Assignment 1

The following set of tasks relate to training a language model from scratch on the movie review dataset and using it for creating movie reviews, which would be treated as the test data for classification as per the classes (labels) assigned in the original dataset.

Text Prediction

Imagine a scenario where you are working on a text prediction feature for a messaging application. The n-gram models can be employed to predict the next word based on the preceding words, thus enhancing the user experience by suggesting contextually relevant words.

N-gram models for prediction:

Unigram Model: Predicts the most frequent word in the corpus.

Bigram Model: Predicts the next word based on the previous word.

Trigram Model: Predicts the next word based on the previous two words.

By implementing these models, you shall gain a deeper understanding of language modelling and its applications in natural language processing.

Objective:

To predict the next word in a sequence using unigram, bigram, and trigram models without using any off the shelf libraries.

Instructions:

Data Preparation:

Read a text corpus from a file.

Preprocess the text by converting it to lowercase and removing punctuation.

Model Building:

Build unigram, bigram, and trigram models from the corpus.

Prediction Function:

Implement a function to predict the next word based on the previous one or two words.

Unigram Model: Predicts the most frequent word in the corpus.

Bigram Model: Predicts the next word based on the previous word.

Trigram Model: Predicts the next word based on the previous two words.

Movie Reviews Labelling (Classification)

Use the basic structure of the code as implemented for the previous task and build on that to implement the movie review labelling (classification) task.

Objective:

To implement a movie review labelling (classification) system using unigram, bigram, and trigram models in Python without using any external libraries.

Instructions:

Bayesian Classification Function:

Implement a function to suggest labels for movie reviews generated by the language model and evaluate it using standard classification evaluation metrics.

Build Models: Unigram, bigram, and trigram models are built from the corpus.

Suggest Correction: For each generated movie review, the function suggests a label based on labels learned from the training data.