Assignment
Selected Topics in CS

Sem I 2020-2021 **Max.Marks: 20**

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, sentiment analysis, named entity recognition, parsing etc.

This assignment will help you to learn word 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. The assignment will also help you to learn role of word embeddings in downstream tasks such as sentiment analysis.

In this assignment you are required to do the following:

Phase I

Implement the word2vec models and train your own word vectors with stochastic gradient descent (SGD).

Write separate functions for the following:

i. Softmax
ii. Negative sampling
iii. Cost function and gradient function
iv. Normalization function to normalize rows of a matrix
v. Skip-gram model
vi. CBOW
vii. SGD Optimizer

Train your model by selecting a suitable dataset e.g. Reuters Corpus of news articles from the text corpora of nltk (details are given https://www.nltk.org/book/). The training process may take a long time depending on the efficiency of your implementation. An efficient implementation will help.

Experimentation and Results

1. Compare your results obtained by skip-gram model, skip-gram with negative-sampling, CBOW.
2. Vary model parameters and compare results
3. Use SGD optimizers and compare results
4. Vary hyperparameters and compare results

Report

Evaluate the word vectors obtained by the following:

1. Find similarity among the words and plot at least five group of similar words.
2. Show at least two semantic and two syntactic relations/analogies by plotting them.

Plot two dimensional vectors of the words. Apply SVD to reduce dimensions to two.

Phase II

Perform a sentiment analysis on the word vectors you have trained in Phase I.

The details of sentiment analysis will be uploaded by the coming week end.

Due Date: 31st Oct 2020 for Phase-I

10th Nov 2020 for Phase-II

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