**Capstone Project: Machine Translation**

**Milestone1: Project Proposal**

***Problem Statement***

Design a Machine Translation model that can be used to translate sentences from German language to English language or vice-versa

***Scope***

MT is a hot topic in NLP and Deep Learning field, because of the shear number of languages that exists and limited ability of humans to learn different languages. In India alone, there are 22 official languages, but only 2-3 languages are taught in schools. Therefore, using advance technologies and data-driven approaches to develop tools like Google Translate can help remove the linguistic barrier to connect people across the world.

We are going for modular approach. We will develop modules that can be simply imported by user, keeping the complex code abstracted and making it nice and easy to use.

import machine\_translator as mt

conv\_text1 = mt.translate\_eng\_to\_de(“../Data/german.txt”)

conv\_text2 = mt.translate\_de\_to\_eng(“../Data/english.txt”)

print(conv\_text1, conv\_text2)

***Data***

We are provided with total 6 text 3 of which contains statements in German/ Deutsch and other 3 contains the translated version of same statements in English. Thus, there are 3 sets of EN-DE pair. Everything will be combined in a single dataset and shuffled. Respective sources are Europarl corpus, Data Crawl corpus and News commentary corpus. From final set, 70% of the data would be used for training, 25% of the statements would be used for testing and 5% can be kept for validation. For German to English, we will take German sentences as input, map them to the respective English statements and train our model and vice-versa.

***Plan***

Milestone 1 –

* Get to know each other in the project group.
* Go over the project statement, try to understand scope and getting familiar with it.
* Revise the concepts on MT and NLP by going over the provided lecture material and jupyter-notebook in the GL course.
* Prepare project proposal for first milestone submission.

Milestone 2 –

* Start on implementation by importing significant packages, loading dataset in NB.
* Explore on the dataset and segregate the relevant dataset for ENG-GERMAN MT problem. Work on EDA to understand the data better and prepare it to work with ML NLP model. Prepare Training, Test and Validation Dataset.
* Work on literature review by exploring and researching on various works done on the similar problem statement. Try to understand the emerging concepts and new methodologies to solve the similar problems.
* Filter out the parts which are applicable for given problem and can be utilized for our Capstone project, given the scope, bandwidth and timeline available.
* Prepare interim report with the help of literature review done and results achieved.

Milestone 3 –

* Create a basic machine translation model to get an idea how machine translation works.
* Try to improve models with the findings in literature review and integrate them with the implemented model. For Ex: Transfer learning, if applicable, can be used here.
* Study the results obtained and compare them with each other wrt performance parameters.
* Enhance the project report with the work done and results obtained.

Milestone 4 –

* Finalize the model and finish with coding implementation along with code documentation (comments, reasons and conclusions).
* Consolidate and compare all the results and conclude with our research findings along with citations. Prepare the detailed report with step-by-step improvements.
* Prepare the detailed presentation with the key findings and results in an easy-to-understand format.
* Prepare for the final presentation delivery and Report submission.