Pushkin's Poetry and Markov Chains

I came across an interesting anecdote regarding the origins of Markov Chain theory.

Suppose you are given a body of text and asked to guess whether the letter at a randomly selected position is a vowel or a constant. Since consonants occur more frequently than vowels, your best bet is to always guess consonant. Suppose we decide to be a little more helpful and tell you whether the letter preceding the one you chose is a vowel or consonant. Is there now a better strategy you can follow?

In 1913, A.A. Markov was trying to answer the above problem analysed twenty thousand letters from Pushkin's poem *Eugene Origin*. He found that 43% letters were vowels and 57%, consonants. So in the first problem, one should always guess "consonant" and can hope to be correct 57% of the time.

However, a vowel was followed by consonant 87% of the time. A consonant was followed by a vowel 66% of the time. Hence, guessing the opposite of the preceding letter would be a better strategy in the second case. Clearly, knowledge of the preceding letter is helpful.

The real insight came when Markov took the analysis a step further. Markov investigated whether knowledge about the preceding two letters confers any additional advantage. He found that there was no significant advantage to knowing the additional preceding letter. This leads to the central idea of a Markov chain - while the successive outcomes are not independent, only the most recent outcome is of use in making a prediction about the next outcome.