CIS 3362 Homework #2: Code Breaking - Substitution, Vigenere Solutions

1) Decode the following message, which was encrypted using the substitution cipher. Make sure to discuss all the steps you took, the key you arrived at, and the decoded message.

```
lwzzgjwzxayyebxyrelrwhaiwzmwsytfyawbawjyrayrebxletemaywowiey
ryrelfllexbhabyellaysvfvzekajwzirwauazqzafmlhallfxaljezlylwy
ravszvwlawjyrelhallfxaelywyattgwsyrfyjsyszahallfxalietteoaby
ejgirfygwsbaaoywowywlapszayravzekafboirfyyravzekaeleylatj
```

Solution

First, let's take a look at letter frequencies of the most common letters using Cryptool, posted off the course webpage:

A = 13.5%
E = 8.4%
F = 4.6%
L = 10.5%
R = 5.9%
W = 8.4%
Y = 11.8%
Z = 5.9%

Now, let's see what the common ngrams are, looking for ones that appear four times:

REL, YRA, VZEKA, RFY, YREL, ZEKA

The most interesting thing to point out here is the pattern VZEKA. The only reason ZEKA appears four times is because it's a substring of VZEKA. So, VZEKA is very likely a word. Then, note that V and K aren't common, but Z, E and A are quite common. In fact, structure wise, with A = 13.5% and the fact that many words end in 'E', it's likely that the cipher text A maps to the plaintext E.

Now, cross reference this list to see that the only three letter word ending in A (ciphertext) is YRA. Because the frequency of Y is very high, it's extremely likely that this word is "THE". Now we need to find a five letter word that fits the form _ _ _ _ E, where the four blanks are NOT T or H. It's likely that E in the ciphertext maps to a vowel, due to the structure of most five letter words that end in E. Based on the frequency, this vowel is likely to be either A or I.

Now, let's look at YREL. This word starts with "TH" and can NOT be the word "THAT". Thus, YREL is very likely to be "THIS".

Here are the mappings we have so far:

Y → T R → H A → E E → I L → S

Going back to our five letter word, we have _ _ I _ E, where the other three letters are consonants we haven't figured out yet. With K at 1.2% and every occurrence of K being part of VZEKA, it's

likely that K maps to an extremely unlikely letter. If we guess 'Z, then the rest of the matching data is consistent with the word "PRIZE". So, here are our current set of matchings:

Y → T R → H A → E E → I L → S V → P
Z → R K → Z

With just these substitutions in Cryptool, here is what we get:

S-RR---R-ETTI--THISH--E--R---T--TE--E--THETHI--SI-I-ET----IT
HTHIS-SSI---E-TISSET-P-PRIZE--R-H-E-ER-RE--S-ESS--ES-IRSTS-T
HEP-RP-SE--THIS-ESS--EIST-TE-----TH-T--T-RE-ESS--ES-I--I-E-T
I---H-T----EE-T---T-SE--RETHEPRIZE----H-TTHEPRIZEISITSE—

From here, we can start making out words. The first word is probably "SORRY" (few words fit the pattern S _ R R _.) This makes W → O, G → Y. Plug these in. The second word has the second and third letters "OR", so it's likely to be the word "FOR", so we add the substitution J → F. Then we notice that the ciphertext letter 'X' appears twice in the next three blanks. Noting that "ING" is a common word ending, we can guess these mappings: B → N, X → G. Now, our message looks like this:

SORRYFORGETTINGTHISHO-E-OR-O-T--TEONEOFTHETHINGSI-I-ETO-O-IT
HTHIS-SSIGN-ENTISSET-P-PRIZEFOR-HOE-ER-RE--S-ESS-GESFIRSTSOT
HEP-RPOSEOFTHIS-ESS-GEISTOTE--YO-TH-TF-T-RE-ESS-GES-I--I-ENT
IFY-H-TYO-NEE-TO-OTOSE--RETHEPRIZE-N--H-TTHEPRIZEISITSE-F

So few letters are missing, we can continue to fill in words we know. The first one that stands out is "ASSIGNMENT" with "SSIGN" given (in the beginning of line two). This gives us the ciphertext characters that map to both A and M. So F → A and H → M. When we plug this in, the word "HOMEWORK" pops out on line one, giving us these mappings: I → W, M → K. When we read the first few words, "SORRY FOR GETTING THIS HOMEWORK", it's very likely the next two words are "OUT LATE". This gives us the mappings: S → U, T → L. With these substitutions, we have (with spaces inserted):

