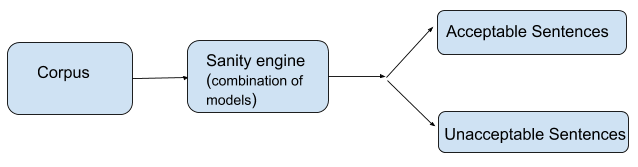
**TARENTO TECHNOLOGIES**

**Corpus Sanity Checker**

**Objective:**

Develop NLP model or combination of models to ascertain the sanity( grammatical acceptability) of sentenes using various NLP/Deep Learning strategies

**Task:**

* Build NLP model which takes english sentence as input and outputs whether it is acceptable or not based upon the grammatical correctness of the sentence
  + Take the following examples:
    - The Secretary General of this Court is directed to depute an Officer to initiate the prosecution in a competent Court having jurisdiction at Delhi.

“For this sentence the model should output as correct (label= 1)”

* + - if the arbitrator wanders outside the contract and deals with matters not allotted to him

“For this sentence the model should output as incorrect( label = 0).

The above sentence has the starting words missing and also the ending is not grammatically complete. Hence the model should output this as incorrect/ unacceptable sentence”

* Sentence will be unacceptable if it is incomplete, has missing words which deteriorates the meaning or has spelling mistakes which again hampers with the meaning/context of sentence.
* The acceptability should be customizable based upon the prediction/logit scores (eg. increase the acceptability threshold to 0.7 probability score)
* For anykind of Training which require GPUs, Google Colab can be used
* Any one of Pytorch/tensorflow/Keras framwrok can be used to code/train/deploy model
* Serve your model on Flask which performs sanity checking at sentence level as well as file level (text file which contains sentences per line)

**Task Evaluation and Criterion:**

This task will be evaluated by the following metrics:

* Approach
  + How do you design your problem before attempting anything.
  + How do you shortlist various architecture/model type to solve your problem.
* Coding
  + Including how you serve your models, your coding style and optimisation of code
* Model Performance
  + The model performance has to be tested against the test set included in the assignment which is a combination of COLA out of domain dev set plus custom scrapped sentences (test\_data folder)
  + Publish your score of correct prediction on above test set, i.e how accurately your model can identify correct as well as incorrect sentence. There is no hardline on accuracy percentage however if your model is able to correctly predict more than 50% of sentences, its good to go. Publishing result in the form of Confusion Matrix is all needed for this assignment.
* Comment on how this model can further be improved. What are current limitations and how you can overcome those.
* Kindly push all your code and models into the google drive and share the link as submission

**Reference and Hints!**

* Pretrained models like BERT/GPTs can be customised to fit your task (need not always train things from scratch ), but you should have the answer of why you used that.
* Avoid readymade online solution. Your approach is what we are looking for.
* Keep sepearate ipynb for training on Google Colab, apart from flask inference module
* Data set from COLA repository can be used to train/fine-tune any NLP model, if required. Use below link to get training dataset
* <https://nyu-mll.github.io/CoLA/>
* Evaluation dataset is attached with this assignment

Good Luck!