How the fitness value changes in different ciphertexts for different n-grams?

Methodology

First I needed to compute how frequent particular n-grams are in the English language. My sample data set was the whole Ulysses book https://www.gutenberg.org/ebooks/4300.

First I generate all possible combinations of n-grams in the English alphabet. The amount of them is as follows:

Digrams: 676

Trigrams: 17576

Quadgrams: 456976

Of course, some of them like quadgram 'xxxx' never occur in natural English language but I will deal with that later.

Then, I extracted all n-grams from the Ulysses book and counted how often their occur. Here are the most popular of them (values in percent):

- Digrams:
  - he: 2.86
  - th: 2.85
  - in: 1.95
  - er: 1.70
  - an: 1.51
- Trigrams:
  - the: 1.84
  - ing: 0.82
  - and: 0.76
  - her: 0.52
  - hat: 0.35
- Quadgrams:
  - ther: 0.27
  - the: 0.27
  - that: 0.24
  - with: 0.23
  - here: 0.18

To calculate the fitness value of a sample text I also divide it into n-grams, sum the frequency of them using the prior generated frequency table and at the end I divide it by the total number of n-grams in the sample text.

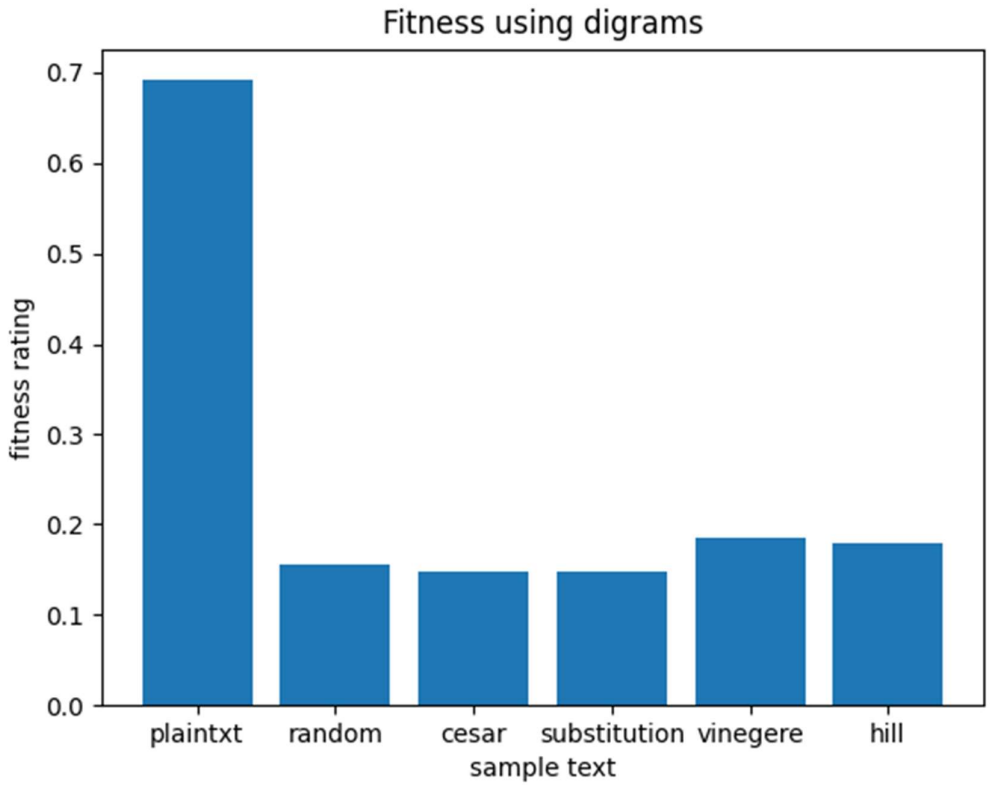
The higher the value is, the more similar sample text is to the English language.

