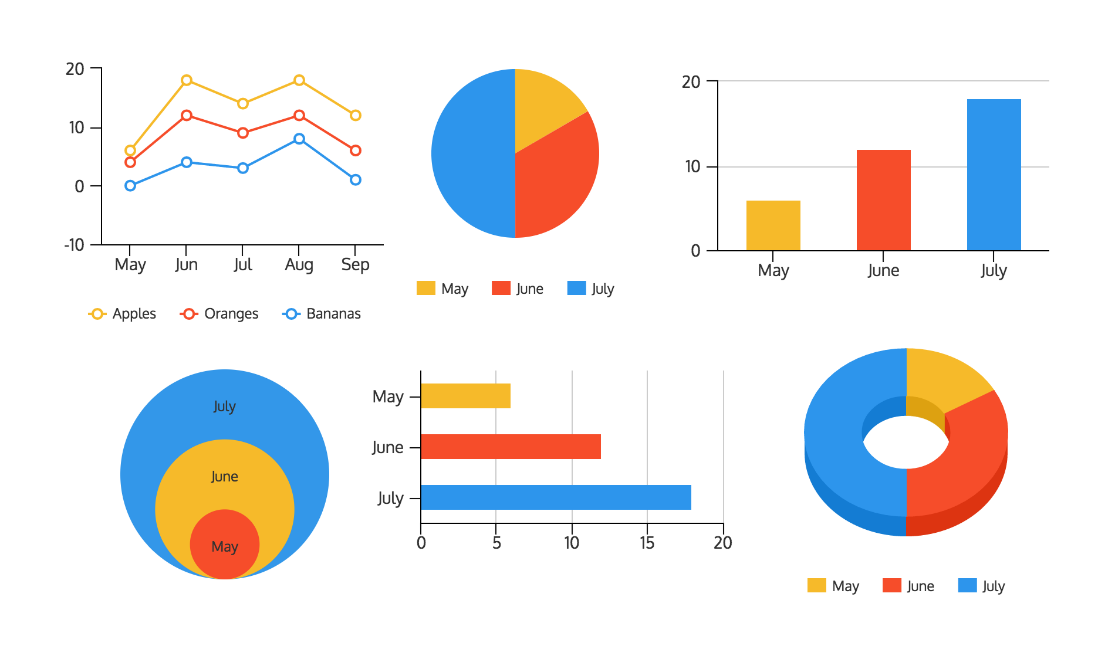


Figure 1.2 Google Form Example

* 1. Proposed Solution

This system is a web-based dashboard which hosts on AWS servers, using the AngularJS as the front-end framework, and Flask as back-end. A standalone background NodeJS server is used to send emails, monitor the corresponding rules in the database and make email sending requests. The firebase is using for data storage.

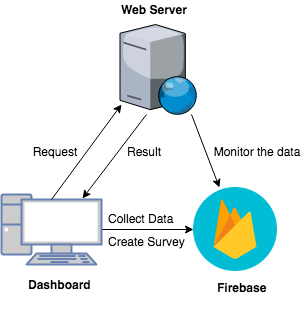


Figure 1.3 System Overview

The frontend is implement by MVC framework and AngularJS, which decrease the difficulties of code changing and reusing for the entire Web Application. Front-end was developed by combining HTML5, Bootstrap, Materialize with CSS3, and JavaScript with jQuery libraries, which make the system powerful and adaptable.

Firebase is a real-time database launched by Google that providing the background management. Firebase not only makes the fetching and management of data more convenient, but also provides high scalability, since the system only needs to read data through the private API.

There are many options for server-side framework selection. Python is the most commonly used languages for data processing. In modern Python frameworks, Flask is a lightweight method that can fulfill our requirements, and easily to complete data processing tasks.[5]

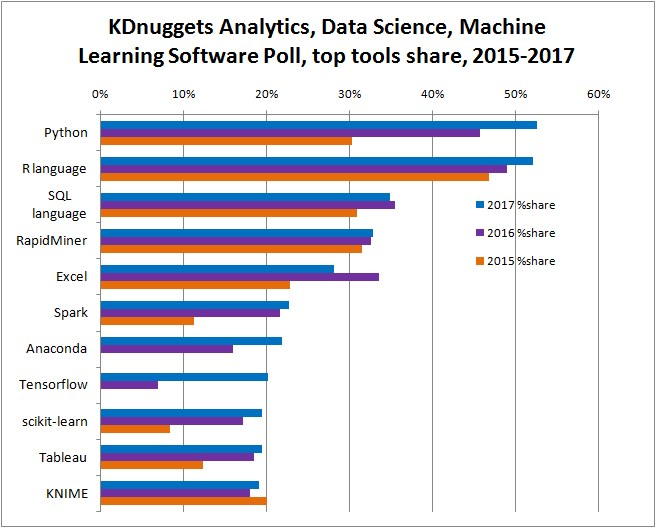


Figure 1.4 Analysis Languages Ranking 2015-2017

The following steps can summarize the entire system operation process:

1. The user completes the user survey at the front end. The firebase reserved API store the data in the database. In the process of completing the questionnaire, the user needs to answer the questions according to the reserved answer specification. The front page will verify the correctness of the input to ensure that the data entered is correct.
2. When the administrator wants to view the user's answer, the detailed page will display the number of users' responses and the answer distribution of each question. In this page, the administrators can customize specific rules (monitored by a Node.JS process) to receive notification emails immediately.
3. At the same page, data filtering is provided before submitting. The submitting will transform data to Flask server through Ajax, with the processed results sending back to the front-end, which refresh the data table and generate clustering result graph.
4. Users are allowed to upload the data and get analysis result.

The report is organized as follows: In Chapter 2, related works are discussed in data management and analysis and their strengths and limitations. The design theory of the web-based analysis and management system are discussed in Chapter 3. Chapter 4 presents the implementation of the system. The future work is discussed in Chapter 5. The conclusion summarized in Chapter 6.

1. Related Work

[6]There are many web-based questionnaire collection applications, such as Google form. Google form is a generic form collection application, with Google Forms, the user can create and analyze surveys right on their mobile or web browser—no special software required. Get instant results as data come in. Summarize survey results at a glance with charts and graphs. And also have some specific application, like Creel Surveys Data Management Application. It consists of two main components: 1) Fish Investigator Mobile Application Hybrid Application Development Use PhoneGap and Apache Cordova, collect survey data on iPad or iPhone and upload them to Amazon Web Services (AWS) MySQL database in EC2 instance. 2) Fish Investigate Web Dashboard adopts LAMP organization And shows biologists and conservationists use the Bootstrap CSS3 framework to view and validate the user-friendly web interface developed above in real time. The Google Form is effective for data collection because Google Form can organize its survey and select different types of problems while the Creel Survey Application cannot design custom questionnaires. Their common disadvantage is that none of them use machine-learned data processing modules and they cannot notify the user based on the user's answer to inform the administrator that the content they are interested in has appeared in the user's response.

Another similar application is Mood Toolkit Dashboard (MTD), which enables analysis and visualization of survey data. MTD provides a method for surveying schemas that outline common application components that are identified in different studies, different metrics such as participants' compliance, responses to average mood changes, and calculations On the days of the study. Some physiological features were analyzed to show the final documents from each study on two levels, study levels and participants. Use different data sets and provide a layered visualization scheme. MTD is a complete stack (LAMP) development project hosted by RESTful web services consumed on Amazon EC-2 and AngularJS front-end frameworks, providing modular design features. Highcharts.js is used to display rich interactive graphics. This application is a safe and easy-to-acquire research data that will benefit researchers due to their analytical capabilities and speed outside of a new pipeline of psychological research. This study is pretty close to what has been designed and implemented in this report. However, it has lost some important functions, such as notification email can only be notified if the registration is successful and the administrator's authorization is successful. And it’s data can only exist in the database in the case of manual import. New modules need to be rewritten to display new data. It cannot dynamically create questionnaires, and complete surveys then analyze the results.

These works were essential in identifying the existing approaches, and how they can be utilized for the proposed solution. They were used for helping build a compatible and powerful web application, which have been further explored in the following chapters.

1. System Design

The proposed system in this paper divided into Client Side and Server Side, shown in Figure 3.1.

The Client Side can be further divided into Survey Module, File Management Module, Data Demonstration Module, Rule Management Module, User Management Module and Data Selection Module.

The Server Side is including Rule-based Notification Module, Data Preprocessing Module, and Unsupervised Learning Module.

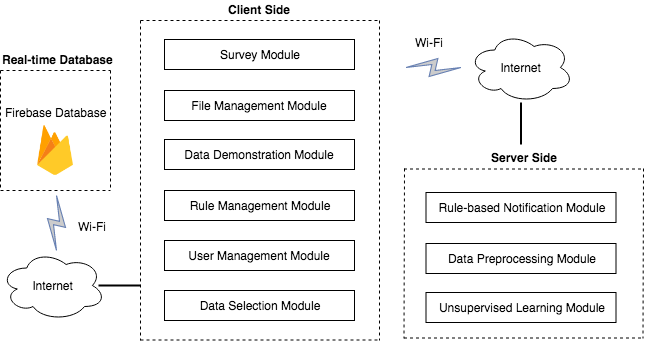


Figure 3.1 Overall System Architecture

* 1. Client Side Module

The client-side application is a user-friendly web dashboard for the user to implement and get an observation of the survey data. This application is composed of six modules: Survey Module, File Management Module, Data Demonstration Module, Rule Management Module, User Management Module and Data Selection Module. The functionality of these modules explains as follows and implementation of these modules discuss in later sections of this report.

* + 1. Survey Module

The survey module is used to collect user's survey answers. All users can fill in questionnaires they can participate in and fill in their responses. After logging in to the home screen, users can select the survey and fill out the survey. All questions will be displayed on the screen one by one. When the user completes each question, it is temporarily saved in the cache until the user has completed all the questions. The answers to all questions will be contained in a JSON object and uploaded to firebase.

In the process of filling out the questions, each question will check if the input is legal. The most basic idea is the standardized survey implementation interface. If it is a single-choice question, it has only one radio option. If it is a multiple-choice question, it can have multiple checkboxes. In this case, we can agree on the answer to the user's input specification. In addition to this, we hope that every response is complete. So when the user completes, it will detect whether the answer has filled completely. If this answer is incomplete, then the answer screen to the next question cannot be reached.

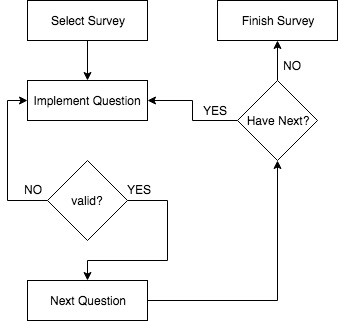


Figure 3.2 Survey Implementation Activity Diagram

* + 1. File Management Module

One of the most useful aspects of the application is allowing administrators and clerks to download records to their local system from the database in a Comma-Separated Value (CSV) file. Besides, import the CSV file from their local system is also allowed, which extends the flexibility of data analysis. Most of the researchers have their own preferred data analysis tools, and this will provide the possibility for them to use their most convenient tools. Also, some researchers want to use this tool to process the data has been collected before. That will need to upload the data to the system and get the analysis result.

The database used by this application is the real-time database firebase, so files should not be stored on the server. Because it is very space-consuming and reusability is not high. The database file has changed a lot when the user wants to download in the other time. The most convenient method is to directly retrieve the latest data from the database and download it when you want to download the file. When the user chooses to download the CSV file, two options will be given. One is to use the index to represent multiple-choice and scale-type answers, and the other is to be represented by text.

Importing files is relatively simple because the import file contains only answers and features. It does not include the detailed information of the questionnaire: for example, details of each problem. Because if you let the user provide the details of the problem, it would be cumbersome to say in the CSV file. So the system treats each column of the CSV file as a feature and treats each row as a single data. The import function only allows the user to use the final data analysis module, not other data observation and monitoring modules. Because the data imported by the user does not contain the detailed information of the user survey, so there is no way to handle it as a survey completed on this system.

* + 1. Data Demonstration Module

For a data collection and analysis application, must have a data demonstration module. So here we use data visualization to show data details. Also, the total number of responses to the questionnaire will be counted and the number of days to answer the survey will be recorded. We have designed a consistent data visualization chart for each question type. All graphs are in real-time mode and designed with an interactive interface to help the researcher zoom in or out the figure, with the choice of hiding or displaying different items on the data graph. Similarly, to demonstrate the results of unsupervised machine learning, the results also need to be visualized. Provide three different charts for user selection.

