# Code demonstration

Student name: Nam Pham

Student ID: 5196268

**Overview**

The program takes input from “input.txt” which includes a ciphertext encoded by Vigenère Cipher with an unknown key and then print out the key and the plaintext to “output.txt”

My program can be divided into 2 steps:

1. Estimate the key length
2. Find the key

**Estimate the key length**

First of all, I exclude any special characters in the ciphertext and only leave alphabet characters.

To estimate the key length, I find out the Index of Coincidence (I.C.) of the ciphertext. It is a statistical technique that gives an indication of how English-like a piece of text is.

I calculate the Index of Coincidence of the ciphertext then estimate the key length based on this formula:

(Christensen, 2015)

, where n is the number of characters in the ciphertext

I is the Index of Coincidence

Demonstration of the above formula can be found in <https://www.nku.edu/~christensen/1402%20Friedman%20test%202.pdf>, which is listed in the references list.

**Find the key**

The Vigenere cipher applies different Caesar ciphers to consecutive letters and the letters at different positions are encoded by the same Caesar ciphers’ s key. For example, if the key length is 4, the sequence of characters 1, 5, 9, 13, 17,… and 2, 6, 10, 14 are encoded with their own Caesar cipher’s key. Therefore, from the key length, I can find the key by applying Chi-squared statistic and frequency analysis. I take sequences of characters which were encoded by the same Caesar cipher’s key, then try to decode each with each of the 25 possible Caesar ciphers, and compare the frequency distribution of the decoded text with the frequency distribution of English for each key by Chi-squared statistics. The correct key will correspond to the decoded text with the lowest Chi-squared statistic. The number of Caesar cipher’s key that I need to find is the estimated length of the key. ("Practical Cryptography", 2020)

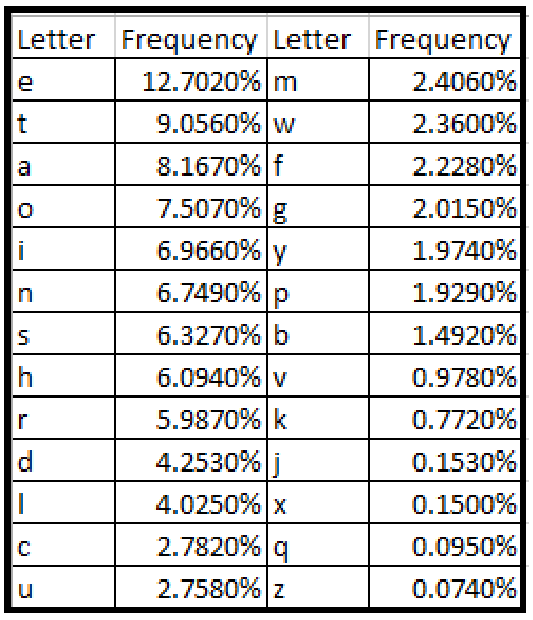


Table 1. Relative Frequency of Letters in the English Language (Agyepong, Enoch & Buchanan, William & Jones, Kevin, 2018)

Chi-squared Statistic is a measure of how similar 2 categorical probability distributions are. If the 2 distributions are identical, the chi-squared statistic is 0. The formula:



("Practical Cryptography", 2020)

, where C(“A”) is the count of letter “A”,

E(“A”) is the expected count of letter “A”

In order to make sure that I find the right key, I check the I.C of the decoded text if it is in range 0.056 - 0.075. This indicates that the decoded text does not differ too much from English which has I.C 0.065. If the I.C of the decoded text is not in this range, I change the key length by adding 1 to it or subtracting 1 from it and find the key again.

After finding the right key, I decrypt the ciphertext to get the plaintext. After that I add any characters that were not alphabet in the original text into the plaintext and capitalize letters appropriately.

**References:**

Christensen, C. (2015). Retrieved 16 April 2020, from https://www.nku.edu/~christensen/1402%20Friedman%20test%202.pdf

Practical Cryptography. (2020). Retrieved 16 April 2020, from <http://practicalcryptography.com/cryptanalysis/stochastic-searching/cryptanalysis-vigenere-cipher/>

Agyepong, Enoch & Buchanan, William & Jones, Kevin. (2018). Detection of Algorithmically Generated Malicious Domain. 13-32. 10.5121/csit.2018.80802.