Assignment III: Language Modeling

Introduction:

The goal of this assignment is to tokenize given list of tweets and generate bigram, trigram, 4-gram and 5-gram out of it.

Approach:

The first part of this assignment which is tokenizing and normalizing given tweets was same as that of first assignment. The only change was in the later part which was generating n-grams out of the tokens generated.

I have used code of the first assignment to tokenize the tweets and tweaked it to generate n-grams.
To generate bigrams, we have to iterate till last but one element of the list and write the value of an iterator and value next to it to the file on a single line separated by spaces. The tweets containing only a single word are printed as they are to the file. To generate trigrams we have to iterate till last but two tokens and write the value of the iterator and next two values on a single line separated by spaces. Tweets containing less that three words are written as they were to the file. To generate 4-grams we have to iterate till last but three elements of the list and write the value of the iterator and its next three values in a single line separated by spaces. Tweets containing less than 4 words are written as they were to the file. To generate 5-grams we have to iterate over last but 4 elements in a list and write the value of the iterator and next 4 values on a single line separated by spaces. Tweets containing less than 5 words are written as they were to the file. In general to generate n-grams, we have to iterate till last but (n-1) tokens in a list and write current as well as next (n - 1) values of the iterator on a single line separated by spaces.

Discussions:

I have discussed my approach with Anurag Kolhe and Vaibhav Gawali. We discussed general strategy to tackle the problem and possible exceptions.

References:

I have referred official python documentation for list functions while solving this problem.