IIITD

CAPSTONE PROJECT

NATURAL LANGUAGE PROCESSING

MACHINE TRANSLATION ASSIGNMENT

PROBLEM STATEMENT

- **DOMAIN:** MACHINE TRANSLATION.

- **CONTEXT:**

  Machine Translation is the automated translation of source material into another language without human intervention. The database comes from ACL2014 Ninth workshop on Statistical Machine Translation. This workshop mainly focusses on language translation between European language pairs. The idea behind the workshop is to provide the ability for two parties to communicate and exchange the ideas from different countries.

- **DATA DESCRIPTION:**

  The database is basically sentences in German/English of various events. Three datasets are obtained from Statistical Machine Translation workshop. Either the dataset can be downloaded from the link or can be used from the shared files. Three datasets are,

  - Europarl v7
  - Common Crawl corpus
  - News Commentary

  Link to download the dataset: https://statmt.org/wmt14/translation-task.html

- **PROJECT OBJECTIVE:**

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

- **PROJECT TASK:** [Score: 100 points]
  - **Milestone 1** [10 Points]
    - Step 1: Project Proposal - Scope, Data and Plan [10 Points]
  - **Milestone 2** [15 Points]
    - Step 2: Import and merge all the three datasets. [2 points]
    - Step 3: Data cleansing and EDA [3 points]
    - Step 4: NLP pre processing - Dataset suitable to be used for AIML model learning [5 points]
    - **Submission**:
      - Initial report along with the notebooks in milestone 2, challenges and proposed solution [5 Points]
  - **Milestone 3** [15 Points]
    - Step 5: Design, train and test simple RNN & LSTM model [4 points]
    - Step 6: Design, train and test RNN & LSTM model with embeddings [5 points]
    - Step 7: Share your insights on the results obtained from steps(5-6) [1 Points]
    - **Submission:**
      - Interim report along with the notebooks in milestone 3 [5 Points]
  - **Milestone 4 (Final Submission)** [40 Points]
    - Step 8: Design, train and test bidirectional RNN & LSTM model [8 points]
    - Step 9: Choose the best performing model and pickle it [2 points]
    - Step 10: Load the pickled model and do the prediction [3 points]
    - Step 11: Share your insights on the results obtained from steps(8-10) [2 Points]
    - **Submission:**
      - Final report along with the notebooks in milestone 1, 2 & 3 [25 Points]
  - **Presentation**: [20 Points]

- Hints:
  - Please refer to the research papers to understand how to Machine Translation: https://statmt.org/wmt14/papers.html
  - Reference: https://www.mygreatlearning.com/academy/learn-for-free/courses/machine-translation

POINTS TO REMEMBER

1. A maximum of 100 points will be awarded for this project
2. Project to be submitted as per the deadlines from the date of release. Late submission will be accepted under genuine situation. Score will be given as per the below formula:

   If the current score is greater than 40 then the final score will be capped at 40.

   Else the current score will be awarded.
3. Any form of plagiarism is strictly prohibited. No score will be awarded in this case.

HAPPY
LEARNING!