NLP PROJECT SYNOPSIS

Creating Continuous Bag of Words (CBoW) Model using Neural Networks & Visualizing them using PCA (Principal Component Analysis)

Natural Language Processing - Dr. Seba Susan

22nd September 2020

Anish Sachdeva
DTU/2K16/MC/013
Delhi Technological University

Overview

Words can't be directly understood or computed by a computer as at the end of the day a computer can understand and work on numbers, or more specifically vectors. In this project I propose to create vectors from words using the Continuous Bag of Words (CBoW) Model where we create a Deep Learning Neural Network which is trained to return the vectors from a given word in the corpus.

We train it by feeding it with the context words of any given word in the corpus, e.g. if we have the sentence "I am happy because I am learning", we will feed it with the context words [I am because I] and we should receive the center word [happy]. Using this neural network we will be able to find vector representations for all our words and use these vector representations to form analytical tasks.

Once we have our vector representations, we can form simple analysis tasks such as finding the relationship between 2 words, e.g. what word is to egypt what paris to France?

Or what is to male what queen is to female?

These are a few questions that can be answered using the vectors that we have computed, we can also perform PCA (Principal Component Analysis) on our 300-dimensional vectors to reduce them down to 2-dimensional vectors and then we can visualize our words after PCA reduction and then visualize them on a 2-dimensional euclidean plane.

By plotting them on a plane we can see which words are clustered close together and identify classes using clustering. So, all in all we will perform the following steps in our NLP Project:

1. Creating the word vectors using the CBoW Model by training a Deep Learning Neural Network.
2. Using the Vectors to identify relationships between countries and Capitals and other such relations.
3. Using PCA (Principal Component) analysis to visualize words on a Euclidean Plane.
4. Performing similarity Metrics like Cosine Similarity and Jaccard Similarity on words.
5. Performing Distance Metrics such as Euclidean, Manhattan (L1) Norms on the word vectors.
6. Studying and introducing standard word vectors such as word2vec by Google.