* + 1. Rule Management Module

To enable researchers to monitor their research process and focus on specific data, the rules management module requires the user to enter the rules that they want to listen to, and save the rules in the database. In the rules management module, it provides users with the ability to manage their own set of rules, add rules and delete rules. Because the rules can be set only after entering the similar survey, and the rules are all corresponding to one survey. The user needs to select the question and select the monitoring rule and the notification text to be notified in the server-side notification module.

* + 1. User Management Module

Before being able to use the dashboard, users will need to register and login. Any user will come across to the login page when they landed. For a first-time user, there will be an option to sign up, where they will be required to fill out some necessary information (name, email address, username, and password).

The details validate through HTML5 tags' required and input types. After completing the registration, their names and other login details will be added to the database. And now they can log in to the application with their information.

Firebase provides the security. Firebase Authentication build around openness and security. It leverages OAuth 2.0 and OpenID Connect, industry standards designed for security, interoperability, and portability. Members of the Firebase Authentication team helped develop these protocols and used their expertise to weave in latest security practices like ID tokens, revocable sessions, and native app anti-spoofing measures to make the app easier to use and avoid many common security problems.

* + 1. Data Selection Module

The role of the data selection module is responsible for selecting data for the final data analysis. For numerical data, the selection is mainly based on data relationships, which is greater than, less than, or equal to the chosen value. If the data satisfies the condition, the data is selected and updated to the selected data set. For text-based data, it only has two types of determination. One of them is that the text contains the chosen keyword. The second type is the text language. The keyword is entirely equal. After making the selection, the result can be sent as a new input to the server for further analysis, and the returned conclusion based on the recalculation of the data. Then encrypt the data and send to the server.

* 1. Server Side Module

The Server Side Module built on a Linux remote server. It composes of three modules: Rule-based Notification Module, Data Preprocessing Module, and Unsupervised Learning Module. This Module supports data monitoring, notification, data preprocessing and analysis on the server.

* + 1. Rule-based Notification Module

Rule-based Notification Module is an independent background process running on the server side. It is used to read each new data that appears in the database under status and adds a new monitoring process based on the results of the analysis to monitor the new data in the survey. Every data that satisfies the rule will be notified to the administrator by mail after being found. The content of the email contains the person who provided the answer, the answer that was answered, and the preset notification content. At the same time, the process will also detect whether the administrator has deleted a monitoring rule. If the rule is removed, the process will stop the listener of the rule.

* + 1. Data Preprocessing Module

The Data Preprocessing Module is performing two tasks sequentially: Data Decryption and Data Preprocessing. The designed flow list is as follows: After the server received the encrypted posted data from the client side, this Module calls a decryption program to decrypt the file. Next, Data Preprocessing Module on the server processes the raw data. For numeric data following this procedure: remove the outliers and Unsupervised Discretization, then normalize the data. For the text data: remove punctuations, remove stopwords and remove digital words. Then all of the words showed in the answers will conduct a dictionary. Then the text input will transform to a 1\*n vector.

* + 1. Unsupervised Learning Module

The primary purpose of the Unsupervised Learning Module is to explore data and tries to identify structures within the data.[7] More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known. Because it is exploratory, it does not make any distinction between dependent and independent variables, like diagnostic clusters. The researcher may devise a diagnostic questionnaire that includes possible symptoms (for example, in psychology, anxiety, depression, etc.). The cluster analysis can then identify groups of patients that have similar symptoms.

We have proposed three clustering methods for the administrator to select, which is K-means cluster, Hierarchical cluster, and Spectral cluster. K-means cluster is a method to cluster large data sets quickly. The researcher defines the number of groups in advance. It is useful to test different models with a different assumed number of clusters. Hierarchical cluster is the most common method. It generates a series of models with cluster solutions from 1 (all cases in one cluster) to n (each case is an individual cluster). The hierarchical clustering also works with variables as opposed to facts; it can cluster variables together in a manner somewhat similar to factor analysis. Besides, hierarchical cluster analysis can handle nominal, ordinal, and scale data;

* 1. Technology Stack

There were multiple tools used for designing and developing. They can be categorized into the front-end and client side, which is using for displaying the interface on the web browsers; the back-end and server side, which the languages are used to analyze the data; and other utilities, which were the software programs and cloud services.

* + 1. Client Side

1. HTML5: Markup Language on which the front-end interface is based for web dashboard.
2. CSS3: Bootstrap and Materialize implements the responsive design on the web dashboard. Offer lots of useful components.
3. AngularJS and jQuery: Used for front-end MVC design. Construct the whole application and send the data to the backend.
   * 1. Server Side
4. Firebase: The real-time database engine for the application
5. Python: Scripting language used server and interaction between the web dashboard and the web server. Handle the data analysis job.
6. Flask: Flask is a micro web framework written in Python. Used to server the application.
7. NodeJS: Node.js is an open-source, cross-platform JavaScript run-time environment that executes JavaScript code server-side.
8. Gunicorn: Python WSGI HTTP Server for UNIX. It's a pre-fork worker model.
9. Amazon Web Services: This is the cloud service provided by Amazon, where the web dashboard and server host on an EC2 instance.
   * 1. Other utilities
10. Atom: A highly capable and customizable text editor
11. FileZilla: A secure and reliable FTP software to transfer files between the local system and the remote server.
12. Github: Collaboration service was used to manage the version of code.
13. System Implementation

This chapter details the implementation of this web-based survey analysis and management system. The Client Side is implemented in JavaScript, HTML, CSS languages with the AngularJS Framework, Materialize and Bootstrap. The Server Side is implemented in Python, JavaScript languages.

* 1. Client Side

This section will include the six modules designed above. Which is Survey Module, File Management Module, Data Demonstration Module, Rule Management Module, User Management Module and Data Selection Module.

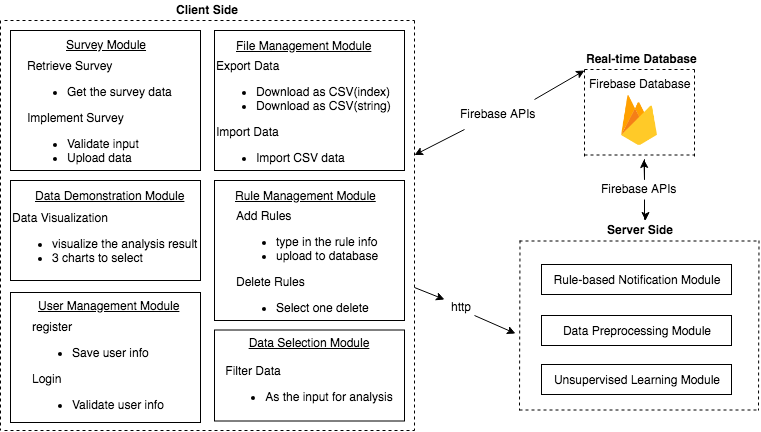


Figure 4.1 Client Side Architecture

* + 1. Survey Module

Survey Module includes data structure design and implementation page.

* + - 1. Data structure

Firebase is a database solution from Google, the real-time database in firebase is a None-SQL database that implements the JSON structure to store data. The data storage structure of this system showing in the figure below.



Figure 4.2 User Information Data Structure

To avoid duplication of data names and data in firebase stored in blueprints. Blueprint generation is based on the time the user created this data, so this ensures that each blueprint is independent and unique. It has the advantage of simplifying the data storage structure and providing independence.

Under the user's Object, there is a personal email address of the user is recorded when the user registered an account, a questionnaire created by the user, and a questionnaire that the user can participate. There is also a public name for this user. The storage structure of the survey also applies encoding. Because it is very likely that users will store new studies together at the same time, the names of the surveys may be duplicated, causing the database to throw exceptions.

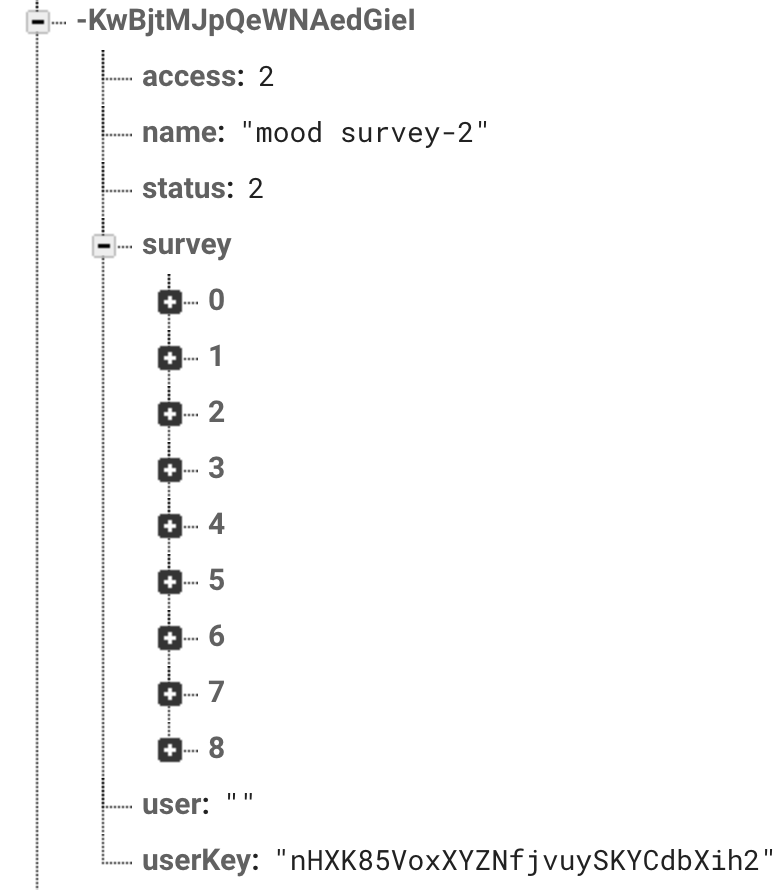


Figure 4.3 User Answers Data Structure

In each survey, the array of questions and the names of the questionnaires were stored. Access and status are used for indicating the access permission of the study and the open/close state of the survey.

* + - 1. Survey Implementation

Select the survey that you want to fill out in the study. The page will jump to the implementation page. The implementation page will start from the first question and will verify when it answers each question. If the verification passes through the next question, the answer will be saved to the database when it completes all the questions in the survey.

When select a survey in Figure 4.4, the survey questions data has been retrieved from the firebase. Then a modal will pop up and ask the user to implement the questions.

