

sddec18-16: Use machine learning to predict relevant support content based on historical user

Week 9 Report

April 1 - April 14

Team MembersErin Elsbernd — *Communication Coordinator and Machine Learning Lead*Ram Luitel — *Project Manager & Software Architect*Faizul Jasmi — *Testing & AWS Tech Lead*Khoa Bui — *Database & Web Master #2*Taizhong Huang — *Testing*Christian Chiang — *Web Master & AWS Tech Lead***Summary of Progress this Report**

This two-week we really focus on performance. We changed our strategies and work to increase the accuracy by grouping class labels and performing classification models in rounds on separate label groups. We also used the component that actually recommends the article based on various textual features. For now, the system gives you the topmost relevant "URL" to the article based on the input text. Based on holdout method, the average nearest match metric or distance is around 0.6 (in another word the accuracy is around 60%) which is quite good given the repetitiveness of the entire data (an only handful of articles/URL). Since accuracy is not a good metric for such system, we have used the individual distance score to evaluate the system's performance. Some of our team members are also working Google Cloud for data cleaning.

Pending Issues

We are expecting some more data from our client sine we really do not have enough data point as of now. Since we have not hit that 70% accuracy prediction requirement for our clients we will continue to work using different machine learning model and different techniques with the different feature to come up with better results. We are also looking alternative data cleaning and processing tool by using Google Cloud.

Plans for Upcoming Reporting Period

Keep working on models to improve performance. We changed our strategies and work to increase the accuracy by grouping class labels and performing classification models in rounds on separate label groups. This shows some improvement in performance so we will keep working on this. We also have our last meeting of the semester with our client and faculty advisor where we will present our result and plan for next semester.

Individual Contributions

Team Member	Contribution	Weekly Hours	Total Hours
Erin Elsbernd	Worked on increasing accuracy by grouping class labels and performing classification models in rounds on separate label groups. Also worked on classification model to group class labels into larger class label groups, and run random forest models on that. Also created ROC plots and metrics to gauge	12	73

	performance of models		
Ram Luitel	Keep working on models to improve performance for article recommendation. I used the component that actually recommends the article based on various textual features. For now, the system gives you the topmost relevant URL to the article based on the input text. Based on holdout method, the average nearest match metric or distance is around 0.6 which is quite good given the repetitiveness of the entire data (the only handful of articles/URL). Since accuracy is not a good metric for such system, I have used the individual distance score to evaluate the system's performance. I also try to predict the title based on event sequences and related articles based on contents and categories.	10	62
Faizul Jasmi	Get together with Christian to read, research and test Google Coud machine learning tool that Alex suggested.	4	45
Khoa Bui	I am learning how to process data, how to clean and extract features. Try writing code for some machine learning models using the features that I extract. I am also exploring more in AWS side of the project which is really useful next semester.	5	49
Taizhong Huang	Tried to add some new features to the improve the model. had some issues with my developing environment and git and tried to fix them.	7	51
Christian Chiang	Suggested by our client Alex, he wishes to use Google Cloud, focusing on the AI side. This feature will allow us to clean the data continuously without the need to run our current python script. More research into this platform is needed. Testing and small models have been done on my personal computer	5	47