**Assignment -1**

**Due Date: 22nd March 2023** **Max.Marks:12.5**

Determining suitable vector representations for words is very useful feature extraction step to effectively represent a document in semantic space. It can help in solving problems related to information retrieval, document classification, question answering, named entity recognition, parsing etc.

This assignment will help you to learn two type of word embeddings viz. frequency-based and prediction-based embeddings and to compare both type of embeddings. We have discussed some of the word vector representation techniques like n-gram, skip-gram, TFIDF, co-occurrence etc. from frequency-based and Word2Vec, GloVe etc. from prediction-based embeddings in the class. In this assignment you are required to do the following:

**1. Selection:**

Select a suitable dataset e.g. Reuters Corpus of news articles from the text corpora of nltk (details are given https://www.nltk.org/book/) or SimLex-999 (https://www.aclweb.org/ anthology/J15-4004 ) dataset.

**2. Implementations:**

(a) Implement for generating Co-occurrence based word embeddings with fixed '*k*' size window i.e. include *k* words on both sides. Prepare a word-word Co-occurrence matrix and reduce dimensions using SVD to get *d*-dimensional embeddings. (SVD tutorial is uploaded for reference)

(b) Implement Word2Vec/GloVe word vector representation method (do not use built-in functions).

[40%]

**3. Investigations:**

(a) Illustrate examples for semantic relationships such as school-student, India-Delhi, India-rupee, brother-sister, etc and syntactic relationships such as rupee-rupees, day-night, good-awesome, great-greater, write-wrote, etc. among the words using the embeddings obtained in 2(a) and 2(b).

(b) Experiment with different length word embeddings and find the appropriate size of the vectors.

(c) Plot words using 2*d* embeddings obtained in 2(a) and 2(b). You can reduce the size of the word embeddings to 2 obtained in 2(b) using SVD.

(d) Try to establish analogies such as "leg is to kick then _ to catch" e.g. 'hand'

[40%]

**4. Innovation:**

Take the pretrained word embeddings obtained from Word2Vec/GloVe and improve the embeddings using your own strategy.

Repeat 3 for evaluation of your method.

[20%]

Note: use of libraries is limited only to download databases.