Urjitkumar Patel

Master of Data Science, NYU

Machine Intelligence Research Intern

Summer 2016

ForwardLane

Powered by
IBM Watson™

Forwardlane Inc.
43 West 23rd Street, New York, NY 10010

Supervisor
Nathan Stevenson
CEO of Forwardlane Inc.
nathan.stevenson@forwardlane.com

ABSTRACT:

During my 3 months of internship period, I got a chance to work with two New York based startups, Forwardlane Inc. and Speerit Co.

At Forwardlane, my specific assignment was to enhance the Artificial Chat Agent interface which can understand Natural Financial Language. I specifically focused on how we can use abundance of natural language data for ennhancement of contextual, conversational Natural Language Interface using clustering techniques on raw data, Series of Machine Learning/Deep Learning Natural language classifiers and Big Data Similarity Algorithms.

At Speerit, I had to build everything from scratch as they did not have any Machine Learning or Big Data Platform. I started from AWS different server setups for Mongo, SQL, Python and Web Applications. I did the whole setup of Data Flow from various Health Trackers to user's application screen. I developed the algorithms to learn Runner's Profile Pattern, to find the Similarity Scores. They are using my scripts and algorithms in their Beta version's Production and soon going to it by end of this year. Due to size limitation of the report, I have not included any further details in this report on work done at Speerit.

COMPANY INTRODUCTION:

❖ FORWARDLANE INC., (http://forwardlane.com/)

ForwardLane scales personalized, ultra-high net worth investment advice to mass affluent investors using machine intelligence by empowering their trusted advisor. They have a passionate team of quantitative strategists, wealth management specialists and artificial intelligence experts with over 60 years of combined expertise who came together to change the shape of modern financial services for the benefit of both the client and the advisor. Their solution empowers the financial advisor – improving the quality, value and speed to access key data in turn enriching the level of advice and service they can provide. Through automation, aggregation and synthesis of data their solutions allow financial advisers to deliver a personalized and differentiated service.

Forwardlane Demo You Tube Video Link :
https://www.youtube.com/watch?v=zNXUyb3PB4A
https://www.youtube.com/watch?v=q_4fe9zJjBA

❖ SPEERIT CO. (http://www.speerit.co/)

Speerit's main objective is to help you get in touch with other runners in your area. They will also match your skill with other runners in order to develop your skills. If you are a novice runner looking for someone who can help you train, you can find one in Speerit's community. If you are an expert runner looking for other expert runners to train with, their community of runners will help you find one. If you are looking for local marathons and races to join, their community will match you up with the right people.

WORK DONE AT FORWARDLANE:

Introduction:

Our aim was to enhance the Contextual Chat Interface for the user. I scraped the Financial Questions data from various web resources and also used Yahoo Question Answer data set for our experiments. With these questions set we did our experiments on Clustering raw data, on building Natural Language Classifiers using Machine Learning techniques and Big Data Similarity Algorithms, which are the key tech modules of an AI Chat Agent.

●	Detailed Technical Report of my work	forwardlane-chat-agent.pdf

Business Problem:

Having an Artificial Chat Agent Which can interact with users can be very useful for any Financial domain company. If a company is able to provide an interface where user can just throw his/her query and get appropriate answer, it will surely increase user ease of access and will also increase user satisfaction. Just imagine yourself as a user, how happy you would be if you have a AI system, where you can ask, "Where should I invest?" and It will search through current market and on all current available policies and finds a best investment policy for you.

Dataset Description:

As there was no cooked data available, we had to scrap data from various web sources using web scrapping.

- Data Files Used:

1. CSV file containing 50000 questions and more than 69 different labels (scrapped from web)
2. Yahoo Question Answer set having more than 4.5 Million Question and Answers

Development Problem:

Development of Chat Agent is basically divided into four sub problems.

1. Clustering of Raw Data
2. Training of Natural Language Classifier
3. Similarity Algorithm
4. Answers Fetching Module

- ***Clustering of Raw Data:***

This was the most difficult and Important part of the whole development cycle. Initially we had 50000 questions with 69 different categories. We tried to separate out main categories such as Investments, Taxes etc. by applying K-Means, but we ended up with high overlapping between those clusters. As a work around, we found the important words from these small clusters using TF-IDF techniques and word cloud analysis. Once we filtered out important words, we grouped all those words into one corpus and performed the TF-IDF cosine similarity between these clusters. With the help from Finance domain expertise, we were able to merge down those 69 clusters into 7 main categories.

- ***Training of Natural Language Classifier:***

As part of Natural language classifiers, we mainly tried variants of two Machine Learning Algorithms Naive Bayes and SVM with different optimization techniques (Batch/Mini/Stochastic Gradient Descent), hyper parameters and vectorizer techniques (TF-IDF / Binary / Count Vectorizers). We improved our result with ensemble method using Random Forest on the top of hierarchical Natural language classifiers on first layer. We are currently trying to improve the accuracy using Data Augmentation techniques and Deep Learning Convolutional Neural Networks/Recurrent Neural Networks for text classification.