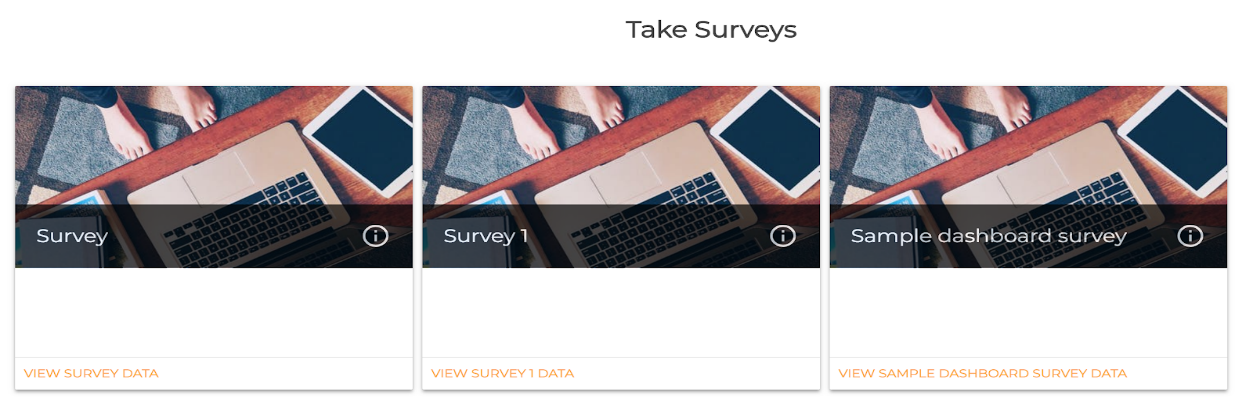
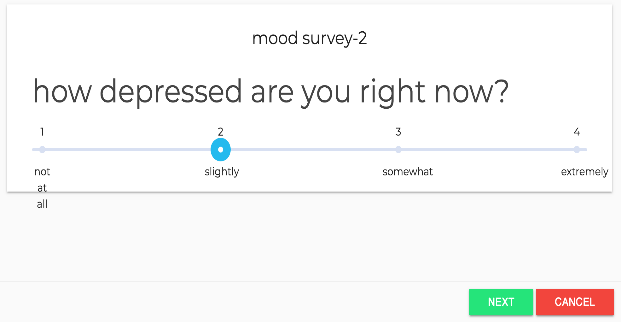
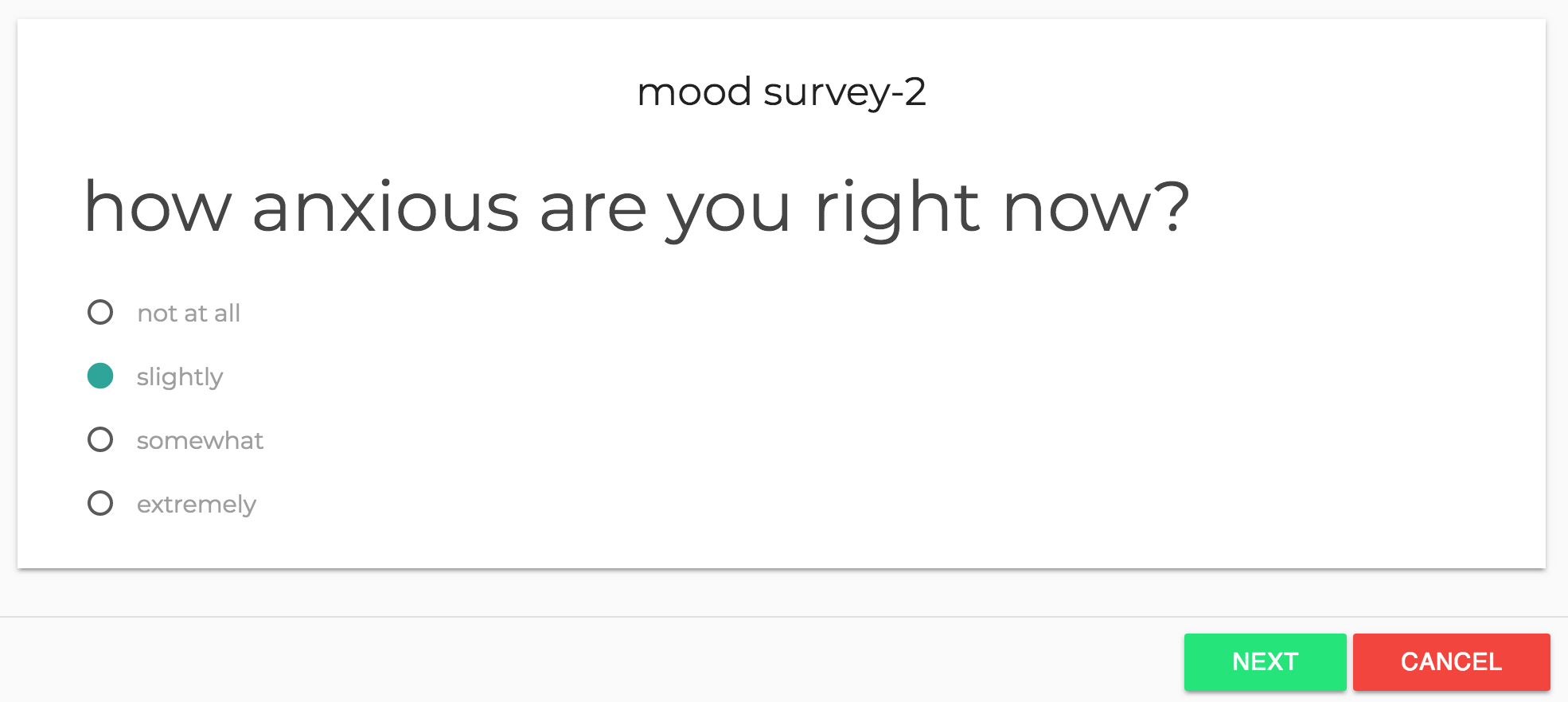
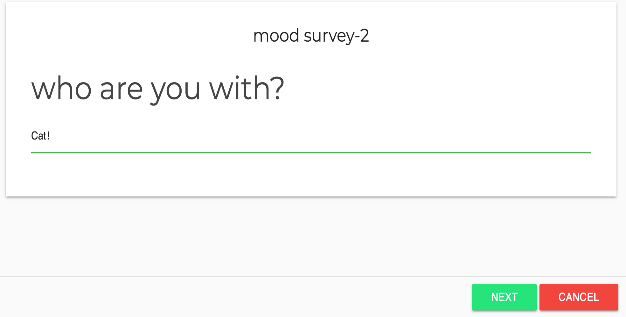
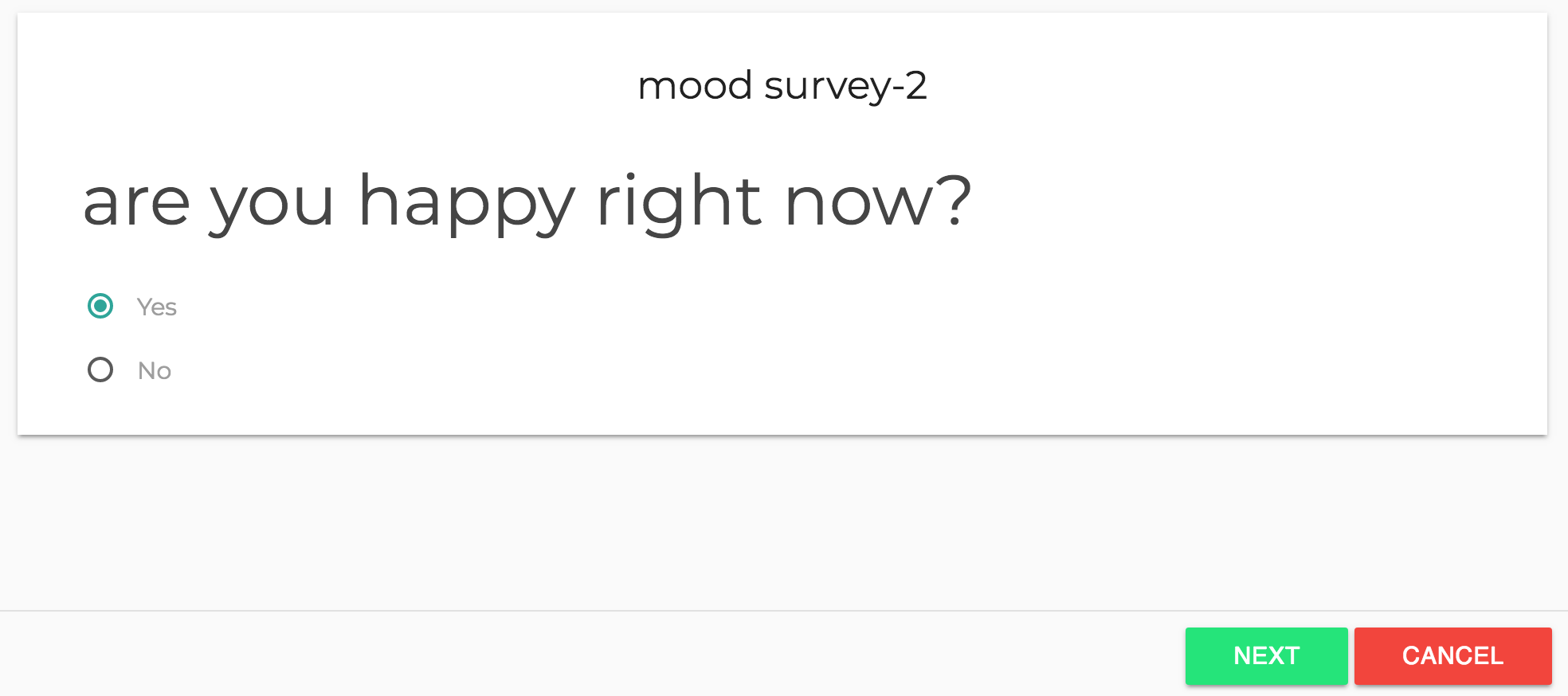


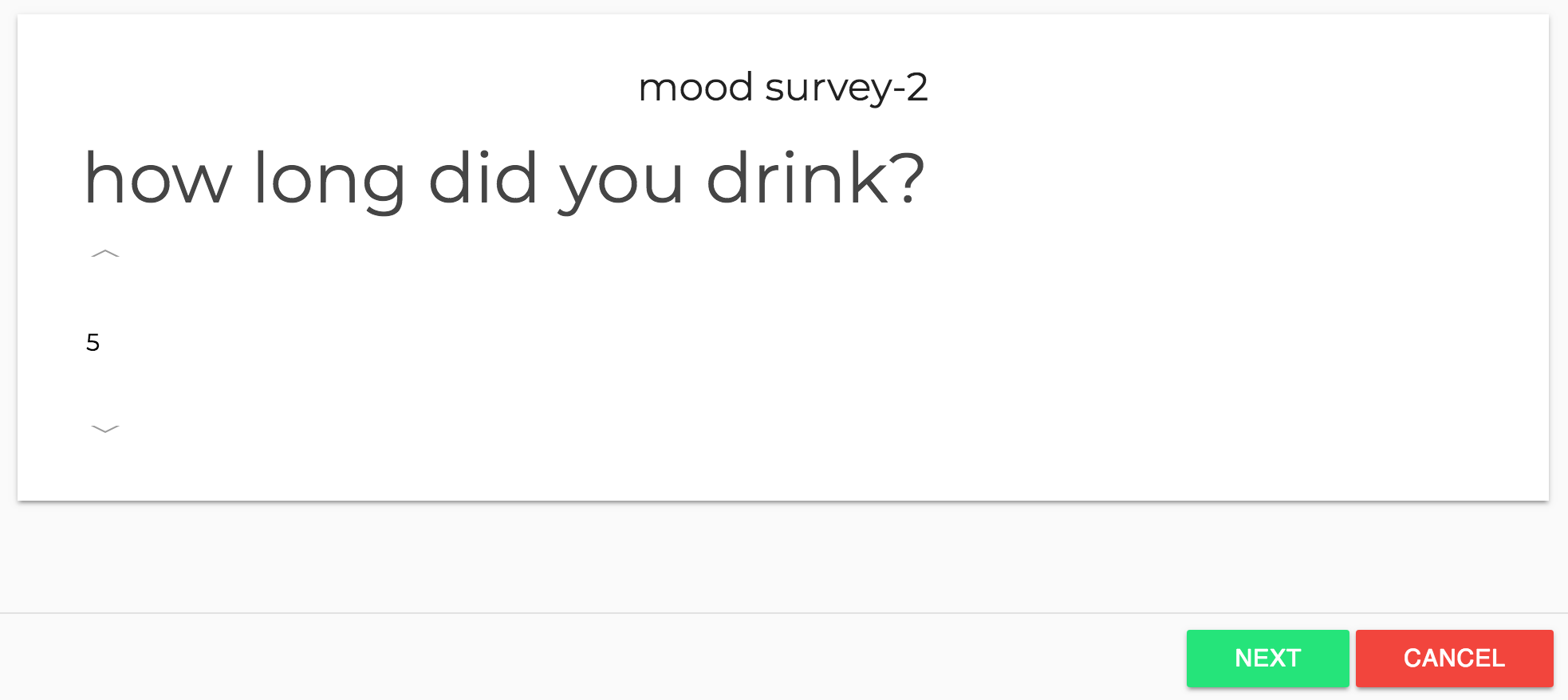
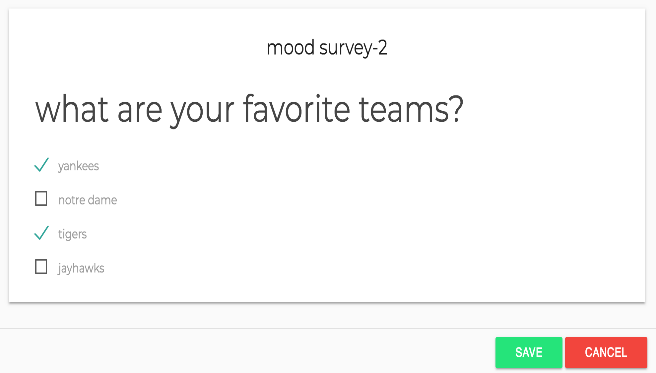
Figure 4.4 Implementation Page Select Survey



(a) (b)



(c) (d)



(e) (f)

Figure 4.5 (a) single choice questions (b) scale type questions (3) yes or no questions,(4) Text questions (5) time interval questions (6) multiple-choice questions

As the types and order of questions set in advance are different, these types of questions displayed in the order of the designed questionnaires in advance. There are six question types here. The Figure 4.5(a-f) shows these 6 types implementation page. They are single choice questions, scale type questions and, yes or no questions, Text questions, time interval questions and multiple-choice questions.

After all the answers have been completed, all the answers will be stored in a js object uploaded to firebase and the answer list will be updated.

* + 1. File Management Module

There are two essential parts in this module, one is to import data, and the other is to export data.

