NLP Assignment 2

Om Khare
112003066
Div 1, B4

Part A

Intro

The provided code is a script for generating and counting n-grams from an input text file. An n-gram is a contiguous sequence of n items from a given sample of text. This model is often used in language modeling and text mining.

Code Explanation

1. Function: generate_ngrams

Explanation:
tokens: The text is split into a list of words.
ngrams: This creates a list of n-grams using the zip function. The zip function combines the tokens, offset by positions from 0 to n-1. This creates n-length tuples of words.
The function returns these tuples as space-separated strings.

2. Function: ngram_model

Explanation:
ngram_counts: This dictionary holds counts for n-grams of sizes 2 through 5. Each entry is a defaultdict with an integer default value, allowing for easy counting.

The code then reads the input_file line-by-line.

For each line and for each n from 2 to 5, the code generates n-grams using the generate_ngrams function and updates the count in the ngram_counts dictionary.

After processing the entire file, the code saves the n-gram counts into separate files. Each file is named based on the n-gram size (e.g., "MIS-No_2-gram-output.txt").

Conclusion

This script efficiently generates and counts 2-grams to 5-grams from an input text file.

Part B

Intro

This code provides the auto complete feature by considering ngram frequency generated from part A.

Code Explanation

1. load_ngrams_from_file()

Load n-grams and their counts from a given file.

2. predict_next_word():
Predict the next word based on the provided sequence of words using available n-gram data.

Conclusion

This code predicts the next word in a given sequence using n-gram models sourced from files and then writes the prediction to an output file. It supports up to 5-gram predictions based on the presence of corresponding n-gram data files.