BD_054_122_543

Team Members:

Amit Patti – PES1UG19CS054

Chandan M – PES1UG19CS122

Tushar Kalaskar - PES1UG19CS543

Spark Streaming using Machine Learning

Dataset Chosen: Sentiment Analysis

Design Details:

Cleaning, Preprocessing, training, and storing the model. 2 models: SGD Regressor and Random Forest Classifier.

Surface Level Implementations:

In our design we cleaned the dataset using regular expressions i.e., regex. Preprocessed the data using tokenizer, we then removed stop words and did hash vectorization to the dataset.

We Built 2 models for our implementation: SGD Regressor and Random Forest Classifier. We trained the model and stored it in a pickle file so that it can be used for the testing dataset.

SGD Regressor:

We start by tokenizing our tweet and then removing the stop words from our tweet, since the stop words aren't important. We then calculate the hash values and train our model using the train.csv file.

We store our model as .pkl file and run our model on test.csv file and obtain the accuracy of our model.

Random Forest Classifier:

We start by tokenizing our tweet and then removing the stopwords from our tweet. We then calculate the hash values and train our model using the train.csv file.

We store our model as .pkl file and run our model on test.csv file and obtain the accuracy of our model.

Here we kept the random_state doesn't affect the dataset because we are training the whole dataset.

Reasons behind Design Decisions:

The data given are in batches and needed to run each batch and learn from each batch. Hence, we needed incremental learning models to process and analyse the data.

Therefore, we have implemented incremental model using partial fit method of scikit learn.

Our dataset is Sentiment, therefore we used regex cleaning to remove unwanted characters from the tweets.

We then pre-processed using tokenizer i.e., splitting the words in a line and removed stop words as they do not affect the sentiment of the tweet.

We then used Hash Vectorization to scale words to numeric data to use it on logistic regression model.

The model takes these numeric vectors and performs the analysis.

Takeaway from the project:

We needed to learn how incremental models work and implement them for the data given. This was a tough experience but definitely learnt a lot about the amalgamation of Data Analysis of huge data which we believe is very fascinating and important in future times.