**OPTISOL DATALABS TASK**

**NAME:** Manassa Varshni S

**POSITION:** Machine Learning Engineer

**COMPETITION CHOSEN:**

Question Answering – https://www.kaggle.com/ananthu017/squad-csv-format

**ABOUT THE DATASET:**

SQUAD, Stanford Question Answering Dataset, is a reading comprehension dataset, consisting of questions posed by crowd workers on some Wikipedia articles. The answer for every question is a segment of text or span, from the corresponding passage. There is a context (paragraph text), question and text (answer text) for each observation. This dataset is a closed dataset that is the answer to a question is always a part of the context.

**PROBLEM:**

The main goal of the problem is to find the text (answer) for any question and context provided. For solving this problem I have taken two major steps that is to be followed, that is:

* Extracting the sentence that is having the correct answer from the context.
* Getting the correct answer from this sentence.

**METHODOLOGY AND REPORTS:**

1. ***Infersent sentence embedding method* :**

* *Infersent sentence embedding* technique developed by facebook is used.
* Just like Sentence-BERT, it uses a siamese network, but instead of BERT, it utilizes a bi-LSTM, a neural network with memory, to remember the whole sentence to encode.

1. ***Unsupervised Learning:***

* *Euclidean distance* was first used to detect the sentence having minimum distance from the question. The accuracy of this model came around 45%.
* Then, *cosine similarity* is used and the accuracy improved from 45% to 63%.

1. ***Supervised Learning:***

* The target variable form text to the sentence index having that text is transformed. The paragraph length is restricted to 10 sentences (around 98% of the paragraphs have 10 or fewer sentences), thus retaining 10 labels.
* For doing *Dependency Parse Tree* feature, Spacy tree parsing has been used since it has a rich API for navigating through the tree. This improved the accuracy of the model by 5%.
* Multinomial logistic regression, random forest & gradient boosting techniques have been used.
* The accuracy of *multinomial logistic regression* is 65% for the validation set.
* The *random forest* gave an accuracy of 67%.
* *XGBoost* worked best with an accuracy of 69% on the validation set.

**NOTE:**

The first file Sentence\_Embedding.ipynb takes care of creating a dictionary of sentence embedding for all the sentences and questions in the Wikipedia articles of training dataset.

The second file Unsupervised\_Learning.ipynb calculates the distance between sentence & questions basis Euclidean & Cosine similarity using sentence embeddings. It extracts the sentence from each paragraph that has the minimum distance from the question.

The last file treats this problem as supervised learning problem where multinomial logistic regression, random forest and xgboost are fitted.