Project V3
Vigenere's Cipher Analysis

Project V3 team

Aaron Crouch

July 22, 2022

Introduction - Vigenere Cipher

Project V3 is an implementation & analysis on the Vigenere cipher. The Vigenère cipher is a method of encrypting alphabetic text by using a series of interwoven Caesar ciphers, based on the letters of a keyword. It employs a form of polyalphabetic substitution which is a cipher where multiple alphabets are used as a key. The Enigma machine is a complex version of the polyalphabetic substitution cipher.

Algebraic Description of Vigenere Cipher

Vigenère can also be described algebraically by considering the letters [A–Z] as numbers 0–25, and addition is performed modulo 26, Vigenère encryption E using the key K can be written as

$$C_i = E_K(M_i) = (M_i + K_i) \bmod 26$$

and decryption D using the key K as

$$M_i = D_K(C_i) = (C_i - K_i) \bmod 26,$$

in which

$M = M_1 \ldots M_n$ is the message,

$C = C_1 \ldots C_n$ is the ciphertext and

$K = K_1 \ldots K_n$ is the key obtained by repeating the keyword $\lceil n/m \rceil$ times in which m is the keyword length.

CryptAnalysis of Vigenere's Cipher

The idea behind the Vigenère cipher, like all other polyalphabetic ciphers, is to disguise the plaintext letter frequency to interfere with a straightforward application of frequency analysis. For instance, if P is the most frequent letter in a ciphertext whose plaintext is in English, one might suspect that P corresponds to *E* since *E* is the most frequently used letter in English. However, by using the Vigenère cipher, *E* can be enciphered as different ciphertext letters at different points in the message, which defeats simple frequency analysis.

The primary weakness of the Vigenère cipher is the repeating nature of its key. If a cryptanalyst correctly guesses the key's length n, the cipher text can be treated as n interleaved Caesar ciphers, which can easily be broken individually. The key length may be discovered by brute force testing each possible value of n, or using the index-of-coincidence (or Friedman test) can help to determine the key length.

The following table describes the 2 important steps in the CryptAnalysis of the Vigenere's cipher

Step	Description	Reason	How	Comments
1	Key Length Determination.	PolyAlphabetic ciphers use a repeating key. Therefore, the first step in cryptanalysis is to determine the key-length	• Index of Coincidence can be used to accurately estimate the key-length. • The repeating nature of the key lends itself to be a good candidate for IC based key length estimation	• Refer section-AAA(to be filled-in) for introduction and analysis of Index of Coincidence. • • Refer Appendix-X(to be filled in) for details on the algorithm used to estimate key-lengths.
2	Key Determination	The 2nd step in CryptAnalysis is the determination of the actual key. Polyalphabetic based cipher texts can be treated as individual cipher texts and hence lend themselves to good candidates for frequency analysis (when grouped correctly based on the key length)	• Determine the key length in step-1 • Group cipher text characters based on the key length • Perform brute frequency analysis on each group of cipher text by calculating the chi-squared value function	• Refer section-CCC(to be filled-in) for introduction and analysis of Chi-Squared Error for CryptoAnalysis. • • Refer Appendix-Y(to be filled in) for details on the algorithm used to estimate key-lengths.

Section AAA - What is Index of Coincidence?

Coincidence counting is the method (created by William F. Friedman in 1922) of putting two texts next to the other and checking the number of times the same letters appear in the same place across two texts. This ratio is known as the Index of Coincidence or IC for short. In other words, given a content string, the Index of Coincidence is the probability of two arbitrarily chosen letters being the same.

Analysis of Index of Coincidence

- Consider a string of length N.
- Let the string be composed from an alphabet of size θ.
- Assume an alphabet a_i appears T_i times.

The number of ways to pick the 1st instance of the alphabet a_i is -

$$\frac{Fi}{N} \quad \text{- - - - - - - - (equation 1)}$$

The number of ways to pick a 2nd instance of the alphabet a_i is -

$$\frac{(Fi-1)}{(N-1)} \quad \text{- - - - - - - - - - (equation 2)}$$

Note: This is due to the fact that we have already picked the 1st instance of a_i in equation (1) or they are dependent probabilistic events.

