NLP Assignment 3
Deadline: 7th Sept 23:59 hr, 20 marks

Please don't copy from net or other students. Plag will be checked.
Allowed Programming language : Python
(Application of any NLP library is only permitted for preprocessing part)

Your task in this assignment is to write a python program that will be able to generate and classify sentences based on some corpus .

Data set link: https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

Please save all the trained models as .pkl file to avoid training during demo.

Part 1: Generative Model

1. Consider all the news .txt files from the folder "comp.graphics" as single corpus and perform the following

    1. Generate 1 sentences using unigram model. [5 Marks]
    2. Generate two different sentence using bigram model.
    3. Generate two different sentence using trigram model.
    4. Mention your observation from the above set of operations in the report file.

2. Perform the same set of tasks using "rec.motorcycles" as the corpus. [5 Marks]

    Sentences should be different and use some assumption so that the sentences make maximum sense. (obviously they will not be correct sentences)

Part 2: Discriminative Model

3.1
In this part you will be designing a discriminative model using bi-gram, so that given any arbitrary sentence as input, it will be able to predict which news group it's most probable to belong to. For the model use **comp.graphics** and **rec.motorcycles** news class only. Consider smoothing for handling unknown words. [5 Marks]

3.2
Rebuild the same model using <UNK> to handle unknown words. (you can use tf idf to decide <UNK>in the training corpus. smoothing will also be required)

[5 Marks]

Submission format: Upload your project folder containing the code file named **Assignment3_nGram_MT170XX.py**
Submit a pdf report file with your implementation details and clearly mention your assumption if any and generated sentences.