Diagrams

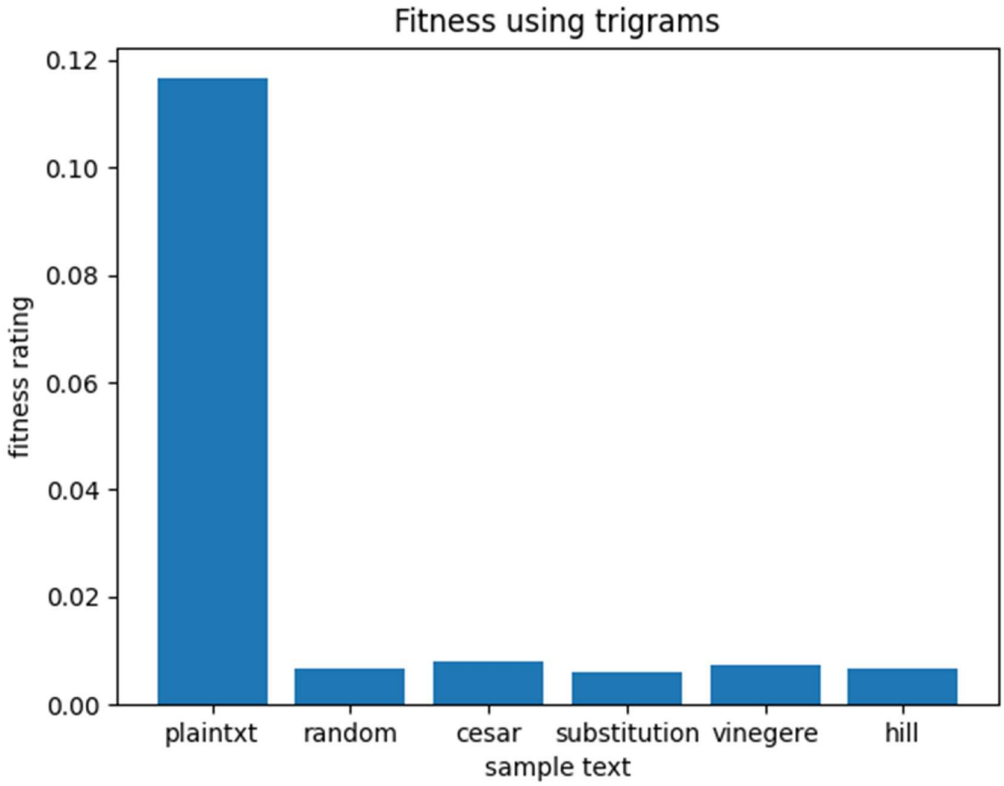
BBC article

In the diagrams, as plaintext I took a BBC article of length 3963 characters (after cleaning the whitespaces, commas, dots, etc.). https://www.bbc.com/news/science-environment-65848872

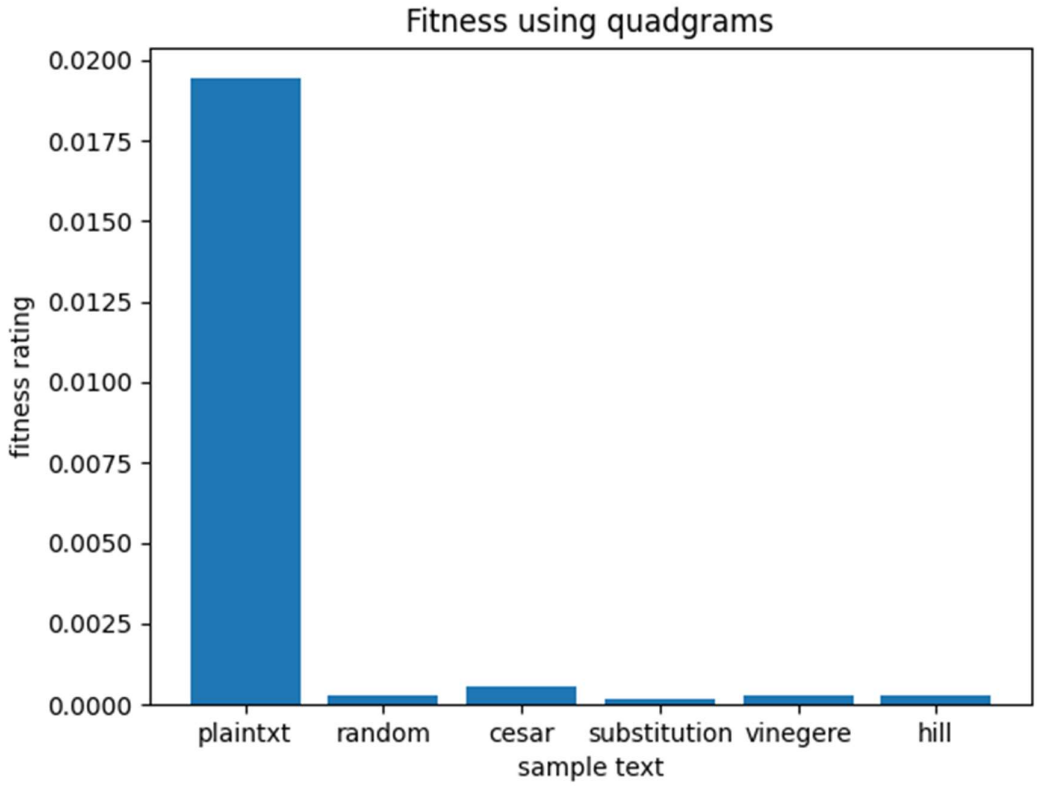
As a control data set I generated a string of random letters of the same length as the plaintext.



As we can see, the fitness of random letters is the same as fitness of Cesar cipher and substitution cipher. Vinegere and hill ciphers seem to have just a little bit higher fitness.



While using trigrams, we can see, random letters and all ciphertext have similarly low fitness rating. Here the difference between plaintext and everything else is bigger that using digrams.



With quadgrams, the fitness value difference between plaintext and random as well as ciphertexts sinks even lower. Only cesar cipher seem to have just slightly higher fitness value.

Conclusion

The fitness value of different monoalphabetic ciphertexts is no different from the fitness value of random letters. It shows, that the ciphertexts are not similar to English language at all.