Therefore the probability of picking 2nd consecutive a_i is the product of equation (1) and equation (2).

$$\frac{Fi\,(Fi-1)}{N\ (N-1)} \text{- - - - - - - - - (equation 3)}$$

Given that there are θ letters in the alphabet, the index-of-coincidence of picking 2 consecutive letters in the alphabet is

$$IndexofCoincidence = \frac{\sum_{i=1}^{i=\theta} Fi(Fi-1)}{(N)(N-1)} \text{ - - - - (equation 4)}$$

As a rule, the IC for any text from the English language is in the range 0.060 to 0.068.

Application of Index-Of-Coincidence

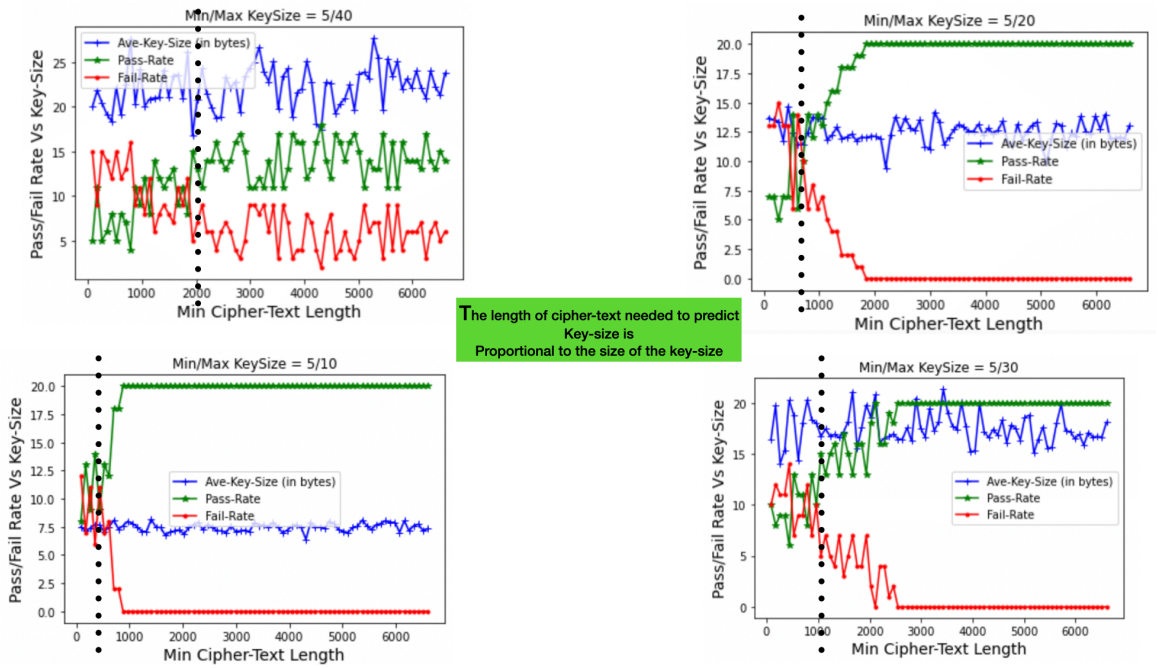
The Index of Coincidence can be used in the analysis of common language in plaintext and in the examination of cryptanalysis.

Vigenere Cipher is a method of encrypting alphabetic text using a polyalphabetic substitution. Given that a repeating-key (of size α) is used, the coincidence rate after every α letters, as a rule, will indicate the size of the key-length. This reality can be utilized to decide the key length, which is initial phase in cracking the system, with a very high probability. There is a strong correlation between the size of the cipher-text available for crypt-analysis and the accuracy with which we can predict the key-size.

The following graphs indicate the minimum packet size (or the minimum length of cipher text) needed to accurately predict the key-size. As the size of the key increases, the number of cipher-text that is needed to accurately predict the key-size also increases.

Reason: Polyalphabetic ciphers disguise the plain text letter frequencies to prevent normal frequency analysis. Therefore, given a large size of cipher-

text, it is still possible to recover the key-size as the main weakness of the Vigenere cipher is the repeating nature of the key



Appendix-X(to-be-filled-in) of this report will describe a detailed algorithm to recover the key-length from a cipher-text encrypted using Vigenere system.

