Innoplexus Online Hackathon

1. A brief on the approach, which you used to solve the problem.

As given in the problem statement this is a Named Entity Recognition problem. I have worked on problems with text/unstructured data before but never in a classic NER problem. You can either ignore the sequential aspect of words in a sentence and train a classifier based on the labels but I got really low score since ‘O’ label is predominant. So decided to use CRF(sequential version of logistic regression) or LSTM to solve the problem. I came up with the LSTM model but spend most of the time reshaping the data to desired input format and couldn’t get any value out of it. Used just CRF model in the end

1. Which data-preprocessing / feature engineering ideas really worked? How did you discover them?

First I joined the words to get sentence for each sentence ID. Then added a POS\_TAG feature using nltk for each word. Then preprocessed the dataframe into tuples of (word,pos\_tag).

To capture the sequential information added (word-2,word-1,word+1,word+2) and (pos\_tag-2,pos\_tag-1,pos\_tag+1.pos\_tag+2)

To get more features of the current word added features like is the word capitalized, punctuated, camelCase, contains hyphen, contains dot , all uppercase, all lowercase, is titled etc..

Final submission include a set of features from the above

1. How does your final model look like? How did you reach it?

I prepared my data as described above and used crf\_suite from sklearn package. I pretty much followed the tutorial given in the package documentation.

I checked the distribution of train and test set using simple analysis like number of sentences, length of sentences and also words. All of them pointed that both were similarly distributed. I used 5-fold cross-validation and flat\_f1\_score metric from crf\_suite to validate my model.

I tuned manually the regularization parameters of the model

1. What are the key takeaways from the challenge, if any?

As I said, this was the first time I was solving a NER problem. I learned more about text processing and corresponding feature engineering. It was a good challenge and if there was time I could have explored LSTM.

1. According to you, what are the 5 things a participant must focus on while solving such problems?
2. Understand the problem and quickly identify the domain.
3. Understand the evaluation criteria
4. Analyse train and test distributions and come up with good validation set that matches test set. Timeseries CV , grouped CV or KFold
5. Feature engineering. In competitions often the differential is a feature based on domain knowledge. Should spend most of the time to extract such features
6. Choosing appropriate modeling technique and also getting more diverse models if time permits.