1) Ukrainian language structure

Language structure analysis is based on Baum-Welch algorithm of Hidden Markov Models. As based text there is "Борислав сміється" Івана Франко.
Dataset consists of 100000 letters of Ukrainian alphabet:
"*а б в г ґ д е є ж з и і ї й к л м н о п р с т у ф х ц ч ш щ ь ю я*".
Alphabet length is 33 and there is 2 hidden states in the model.

The result of learning model is matrix A transition matrix of hidden states, B transition matrix to observed states and m initial distribution.

$$A = \begin{pmatrix} 0.269 & 0.731 \\ 0.954 & 0.047 \end{pmatrix}$$

$$B = \begin{pmatrix} 0.00 & 0.04 & 0.10 & 0.03 & 0.00 & 0.07 & 0.00 & 0.004 & 0.02 & 0.04 & 0.00 & 0.00 & 0.007 & 0.03 & 0.06 & 0.07 & 0.06 & 0.10 & 0.00 & 0.05 & 0.07 & 0.07 & 0.09 & 0.00 & 0.002 & 0.02 & 0.01 & 0.03 & 0.02 & 0.01 & 0.000 \\ 0.19 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.11 & 0.000 & 0.00 & 0.00 & 0.16 & 0.12 & 0.002 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.23 & 0.00 & 0.00 & 0.00 & 0.00 & 0.08 & 0.000 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.035 \end{pmatrix}$$

$$m = (1, 0)$$

By using matrix B we can cluster alphabet symbol this way:

- state 1 - [*б, в, г, д, є ж, ї, й, л, м, н, п, р, с, т, ф, х, ц, ч, щ*]
- state 2 - [*а, е, и, і, о, у, ю, ь, я*]

As we can see from matrix B our HMM model divided letters almost into *vowels* and *consonants* (letters ї, є arent common and they have pretty equal probailities to belong to both classes).

So, roughly speaking, our model gives us the probailities of transitioning between vowel and consonant letters. In Ukrainian language it is more like for consonant to meet another consonant, words are full of consonant sounds, like *"припустимо, пробачте"*, especially in prefixes, suffixes. For vowel it is mostly like not to meet a vowel, there is really rare in Ukrainian words to have two or more vowel together.

3 hidden states:

- state 1 - [*а, е, и, і, о, у, ю, ь, я*]
- state 2 - [*б, г, ґ, д, ж, к, л, м, н, п, р, т, ц ч, ш, щ*]
- state 3 - [*в, є, з, ї, й, с, ф, х, ю*]

$$A = \begin{pmatrix} 0.0003 & 0.62 & 0.378 \\ 0.988 & 0.012 & 0 \\ 0.008 & 0.828 & 0.164 \end{pmatrix}$$

4 hidden states:

- state 1 - [*б, г, д, к, м, п, т, ч, ш*]
- state 2 - [*ґ, л, н, р, ц, щ*]
- state 3 - [*в, є, ж, з, ї, й, с, ф, х, ю, я*]
- state 4 - [*а, е, и, і, о, у, ь*]

$$A = \begin{pmatrix} 0.00129 & 0.17810 & 0.00000 & 0.82060 \\ 0.00000 & 0.00000 & 0.00000 & 1.00000 \\ 0.78124 & 0.12047 & 0.09824 & 0.00005 \\ 0.51789 & 0.20138 & 0.28041 & 0.00031 \end{pmatrix}$$

With 4 hidden states there is groups of consonants that almost certainly transition to vowel or which are likely to go after a vowel.

(If we add whitespace or $'$ to alphabet nothing would change thease characters are recognized as consonant letters, but of course $'$ is a letter which likely meet a wovel)

2) Text decodng

Matrix A consists of frequencies of two neighbor letters (i, j) of the alphabet from $n = 200000$ sample.
As initial value we take $\mu = (\sim \frac{1}{33}, \dots, \sim \frac{1}{33})$
Initial B is also filled with $\sim \frac{1}{33}$ values

B and μ are reestimated with Baum-Welch algorithm on encrypted text by Ceasar cipher (first 2000 symbols of encrypted text, 800 iterations).
Then A, B and μ parameters are used in Viterbi algorithm to decrypt whole text.

Results: 92.252% of symbols were successfully decrypted, decrypted text is written in decrypted.txt file.