Project Status Report

- **Project Topic:** Recurrent Neural Network for English-to-Spanish Translation

- **Team Members**

Aman Bommena	(axb220065@utdallas.edu)	Namrata Nag	(nxn230019@utdallas.edu)
Christopher Fan	(clf170000@utdallas.edu)	Shaan Sekhon	(sss152530@utdallas.edu)
Aditya Kulkarni	(axk230069@utdallas.edu)	Dat Nguyen	(dtn190004@utdallas.edu)

- **Technique and Algorithms**

  a. Main Algorithm: Recurrent Neural Network
  b. Text encoding/ embeddings generation algorithms
  c. Text preprocessing techniques

- **Dataset Details**

  - Anki Tab-delimited Bilingual Sentence Pairs: English to Spanish
  - This dataset is part of Tatoeba Project, Tatoeba is a large database of example sentences translated into many languages by its members who volunteer their time. Many of the sentences came from the Tanaka Corpus which was imported into the Tatoeba Corpus.
  - Dataset Link: https://www.manythings.org/anki/
  - Format: English + TAB + Spanish + TAB + Attribution
  - Sentence count: 141543.
  - Maximum length of English Sentence: 70
  - Maximum length of Spanish Sentence: 68
  - Minimum length of English Sentence: 1
  - Minimum length of Spanish Sentence: 1
  - Known Issue: The accuracy of translation is dependent on the distribution of words which is not the same due to languages favoring different translations. More common words are translated properly, while rarely used words may not be translated with same accuracy.

- **Languages and Libraries**

  a. Programming Language: Python
  b. Libraries: NLTK, scikit-learn, numpy, pandas, matplotlib, seaborn, keras, tensorflow