SORRY FOR GETTING THIS HOMEWORK OUT LATE. ONE OF THE THINGS I LIKE TO -O WITH THIS ASSIGNMENT IS SET UP A PRIZE FOR WHOE-ER -REAKS MESSAGES FIRST. SO THE PURPOSE OF THIS MESSAGE IS TO TELL YOU THAT FUTURE MESSAGES WILL I-ENTIFY WHAT YOU NEE- TO -OTO SE-URE THE PRIZE AN- WHAT THE PRIZE IS ITSELF

The key mapping missing is D, which is the first blank to make the word "DO". This mapping is O → D. The next blank is V, so we have: U → V. The next blank is B, so we have: Q → B. The last blank not filled in is a 'C', to complete the word secure. Here is the decrypted message:

SORRY FOR GETTING THIS HOMEWORK OUT LATE. ONE OF THE THINGS I LIKE TO DO WITH THIS ASSIGNMENT IS SET UP A PRIZE FOR WHOEVER BREAKS MESSAGES FIRST. SO THE PURPOSE OF THIS MESSAGE IS TO TELL YOU THAT FUTURE MESSAGES WILL IDENTIFY WHAT YOU NEED TO DO TO SECURE THE PRIZE AND WHAT THE PRIZE IS ITSELF

2) Decode the following message, which was encrypted using the substitution cipher. Make sure to discuss all the steps you took, the key you arrived at, and the decoded message.

```
cubphgtociplcncdllczgajxhbztxsxcepbudgcocxpjocephgmjgzczgfsp
xapdgtxrtxptzgmprltudubphwmcamrpjdbgmjgczftuvufjlczgajxhftuj
xplcncdlmjoztzgmprtdpfgtgmpatrsjdfztxzxppjdhgmptgmpxmjozgtft
uxcdgpdhphxpacscpdggmcbbtudhbcdpzzcacpdggtrpbtgmpsxcepztxvxp
jicdljoogmpathpbzcxbgztxgmcbmtrpwtxicbgwpdgfhtoojxbcdajbmgtv
pbsocgcdgtgwtgpdhtoojxajbmsxcepbztxvtgmsptsopcdgmplxtus
```

Solution

Looking at repeated n-grams, we see that "GMP" appears 27 times and the corresponding letter frequencies are similar to what we might expect for "THE". Put in these matchings: G → T, M → H, P → E.

The next most common n-gram is "ZTX", which has no letters in common with "GMP". So, we are looking for a common trigram that does NOT contain T,H or E. Some of these are: ING, AND, WAS and FOR. The letter frequencies for ZTX are 5%, 9.2% and 7%. Based on the frequency data, we expect a vowel in the middle, so "WAS" or "FOR" are out best candidates from this list. Plugging in "ZTX" for "WAS" definitely doesn't work as it creates the string "ASEAWTHE". Now, try "FOR". At first I thought this wouldn't work because it creates the string "FOFTHE". But, then I realized that this could just be "OF THE" and the F was the end of a previous word.

```
---E-TO---E--------FT--R--FOR-R--E---T---RE----E-TH-TF-FT--E
R-E-TOR-OREOFTHE--O----E--H--H-E---TH-T-F-O-------FT--R--O--
RE------H--FOFTHE-O-E-TOTHE-O-----FORFREE---THEOTHERH--FTO-O
-R--TE--E-RE----E-TTH---O------EFF---E-TTO-E-OTHE-R--EFOR-RE
--------THE-O-E-F-R-TFORTH--HO-E-OR---T-E-T--O---R------HTO-
E----T--TOT-OTE--O---R---H-R--E-FOR-OTH-EO--E--THE-RO--
```

There looks to be one word after the last the and a word right before it that might just be two letters long. The ciphertext for these two letters is CD and the letter frequencies are 9% and 5.9%, respectively. A strong candidate here is a common vowel, consonant digram. Knowing that we already have T,H,E,F,O, and R mapped, if we remove the common digrams that have these letters, we are left with: IN, AN, ND, NG, AS, and IS. We can throw out both ND and NG based on letter frequencies. Of these, IN seems promising, so let's try this, using C → I and D→ N. This yields:

```
I--E-TO-I-E-I-IN--IFT--R--FOR-RI-E--NTI-IRE--I-E-TH-TFIFT--E
R-ENTOR-OREOFTHE--O-N--E--HI-H-E-N-TH-TIF-O------IFT--R--O--
RE-I-IN-H--FOFTHE-ONE-TOTHE-O---N-FORFREE-N-THEOTHERH--FTO-O
-RINTEN-E-RE-I-IENTTHI--O-N--INEFFI-IENTTO-E-OTHE-RI-EFOR-RE
--IN----THE-O-E-FIR-TFORTHI-HO-E-OR-I-T-ENT--O---R-IN---HTO-
E---ITINTOT-OTEN-O---R---H-RI-E-FOR-OTH-EO--EINTHE-RO--
```

Reading towards the end of the first like, we can make out "FIFT—ER-ENT" which looks like "FIFTYPERCENT". So, we can get mappings for Y, P and C: F → Y, S → P, A → C.

