Assignment 2

Task 1:

1. For this task, Grid Search based automatic hyper parameter search was used. Considering we have limited parameters and a fixed range of values to search upon for each parameter, I believe Grid Search is fast and efficient. The hyper parameters considered for tuning were C, max iterations, solver and Penalty. As suggested in Lab, 3 to the power of K with K values ranging from -5 to +5 was used as range for C. C should be a positive value and with this range, we were able play with values less than 1 till 243.
2. The best performing C value was 27
3. The final accuracy was approximately around 0.85

Task 2:

1. As first step, the data is read from json file and split into train and validation dataset in the ratio of 9:1. From my observation, the data is unbalanced. In order to prevent its effect overfitting towards the data with high frequency, while splitting the data the data distribution was maintained. Now each string from the data is read individually, and tokenized using TreebankWordTokenizer library. Once tokenized, each string is now vectorized using TfidfVectorizer with max features in a vector set to 10000. The class labels were One hot encoded in order for training purposes. Now with the vectorized data and encoded class labels we train the logistic regression for multiclass classification. This is a one vs rest model. And the parameters for these model were C=1000, max_iter = 250, solver= saga.
2. We followed different split strategies for dataset – 9:1, 8:2, 7;3. The hyper parameter tuning was carried away using Grid Search, and the best set of parameter were found and fixed.
3. We carried testing on several methos which were different parts of the pipeline of this task. I categorized them, and discussed them as separate topics underneath.
   - Pre-Processing: Several pre-processing methods were used for implementation purposes and unfortunately nothing worked best. Few of them include lower case conversion, removal of digits, removal of stop words, removal of punctions and special characters.
   - Vectorization: Two different vectorization methods were used for testing. One is TF-idf vectorization and the other one is Word2Vector. In Word2Vec, two methods were used. The first one uses selected vocabulary provided in Lab which were extracted from ‘en_core_web_md’, and the second one was using vectorization created using ‘‘en_core_web_lg’ with the help of spacy library.
   - Classification models: Three different models were used for testing. First one is the regular logistic regression which is a linear model. The second one is a sequential non linear model created using keras library which uses Neural Network. The third on uses LSTM which is also a non linear model
4. Major reason would be the data availability. Considering the range of features that we are training and the amount of data we are testing it with, the less accuracy infers that neural networks aren’t efficient enough when trained with less data and this is a evident fact.

Task 3:

Argmax:

John Brenner

Random:

Valer

Jolam Rober

Dann Jlvenk Mikehhes

Pais Kyrinn

Gatk Kart

Arn Setzer

Woskian Rendomo

D T Cohndere H Celbe

JoVere Lerr

Danddept Tonbittan