Assignment 1

NLP Applications

Deadline : 30th Jan 2018 11:59 PM

Modern distributional semantic algorithms (also known as word embedding algorithms) can be found inside many intelligent systems dealing with natural language. They became especially popular after the introduction of prediction based models based on artificial neural networks, like Continuous Bag-of-Words and Continuous Skip-gram algorithms, first implemented in word2vec tool. Their ultimate aim is to learn meaningful vectors (embeddings) for words in natural language, such that semantically similar words have mathematically similar vectors

The current assignment aims to make you familiar with the above algorithms. Your task is to generate the word2vec embeddings over the given dataset using the skip-gram & CBOW algorithms.

Dataset:

https://drive.google.com/open?id=1aYCZPx36ojGxvvvQpTMousrobXY8HhgB

Evaluation:

Evaluation will be based on the quality of the learnt word embeddings. We'll provide a list of words during evaluation and you will have to print the 10 nearest neighbouring words based on cosine distance.

Submission Format:

<rollno>_assignment1.zip

----- <rollno>_skipgram.py

----- <rollno>_cbow.py

NOTE: You cannot use the existing libraries (like gensim) to generate the word embeddings. Although, you can use gensim to get the nearest neighbours based on cosine distance.