The importance and necessity of importing data have already introduced before, so now I will explain the specific process of how to import data. When the user has a prepared CSV file, he can choose to import the data by pressing the import button. HTML5 has File API that allows the browser to interact with files on the file system directly. Most modern browsers that support HTML5 can use this API to perform client-side only processing without the need for a round trip to the servers or the use of Ajax. After selecting the file, use the HTML5 File API to upload the file for parsing. The parsed Js object can be used directly for transmission to the back-end server for later analysis. At the same time, the data table of the front page will also update.



Figure 4.6 CSV File Upload

Data export is a more important function, because only data can be exported to provide researchers with more choices for data analysis. Since the database-side record structure determines the data can be exported with text options, you can also directly select the export with index labels.



Figure 4.7 Index/String Switch

Multiple choice questions, Yes/No questions and Scale type questions need to transform into text-based answers by performing these three converters. Then collect all survey questions from the database and then create the DOM element for download.

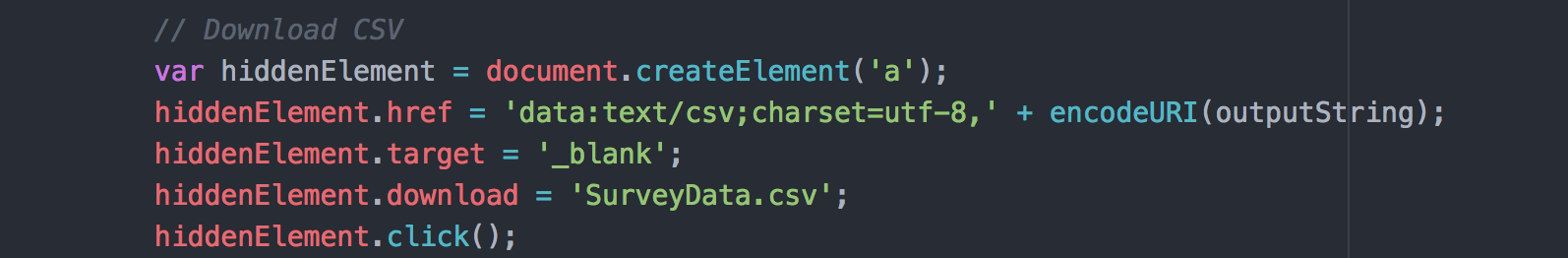


Figure 4.8 CSV File Download

* + 1. Data Demonstration Module

Data Demonstration Module Contains three different charts for the user to choose for showing the data analysis result. The three types of charts are bar chart, pie chart, and column chart respectively. Charts are showing by using Highcharts library. Highcharts is a SVG-based, multi-platform charting library. It makes it easy to add interactive charts to the web project. Numerous events supply hooks for programming against the charts, making it easy to demonstrate complex relationships between data with live, dynamic updates of data and customizable animations.

Data can be handled over to Highcharts in any form, even from a different site, and a callback function used to parse the data into an array.

Users can export the chart to PNG, JPG, PDF or SVG format at the click of a button, or print the chart directly from the web page.

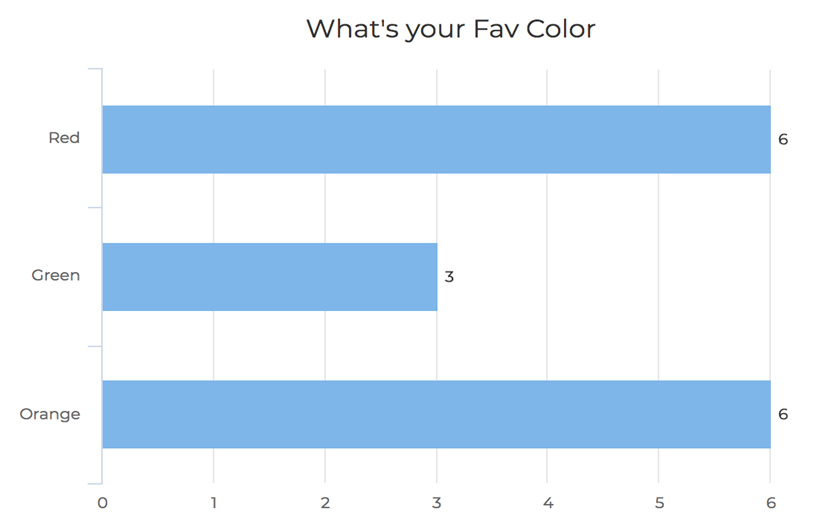


Figure 4.9 Data Visualization Sample

And the result page provides a data table that supports filter and sort. The whole module is: when the administrator created the survey, and some user implemented the questionnaire. The administrator can navigate to the survey detail page to see the summary of the result.

The data table shows the whole answers answered by the users. And can select a question like Yes or no question to filter. If the administrator only wants to see the answer with “Yes”. The data table will update and show the result.

* + 1. Rule Management Module

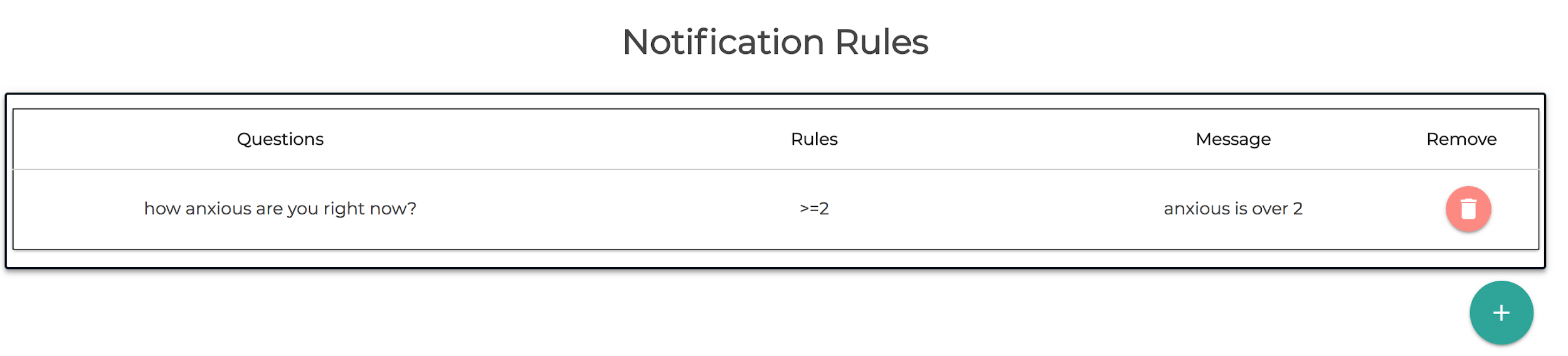


Figure 4.10 Rule Management Panel

The Notification Rules Panel used for managing user’s monitoring rules. The administrator can create a new listener rule by clicking the plus button in the lower right corner.

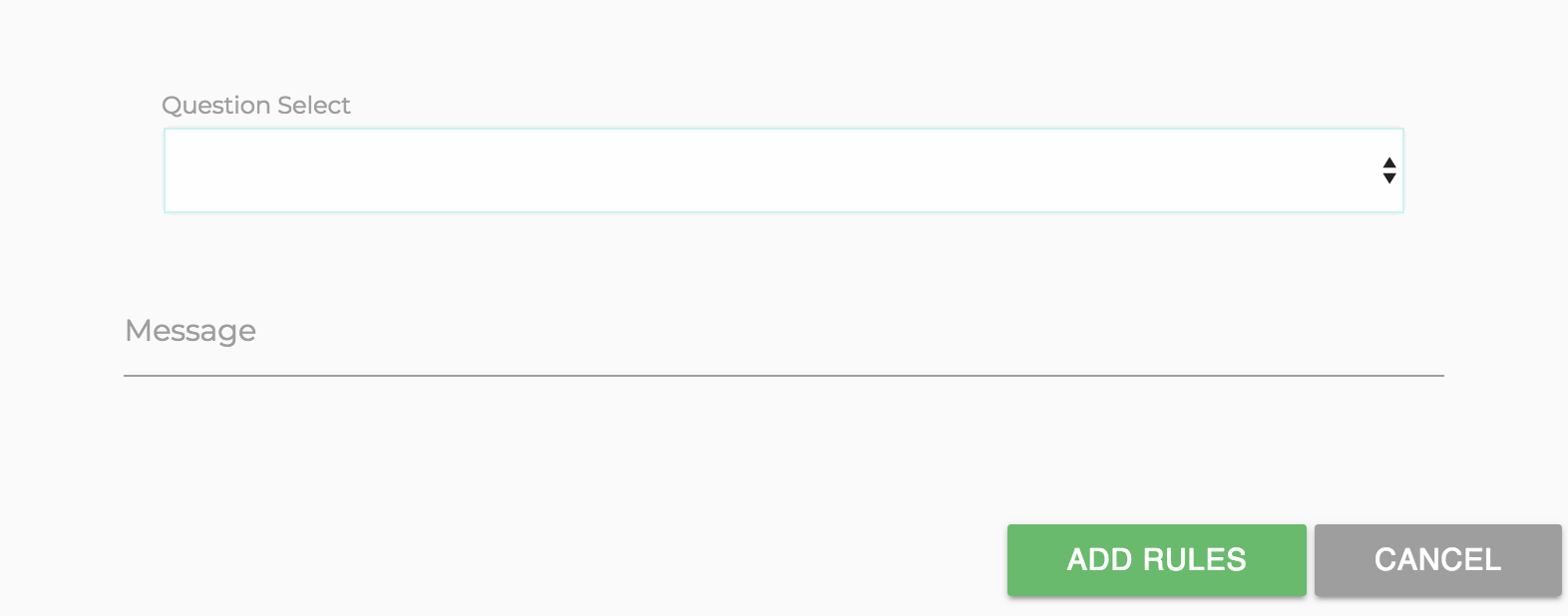


Figure 4.11 Rule Panel Question For Select

After selecting the question in the questionnaire, a new condition box will appear. The administrator needs to pick the conditions that they want to monitor in the new box. And add a notification message to remind yourself that you have received this notification.

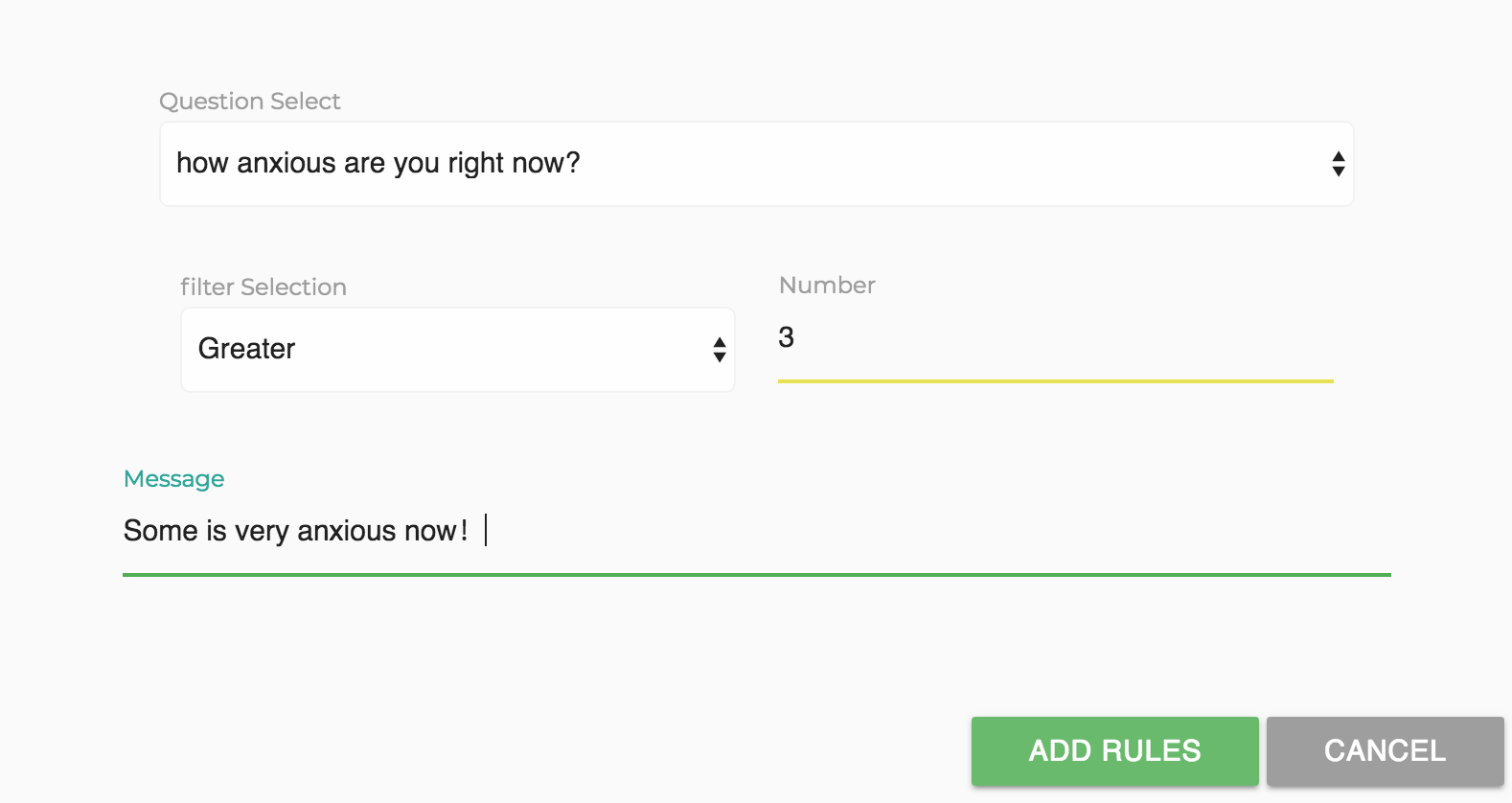


Figure 4.12 Select the rule

After Clicked the Add button, the new rule will be written to the database to wait for the background response.

