Cryptology (course 1DT075) - Uppsala University
Cryptanalysis of the (modified) Vigenère

Group 10: Deniz Kücükahmetler och Erik Krantz

February – 2020 Spring

1 B

1.1 Compiling and running the code

In order to try different ciphertexts, the content should be changed in the local file named ”text.txt”.

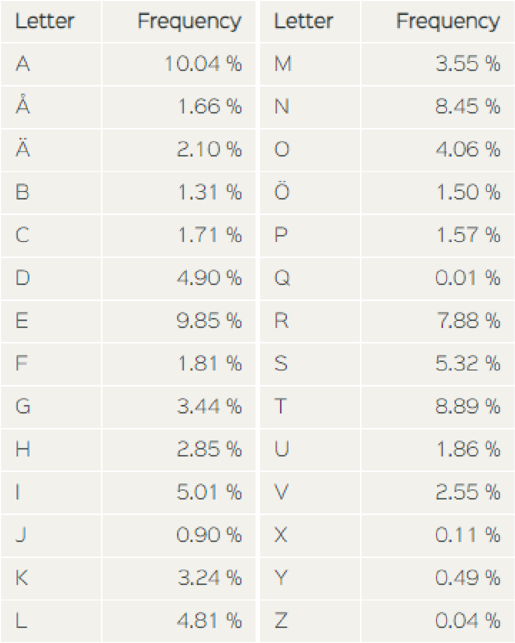
```
> python3 crypto_1a.py
```

1.2 Techniques used

Firstly, Friedman test used to find the key length. The starting point was calculating the index of coincidence of Swedish, using the frequencies of each letter obtained from reliable sources. The ciphertext was split into separate strings according to the changing key length. Then, index of coincidence is calculated for different key lengths. The best length (providing the closest value to the Swedish index of coincidence) taken into consideration. After obtaining the key length, frequency analysis is done for each position of the key. One by one, characters of the key is identified. In the last step, because the key no longer unknown, cipher text is deciphered by shifting the letters back to their actual place in the plaintext.

1.3 Frequency of Swedish letters

The frequencies were taken from *www.sttmedia.com*, ”Alphabet and Character Frequency: Swedish”, as can be seen in figure 1. Since the table did not cover a value for ’W’ we based its value on the position in a list, sorted by frequency of the letter, from *letterfrequency.org*, ”Letter Frequency by Language”. Due to being more frequent than ”z” but less frequent compared to ”x” we its value to be within that range in the previously mention table. Furthermore, we set the value to 0,05%.

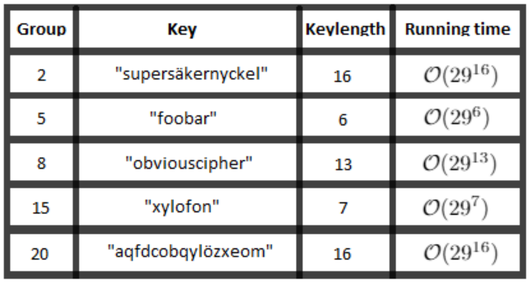


Letter	Frequency	Letter	Frequency
A	10.04 %	M	3.55 %
Å	1.66 %	N	8.45 %
Ä	2.10 %	O	4.06 %
B	1.31 %	Ö	1.50 %
C	1.71 %	P	1.57 %
D	4.90 %	Q	0.01 %
E	9.85 %	R	7.88 %
F	1.81 %	S	5.32 %
G	3.44 %	T	8.89 %
H	2.85 %	U	1.86 %
I	5.01 %	V	2.55 %
J	0.90 %	X	0.11 %
K	3.24 %	Y	0.49 %
L	4.81 %	Z	0.04 %

Figure 1: Frequency table as presented on *www.sttmedia.com*

1.4 Running times of the different ciphertexts

The basic principle is that for a Vigenére we can say that for l = length of the key, and A = letters in the alphabet, by using a brute-force method we would achieve a time complexity of $\mathcal{O}(A^l)$. For a regular Ceasar cipher it would be in constant time since the length of the alphabet is constant: $\mathcal{O}(1)$. For the Vigenére cipher, due to the repeated shifts, we have to take the length of the key into account. Additionally, since we know the number of letters in the Swedish alphabet we can set the value to 29, i.e $\mathcal{O}(29^l)$. In order to calculate the running times for each group we can simply change the value of l, based on the length of the key found as can be seen in figure 2.



Group	Key	Keylength	Running time
2	"supersäkernyckel"	16	$\mathcal{O}(29^{16})$
5	"foobar"	6	$\mathcal{O}(29^{6})$
8	"obviouscipher"	13	$\mathcal{O}(29^{13})$
15	"xylofon"	7	$\mathcal{O}(29^{7})$
20	"aqfdcobqylözxeom"	16	$\mathcal{O}(29^{16})$

Figure 2: Running times based on time complexity

We were unsuccessful in breaking the ciphertexts that was not provided by other groups. There are several factors that could cause this.

- The main cause, which would be the lengths of the keys, would in the worst case scenario be as long as the plaintext and therefore more secure and difficult to break.

- The plaintext was not sufficiently long enough so that it could properly represent the frequency distribution.
- The frequencies the texts were based of differed from the ones we used.
- The length of the alphabet used was greater, i.e it contained more characters, which would in return change the shifts.

1.5 Evaluation

main difficulty for us was applying the code to finding keys of lengths greater than 16, i.e the ones not provided by other groups, as explained in section 1.4. Furthermore, in question 4, turned out to be quite difficult to create a formal proof.

The easiest part was finding out the frequencies since they were easily accessible from a reliable source. Also, while writing the report we had less difficulty for questions 2.1-2.3. The most fun part was the problem solving and the satisfaction of successfully deciphering another groups message. The least fun part was not being able to decrypt the messages provided by the instructor, i.e Text 1-5.

The communication in our group worked great and we were constantly updating one another. Although we were only two in our group the assignment progressed smoothly and both of us contributed to some extent in all of the problems.

2 C

2.1 You have been asked to break a modified version of the Vigenère cipher, with 29 characters instead of 26. Do you think this version is easier to break, more difficult, or of the same level of difficulty?

Since an increase in the amount of characters means there are more possible options there will be a larger dependency on the plaintext being longer. Therefore one can argue that the 29 letter alphabet will be harder to break. Furthermore, the letter frequencies in the alphabet must be taken into account. With more characters the differences between frequencies will also be lowered making it even more difficult to distinguish specific characters. A counter argument however, can be that with more characters the composition of different character alignments is more unique, therefore easier to find a good match, due to a longer alphabet when comparing the frequencies. However, ıt is more relevant with a more significant difference in amount of characters.

2.2 How does the length of the key affect the security of the Vigenère cipher? Are there other characteristics that can impact the security of a particular key?

The security of the Vigenère cipher is highly dependant on the length of the key. With a shorter key the security is much more exposed. This due to the fact that a longer key will make the text appear more random. Since finding a periodicity, i.e repeating keyword occurrences, is important this will be much more difficult with a longer keyword. If the key length is equal to the plaintext, the ciphertext letters would be randomly distributed. So, a pattern between letters can not be recognized theoretically. In addition to key length, also the variety and randomness of the letters in the key have an impact on the security. Because varied letters in the key means, variety of shifts in the plaintext. It will increase the randomness of the key, so shifts of letters in the plaintext will me more random. Thus, it will be more secure. In conclusion, length and diversity of letters have an impact o the security of the key.

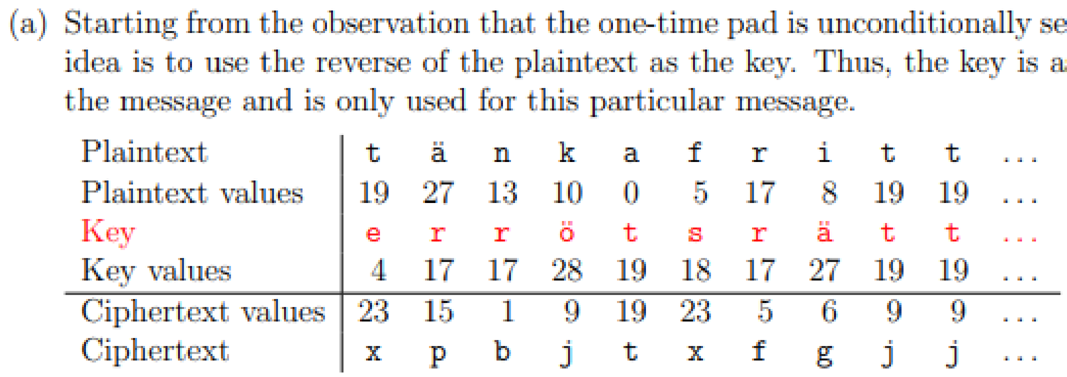
2.3 How could you break the Vigenère cipher if the language of the plaintext is (one of the world's major languages, but precisely which one is) not known?

Although a first thought would be to seek out common themes such as names or name of places this is not a consistent approach since there is no guarantee of them existing in the plaintext nor that the name of a place is spelt in an identical way. It all comes down to frequencies of specific characters. Although, they will most likely not be identical an extreme example would be that characters such as "a" and "e" should have a much high frequency in most known languages compared to a character such as "q" or "z". Given a sufficiently long plaintext, by basing your deciphering on the English alphabet, there is a great chance of finding appropriate frequencies which will match a known major language. So, the frequency analysis can be done using the frequency distribution in English. Also, it depends on the length of the alphabet (it can be estimated by looking the letter variety in the ciphertext, if the ciphertext is long enough, probably all letters of the alphabet can be spotted). Most of the time, shorter alphabet would ease the process. Therefore, it would be a better to determine frequency distribution according to alphabet length. For example, if there are 29 characters spotted in the ciphertext, it is preferable to use Swedish letters frequency distribution in the first place instead of English.

2.4 Examples below are based on the Vigenère cipher. Our task was to determine whether they can be defined as a cryptosystem. Furthermore, if they were more or less secure compared to the Vigenère cipher

To determine if something is defined as a cryptosystem you have to verify that a decryption of an encryption of a message is equal to the message itself for all keys. Basically, you want to confirm that the plaintext is protected going from encrypted to decrypted.

2.4.1 Example: A



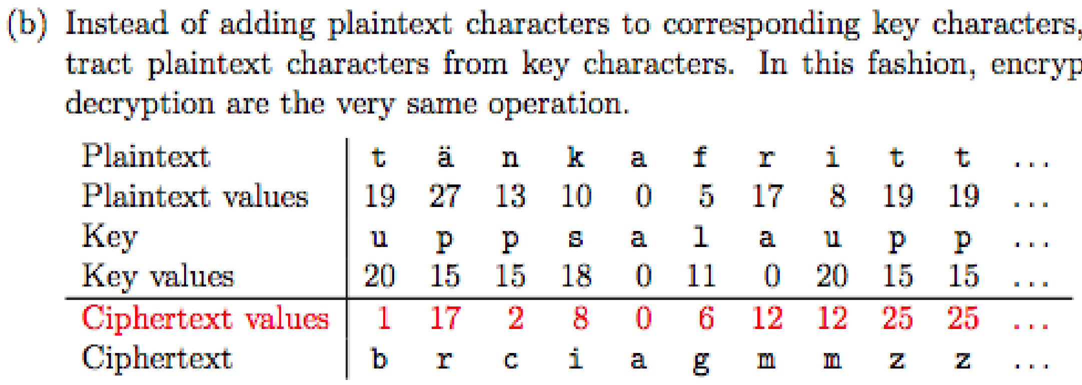
(a) Starting from the observation that the one-time pad is unconditionally secure, the idea is to use the reverse of the plaintext as the key. Thus, the key is as long as the message and is only used for this particular message.

Plaintext	t	ä	n	k	a	f	r	i	t	t	...
Plaintext values	19	27	13	10	0	5	17	8	19	19	...
Key	e	r	r	ö	t	s	r	ä	t	t	...
Key values	4	17	17	28	19	18	17	27	19	19	...
Ciphertext values	23	15	1	9	19	23	5	6	9	9	...
Ciphertext	x	p	b	j	t	x	f	g	j	j	...

Figure 3: Example: A

It can be defined as a cryptosystem since the ciphertext is generated from the plaintext, using the key. It is not a insecure way of generating ciphertext but, because the ciphertext will be symmetric, it would give a clue about the technique and thus the plaintext. Contrarily, because the key is not repeating itself, making a frequency analysis is still impossible so it is more secure than repeated-keyed Vigenère cipher. However, since plaintext is reused in the key it gives more information about the plaintext. Therefore, it makes the ciphering less secure to some extent.

2.4.2 Example: B



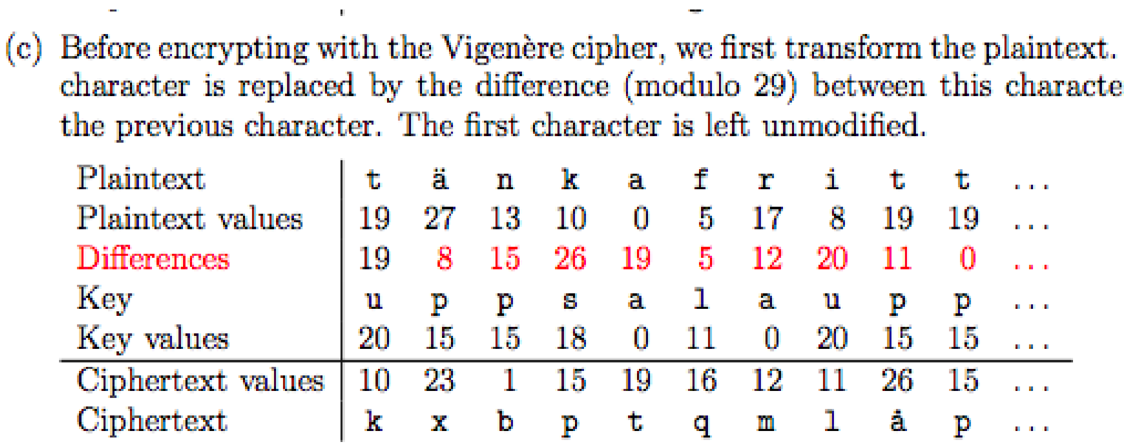
(b) Instead of adding plaintext characters to corresponding key characters, we subtract plaintext characters from key characters. In this fashion, encryption and decryption are the very same operation.

Plaintext	t	ä	n	k	a	f	r	i	t	t	...
Plaintext values	19	27	13	10	0	5	17	8	19	19	...
Key	u	p	p	s	a	l	a	u	p	p	...
Key values	20	15	15	18	0	11	0	20	15	15	...
Ciphertext values	1	17	2	8	0	6	12	12	25	25	...
Ciphertext	b	r	c	i	a	g	m	m	z	z	...

Figure 4: Example: B

It can be defined as a cryptosystem because ciphertext is systematically generated with subtracting key character values from the plaintext character values sequentially. It has the same security with Vigenère cipher since the only difference between them is subtracting the key instead of adding.

2.4.3 Example: C



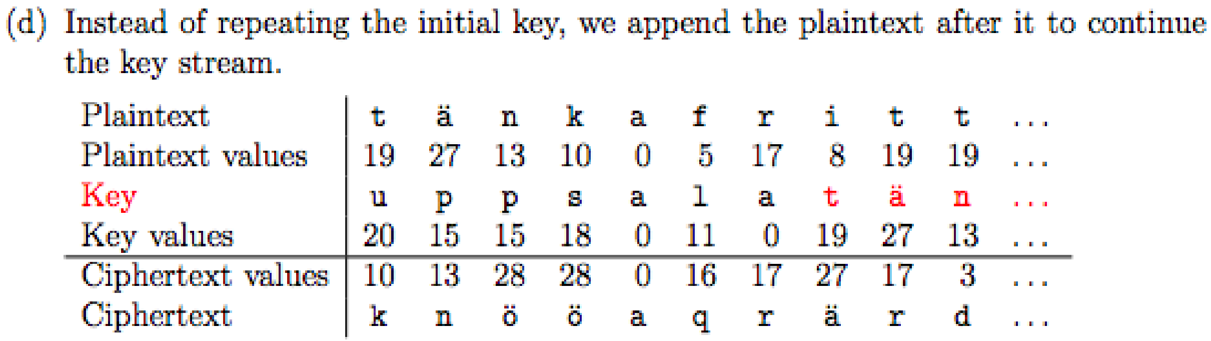
(c) Before encrypting with the Vigenère cipher, we first transform the plaintext. Each character is replaced by the difference (modulo 29) between this character and the previous character. The first character is left unmodified.

Plaintext	t	ä	n	k	a	f	r	i	t	t	...
Plaintext values	19	27	13	10	0	5	17	8	19	19	...
Differences	19	8	15	26	19	5	12	20	11	0	...
Key	u	p	p	s	a	l	a	u	p	p	...
Key values	20	15	15	18	0	11	0	20	15	15	...
Ciphertext values	10	23	1	15	19	16	12	11	26	15	...
Ciphertext	k	x	b	p	t	q	m	l	å	p	...

Figure 5: Example: C

It can be defined as a cryptosystem because the ciphertext is generated with a particular key and using the differences of the plaintext. It is more secure than Vigenère cipher since it is impossible to find the repetitive key because characters of the key is not directly added(except the first character) to the characters of the plaintext. Additionally, since it does not follow a strict one-by-one character comparison, i.e. it ıs reliant on two character values, it is more difficult to find a pattern.

2.4.4 Example: D



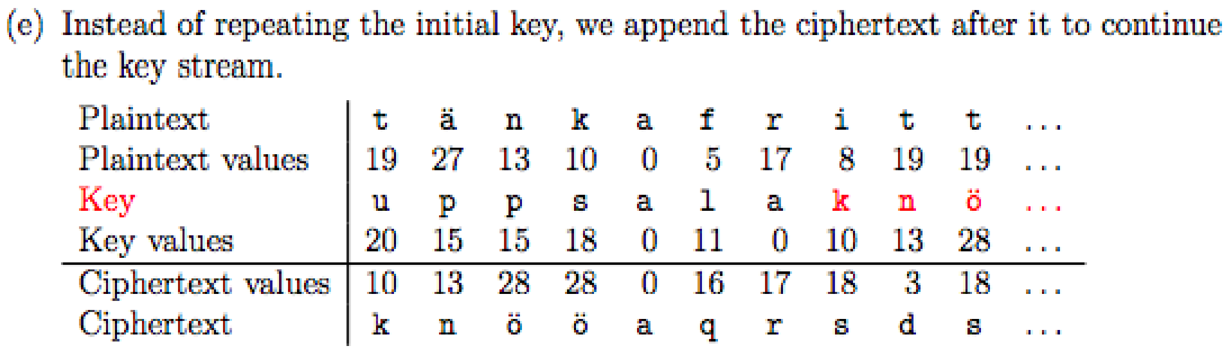
(d) Instead of repeating the initial key, we append the plaintext after it to continue the key stream.

Plaintext	t	ä	n	k	a	f	r	i	t	t	...
Plaintext values	19	27	13	10	0	5	17	8	19	19	...
Key	u	p	p	s	a	l	a	t	ä	n	...
Key values	20	15	15	18	0	11	0	19	27	13	...
Ciphertext values	10	13	28	28	0	16	17	27	17	3	...
Ciphertext	k	n	ö	ö	a	q	r	ä	r	d	...

Figure 6: Example: D

Because the ciphertext is generated systematically adding the plaintext character values to key character values, it can be defined as a cryptosystem. Because key is not repeating itself, it is impossible to catch a pattern for frequencies. Therefore, it is more secure than Vigenère cipher. But, after key ends, plaintext is used as a key. So, key includes information about the plaintext and this may be the weak point of this cryptosystem.

2.4.5 Example: E



(e) Instead of repeating the initial key, we append the ciphertext after it to continue the key stream.

Plaintext	t	ä	n	k	a	f	r	i	t	t	...
Plaintext values	19	27	13	10	0	5	17	8	19	19	...
Key	u	p	p	s	a	l	a	k	n	ö	...
Key values	20	15	15	18	0	11	0	10	13	28	...
Ciphertext values	10	13	28	28	0	16	17	18	3	18	...
Ciphertext	k	n	ö	ö	a	q	r	s	d	s	...

Figure 7: Example: E

It is a cryptosystem with the same reason as example D. Key is again not repeating itself so it is more secure than Vigenère cipher because frequency analysis cannot be done. Different than example D, rather than using plaintext itself, ciphered text is used as a key. Thus, it may be more secure than the previous example. Since the breaker already knows the ciphertext the key provides less new information. Therefore, making it more difficult to break.

References

[1] Alphabet and Character Frequency: Swedish (Svenska), `https://www.sttmedia.com/characterfrequency-swedish`

[2] Letter Frequency by Language, `http://letterfrequency.org/letter-frequency-by-language/`

[3] ”Cryptanalysis” by Tjark Weber, Uppsala University,