Pr?gr?mm?ng?H?m?w?rk 3

In this exercise we model a string of text using a Markov(1) model. For simplicity we only consider letters 'a-z'. Capital letters 'A-Z' are mapped to the corresponding ones. All remaining letters, symbols, numbers, including spaces, are denoted by '.'.

We have a probability table T where $T_{i,j} = p(x_t = j | x_{t-1} = i)$ transition model of letters in English text for $t=1,2 \dots N$. Assume that the initial letter in a string is always a space denoted as $x_0 = \text{'.'}$. Such a model where the probability table is always the same is sometimes called a stationary model.

1. For a given N, write a program to sample random strings with letters $x_1, x_2, \dots, x_N$ from $p(x_{1:N}|x_0)$
2. Now suppose you are given strings with missing letters, where each missing letter is denoted by a question mark (or underscore, as below). Implement a method, that samples missing letters conditioned on observed ones, i.e., samples from $p(x_{-\alpha}|x_{\alpha})$ where α denotes indices of observed letters. For example, if the input is 't??.', we have $N=4$ and $x_1 = \text{'t'}$ and $x_4 = \text{'.'}$, $\alpha=\{1,4\}$ and $-\alpha=\{2,3\}$. Your program may possibly generate the strings 'the.', 'twi.', 'tee.', etc. Hint: make sure to make use all data given and sample from the correct distribution. Implement the method and print the results for the test strings below.
3. Describe a method for filling in the gaps by estimating the most likely letter for each position. Hint: you need to compute

   $$x^*_{-\alpha} = \arg\max_{x_{-\alpha}} p(x_{-\alpha}|x_\alpha)$$

   Implement the method and print the results for the following test strings along with the log-probability $\log p(x_{-\alpha}^*,x_{\alpha})$.
4. Discuss how you can improve the model to get better estimations.