Advanced Technology Attachment Programme (ATAP)

Interim Project Report

at

Intra Technologies Private Limited

Reporting Period
6 Jan 2020 to 19 June 2020

by

Abhinav Ramnath

Department of Computer Science

School of Computing

National University of Singapore

2020/2021

Project Title: Web Dashboard Designing

Project ID: AY19143848

Project Supervisor: Tan Tiow Seng

Summary

My internship focuses on Full-Stack Software Development to build interactive Web Dashboards. The purpose of these Dashboard's is to serve as administrative systems that administrators can use to interact with the Database and perform complex operations with ease. This involves me following the Requirements/Specifications for the software and implementing it using software Design Patterns. I also gain a deeper undertanding of Software Architectures of Web Applications, by working on all layers of the application from Database modelling using firebase, backend using Node and the User interface using VueJS.

Subject Descriptors:

    D.2.1 Requirements/Specifications

    D.2.10 Design

    D.2.11 Software Architectures

Keywords:

    Web Dashboard, Databae, Testing and Debugging, Web Applications

Implementation Software and Hardware:

    VueJS, Node.JS, Javascript, Firebase

Contents

1 Introduction

1.1 Background and Organizational Structure of Host Organization

I am doing my ATAP at Intra Technologies Private Limited, a software development house start-up that has a few self-initiated projects. The CEO, COO, CTO work together to manage the timelines, business development of these projects. The projects the company has initiated include Lemon and Rest. I am currently focused on developing software for Lemon, with the CTO as my supervisor.

1.2 Principal Activities of Host Organization

Intra Technologies is a start up that has projects that capitalise on opportunities in various markets.

Lemon serves to offer additional services to customers who hail ride sharing services such as Grab or Gojek. In their pilot trial, customers can buy various items from within the vehicle. Lemon is the main focus of the company now.

Rest is a service that allows customers to book Hotel Rooms in advance or with short notice from its platform to maximise convenience. Additionally customers can pay by the hour for said rooms, and can choose from many established Hotel partners that Rest has.

1.3 Training Programme within Host Organization

Intra Technologies provided various Udemy courses on Software Development for interns to learn using. For my projects I used the VueJS and Firebase courses.

2 Training Schedule And Assignments

2.1 Training Schedule By Month For The Entire Training Period

My training schedule is depicted in Table 1. My training schedule is not fixed as there are random task that popped up during my training period.

Table 1: Training Schedule

Task	Metric	Month
Create NLP pipeline to identify conflicts in chatbot dataset	RAM usage and computation time	Jan
Create mockups of dashboard and conduct user testing	Intuitive designs	Feb
Develop dashboard and integrate NLP pipeline in the backend code	Robustness and latency	Mar
Roll out first version of dashboard for government agencies to use	Robustness and latency	Apr
Collect user feedback and work on improving the dashboard	Speed and responsiveness	May
Deploy using Kubernetes and inhouse CI/CD system	Robustness	Jun

2.2 Training Assignments Completed in 1st Month

Due to the recent outbreak of Coronavirus in Singapore, my team was tasked to built a chatbot to disseminate information on the Coronavirus situation in Singapore as shown in Figure 1. It was done in 2 days. I was involved in training the chatbot on Dialogflow.