Section BBB - Chi-Squared Error Statistic

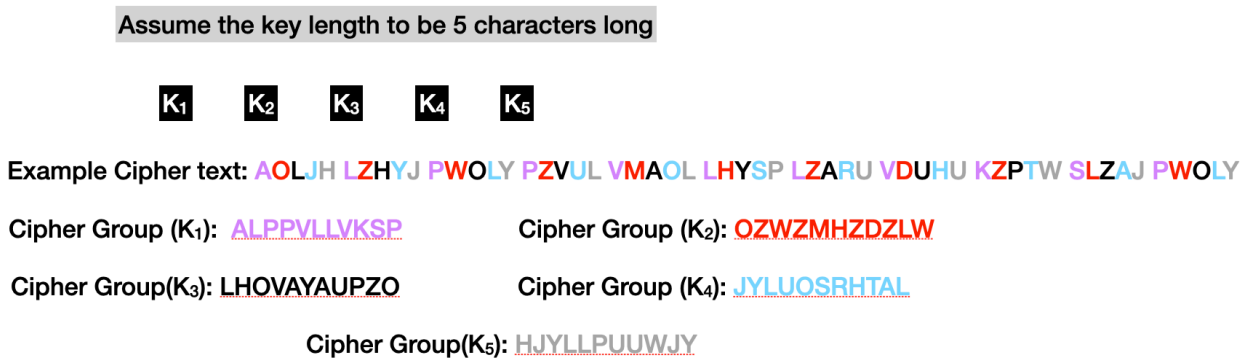
The Chi-squared Statistic is a measure of how similar two categorical probability distributions are. If the two distributions are identical, the chi-squared statistic is 0, if the distributions are very different, some higher number will result. The formula for the chi-squared statistic is:

$$\chi^2(C,E) = \frac{\sum_{i=A}^{i=Z} Ci^2 - Ei^2}{Ei} \text{ - - - - - - - (equation 5)}$$

Where

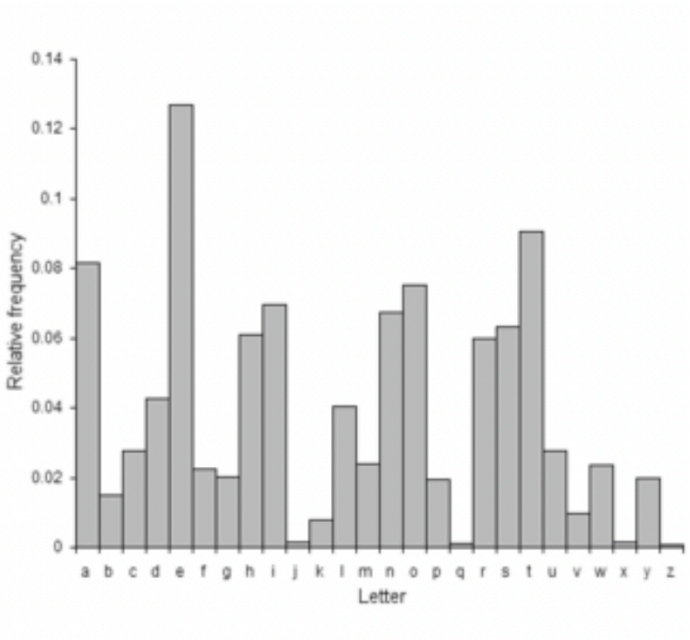
- $C_{A..Z}$ is the count of the letter A..Z
- $E_{A..Z}$ is the expected count of the letter A..Z

Polyalphabetic ciphers disguise the letter frequency by using a different (but repeating key). Therefore, by classifying cipher texts into as a mono alphabetic ciphers groups, it lends itself to be able to perform the chi-squared error based cryptanalysis. The following figures gives a good illustration of 'grouping' a polyalphabetic cipher text into a 'monoalphabetic-based' cipher text group



The following table provides a good indication of the frequency distribution of the letters in the 9th edition of the Oxford English language dictionary(source: put-the-source-here).

