A Hidden Markov Model for Optical Character Recognition (OCR)

Due Date: 23.12.2023

The objective is to find the most likely character sequence via a 1st order Hidden Markov Model (HMM), using the optical sensor outputs as the observable and the actual letters as hidden states. For all states, the only possible values are upper-case English letters.

In two separate files (data_actual_words.txt and data_ocr_outpus.txt), you are given a list of actual words and their OCR readings. Using this data, you are expected to compute approximate transition and emittance probabilities. Finally, using those probabilities, you are expected to compute the most likely character sequences given OCR observations.

- The files include 60244 actual words and their OCR outputs. You should use the first 50000 of those for estimating the transition and emittance probabilities.
    - You should print the initial state probabilities.
    - You should print the transition probabilities.
    - You should print the emission probabilities.
- The rest of the data is to be used for performance analysis:
    - 10244 most likely word sequences should be estimated and compared with the actual words.
    - You should print the words for which a different sequence is estimated than the original OCR output.
    - You should print the number of corrected letters for which the OCR output is erroneous, but the HMM estimation is correct.

BONUS:

As a bonus, you can additionally implement a higher-order Markov chain and improve the correction performance.