horizontal line

Sprint 3 Results

Sprint End: 11.03.2019

Sprint Captain: Mianchun Lu

Sprint Co-captain: Hongying Liu

**─**

# Sprint Goal

## Optimize Neural Nets and the Train Function:

* Modify the dataset and train function for better performance:

1. Split the validation dataset into validation and test dataset
2. Add early stopping method in the train function based on validation loss to avoid overfitting
3. Add test function to test the model on test dataset and calculate F1 score and auc-roc curve

* Generate additional features such as POS tags and add into the model
* Add more layers into the neural nets

## Snorkel Optimization:

* Continue adding more labeling functions to Snorkel to see if we can improve its performance. This week we will focus on clustering, topic modeling, and POS tags.

# Results

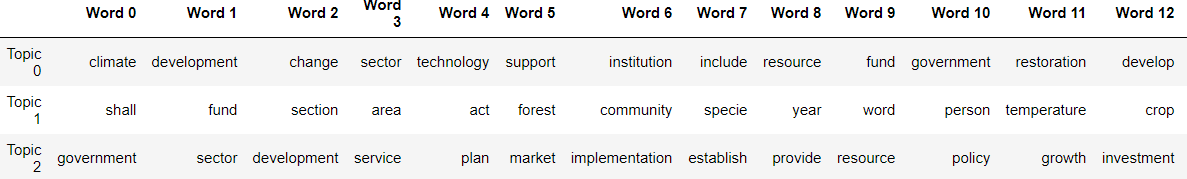
* Neural Network Model:

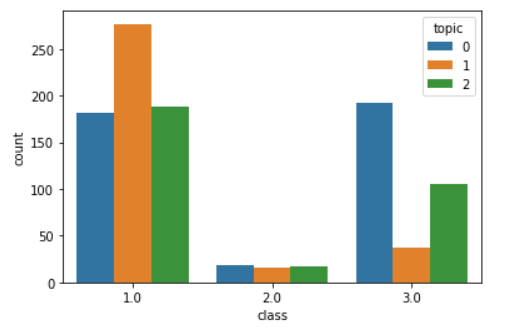
We split the original validation dataset into **new validation dataset** and **test dataset**. The new validation dataset was used to test the overfitting problem, and the test dataset was used to calculate the F1 score. And a simple stopping method, which stops the calculation when the loss of validation data is not improving, was added to avoid overfitting.

Then, we tried two neural network models. In one model, we added an additional LSTM layer. In the other model, we trained the model with the new feature after topic modeling(dominant\_topic). The F1 score was improved. The current accuracy score is 0.586.

* Feature Engineering

We did LDA(latent Dirichlet allocation) with sklearn package, and grouped the sentences into three topics/clusters after grid search. The dominant\_topic column was added to neural network model as a new feature. The top 13 most important words in each topic, and the distribution of sentences in each class are shown as follows.





* BabbleLabble implementation

We have figured out the rules of writing certain babbleLabble explanations, but got stuck on how to transfer our csv data to pkl file, as Babble only takes in the pkl format data. We tried many ways but still not sure how to achieve this transformation. The pkl file in the babble tutorial contains specific data structures and need further explanations and help on this.