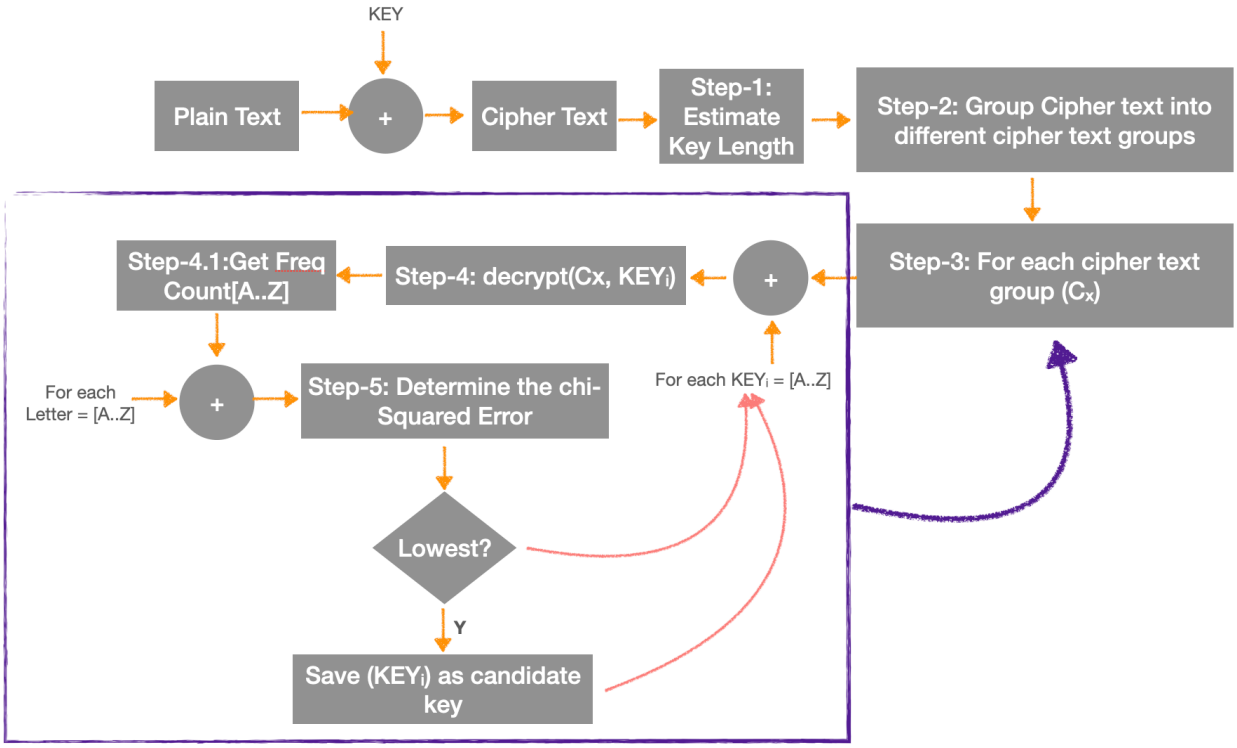
E	11.1607%	56.88	M	3.0129%	15.36
A	8.4966%	43.31	H	3.0034%	15.31
R	7.5809%	38.64	G	2.4705%	12.59
I	7.5448%	38.45	B	2.0720%	10.56
O	7.1635%	36.51	F	1.8121%	9.24
T	6.9509%	35.43	Y	1.7779%	9.06
N	6.6544%	33.92	W	1.2899%	6.57
S	5.7351%	29.23	K	1.1016%	5.61
L	5.4893%	27.98	V	1.0074%	5.13
C	4.5388%	23.13	X	0.2902%	1.48
U	3.6308%	18.51	Z	0.2722%	1.39
D	3.3844%	17.25	J	0.1965%	1.00
P	3.1671%	16.14	Q	0.1962%	(1)



The above table forms one of the key inputs for recovering the key in a polyalphabetic cipher like the vigenere's cipher.

The frequency distribution of each of the letter in cipher text groups can be compared against the 'expected' frequency of those letters to determine the chi-squared error. The lowest chi-squared-value indicates (or correlates) that the 2 distributions match and hence can be used to recover the key.

The following figure illustrates at a high level the method used to recover the key. Appendix-Y (to-be-filled-in) will describe a detailed algorithm to recover the individual key from the cipher text provided.


KEY
Plain Text
+
Cipher Text
Step-1: Estimate Key Length
Step-2: Group Cipher text into different cipher text groups
Step-3: For each cipher text group (Cx)
+
For each KEYi = [A..Z]
Step-4: decrypt(Cx, KEYi)
Step-4.1:Get Freq Count[A..Z]
For each Letter = [A..Z]
+
Step-5: Determine the chi-Squared Error
Lowest?
Y
Save (KEYi) as candidate key