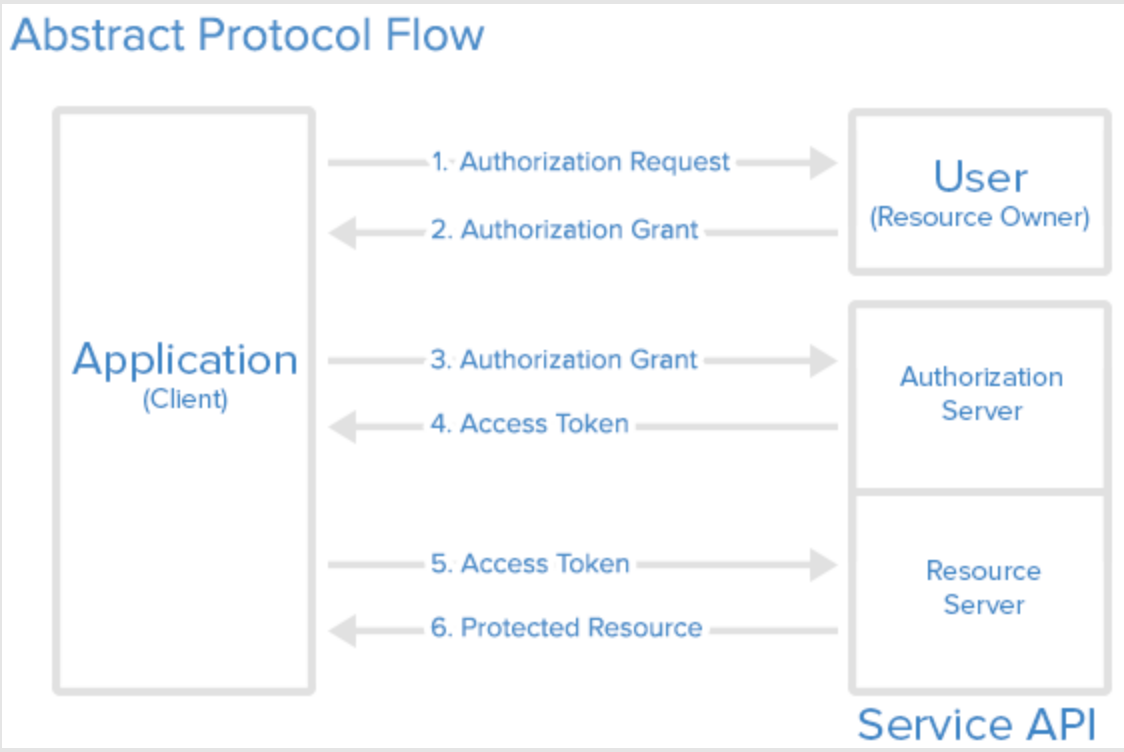
When admin needs to remove a rule, only need to click the delete button next to the corresponding rule to delete it. The relevant data will be eliminated in the database.

* + 1. User Management Module

User Management Module designed for all users: administrator and subject all need to sign in the system to verify their identification. We don’t differentiate the administrator and users from the very beginning. When subject create a new survey, he is the administrator of that survey. Firebase provides the security.



Figure 4.13 Login and Register



[8]Figure 4.14 Oath Protocol

Here is a more detailed explanation of the Oath protocol:

The application requests authorization to access server resources from the user

If the user authorized the request, the application receives an authorization grant

The application requests an access token from the authorization server (API) by presenting authentication of its own identity, and the authorization grant

If the application identity is authenticated and the authorization grant is valid, the authorization server (API) issues an access token to the application. Authorization is complete.

The application requests the resource from the resource server (API) and presents the access token for authentication

If the access token is valid, the resource server (API) serves the resource to the application.

* + 1. Data Selection Module

Data Selection Module is using for selecting data for data process. Data has been filtered as follow, then selected data will be saved for further process. As mentioned before, for numerical data, the selection is mainly based on data relationships that are greater than, less than, or equal to the chosen value. If the data satisfies the condition, the data is selected and updated to the selected data set. For text-based data, only two types of determination were used. One of them is that the text contains the chosen keyword. The second type is the text language. The keyword is entirely equal.

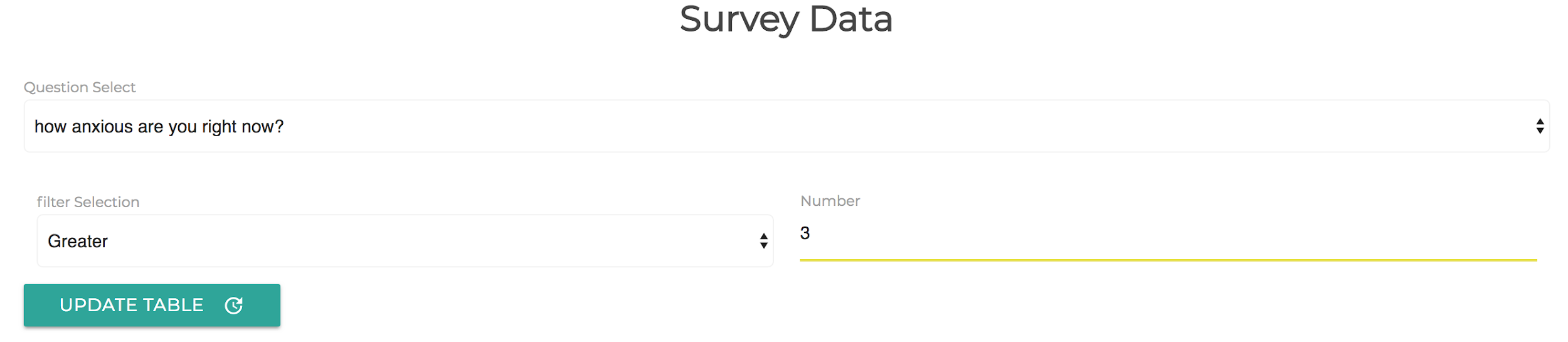


Figure 4.15 Data Selection

After the data has been filtered and selected, it needs to send process request to the back-end. But if someone has hacked the data, that will be a bad thing. Because most of the research is private, no one wants to leak information about clients. So we want to secure the data by hash the data using JSON Web Tokens RS256 method. Then using Ajax.post to request the response on the server-side.

RSA algorithm is asymmetric cryptography algorithm. Asymmetric means that it works on two different keys, i.e., Public Key and Private Key. Like the name's description, the Public Key will give to everyone, and the Private key is private. Since this is asymmetric, nobody else except browser can decrypt the data even if a third party has public key of browser. The idea of RSA is the fact that it is challenging to factorize a large integer. The public key consists of two numbers, where one number is the multiplication of two large prime numbers. And the private key is also derived from the same two prime numbers. So if somebody can factorize the large number, the private key is compromised. Therefore, encryption strength entirely lies on the key size, and if we double or triple the key size, the strength of encryption increases exponentially. RSA keys can be typically 1024 or 2048 bits long, but experts believe that 1024 bit keys could break soon. But till now it seems to be an infeasible task.

* 1. Server Side

This section will introduce three modules designed above. Which is Rule-based Notification Module, Data Preprocessing Module, and Unsupervised Learning Module. The structure of the server side is showing in the figure. Gunicorn works by internally handling the calling of your flask code. It is done by having workers ready to handle the requests instead of the sequential one-at-a-time model that the default flask server provides. Thus the app can handle more requests per second. Also, Nginx in front of a pre-forking server is a great combination. Nginx handles communications with clients and doesn't suffer a penalty for handling slow clients. It sends requests to the backend as fast as the backend can handle those requests, enabling the backend to be as efficient with server resources as possible. The backend returns the result as soon as it calculates it, and Nginx buffers that response to feeding it to slow clients at their own pace. Meanwhile, the backend can move on to handling another request even as the slow client is still receiving the result.

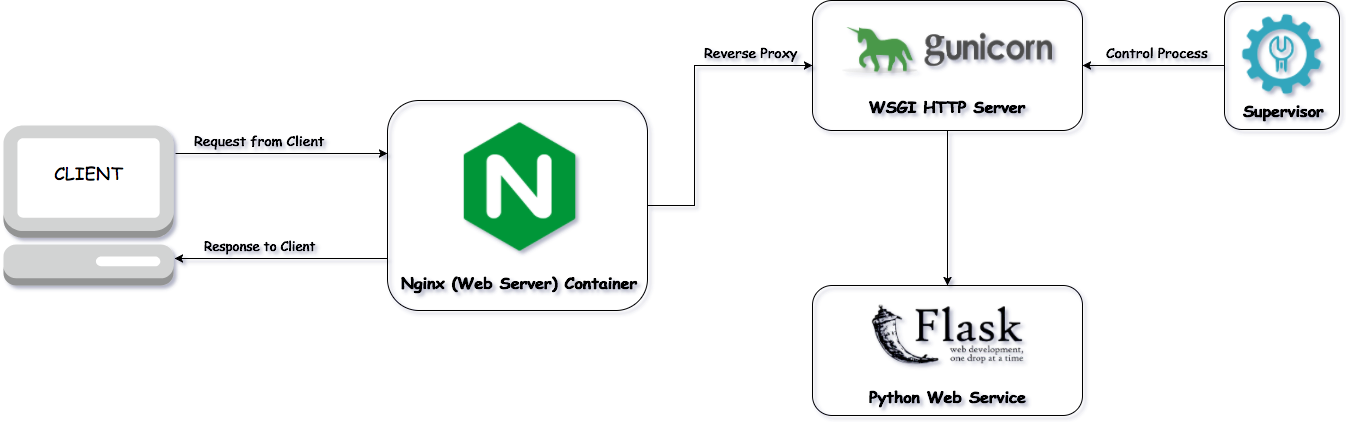


Figure 4.16 Structure Of The Server-Side

* + 1. Rule-based Notification Module

Rule-based Notification Module can perform long-running processes in the background. This module is written in NodeJS language because firebase offers Admin SDK for the JavaScript. And the server side JavaScript is NodeJS.