After which, I helped to redesigned the chatbot to give it a more "Telegram" feel. I also scraped information from MOH website to provide end users latest updates on Coronavirus situation in Singapore. Besides, to ensure that adding new training content does not reduce the chatbot's performance, I helped automated regression testing. Test cases are utterances that my team prepare on Google Sheet and the expected response are compared against the actual response obtained from DialogFlow APIs. The automation tool will pull these test cases from Google Sheet, tests the response with Dialogflow APIs, and populated the results back on Google Sheet.
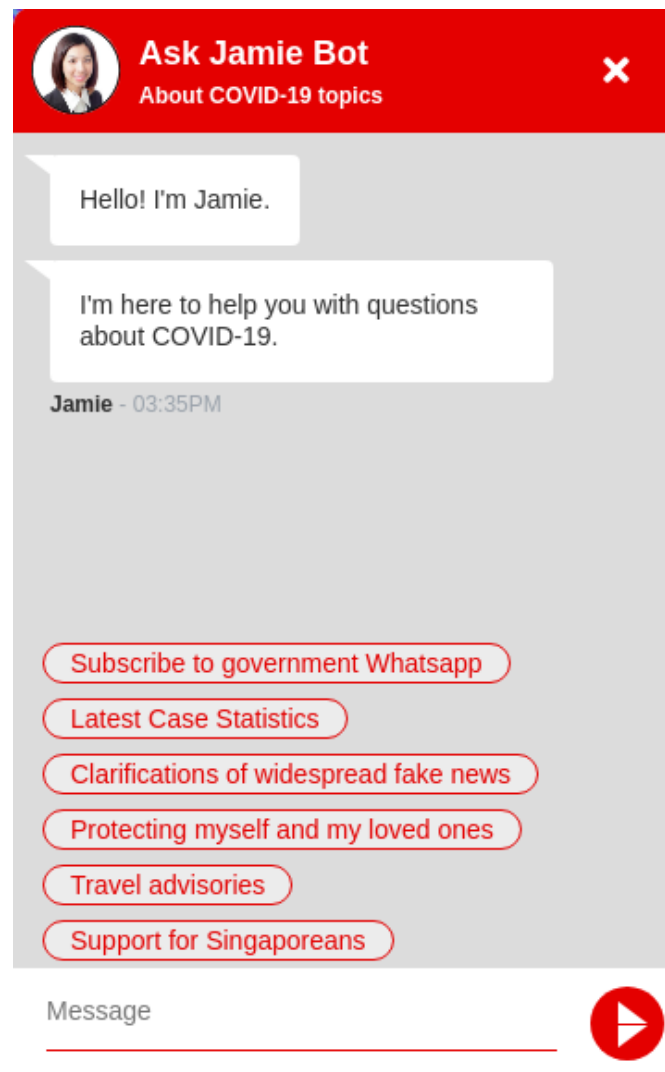

Figure 1: GovSG Chatbot To Disseminate Information On nCov

2.3 Training Assignments Completed in 2nd Month

One of my main task is to create a platform to provide end users an overview of the chatbot dataset and identify areas of improvement in the knowledge base. This platform is a web-based dashboard where user only uploads their dataset. After the dataset is processed, the user can go to the dashboard to view and understand the different issues in the chatbot dataset. Before commencing the software development works, I started by doing up mockups using AdobeXD for the dashboard and did user testing to ensure the dashboard is intuitive enough for end user. The mockup for the dashboard page is shown in Figure 2.
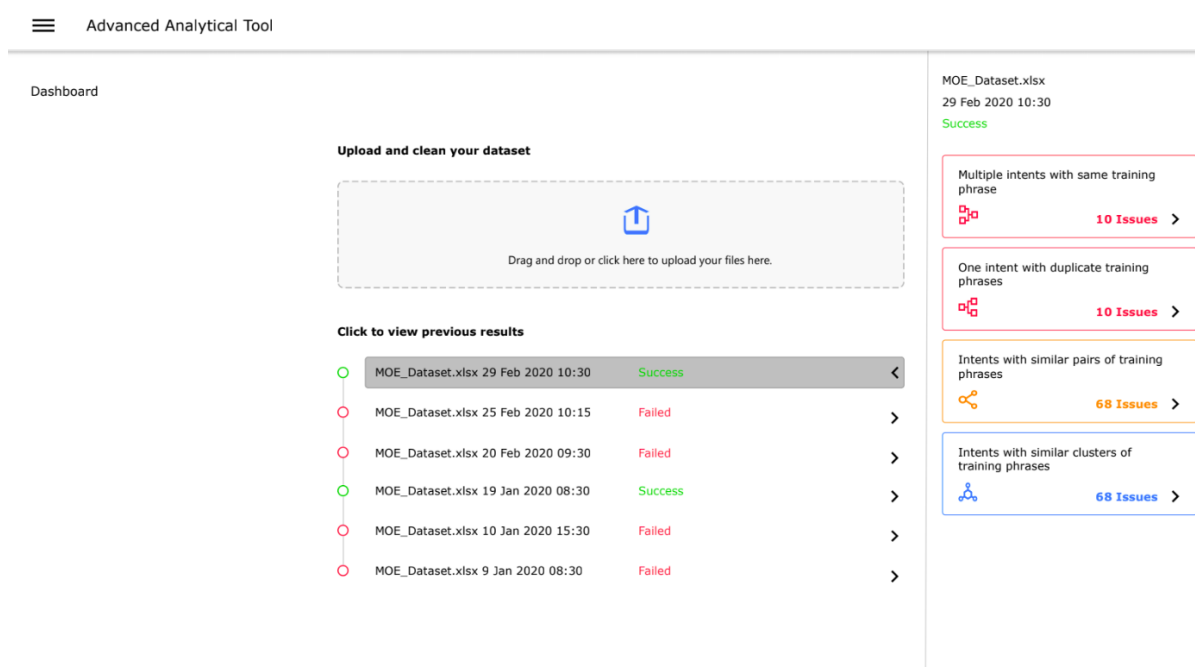

