Muhammad ‘Atif bin Mustaffa 1429619

CSC 4604 Cryptography

Assignment 1: Monoalphabetic Substitution

Documentation

Programs:

1. CharFrequency.py – Count the frequency of each letters in the cipher text.
2. EncDec.py – Encrypt and decrypt the text based on a key.
3. Guess.py – Swap letters in key, generate new key and instantaneously show decryption result.

Text files:

1. ciphertext1.txt – Stores the full cipher text.
2. guess.txt – Stores all the guesses made on Guess.py program.
3. key.txt - Stores a key which will be used with EncDec.py program.

Firstly, I used CharFrequency.py to analyzed the cipher text. I noticed that it contains both uppercase and lowercase letters. However, I wanted to focus only on the lowercase letters first before guessing uppercase letters, so I rewrote the code to convert all letters to lowercase to count frequency. Based on the result, top 10 characters are:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Character | C | U | S | I | M | O | J | K | A | B |
| Frequency | 147 | 106 | 85 | 82 | 79 | 76 | 67 | 63 | 60 | 49 |

From the analysis on letters used in English, top 10 characters are:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Character | E | T | A | O | I | N | S | H | R | D |
| Frequency Percentage | 12.7 | 9.1 | 8.2 | 7.5 | 7.0 | 6.7 | 6.3 | 6.1 | 6.0 | 4.3 |

Based on the both results, letter ‘c’ and ‘u’ are most likely to be letter ‘e’ and ‘t’ respectively because the other letters have small difference in frequency count. So, by using the Guess.py program, I managed to generate a key for letter ‘e’ and ‘t’ as well as the decryption result.

Initial key: abcdefghijklmnopqrstuvwxyz

Guess key: abedcfghijklmnopqrsutvwxyz

Below is the letter analysis for English common letter patterns.

|  |  |
| --- | --- |
| Common pairs | TH, EA, OF, TO, IN, IT, IS, BE, AS, AT, SO, WE, HE, BY, OR, ON, DO, IF, ME, MY, UP |
| Common repeated letters | SS, EE, TT, FF, LL, MM and OO |
| Common triplets | THE, EST, FOR, AND, HIS, ENT or THA |

From the pattern analysis and the decryption result, I looked for common triplets ‘the’ and assumed letter ‘a’ as ‘h’.

Guess key: hbedcfgaijklmnopqrsutvwxyz

Found letters: t h e

From the decryption result, I noticed a repetition of pair letters ‘getteo’ and ‘getteoi’ based on the letter ‘e’ and ‘t’ found. Luckily, I found a website that could list the words that contain certain letters. Website: <http://www.wordplays.com/en/words-that-contain>. I searched on the 6 letters word that contains ‘ette’. There are 4 promising words: ‘better’, ‘letter’, ‘getter’ and ‘setter’. But, since there were two different repetition ‘getteo’ and ‘getteoi’, I could assume that letter ‘i’ is ‘s’. Thus, the only word left out of 4 words is ‘letter’ and ‘letters’ for plural. I managed to swap ‘g’ with ‘l’, ‘o’ with ‘r’ and ‘i’ with ‘s’.

Guess key: hbedcflasjkgmnrpqoiutvwxyz

Found letters: t h e l r s

Then, I found repetition of ‘thmt’ twice, so I looked for common triplets ‘tha’ and swapped letter ‘m’ with ‘h’. So far, I have gotten words of ‘the’ and ‘that’.

Guess key: mbedcflasjkghnrpqoiutvwxyz

Next, I noticed repetition of letters ‘bertaij’ twice, but one is followed by ‘letters’. Again, I used the same method by using the website. The website gave me only 2 most probable words: ‘certain’ and ‘pertain’. I managed to guess the 7 letters word is ‘certain’ because of word ‘letters’ that came after ‘bertaij’. I swapped letter ‘b’ with ‘e’ and ‘j’ with ‘n’.