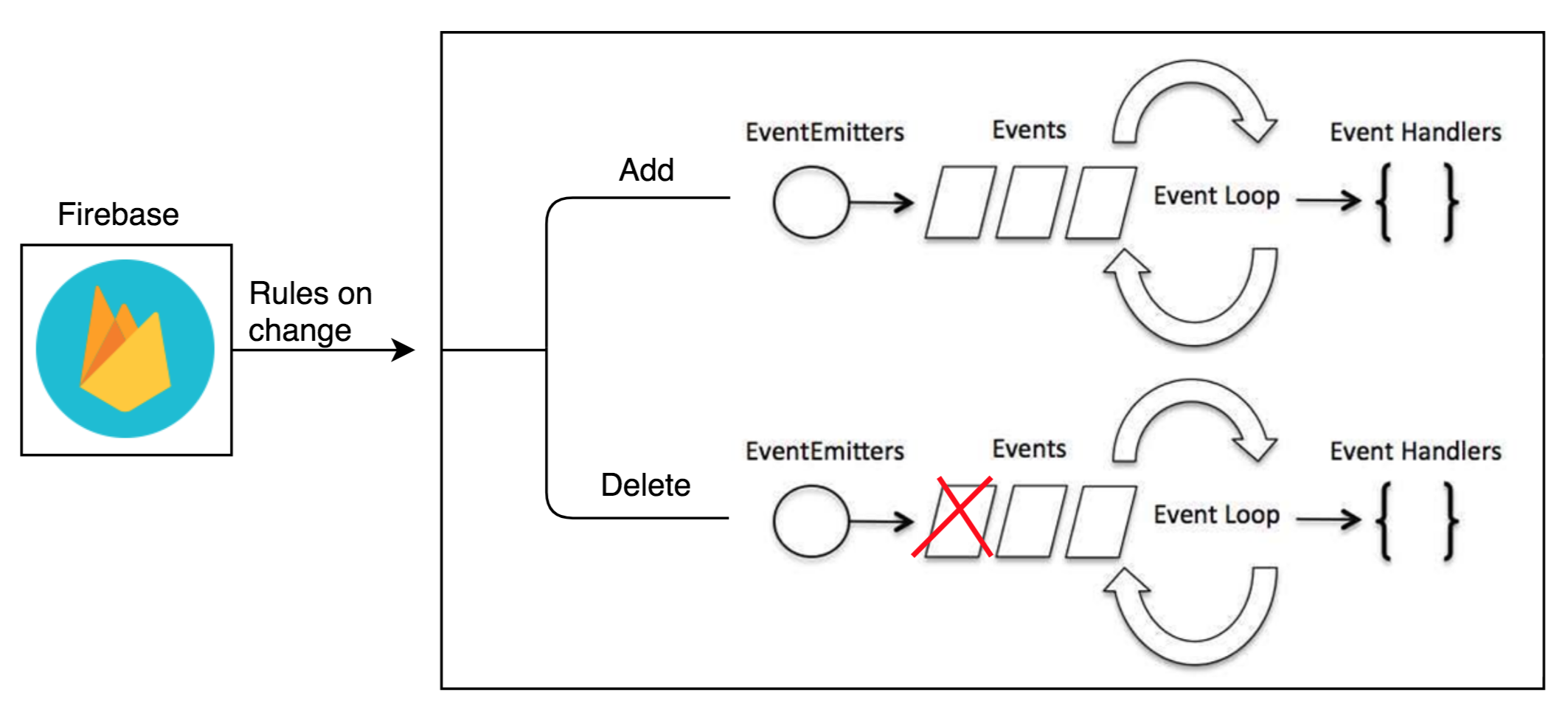


Figure 4.17 Notification Module Pipeline

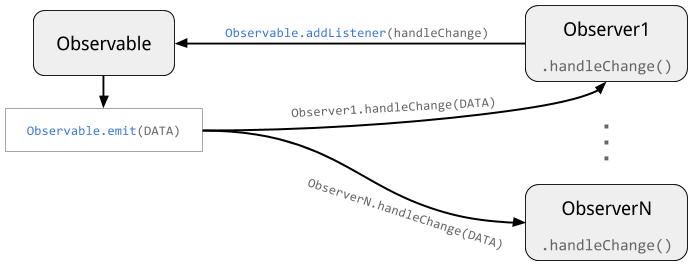


Figure 4.18 Event Structure

After the data in the database changes, the process gets changed data. Parsed this piece of data, the data obtained has the index of this rule and the survey information. The name of this listener event is determined using a combination of the name of the survey and the index to make sure it is unique id.

Each event will only focus on one survey and only one rule. Every event will get the new answer when a user has implemented a similar survey. The event has some rule-based methods to judge whether the answer is meet the rule requirement.

All the data will send to Nools rule engine, and only if it met the rule, the message would submit to the administrator. The mail is sent by Nodejs library: Nodemailer. Email sent based on the protocol. The content of the email contains the person who provided the answer, the answer details, and the preset notification content.

* + 1. Data Preprocessing Module

Data Pre-processing Module mainly has three functionalities: decrypt data, pre-cleaning the data and preprocessing the data. First, using the same package to decrypt the data. The second is to clean the data.

For the missing data, we have three processing methods:

1. Through the maximum and minimum values in the data, we assign the intermediate value to the missing data;

2. Through the average value of all data attached to the missing data;

3. Delete vacant data items;

At here, for the data collected in the front-end because we have validated the data in the front-end. So, we don’t need to handle null value exception. But for the user imported data, they may have the records lack value. So use the first method to process, we assign the intermediate value to the missing data.

Then for numeric data, I will do Min-Max Normalization. Which is:

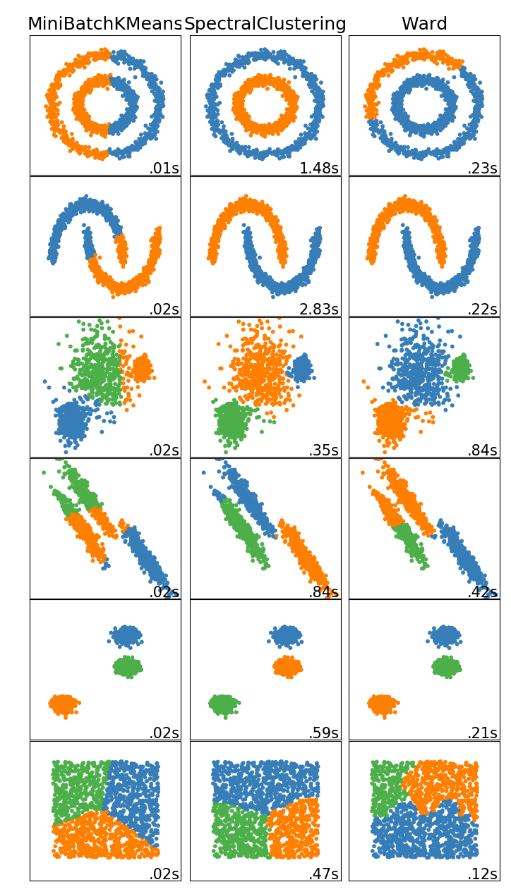
The date and time data will translate to the seconds of a day and then normalize as the standard numeric data.

Numeric data is easy to preprocess, but for the text data. We need to make more efforts to make it possible as a feature in the module. The process can be described as follow.

1. Punctuations Removal: We do not need punctuations as the features, they are just symbols to separate sentences and words.
2. Stopwords Removal: There is a kind of word called stopword[9]. They are some common function words in a sentence, like ‘a’, ‘the’, ‘and’, ‘to’, ‘at’, etc. These words seem to be useless for sentiment analysis.
3. Digital words Removal: Some words start with a digit, like a year ‘2018’, ‘2:00 pm’. These words also have no relationship with attitudes or feelings. So these words should be removed.

After the clean and remove these words. The text data contains the word we can treat as the feature. Then use one-hot decoding to decode the text words as one feature. After all the preprocessing methods, the data is ready for clustering method to do cluster analysis.

* + 1. Unsupervised Learning Module



[10]Figure 4.19: A comparison of the clustering algorithms in scikit-learn

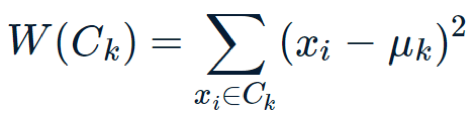
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method name** | **Parameters** | **Scalability** | **Use case** | **Geometry (metric used)** |
| K-Means | number of clusters | Very large n\_samples, medium n\_clusters with n\_samples | General-purpose, even cluster size, flat geometry, not too many clusters | Distances between points |
| Spectral clustering | number of clusters | Medium n\_samples, small n\_clusters | Few clusters, even cluster size, non-flat geometry | Graph distance (e.g. nearest-neighbor graph) |
| Ward hierarchical clustering | number of clusters | Large n\_samples  and n\_clusters | Many clusters, possibly connectivity constraints | Distances between points |

Table 1: A comparison of the clustering algorithms

* + - 1. K-means Method

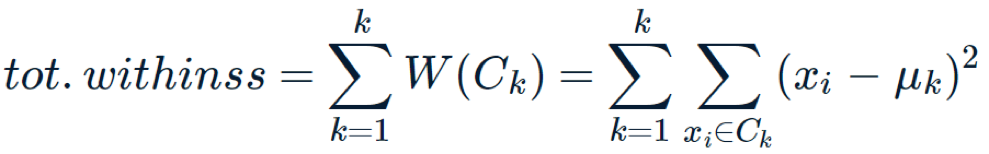
The basic idea behind k-means clustering consists of defining clusters so that the total intra-cluster variation (known as total within-cluster variation) is minimized.[11]

There are several k-means algorithms available. The standard algorithm is the Hartigan-Wong algorithm (Hartigan and Wong 1979), which defines the total within-cluster variation as the sum of squared distances Euclidean distances between items and the corresponding centroid:



* design a data point belonging to the cluster
* is the mean value of the points assigned to the cluster

Each observation () is assigned to a given cluster such that the sum of squares (SS) distance of the observation to their assigned cluster centers is a minimum. We define the total within-cluster variation as follow:



The first step when using k-means clustering is to indicate the number of clusters (k) that will be generated in the final solution. The algorithm starts by randomly selecting k objects from the data set to serve as the initial centers for the clusters. The selected objects are also known as cluster means or centroids.

Next, each of the remaining objects is assigned to it’s closest centroid, where closest is defined using the Euclidean distance between the object and the cluster mean. This step is called “cluster assignment step”. Note that, to use correlation distance, the data are input as z-scores.

After the assignment step, the algorithm computes the new mean value of each cluster. The term cluster “centroid update” is used to design this step. Now that the centers have been recalculated, every observation is checked again to see if it might be closer to a different cluster. All the objects are reassigned again using the updated cluster means.

The cluster assignment and centroid update steps are iteratively repeated until the cluster assignments stop changing (i.e. until convergence is achieved). That is, the clusters formed in the current iteration are the same as those obtained in the previous iteration.

