Natural Language Processing

Assignment-1: Sentence Generator using N-gram

Instructions:

1- Students will form teams of 3 students (**Can be from different groups).**
2- Deadline of submission is **Monday 15 / 04 / 2024 @11:59 PM.**
3- Submission will be on google classroom.
4- No late submission is allowed.
5- No submission through e-mails.
6- File Naming (team ids) → No grade for wrong ids and missing ids.
7- **In case of Cheating, you will get a negative grade whether you give the code to someone or take the code from someone/internet.**
8- You have to write clean code and follow a good coding style including choosing meaningful variable names.

Assignment Details:

Write a python program to build a sentence generator program using bi-gram and tri-gram model. You will generate M sentences by starting with random (n-1) gram from the model and then randomly select the next word until reaching a max length of sentence. The corpus used in this assignment is Brown Corpus in NLTK library. Get all sentences from brown corpus. **Submit your code .py or .ipynb**

Input:

- m -> Number of sentences to be generated
- n -> 2 for bigram – 3 for trigram
- maxLen -> Max number of words in sentence generated. Used as stopping condition in generating a sentence.
- Corpus -> list of sentences to generate n-gram model from it.

Output:

- m sentences each sentence contain maxLen words.

Steps:

1-Apply data preprocessing:

- Tokenize sentences into words (Word Tokenization).
- Remove punctuation marks from tokens.
- Convert all tokens to lowercase.

2-Build N-gram model:

Build n-gram dictionary, where the keys are tuples of n-1 words (n-1 gram) and the values are lists of possible next words.

3-Sentence Generator:

Sentence generation process start with a random n-1 gram from the model and then randomly select the next word until reaching maxLen or unable to find next word.