- ***Similarity Algorithm:***

We wanted to show top three matched questions from our questions database, once user asks the question to Chat agent. If we got user’s desired question in these top three matches, user can select that question. To get the top match, we used TF-IDF cosine similarity score and sorted the questions into Descending order. Initially as we won't be having all the important questions, there may be some cases where user will not find the desired question into the top three matches. That is why we have added 4th option for the user which is "No match found". If user does not find any good match, then we simply take this question and put it into the separate set of new questions. Business person can now work on these type of questions and update our current question answer database. One benefit here is that we don't need to fetch all the questions into our main memory, we just need to fetch particular cluster questions.

- ***Answers Fetching Module***

Once user selects a question, next step is to fetch the answer for it and display on the user screen. Algorithm for this is attached in the other paper for your reference

❖ Performance Evaluation:

For us, it was really very important that we don't misclassify user's question with wrong category. Otherwise we wound never able to find a good match questions from our database. Hence, we mainly focused on increasing precision rather than focusing on Recall. We used confusion Matrix to get the idea of TPs, FPs, FPs, and FNs. Our Main aim was to increase TPs/Decrease FPs.

Baseline Model Accuracy: 45.0 +- 2.5

Enhanced Model Accuracy: 85 +- 2

Ensemble Model Accuracy: 85+-2 (less Overfitting as I got approx. 88 accuracy on Train set with Ensemble)

❖ End Product Details:

Programming Language: **Python**, SQL
Web Scrapping: **Scrapy Python**
Approximate Number of code lines: **4000 - 5000**

All my work with Forwardlane has been done mainly in Python have created an interface where Anyone from Forwardlane can Train the classifier of their choice (i.e. NVB, Multinomial NVB, SVM, Linear Classifier, SVM with SGD etc.). They can check the accuracy of these classifiers and also use this classifier to chat with Chat Agent. There is another team working on absorbing my code as Micro Service and use it in their Node.js code for attractive User Interface.

How my work will contribute to Forwardlane?

When I started, Forwardlane was using only Third Party APIs and Services Such as IBM Watson, Thomson Reuters. They were highly dependent on those services which were also creating some limitations on their end. In my assignment, I have developed a code in Python which can replace these third party services, which are costly at the same time. Hence, with my code, Forwardlane now has full control on the Natural Language Processing Interface and also would save money by not using Third Party Services.

Challenges faced:

Till now, in our university projects, we had worked directly on the data available somewhere on web. But through this internship, I realized that in real world domain, getting data and make it compatible for the machine learning or deep learning algorithm is most important task.

When I started, Main problem was to prepare good training set of questions and answers for Natural Language Classifier. I searched around on web, found potential web sources and scrapped the data from web. Once we had this data, we now had to deal with data clustering. We potentially had data from more than 100 different categories. We had to cluster it down to main financial domain 7 to 8 categories. We tried K-means to separate these 7 to 8 categories, but we observed too much overlapping between K-means clusters. Finally, we solved this issue by finding TF-IDF based important words and then applied cosine similarity to find close clusters and then merged the clusters into main 7 categories.

FUTURE WORK AND CONCLUSION:

Currently Forwardlane does not have that much proprietary data and hence It would be really difficult to train Deep Learning's Neural Networks for the Classification. The data they have currently is more suitable for Machine Learning Methods. In fact, I am currently working on creating Deep Learning Platform and Code in Torch for them, which can be used later on once they have sufficient data to train a deep and more accurate networks such as CNNs and RNNs. I believe, Yahoo Question Answers data can prove really very useful for this task. Apart from that, with the ontology knowledge and abundance of data, it would be possible for them to deploy hierarchical classifiers which would be able to classify questions with category tree kind of structure.

In Conclusion, during this internship, I had been exposed to a data scientist working life in real world domain. I could understand more about the definition of a data scientist and challenges one data scientist can face in real world domain. I realized the importance of Data Gathering and Preprocessing for any data science problem. I got chance to connect and work with the Quantitative Analytics teams, Business side teams, and Software Engineering teams. I realized how important it is to have proper communication between teams and could understand how Data Science can contribute to all these domains within the organization. Through this internship, I developed leadership skills as well. I am happy that my proposed Chat Agent Algorithm was selected for the first prize in their Hackethon Competition. Technically, I gained a lot of experience in the domain of Natural Language Processing. I explored and read lot of research papers in this specific domain and educated myself with the latest techniques and methods in this specific domain. With the Machine Learning and Programming knowledge gained during my first two semesters at NYU, I was able to deliver them a Data Science Platform built in Python, where they can now train new Natural Language Classifiers with the new datasets in future and use them for the Chat Agent. Although I would like NYU Classes to focus more on real life data gathering aspect and Unsupervised Learning methods. In Sum, this internship had provided me the best platform to explore real data science work domain and would undoubtedly help me in future to face challenges in a working environment and would also help me thrive in my future organization.