With these mappings, we get really close:

```
I--E-TO-I-E-I-IN--IFTC-R--FORPRI-E--NTI-IRE--I-E-TH-TFIFTYPE
RCENTOR-OREOFTHE--O-N--E--HICH-E-N-TH-TIFYO---Y--IFTC-R-YO--
RE-I-IN-H--FOFTHE-ONEYTOTHECO-P-NYFORFREE-N-THEOTHERH--FTOYO
-RINTEN-E-RECIPIENTTHI--O-N--INEFFICIENTTO-E-OTHEPRI-EFOR-RE
--IN----THECO-E-FIR-TFORTHI-HO-E-OR-I-T-ENTY-O---R-INC--HTO-
E-P-ITINTOT-OTEN-O---RC--HPRI-E-FOR-OTHPEOP-EINTHE-RO-P
```

From here, we can see the word "CO-P-NY", which is COMPANY and allows us to get the mappings for M and A: R → M, J → A. Reading the last line, it looks like the ending is "FORBOTHPEOPLEINTHEGROUP". This gives us mappings for B, L, G and U: V → B, O → L, L → G, U → U. This gives us:

```
IU-E-TOLI-EGI-INGGIFTCAR--FORPRI-E-UNTILIREALI-E-THATFIFTYPE
RCENTORMOREOFTHEMGOUNU-E--HICHMEAN-THATIFYOUBUYAGIFTCAR-YOUA
REGI-INGHALFOFTHEMONEYTOTHECOMPANYFORFREEANTHEOTHERHALFTOYO
URINTEN-E-RECIPIENTTHI--OUN--INEFFICIENTTOME-OTHEPRI-EFORBRE
A-INGALLTHECO-E-FIR-TFORTHI-HOME-OR-I-T-ENTY-OLLAR-INCA-HTOB
E-PLITINTOT-OTEN-OLLARCA-HPRI-E-FORBOTHPEOPLEINTHEGROUP
```

From here, we can fill in the rest of the letters, just but filling in words. The message begins, "I used to like giving gift cards"…Here is the full message and all of the mappings:

I USED TO LIKE GIVING GIFT CARDS FOR PRIZES UNTIL I REALIZED THAT FIFTY PERCENT OR MORE OF THEM GO UNUSED WHICH MEANS THAT IF YOU BUY A GIFT CARD YOU ARE GIVING HALF OF THE MONEY TO THE COMPANY FOR FREE AND THE OTHER HALF TO YOUR INTENDED RECIPIENT THIS SOUNDS INEFFICIENT TO ME SO THE PRIZE FOR BREAKING ALL THE CODES FIRST FOR THIS HOMEWORK IS TWENTY DOLLARS IN CASH TO BE SPLIT INTO TWO TEN DOLLAR CASH PRIZES FOR BOTH PEOPLE IN THE GROUP

3) Decode the following message, which was encrypted using the Vigenere cipher. Make sure to discuss all the steps you took, the key you arrived at, and the decoded message.

```
xpvdetmwmevtebzqwbriprqhmbptgvhfhmayeyvrxhgaafaaiexhbmhuwlsdt
vngdaqiiihsafislmjzchvmizwmirswbbnavwiiuserogxhgcpmzrogvpmecm
azsksnqlhxthfurweexrfmrggmsidrptmbylaytrlhussmskwegbnexkytise
rogqnwkiiyehuqhvytvehsqwtvvgiqoiapxxjpvrulgiblvpvqrogltzvrefa
favnwzhmgmrgimlvaefmmqwlextavelurwmtpxrphaxaitwsfxnurlxijgghr
uw
```

Solution

We can first use Cryptool and look at the Index of Coincidence for several different possible keyword lengths.

Here are the results we get:

n = 2:	4.37%	4.57%							
n = 3:	4.81%	4.01%	5.48%						
n = 4:	3.96%	4.35%	4.35%	4.97%					
n = 5:	5.79%	6.44%	4.51%	5.25%	4.97% (note these are higher, but not high enough)				
n = 6:	4.47%	3.59%	5.55%	4.24%	3.83%	6.93%			
n = 7:	5.98%	4.20%	5.09%	4.09%	5.42%	5.98%	5.11%		
n = 8:	4.97%	4.97%	4.69%	4.35%	4.20%	3.30%	3.90%	5.85%	
n = 9:	5.52%	4.35%	5.68%	4.92%	5.87%	7.00%	3.97%	3.78%	5.30%
n = 10:	7.12%	6.66%	6.20%	8.27%	5.74%	11.72%	7.12%	7.63%	5.91%	8.12%

BINGO! Notice how much higher the Index of Coincidence is, on average, for each bin, with n = 10. This is almost definitely the keyword length.

