TEAM - *PERO_CODERS*

Data Cleaning

The sentences in ‘TITLE’, ‘DESCRIPTION’ and ‘BULLET_POINTS’ were cleaned. The word2vec google pretrained model was used which had a vocab size of 300000. The words which do not come under this vocab were deleted from the dataset. Some additional cleaning was done to work with special characters. These words were tokenized using keras tokenizer. Different length of sequences were taken for different columns.
‘TITLE’ column was padded to a length of 25 (99 percentile).
‘DESCRIPTION’ column was padded to a length of 125 (90 percentile).
‘BULLET_POINTS’ column was padded to a length of 90 (90 percentile).
They were all concatenated to get a sequence of 240 length which is the input.

Modelling

Due to the lack of computational power we were unable to train the state-of-the-art NLP models. What we tried out is a simple LSTM based network followed by dense layers.
Model 0 - Word2Vec pretrained embedding + 1 LSTM 250 non-sequence output + dense layer (10000 output units, with relu activation)+ dense output layer of 9919 units and a softmax activation. 3500 batch size. Got validation accuracy of 75.4.
Model 1 - Word2Vec pretrained embedding + 2 LSTM layers (one with sequence output 200 dims and another without sequence output of 200 dims) + Dense layer (1000 output and relu) + Dense output layer of 9919 outs and a softmax activation. 2048 batch size was used. Got validation accuracy of 73.8.
Model 2 - Same as model 1 but used 250 dims as output for the LSTM layers instead of 200. Batch size of 1024 was used. Got validation accuracy of 75.0.
Model 3 - Same as model 0 but the LSTM layer was followed by an attention layer as well. Used a batch size of 1024. Gave a validation accuracy of 76.2.

Ensembling

From 4 different models top models were selected for ensembling. First ensembling was done within best solutions of one model and them amongst the model. Based on validation accuracy, top 6 outputs from model 0, 1 output from model 1, 5 outputs from model 2 and 3 outputs from model 3 were selected.
For ensembling we used mode of all the browse node ids as answer, if all values are different then values of output with highest validation accuracy was used.
This was carried out within model 0,1,2 and 3 to get one best output from them. Then on these output another ensembling was done to get final output.