Lab report: Text segmentation using Hidden Markov Models

Author: François Amat

Contact : francois.amat@telecom-paristech.fr

Abstract

In this lab we want to separate the body, the header of a email. In order to do that, we will use an implementation of the Viterbi algorithm to compute the best segmentation between header and body.

Task : automatic segmentation of mails

Q1 : Give the value of the π vector of the initial probabilities

The initial probability is $ \pi $ = $(1, 0)$. We always start an email by a header, so the probability to begin with state one is 1.

Q2 : What is the probability to move from state 1 to state 2 ? What is the probability to remain in state 2 ? What is the lower/higher probability ? Try to explain why

The probability to move from state 1 to state 2 is 0.000781921964187974.

The probability to remain in state 2 is 1.

It seems to be unlikely to move from state 1 to state 2, and very likely to stay in its own state.

The lower probability is to move from state 2 to state 1 : 0.

The higher probability is to move frome state 2 to state 2 : 1.

I suppose that is because there are few specific words that make the transistion from a header to a body in a email.

Q3 : What is the size of the corresponding matrix ?

The size of the corresponding matrix is by convention $(N, 2)$.

To implement

Q4 : print the track and present and discuss the results obtained on mail11.txt to mail30.txt

This is done in the notebook file. The results seems to be accurate, there are some errors but they do not exceed one or two line around the expected separation between the header and the body.

Q5 : How would you model the problem if you had to segment the mails in more than two parts (for example : header, body, signature) ?

I would use an other transistion matrix with a dimension of (3,3), still triangular superior. And I would get a third row to the emission matrix. With these elements I would use the same algorithms to get the three parts.

Q6 : How would you model the problem of separating the portions of mail included, knowing that they always start with the character ">".

I would change the emission matrix and associate a very high number for the character ">", That change will help to separate the portions of mail included.

Thoughts on the python implementation of the Viterbi algorithm

In the process of computing probabilities, with the data given, we had value under 10^-308 which is the double limit in python. This limit gave me a lot of debugging, the simplest solution i have found, discussing with my pairs was to multiplicate by a magic factor of 40. This factor

will make this algorithm works only on this type of problems.
In order to make this algorithm works to all types of problems, a transformation with a logarithm would work (but this necessitate handling the 0 values).