IT412: Natural Language Processing

Assignment 10: Word Embeddings

Instructor: Prasenjit Majumder and Thomas Mandl

Learning Outcome: At the end of this assignment you will learn training a word2vec model and using pretrained word2vec model for classification

1 Problem description

Word embedding is representation of words in latent space such that similar words are grouped together. In our case we will deal with distributed representation where we will learn the vector representation of a word based on words surrounding it. One method that uses distributed representation for words is Word2Vec introduced by Mikolov et al. Word2Vec has two variations: Continuous Bag of Words (CBOW) and Skip Gram. In this assignment we will use CBOW and Skipgram for obtaining the distributed representation.

2 Implementation

2.1 Dataset

- For Part A of the assignment we will use Telegraph news articles, which is in XML format. It contains news on different categories for the year 2004 to 2007. You can download the dataset from this link: https://drive.google.com/open?id=1JuawXQmYVkjpfL3HOblqjDrqw8V1lHrC
-
- For Part B of the assignment use the "opinion lexicon" dataset to classify the token to their corresponding sentiment use the following lines: import nltk
nltk.download('opinion_lexicon')
from nltk.corpus import opinion_lexicon

2.2 Exercise

- Implement CBOW and Skipgram approach using Gensim library use the notebook shared with you in the classroom.
- Split the opinion tokens into training, testing set. Use the pretrained word2vec vectors to represent the tokens. Use SVM to perform classification. Train a deep neural network to perform the classification. Report the Macro F1 score for both the approaches.

3 References

- Tomas Mikolov et al.,Distributed Representations of Words and Phrases and Their Compositionality, NIPS, Volume 2, 2013, 31113119
- https://pytorch.org/tutorials/beginner/nlp/word_embeddings_tutorial.html
- https://kavita-ganesan.com/gensim-word2vec-tutorial-starter-code/#.XlOEPigzZPY
- https://radimrehurek.com/gensim/models/word2vec.html
- http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

4 Submission

- The submission deadline for this assignment in **7th December 2021 at 11 PM**