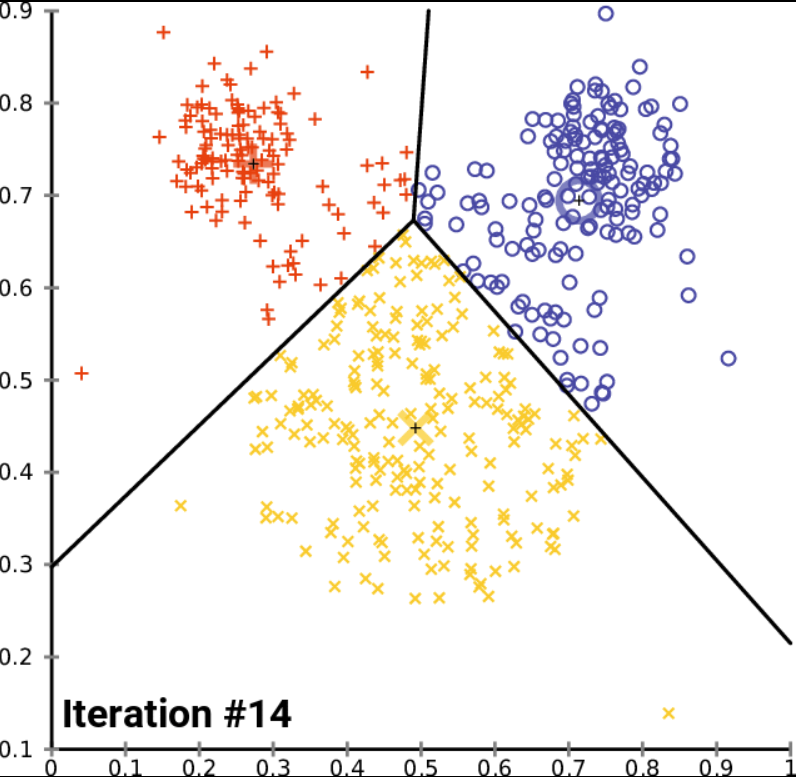
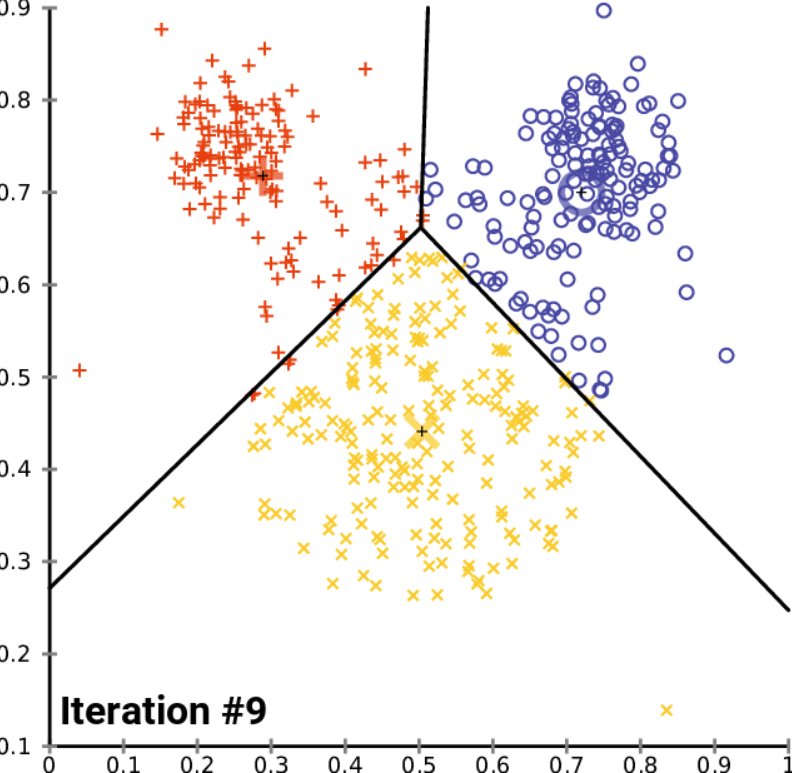
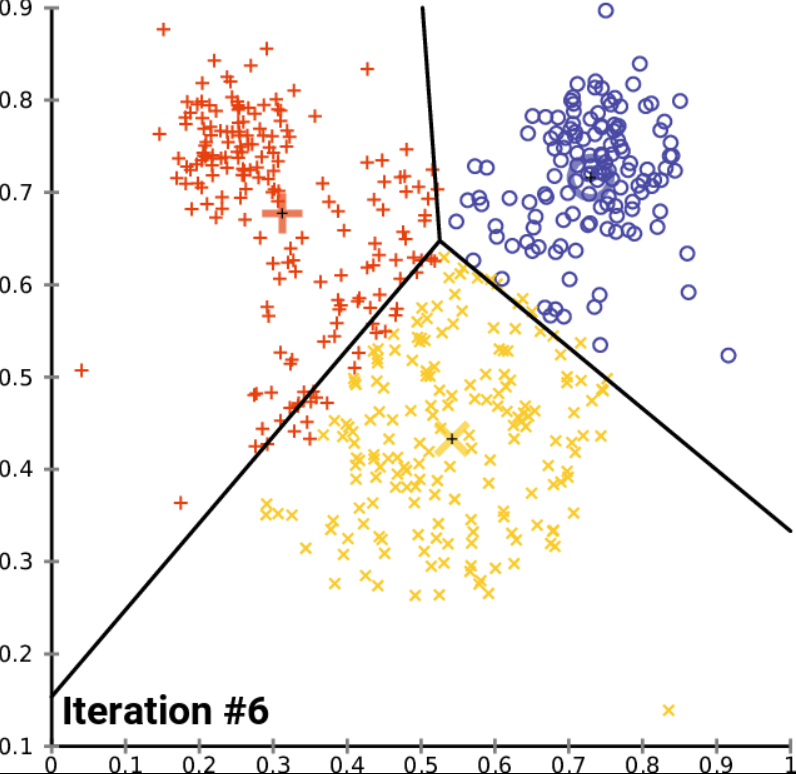
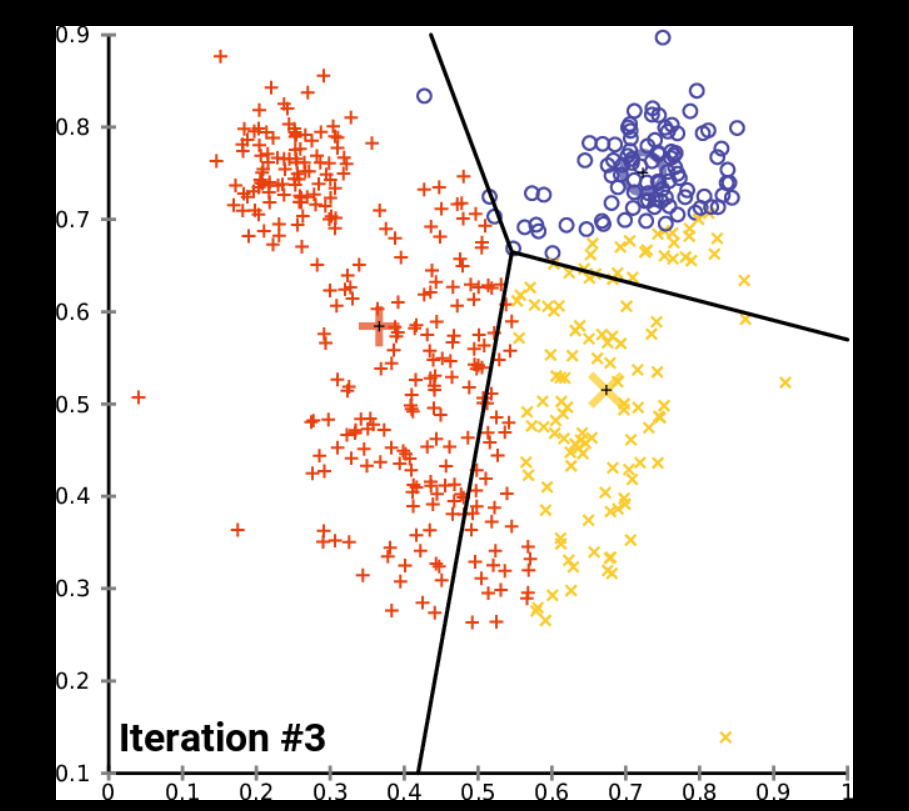
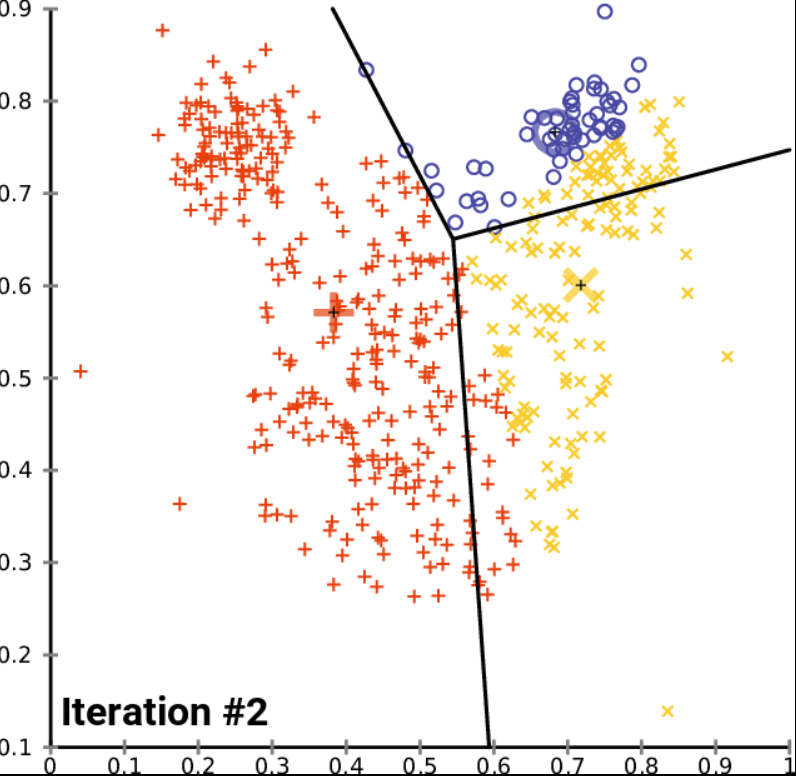
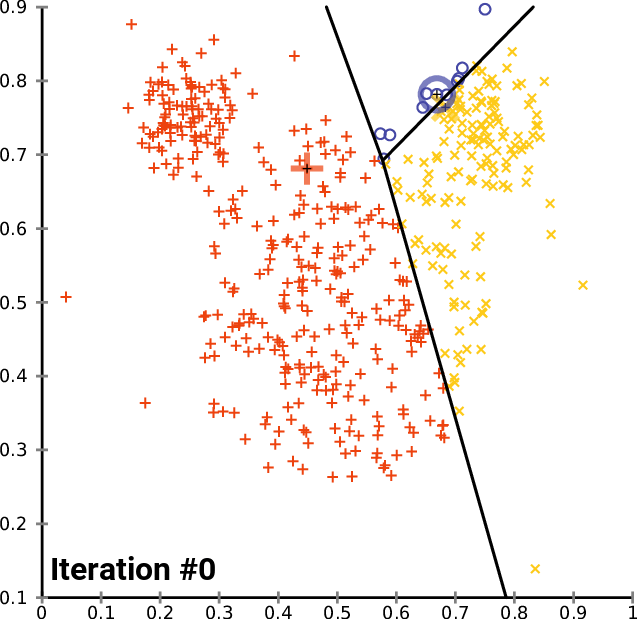
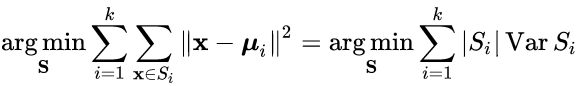
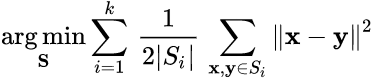


Figure 4.20 K-means Result By Iteration

Formally, the objective is to find[12]:



where is the mean of points in . This is equivalent to minimizing the pairwise squared deviations of points in the same cluster:



The Equivalence can be deduced from identity:



Because the total variance is constant, this is also equivalent to maximizing the squared deviations between points in different clusters (between-cluster sum of squares, BCSS).

* + - 1. Spectral Clustering Method

Now I would like to state the most common spectral clustering algorithms.[13] We assume that our data consists of “points” which can be arbitrary objects. We measure their pairwise similarities by some similarity function which is symmetric and non-negative, and we denote the corresponding similarity matrix by .

1. Unnormalized spectral clustering

Input: Similarity matrix , number of clusters to construct.

* Construct a similarity graph. Let be its weighted adjacency matrix.
* Compute the unnormalized Laplacian.
* Compute the first eigenvectors of.
* Let be the matrix containing the vectors as columns.
* For , let be the vector corresponding to the-th row of .
* Cluster the points in with the -means algorithm into clusters .

Output: Clusters with .

There are two different versions of normalized spectral clustering, depending which of the normalized graph Laplacians is used.

1. Normalized spectral clustering according to Shi and Malik (2000)

Input: Similarity matrix , number of clusters to construct.

* Construct a similarity graph. Let be its weighted adjacency matrix.
* Compute the unnormalized Laplacian.
* Compute the first generalized eigenvectors of the generalized eigen problem .
* Let be the matrix containing the vectors as columns.
* For , let be the vector corresponding to the-th row of .
* Cluster the points in with the -means algorithm into clusters .

Output: Clusters with .

Note that this algorithm uses the generalized eigenvectors of , which according to Proposition 3 correspond to the eigenvectors of the matrix . So in fact, the algorithm works with eigenvectors of the normalized Laplacian , and hence is called normalized spectral clustering. The next algorithm also uses a normalized Laplacian, but this time the matrix instead of . As we will see, this algorithm needs to introduce an additional row normalization step which is not needed in the other algorithms.

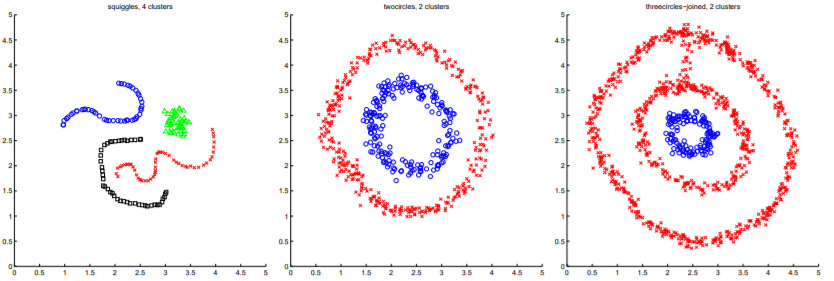
1. Normalized spectral clustering according to Ng, Jordan, and Weiss (2002)

Input: Similarity matrix , number of clusters to construct.

* Construct a similarity graph. Let be its weighted adjacency matrix.
* Compute the normalized Laplacian .
* Compute the first eigenvectors of .
* Let be the matrix containing the vectors as columns.
* Form the matrix from by normalizing the rows to norm 1, that is set
* For , let be the vector corresponding to the-th row of .
* Cluster the points with the -means algorithm into clusters .

Output: Clusters with .

All three algorithms stated above look rather similar, apart from the fact that they use three different graph Laplacians. In all three algorithms, the main trick is to change the representation of the abstract data points to points . It is due to the properties of the graph Laplacians that this change of representation is useful.



*Figure 4.21: Clustering examples, with clusters indicated by different symbols (and colors where available).*[14]

The relationship between K-means and spectral clustering is showing below.

