The objective of this lab is for you to explore the behavior of a discrete Markov process by training it on sequences of characters, and sampling sequences from it. The TA will first demonstrate the results of this on several datasets of sequences, and then you will replicate those results, and further explore the datasets.

We provide you with the following:

- The script `lab11.m` sets up the problem (toy dataset) and plots various figures.
- The functions `dmptrain.m` and `dmpsample.m` train the discrete Markov process and sample from it, respectively.

I Datasets

Use sequences of characters defined on N states, say $N = 5$ with states $\{\texttt{a}, \texttt{b}, \texttt{c}, \texttt{d}, \texttt{e}\}$, and sequences such as `cbccd`, `edcbeddeea`, etc. Once this works, use English sentences as training sequences, where the states correspond to the English letters and punctuation, e.g. `a`–`z`,`.`,`;`, etc. You can also consider as states whole words rather than individual characters, and have the discrete Markov process generate sequences of words.

II Using discrete Markov processes

The functions `dmptrain.m` and `dmpsample.m` implement the following, respectively:

- *The learning problem*: given a dataset of training sequences over a set of N states, estimate the parameters:
  - The transition matrix $\mathbf{A}$ of $N \times N$.
  - The initial distribution $\boldsymbol{\pi}$ of $N \times 1$.
- *The sampling problem*: given parameters $(\mathbf{A}, \boldsymbol{\pi})$, generate a sequence of length T by sampling from the discrete Markov process.

The following Matlab functions are useful in the implementation (among others): `randi rand cumsum unique cell2mat`.

We visualize the results with the following:

- Plotting the parameters $(\mathbf{A}, \boldsymbol{\pi})$ as an image.
- Printing the sequences sampled from the discrete Markov process.

Consider the following questions:

- Given the parameters $(\mathbf{A}, \boldsymbol{\pi})$ learnt from a set of English sequences:
  - Inspect their values and try to guess what kind of sequences will be generated with high probability.
  - How will sequences sampled from these parameters $(\mathbf{A}, \boldsymbol{\pi})$ differ from:
    - True English text sequences.
    - Random sequences of English characters. (We call "random sequences" sequences generated uniformly at random, i.e., picking each element independently from each other at random from the states.)
- If you train a discrete Markov process on a set of random sequences, how will $(\mathbf{A}, \boldsymbol{\pi})$ look like?