IT412: Natural Language Processing

Assignment 5: Word Embedding

Instructor: Prasenjit Majumder

Lab Outcome: After performing this assignment you will be able to create and use pre-trained word embedding.

1 Word Embedding

A word embedding is a learned representation for text where words that have the same meaning have a similar representation. Word embeddings are in fact a class of techniques where individual words are represented as real-valued vectors in a predefined vector space. Word embedding is a technique to represent words into dense vector. There are many Word embedding techniques i.e Word2Vec, GloVe, FastText, ... BERT. Pre-trained word embedding for each of them are available.

2 Implementation

Implementations are to be done using Gensim Library.

2.1 Dataset

For this assignment you are provided with twitter sentiment analysis dataset, which will be used for task 3 and 4. Divide the given dataset into 80:20 ratio, 80% for training and 20% for testing,

2.2 Exercise

Note: Text pre-processing can be done as per requirement.

1. Show word similarity on pre-trained embedding by taking 5 examples
2. Show word analogy on pre-trained embedding by taking 5 examples.
3. Perform sentiment analysis by using pre-trained embedding by using classifier of your choice.
4. Gensim provides facility to train task specific embedding, so create embeddings using given dataset and perform sentiment analysis on it.

3 References

- Efficient Estimation of Word Representations in Vector Space
- Distributed representations of words and phrases and their compositionality