**Kaggle Project 11**

**US Patent Phrase to Phrase Matching**

**Prepared by: Dwaipayan Mukherjee (2211569, AIA – AIML)**

About the Project:

In this dataset, we are presented pairs of phrases (an anchor and a target phrase) and asked to rate how similar they are on a scale from 0 (not at all similar) to 1 (identical in meaning). This challenge differs from a standard semantic similarity task in that similarity has been scored here within a patent's context, specifically its CPC classification (version 2021.05), which indicates the subject to which the patent relates.

Steps that have been followed:

1. Necessary imports of NumPy and Pandas have been done.
2. Textacy and Spacy has been downloaded.
3. All necessary libraries have been imported.
4. After loading the data, some columns have been removed. These include the **id** and the **context** column.
5. All words are then made into lowercase letters.
6. Stopwords, which do not add any semantic meaning to the text, are then removed.
7. Hashtags and punctuations are then removed.
8. Next, the Snowball Stemmer is used to stem the words to their roots.
9. The text is then converted to vector form using the TF-IDF vectorizer algorithm.
10. Next, the cosine similarity is calculated between the target and the anchor.
11. Then the Levenshtein distance has been calculated between the target and the anchor.
12. Next, the Jaccard score is calculated between the target and the anchor.
13. Lastly, the Euclidean distance is calculated between the target and the anchor.
14. At the end, all these scores have been given alongside the target and the anchor.