Guess key: mebdcflasnkghjrpqoiutvwxyz

Found letters: t h e l r s a c i n

Then, I noticed a lot repetition of letters:

1. ‘xrepyen’ 7 times, where some places are ‘xrepyencw’ and one is ‘xrepyencies’..
2. ‘anv’ 6 times.
3. ‘tebt’ 4 times.
4. ‘cigher’ 13 times

Some of the cipher text now can be read and guessed even though some words are not clear. So, based on the findings, I easily guessed that:

1. ‘xrepyen’ is ‘frequen’ because ‘xrepyencw’ is fit for ‘frequency’ and ‘xrepyencies’ fit for ‘frequencies’. Letter ‘x’, ‘p’, ‘y’ and ‘w’ are 'f', 'q', 'u' and 'y' respectively. The letters were swapped with ‘f’, ‘q’, ‘t’ and ‘t’ respectively.

Guess key: mebdcxlasnkghjrqpoiuyvtfwz

Found letters: t h e l r s a c i n f q u y

1. ‘anv’ is ‘and’. Letter ‘v’ was swapped with ‘d’.

Guess key: mebvcxlasnkghjrqpoiuydtfwz

Found letters: t h e l r s a c i n f q u y d

1. ‘tebt’ is ‘text’. Letter ‘e’ was swapped with ‘f’.
2. ‘cigher’ is ‘cipher’. Letter ‘l’ was swapped with ‘q’.

Guess key: mfbvcxqasnkghjrlpoiuydtewz

Found letters: t h e l r s a c i n f q u y d x p

After that, most words in the cipher text became predictable. I easily predict 'lanzuazes' as 'languages', the letter 'z' is 'g' so I swapped 'q' with 'z'.

Guess key: mfbvcxzasnkghjrlpoiuydtewq

Found letters: t h e l r s a c i n f q u y d x p g

To speed things up, I kept on swapping letters in the key until i get the actual key. Letters:

'k' is 'o', swapped 'k' with 'r'.

Guess key: mfbvcxzasnrghjklpoiuydtewq

Found letters: t h e l r s a c i n f q u y d x p g o

'k' is 'w', swapped 'r' with 't'.

Guess key: mfbvcxzasntghjklpoiuydrewq

Found letters: t h e l r s a c i n f q u y d x p g o w

'j' is 'b', swapped 'f' with 'n'.

Guess key: mnbvcxzasftghjklpoiuydrewq

Found letters: t h e l r s a c i n f q u y d x p g o w b

'j' is 'k', swapped 'f' with 't'.

Guess key: mnbvcxzastfghjklpoiuydrewq

Found letters: t h e l r s a c i n f q u y d x p g o w b k

'j' is 'v', swapped 'd' with 't'.

Guess key: mnbvcxzasdfghjklpoiuytrewq

Found letters: t h e l r s a c i n f q u y d x p g o w b k v

The ciphertext is finally readable, except for uppercase letter. However after reading the decrypted ciphertext, uppercase letter key is the same as the lowercase key. Therefore, the final key is:

**mnbvcxzasdfghjklpoiuytrewq**

Finally, I used the EncDec.py program to test the key. The key is correct. All 25 alphabets have been found. Letter 'j' was not used in the plain text.

Plaintext:

Inmostlanguagescertainletterswordsorsymbolsappearatcertainfrequenciesifthetextislong

enoughFrequencyanalysisisbasedonthisideaForExampleinEnglishtexteisthemost

frequentlyusedletterthatmeansitappearsathighestfrequencyThedifferencesbetweenthehigh

frequencylettersandthelowfrequencyletterscanbeusedtoanalyzetheciphertextInthe

appendixtherearestatisticdataformostcommonusedlettersanddigramsandtrigrams

InsubstitutionciphersoneletterisreplacedbyanotherletterTherearemanycategoriesof

substitutionciphersInthissectionwearegoingtodiscussmonoalphabeticsubstitutionciphers

homophoniccipherspolygraphiccipherspolyalphabeticsubstitutionciphersandtheonetime

pad

TheMonoalphabeticSubstitutioncipheralsocalledasSimpleSubstitutioncipheristheonein

whicheachcharacterintheplaintextisreplacedbyacorrespondingonefromacipheralphabet

ThecipheralphabetcanbereversedorshiftedorscrambledAlthoughthenumberofpossible

keysisverylargethiscipherisnotverystrongandconsideredeasily

breakablebyfrequencyanalysisHowevertheadvantageforthiscipheristhatitcanbe

performedbydirectlookupandthetimetoencryptmessageofncharactersisproportionalton