Eckovation AI&ML: Programming Exam

Natural language processing (NLP)

Date: 07-02-2021

Time: 3 hours
Max Marks: 50

Question

The first and the most important step for any NLP related task is data preprocessing. For this question, we will try to implement a simple sentiment analysis movie review model. We will be using the first 10,000 reviews from **imdb dataset** for this task. Download the dataset and perform the following steps:

1. Preprocess the text (i.e., the steps required prior to converting the sentence into a vector) using any library of your choice. [15 Marks]

2. Given any sentence, perform vector semantics i.e., convert the given dataset into vectors using Bag of Words approach. [15 Marks]

3. Train a simple classifier (using Scikit-learn e.g.: SVM) to perform sentiment analysis on the generated dataset. [20 Marks]

4. BONUS: Repeat the above process again but now using word2vector. (for any sentence, take average of the vector representation of all the tokens to get the vector representation)