Assignment on TXTA

There are three problems in this assignment. You are free to take helps from the Internet but you are also required to mention the source. If I find any copied material without the source, there will be no marks for the same. Avoid sharing codes with each other. **If I find a copy which is very similar to another copy which I checked earlier, the current copy will be penalized**. So, keep this point in mind while helping your friend with your own codes. The problems are mentioned below:

Problem 1 [50]

There is a sample of IMDB movie reviews. You need to do sentiment analysis using

1. A dictionary of word score (named 'word_score.txt). In this method, the sentiment score is required to be calculated as (*Sum of scores of positive words in the review*)+(*sum of scores of negative words in the review*) / (*total number of positive and negative words together*).
   IMP: You cannot use any other dictionary except the word_score.txt file. In this file, beside each word, the respective score is provided which is going to be used for getting sentiment score.
2. Either XgBoost or LightGBM with hyperparameter tuning
3. Deep learning method

Problem 2 [30]

You need to scrape data from websites dealing with real-estate properties. Try to extract information related to first 200 properties. You can extract as much info as possible but there should be info of:

a) BHK
b) Price
c) Location

The websites are https://www.99acres.com/, https://www.makaan.com/, https://housing.com/

Rule for selecting the website:

Sort the enrolment number is ascending order and follow the sequence as mentioned below:

Roll No 1 ---> https://www.99acres.com/

Roll No 2 ---> https://www.makaan.com/

Roll No 3 ---> https://housing.com/

Roll No 4 ---> https://www.99acres.com/

Roll No 5 ---> https://www.makaan.com/ and so on

Problem 3 [20]

Take the bbc-fulltext.zip dataset, extract the news articles and do the following:

1. Create word vectors using both LSA and Word2Vec
2. Using the above vectors predict the news classes and compare the performances