Figure 2: Mockups for Analytics Dashboard

My second main task was to setup and refine the NLP pipeline. The NLP pipeline is built to identify overlapping intents and ambiguous ones that affect the chatbot performances. I setup BERT and ALBERT to convert text into vectors of numbers (embeddings) that the computer can understand. The embeddings, theoretically, accounts for the semantic meaning. Cosine similarity and hierarchical clustering are performed on these embeddings to identify the problematic intents in the chatbot dataset.

As the chatbot dataset can be huge, I optimize one part of NLP pipeline so that the RAM usage

will not explode. I also provisioned a AWS EC2 Memory and GPU optimized instances to run my NLP pipeline.

The NLP pipeline helps to identify problematic intents in the training dataset. The source of this problematic intents is mainly a result of end users not understanding that a chatbot cannot handle ambiguous intents. Therefore, training phrases need to be specific. Secondly, end users may not have realized the intents already exist in the training dataset, and added new ones that results in a conflict. Table 2 shows an example of overlapping intents. My clustering tool in NLP pipeline grouped these set of training phrases together. It is obvious by looking at the training phrases alone, it is difficult to guess the intent. Similarly the chatbot would face the same difficulty that could have resulted in the chatbot's poor performance.

Table 2: Sample results of NLP pipeline

Training Phrases	Intents
Tax reference No Fxxx	How do I contact IRAS?
ref no Fxxx	GPF 58 What is the User ID? What is my tax reference number?
Tax reference no Gxxx	What is the User ID?

2.4   Training Assignments Completed in 3rd Month

One task I did was Topic Modeling. I use Gensim and Mallet to identify topics in positive and negative sentiments. These topic helped Product Owners understand the strength and weakness of current digital products for areas of improvements. A sample result of the topic modeling for negative sentiment is show in Table 3.

Topic modeling identifies a set of most representative keywords that could help Product Owners deduce the topic. I also find the most representative sentiment for each topic to provide context to these keywords. For example, Topic 1 in Table 3 suggests caching of user details so that user does not have to re-input the same details subsequently.

Another task I completed was to setup the backend code in Flask for the analytics dashboard. All the endpoints for the backend have been exposed. Sentry is setup to monitor for errors during

Table 3: Sample results of topic modeling

Topic	Keywords	Text
1	card, credit, detail, key, info, save, store, scan, regular, future	Re-input of credit card payment details each time. Maybe ca scan or take photo of credit card to capture details or save last transaction details.
2	card, detail, credit, key, remember, input, type, everytime, color, product	Remember my credit card number except my ccv number expiry date. At least I can input my card number w/o going to wallet downstairs to key in again.

runtime. PyTest was setup for unit and integration test. SwaggerHub was also setup for document the RestAPIs. RQ and Redis was also setup for the queuing system to run long computation tasks in the background. MongoDB was also setup to store results in the NLP pipeline. The backend code is also dockerized and deployed on AWS.

3 Knowledge And Experience Gained

3.1 Technical Knowledge Gained From Assignments

The first technical knowledge I gained was learning about NLP. I had the opportunity to learn different unsupervised clustering techniques, such as HDBSCAN, Agglomerative, and DBSCAN. I also had hands-on experience with Tensorflow and PyTorch to use BERT and ALBERT to convert text into embeddings. Another opportunity I had was to do Topic Modeling for positive and negative sentiments with one of the datasets using Gensim and Mallet. I later learned about using TextRank as an alternatives for Topic Modeling. One of the biggest challenges I faced is the memory usage and computational time. Currently, I have a fairly large dataset that consumes over 200GB of RAM for the unsupervised clustering and overs 6hrs of computation time for one of the NLP steps. I am still investigating how I can optimize the pipeline to handle large dataset more efficiently.

