1. As already discussed in the class, create data splits in your data. Use random sampling to choose sentences and create the validation and test sets.
    a. Validation Set - 1000 sentences
    b. Test Set - 1000 sentences
    c. Training Set - Remaining Sentences
2. By now, you have already built the following models,
    a. Unigram Model
    b. Bigram Model
    c. Trigram Model
    d. Quadrigram Model

    Implement Good Turing Smoothing for all the models.

For unseen n-grams, the cumulative probability to be distributed to all unseen n-grams = N_1/N, where N = total number of seen n-grams and N_1=number of times a n-gram has occurred only once

Good Turing smoothing for individual unseen n-gram P_{unseen} = (N_1/N)/(number of unseen n-grams) = (N_1/N) / (V^n - N) = N_1 / N * (V^n - N) where V = vocabulary size, where n >= 2 for unigram model it can be = V - unique seen unigrams (U), P_{unseen} = N_1 / N * (V - U)

Compute the probability of each sentence in the validation and test sets using the smoothed models.

3. Show a table with top 100 frequencies as the following:

C (MLE)	N_c	C*
0		
1		
…		

4. Implement deleted interpolated smoothing technique for the quadrigram model and find the best parameters.