# NLU Course Project [01/04/2022]
# CSL7340

## Instructions:

1. **Evaluation will be done through viva where you have to demonstrate your model live, along with reviews/code.**
2. **Total marks - 20 marks, Deadline:- 28-04-2022 05:00 PM,**
3. **ANY ONE MEMBER OF THE GROUP CAN SUBMIT AND TURNIN. No need for other people to submit/turn-in.**
4. **All libraries and tools are allowed.**
5. Submit a **.zip** file containing all the working codes (.py files and .pdf file). The zip file should be named in the format <RollNo1_RollNo2_RollNo3_NLU_A1>.zip.
   **[Do not include model weights and data in the zip file]**
6. Submit a **.pdf** report which should contain:
   a. A detailed description of what all you have done,
   b. A description of what else could be done to improve results - challenges faced
   c. Links to the Google-Colab files (if any),
   d. Clearly mention the contribution of each group member.
7. **Copying from the Internet and/or your classmates is strictly prohibited. Any team found guilty will be awarded a suitable penalty as per IIT rules.**
8. **Models have to be saved and run during the demonstration - model training should be complete and no training should be initiated during the demonstration.**

# Project - Choose any one of the following projects

## 1. Relation Classifier:

**Dataset:-** KnowledgeNet (data description and dataset can be found here:- link), use train.json inside dataset as your source of data.

### Tasks:-

1. Review 3 recent papers (that use DL techniques) for relation extraction and make a comparative study containing approach differences, results and evaluations, dataset etc.

2. Building a relation classifier using any one of the above methods. The classifier will take in a sentence (maybe along with additional markers like Named Entities etc.) and predict the relation in a triple format. The triple format relation associated with each sentence is to be considered as the expected output of the classifier.
3. Create a subset of the KnowledgeNet data using sentences containing the following relations only: (make a subset of train.json with these relations only)
    a. DATE_OF_BIRTH (PER–DATE)
    b. RESIDENCE (PER–LOC)
    c. BIRTHPLACE (PER–LOC)
    d. NATIONALITY (PER–LOC)
    e. EMPLOYEE_OF (PER–ORG)
    f. EDUCATED_AT (PER–ORG)
4. Make a split of 80:20 of this data for training and test purposes.
5. We have held-out test sentences at our end. We will share this test set 48 hrs prior to the submission deadline. You have to run your relation extraction model on this held out test set, predict the relation triplets (subject, relation, object) and submit a .csv file for the same containing these triplets. (.csv file guidelines will be shared later)

## 2. Translation from English to Indic language

**Dataset:-** Choose any corpus (en-x, x can be any Indic language) from here. (if you are choosing this project, you have to add your dataset choice in this sheet) not later than 15th April 2022.

**Tasks:-**

1. A detailed review of at least three papers presented in NIPS / ACL / KDD / COLING / NAACL / conference of similar tier over the last 3-4 years - that addresses the task using a DL architecture.
2. Implement a transformer-based encoder-decoder architecture for solving the task.
3. Discuss the evaluation metrics used to judge the performance of the model, and show the model performance using these metrics. Comment

on the model's performance. Compare your results with the papers reviewed.

4. Make clear documentation of the same along with model-related information like architecture, training, validation and test splits, hyperparameters choice (and appropriate reasoning), and any other design considerations made, shortcomings of the model, limitations etc.
5. Show some examples where the model has given correct translations as well as some wrong ones.
6. The model will be tested with sentences given during the demonstration.