The second technical knowledge I gained was developing the web-based platform from scratch. I also had to do mock-ups and conduct user testing to check whether my designs would be intuitive for end users. Having taken CS3216, I wanted to apply the technologies and practices I learned from my seniors when I worked on projects with them. I am current using ReactJS with Typescript, Flask, and MongoDB. The application is deployed over Docker. As the platform is still in development, I may be able to explore Kubernetes and GraphQL later in the project for better support. Finally I also integrated a queue based system using RQ to managed processes that requires long computation time in the background.

3.2 Organizational/Industry Experience Gained From Assignments

The first industrial experience I gained was sharing my NLP results with government agencies. The very first presentation I gave was to MOE, but that was a last minute notice and I did not have any slides prepared. During the meeting, I could tell the audience could not follow. After that incident, I prepared a PowerPoint template to share my results that is easy to understand for end users. When I finally presented my results to MOE and IRAS again, there was a huge difference and the audience were asking questions that showed they understood my sharing. Most importantly, the

government agencies know how to use the CSV files to clean their chatbot's dataset.

The second industrial experience I gained was working with the team. My team has daily stand-ups which I was unused to at first, as I was clueless the first few weeks and did not know how to report my activities in NLP. After asking around and observing how others report, I was able to. My team also uses JIRA to track issues and also uses it for sprint planning. It took me about a week to adjust because I do not use such platform often in school. The most exciting part was emergency request to develop a chatbot within 2 days to disseminate information about masks due to the Coronavirus situation in Singapore. I first-hand saw how a single developer was able to build it in 2 days. Even though I training the chatbot through DialogFlow that 2 days, I helped out with redesigning the chatbot interface, scraping MOH web contents to populate the statistics into the database, and automating the regression testing for the training dataset.

The third industrial experience I gained was understanding the terminologies that government agencies used. In my first team meeting, the Director was present to supervise the team's progress. During the team meeting, there were many acronyms thrown that left me confused. I was not familiar with many of the government agencies and little did I know the purpose of each of these agencies. However, I was able to follow the conversations in subsequent meetings as I became more familiar with the terminologies.

3.3   Areas of Applicability of Knowledge And Experienced Gained

NLP is a niche field in Machine Learning and Deep Learning, but is growing rapidly. NLP has directly applications to any projects related to developing chatbot and understanding consumers' sentiments. The concepts and skills in Machine Learning and Deep Knowledge, however, is applicable to other fields, such as Computer Vision. As I hoped to develop a talking robot that can conversational capabilities of a human one day, I believe the NLP skills I gained at GovTech would help me in future.

4 Conclusion

4.1 Summary Of Work Completed And Training Received

The training I received was learning about the project's background and the skills in NLP.

One of the major work I completed is setting up the NLP pipeline that can easily integrate into the backend code.

The second major work I completed is the setting up the backend endpoints to fetch data from database.

Other works I have done include topic modeling, web scraping, user testing with mockups, queue system for pipeline, sharing NLP results, redesigning chatbot interface, automating regression testing, and training chatbot with DialogFlow.

4.2 Problems Faced

One problem I am investigating is how I can optimize the NLP pipeline to handle large datasets more efficiently as the current computation time and memory usage is unfeasible for long term usage.

The second problem I am exploring is to optimize the platform for production-ready software. I plan to integrate Kurbenetes and GraphQL into the platform in future.

4.3 Assessment Of Training Experience And Concluding Remarks

I enjoyed learning about NLP. In fact, I am learning how to use NLP for foreign languages, since I am also learning a foreign language currently. I hope to have opportunities to learn Natural Language Understanding (NLU) in chatbot if the project caters for it. Besides my current team, I would be joining another team to work on Robotics and Artificial Intelligence. I cannot wait to learn more about Machine Learning and Deep Learning in the other team.