NICTA - Statistical Machine Learning advanced course

First assignment: probability and conditional probability

The goal of this assignment is to write a Matlab code

1. Go on the web and find a huge text file. See for instance:

   `http://www.mechon-mamre.org/e/et/et01.htm`

2. Write a Matlab routine estimating the probability for a character to begin a sentence,
3. Write a Matlab routine estimating the probability for any character to appear in the text,
4. Write a Matlab routine estimating the probability of any sequence of 2 characters to appear in the text.
5. Write a Matlab routine generating a 100 character long sentence. The first character is picked randomly according to the probability for a character to begin a sentence. The following character will be randomly generated acording to the conditional probability for a character to appear knowing its predecesor.
6. Write a Matlab routine estimating the probability for any sequence of 3 characters to appear in the text.
7. Write a Matlab routine generating a 100 character long sentence acording to this estimate of the joint probabilities. Explain in coments in your code how you deal with zeros and low probability events.
8. Write a readme file describing how to run the code and briefly summarize your results.

For instance, this is a kind of sentence generated by the 2 characters based random generator estimated based on a french text.

```
daieus Ilesora mprs de cer die, ve a t me pluistaus pavesa Etts voiona! Apauisube, e qu'hai c
```

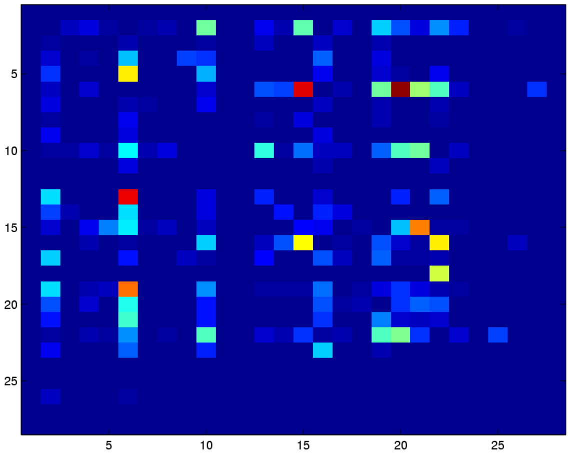

Figure 1: example of the estimation of the joint probability for the same french text.

If you wish to use any other high level language feel free to do so. Your code should run and compile on a standart linux system.

some Matlab tricks :

- j = fopen('esc.txt')
- while(not(feof(j))); line=fgetl(j); ind = double(line); C(ind) = C(ind) + 1; end;
- fclose(j);
- num2str
- char
- help iofun; help fileformats; help xlsread; help dlmread; help sprintf
- $i = \text{rand}(1); \text{debut} = \max(\text{find}(\text{cumsum}(pC) < i));$
- image; imagesc;