The kernel -means problem is an extension of the -means problem[15] where the input data points are mapped non-linearly into a higher-dimensional feature space via a kernel function

The weighted kernel -means problem further extends this problem by defining a weight for each cluster as the reciprocal of the number of elements in the cluster,

Suppose is a matrix of the normalizing coefficients for each point for each cluster if and zero otherwise. Suppose is the kernel matrix for all points. The weighted kernel -means problem with n points and k clusters is given as,

such that

,

such that . In addition, there are identity constrains on given by . This problem can be recast as,

This problem is equivalent to the spectral clustering problem when the identity constraints on are relaxed. In particular, the weighted kernel k-means problem can be reformulated as a spectral clustering (graph partitioning) problem and vice versa. The output of the algorithms are eigenvectors which do not satisfy the identity requirements for indicator variables defined by . Hence, post-processing of the eigenvectors is required for the equivalence between the problems. Transforming the spectral clustering problem into a weighted kernel k-means problem greatly reduces the computational burden.

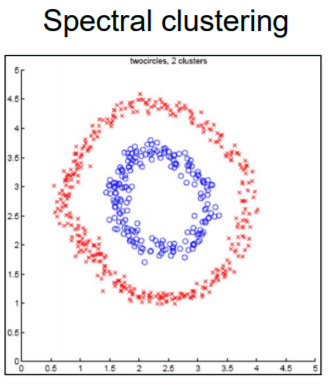
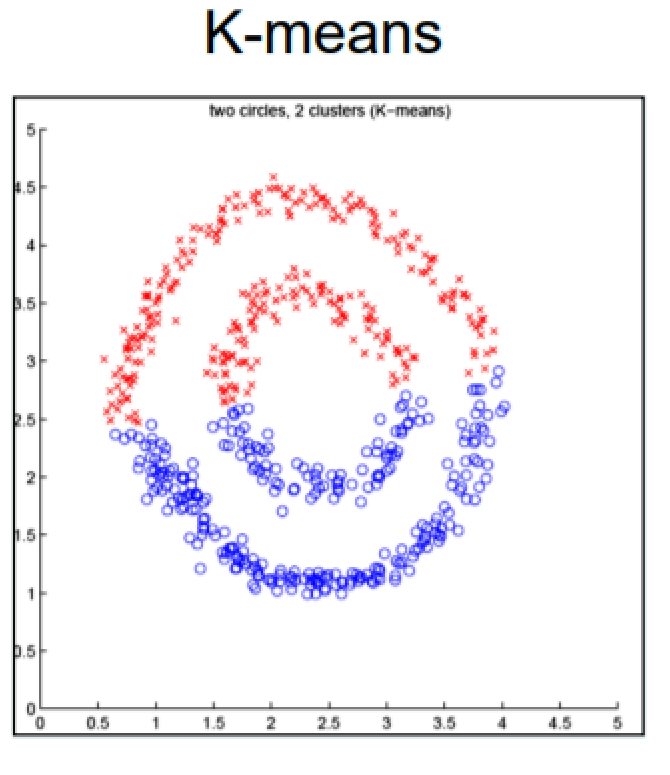
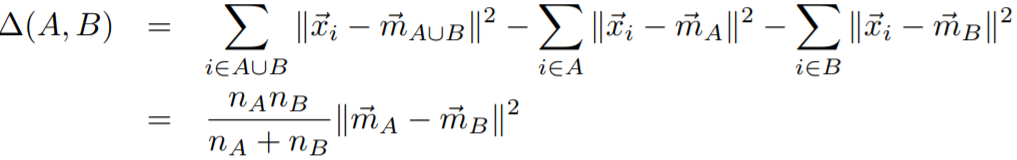


Figure 4.22: The example of K-means and Spectral clustering

* + - 1. Ward Hierarchical Clustering

Ward’s method says that the distance between two clusters, A and B, is how much the sum of squares will increase when we merge them[16]:



where is the center of cluster , and is the number of points in it. is called the merging cost of combining the clusters and .

With hierarchical clustering, the sum of squares starts out at zero (because every point is in its own cluster) and then grows as we merge clusters. Ward’s method keeps this growth as small as possible. This is nice if you believe that the sum of squares should be small. Notice that the number of points shows up in , as well as their geometric separation. Given two pairs of clusters whose centers are equally far apart, Ward’s method will prefer to merge the smaller ones.

Ward’s method is both greedy and constrained by previous choices as to which clusters to form. This means its sum-of-squares for a given number of clusters is usually larger than the minimum for that , and even larger than what -means will achieve. If this is bothersome for your application, one common trick is use hierarchical clustering to pick , and then run -means starting from the clusters found by Ward’s method to reduce the sum of squares from a good starting point.

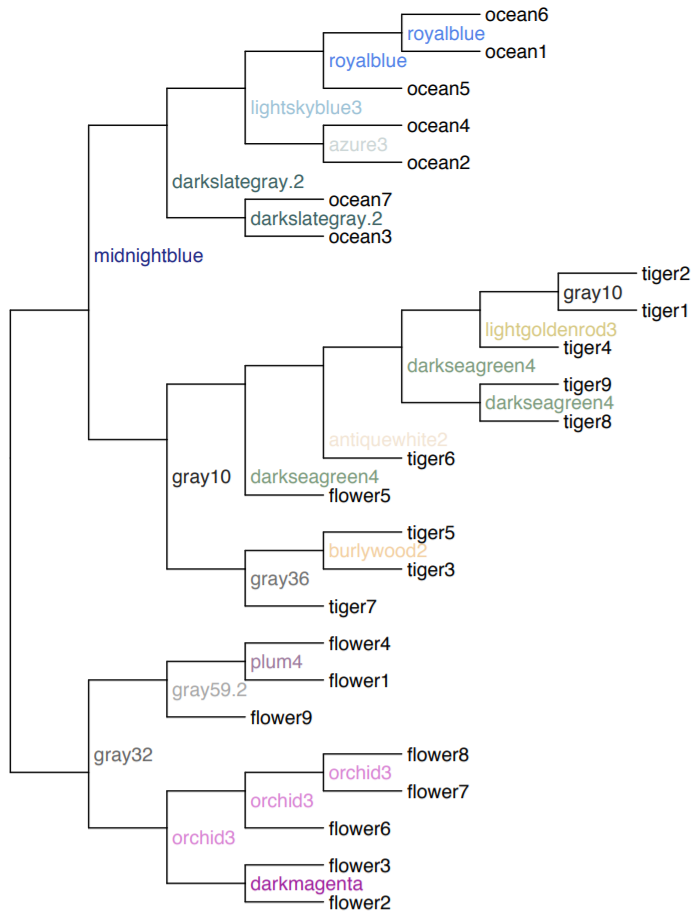


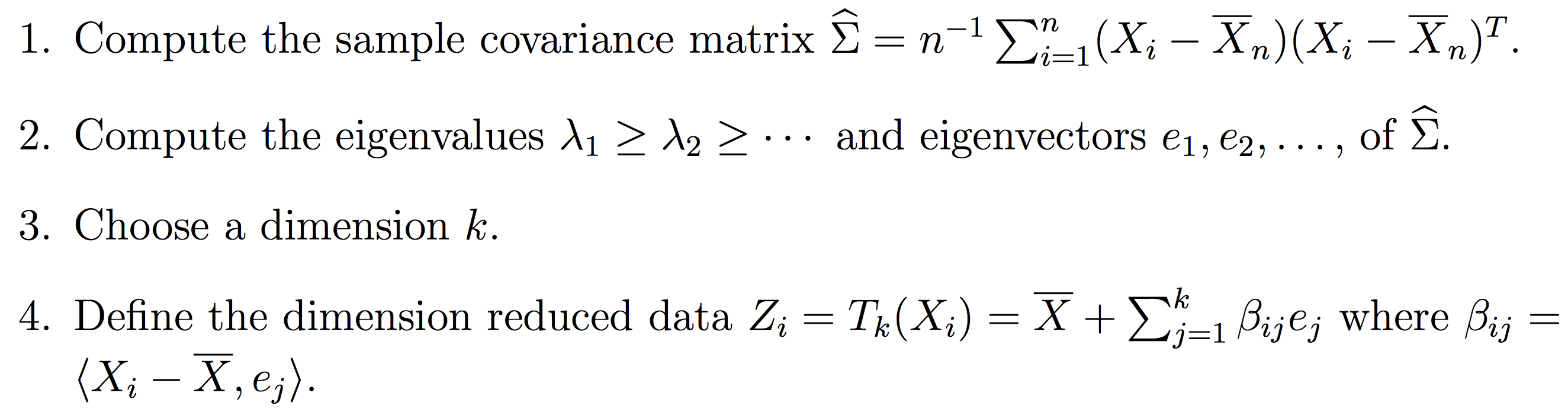
Figure 4.23: Using Ward’s method to form a hierarchical clustering

The k-means algorithm gives no guidance about what k should be. Ward’s algorithm, on the other hand, can give us a hint through the merging cost. If the cost of merging increases a lot, it’s probably going too far, and losing a lot of structure. So, a rule of thumb is to keep reducing k until the cost jumps, and then use the k right before the jump. Of course, this leaves you to decide how big a merging cost is acceptable, and there’s no theory whatsoever to say that this will often or even usually lead to good choices, but it does make a kind of sense. Of course, the same rule of thumb can be applied to other hierarchical clustering techniques: pick the k just before the merging cost takes off.

* + - 1. Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. If there are n observations with p variables, then the number of distinct principal components is min (n-1, p). This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set.

The procedure of the PCA can be explain as:



Such dimensionality reduction can be a very useful step for visualizing and processing high-dimensional datasets, while still retaining as much of the variance in the dataset as possible. For example, selecting k = 2 and keeping only the first two principal components finds the two-dimensional plane through the high-dimensional dataset in which the data is most spread out, so if the data contains clusters these too may be most spread out, and therefore most visible to be plotted out in a two-dimensional diagram.

1. Future Work

For the data management part, firstly, could add more visualization options for the users, and can add the options for the administrator that let administrator see the answers of each user. And then the data map could be inserted into the application, allowing administrator know where the data come from. The rule-based system can add more rules to monitoring the user’s activity. Data selection module, make a combination, query like selection to select data.

For the data analysis part, supervised learning method such as Naïve Bayes and KNN is the step, which may contribute to the model construction. The parameters for selection is important. In recent years, more and more researchers have the experience with machine learning experience, so the interface for the expert can be on the list. Deep learning and image classification is the hottest topic right now, deep learning module can be considered as the new feature.

1. Summary

The report describes the design and development of a web-based survey analysis and management system, combines with data validation and data acquisition capabilities, for the use of general survey study. Dashboards will collect user data, upload it directly to the database, perform real-time analysis, set listening rules, and observe data.

The system provides a new assessment method for survey study, including survey implements, management, as well as real-time rule-based notification. More flexible functionalities on managing the data will be performed by importing and exporting the data. In this system, three unsupervised machine learning methods with basic data preprocessing have been implemented. The security of data transmitting is guaranteed in the whole process.

This system enhances the preciseness and the effectiveness of survey management, monitoring and analysis. This real-time, efficient system is an efficient tool for researchers to collect survey data and automatically utilizing unsupervised machine learning methods to get the profound grasp of data.

References

1. Kristalyn Salters-Pedneault, P: Self-Report Information in Psychology

https://www.verywellmind.com/definition-of-self-report-425267

1. Google Forms. https://www.google.com/forms/
2. Fanning J, McAuley E. A Comparison of Tablet Computer and Paper-Based Questionnaires in Healthy Aging Research. JMIR Res Protoc 2014;3(3):e38
3. How to Visualize Survey Results. https://venngage.com/blog/survey-results/
4. New Leader, Trends, and Surprises in Analytics, Data Science, Machine Learning Software Poll. https://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html
5. G Suite Learning Center. https://gsuite.google.com/learning-center/products/forms/get-started/#!/
6. Conduct and Interpret a Cluster Analysis. http://www.statisticssolutions.com/cluster-analysis-2/
7. An Introduction to OAuth 2. https://www.digitalocean.com/community/tutorials/an-introduction-to-oauth-2
8. Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." Journal of Computational Science 2.1 (2011): 1-8.
9. Clustering. http://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering
10. Partitioning Clustering Essentials-K-Means Clustering Essentials. http://www.sthda.com/english/articles/27-partitioning-clustering-essentials/87-k-means-clustering-essentials/
11. k-means clustering. https://en.wikipedia.org/wiki/K-means\_clustering
12. von Luxburg, A tutorial on spectral clustering. U. Stat Comput (2007) 17: 395. https://doi.org/10.1007/s11222-007-9033-z
13. AY Ng, MI Jordan, Y Weiss. On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems. 2002 - papers.nips.cc
14. Spectral clustering. https://en.wikipedia.org/wiki/Spectral\_clustering
15. Distances between Clustering, Hierarchical Clustering. http://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf