

Assignment No 2

18P-0066 Usman Ahmad

February 24, 2023

1 Task 1

The first step is to load the Poetry Corpus text file and process it using the spaCy library. The code then creates two lists: unigrams and bigrams.

1.1 Generate Unigram and Bigram Function

The generate unigram function randomly selects words for each verse by generating a random integer between 7 and 10. The Bigrams list uses the bigrams list to generate each verse, selecting the next word based on the bigram probabilities in the list. The main loop of the program generates three stanzas, each consisting of four verses, with the first and third verses in each stanza generated using the generate unigram . The print statements at the end of each loop simply print each verse to the console, with an empty print() statement separating the stanzas.

2 Task 2

This code first applies tokenization on the input document doc using the spaCy library. It then filters the punctuation and spaces from the tokens using the conditions not token.is punct and not token.is space. The remaining words are stored in the list words.

2.1 Generate Unigram and Bigram

The first line counts the frequency of each word in the text and creates a dictionary with each key as a word and its frequency as the corresponding value.

2.2 Function Generate Verse

The code defines a function called generate verse that generates a random verse for a poem using a bigram model. The function randomly selects a starting word and adds next words before the length of the verse is between 7 to 10 words. It then loops over 3 stanzas each consisting of 4 verses, and prints the generated verse. After printing each verse, the function prints a blank line to separate the verses. Overall, the code generates a random poem with 3 verses each with a length between 7 to 10 words.