Practicals°3: (trying to) decipher Vigenere

The Vigenère Cipher is a polyalphabetic substitution cipher. The method was originally described by Giovan Battista Bellaso in his 1553. However, the scheme was later misattributed to Blaise de Vigenère in the 19th century, and is now widely known as the Vigenère cipher.

The Vigenère Cipher was considered as "le chiffre indéchiffrable" (French for the unbreakable cipher) for 300 years, until in 1863 Friedrich Kasiski published a successful attack on the Vigenère cipher. Charles Babbage had, however, already developed the same test in 1854. Gilbert Vernam worked on the vigenere cipher in the early 1900s, and his work eventually led to the one-time pad, which is a provably unbreakable cipher.

1. The Algorithm

The **'key'** for a vigenere cipher is a keyword. e.g. 'PASSWORD'

The Vigenere Cipher uses the following array (the **'tabula recta'**) to encipher the plaintext:

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**
A	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
B	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
C	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B
D	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C
E	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D
F	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E
G	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F
H	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G
I	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H
J	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I
K	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J
L	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K
M	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L
N	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M
O	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N
P	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Q	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
R	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
S	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
T	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
U	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
V	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
W	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
X	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
Y	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Z	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y

To encipher a message, repeat the keyword above the plaintext:

```
PASSWORDPASSWORDPASSWORD
DEFENDYOURDATAANDACCOUNT
```

The Vigenere cipher was thought to be completely unbreakable for hundreds of years, and indeed, if very long keys are used the Vigenere cipher can be unbreakable. But if short keys are used, or if we have a lot of ciphertext compared to the key length, the Vigenere cipher is quite solvable.

Cryptanalysis of the Vigenere cipher has 2 main steps:

- identify the period of the cipher (the length of the key),
- then find the specific key.

To identify the period we use a test based on the index of coincidence and to find the specific key we will have to use the CHI² statistic. The first thing to note is that there is no guarantee that the period of key that we find is the actual key used. If the message is very long, we can be almost certain of being correct, but the methods provided here are approximate.

1.1 Finding the Period

The Vigenere cipher applies different Caesar ciphers to consecutive letters. If the key is 'PUB', the first letter is enciphered with a Caesar cipher with key 16 (P is the 16th letter of the alphabet), the second letter with another, and the third letter with another. When we get to the 4th letter, it is enciphered using the same cipher as letter 1. As a result, if we gather letters 1,4,7,10,... we should get a sequence of characters, all of which were enciphered using the same Caesar cipher. The sequence of characters 2,5,8,11,... and 3,6,9,12,... will also be enciphered with their own Caesar cipher. The exact sequence will of course depend on the period of the cipher i.e. the key length.

The **Index of Coincidence** (I.C.) is a statistical technique that gives an indication of how English-like a piece of text is (*see link below*). One of the useful properties of the technique is that the result of the I.C. does not change if you apply a substitution cipher to the text. This is because the I.C. is based on letter frequencies, and simple substitution ciphers do not modify the individual letter frequencies. If text is similar to english it will have an I.C. of around 0.06, if the characters are uniformly distributed the I.C. is closer to 0.03-0.04.

To determine the period of a Vigenere cipher, we first assume the key length is 2. We extract the two sequences 1,3,5,7,... and 2,4,6,8,... from the ciphertext. For the example we are working with we get the following result (note that the I.C. is calculated using the whole sequences, not just the part shown). This procedure of breaking up the ciphertext and calculating the I.C. for each subsequence is repeated for all the key lengths we wish to test.

We'll have probably y rows that have very high values of average I.C. This indicates the key is probably of length y, but could also be of length 2y. Both of these probabilities should be tested.

1.2 Finding the Key

Since we now know the period is x, we only have x Caesar ciphers to break, which is fairly easy. For this task we will use the Chi² statistic, which will compare the frequency distribution of our subsequences to the expected English frequency distribution.

In essence, we try deciphering this sequence with each of the 25 possible Caesar ciphers, and compare the frequency distribution of the deciphered text with the frequency distribution of English for each key (see link below). If we perform this, we get 26 values for the Chi² statistic. The correct key will correspond to the deciphered text with the lowest Chi² statistic (we hope, due to the statistical nature of the problem it may be the second or third lowest value).
We have to repeat this procedure for each of the y key letters.

2 Exercise

Your work will be to

1. write some python code (you can write several pieces of code, functions, etc.) in order to decode the following message cyphered with Vigenere code:
2. write a short report in which you'll detail your process and intermediary results. Please send your report before March, 16th by email to Philippe.Truillet@irit.fr

Nota: you known that the password is between 2 and 8 letters
space letter is not coded and must be ignored for decoding

```
fhw tmd af lai tamhnxvd subieoe xhv rqrgltroe etwkqr vxkiqe al xf qdmvekq
fmmyiq cgftlfej lgzqnlbwke afw qrzayxvj iilame mn agxvdnsmmfzad psiwifz
iehijhrdqnl licqclbrx fadxrkqd kmyuqnll jiam sep rdomgh kte ohvcp il tmde
tg ivfhivx wgqcatpzet zbky xenxp tarw drfilwwkv mnv t yeuqmx wgqcatpkk if
vsdbulxv joiwggv roj tiiashtgv qnuhqgmskbrx nola lrddotvv mnv lswfwski
jwidew smsww me fomeslee lai cqavbrx qujhtvmn ubxp un sxvfzammmte afw wgmcw
bx sqnwymke fjhq yugz eimql uhyieek zmmqn tr pvotmkiie rwlirdczxvj mnv
xrxunwxvj rrgf afdlv eirpifz gfypsgmve afw vveeskgy xathvrfojbij euua ej
uram prms ggiim, aakfle dwyieoe sgh jbaux xymlwl geqs akx jmifm-iogpwkc
ryofz sktejl xyus mgmhge kxxkuny firzs laek fhwki rde kmvfzg dbrbe afw e
cany amjfojr sw oodeesarsmmfz bwmavqn lai lzinxvjutq tru mejhwgmcw
vsdbafbij etmwiefs wgvfxl ag e kio qxei oujkmtglmf gfhejbrx fhw yyepaexrkml
fhxzank hj tdilbgrx sqlxvys wffvpdww wpetwfw jmfwmc jqcmkmkk cwkxzriutxzan
agxvdaummmq sqlxvys vxhzoalxh rdczbxvotmkij zelpsiws sgh jkslxqj dese xzye
krwkqmk bqrse sgecksal eifixbgzml agxvxlazieoe agjfdmsmmfz sqlxvys sgh
umtsuejqs
```

3. Links

- **Index of Coincidence,** https://www.dcode.fr/index-coincidence
- **English letter frequency,** http://pi.math.cornell.edu/~mec/2003-2004/cryptography/subs/frequencies.html
- **Chi² staticitics,** http://practicalcryptography.com/cryptanalysis/text-characterisation/chi-squared-statistic/