Now, what we can try to do to automate the process a bit is write some code that calculates the relative Mutual Index of Coincidence between bin 1 and bin 2, with bin 2 shifted in between 0 and 25, characters. Of these 26 calculations, we can store the two highest. We will then repeat this process for bin 1 and bin 3, then bin 1 and bin 4, and so forth. **The code is a separate file posted with the solutions: micanalysisq3.java.** Here is what it prints out for possible ***relative shifts*** to bin 1:

Bin 2 = 15 or 11
Bin 3 = 24 or 9
Bin 4 = 10 or 23
Bin 5 = 11 or 18
Bin 6 = 13 or 23
Bin 7 = 15 or 3
Bin 8 = 4 or 15
Bin 9 = 15 or 19
Bin 10 = 2 or 15

What we can first try is guessing that the max MIC is correct and list all 26 possible keywords. For example, if the first keyword letter was A, then the keyword would be:

APYKLNPEPC (notice that P is 15 letters after A, Y is 24 letters after A, and so forth.)

Here is the full list:

```
apyklnpepc
bqzlmoqfqd
cramnprgre
dsbnoqshsf
etcoprtitg
fudpqsujuh
gveqrtvkvi
hwfrsuwlwj
ixgstvxmxk
jyhtuwynyl
kziuvxzozm
lajvwyapan
mbkwxzbqbo
nclxyacrcp
odmyzbdsdq
penzaceter
qfoabdfufs
rgpbcegvgt
shqcdfhwhu
tirdegixiv
ujsefhjyjw
vktfgikzkx
wlughjlaly
xmvhikmbmz
ynwijlncna
zoxjkmodob
```

The item underlined seems closest to a keyword. If we try to decrypt with that keyword, we get the following:

```
ilieeridingprczosinrandimzlacesbunawafraidtbadwhenidonhsssomerahdymper
songiqhtfindinixsteadsocnyrderforsoetoclaimsoerwinninasioumustgifsndan
insnrectoroncumzuswhenyiuqotothatcnctructorsoemusttelfhsmorherrisosare
redpiyletsarevleeiheardsoehavecasbfyrusisitnreethelasnmossagewiflsndic
ateqhythechosynsnstructirss
```

This looks pretty close. In fact, it's extremely likely that the first few words are "I LIKE HIDING PRIZES". This would indicate that letter 4 of our keyword guess is wrong, as well as letter 6. Letter 4 of the ciphertext is 'D' and letter 4 of what we think is the correct plaintext is 'K'. This means that the fourth letter of the keyword should actually be 'D' – 'K' = 3 – 10 = -7 ≡ 19 (mod 26). This is the letter 'T'. Letter 6 of the ciphertext is 'T' and letter 6 of what we think is the correct plaintext is 'H'. This means that the sixth letter of the keyword should actually be 'T' – 'H' = 19 – 7 = 12. This is the letter 'M'. Our updated keyword is "PENTAMETER". Now, plug this into get the correct plaintext:

I like hiding prizes in random places but am afraid that when I do this some random person might find it instead. So in order for you to claim your winnings you must go find an instructor on campus. When you go to that instructor you must tell him or her, "roses are red, violets are blue I heard you have cash for us is it true?" The last message will indicate who the chosen instructor is.

4) Decode the following message, which was encrypted using the Vigenere cipher. Make sure to discuss all the steps you took, the key you arrived at, and the decoded message.

```
myekqkosytsndunfyerocelrnvyxsltxlbjaimgwkcvqhwrvhxoxlymxdaipb
orprelrndzlpvfxxhitjannhpkekkmzzmxyecibjnjfgaaejyvwiirtyxtach
lgjfhplpzrzznzqkerroahrtmunieyyphfkpbrkyhdpejezmoavjgekxekthp
srmeiyoatbojlxiukerqxjxszwxkwjoeqiddeftoirbziyochlwdxebmrpvxf
uoiakmedwlfljzzwwjjyvrnyqvncaoxazshqloerzeguiakhqkzoxazsamkpl
pixavpvelsscyqbxhotxhtjjy
```

Solution

Note, in order to save some time, the same process will be used to break this cipher, but not all of the data will be presented. Rather the relevant results will only be listed.

The best choice for number of bins is 11. Here are the corresponding Index of Coincidences for each bin:

7.14% 5.66% 5.66% 6.89% 5.66% 5.17% 5.66% 4.43% 5.91% 5.41% 8.62%

Next, we use a similar program to last time printing out most likely relative shifts to the first bin (just using the max MIC for each pair of bins). **The corresponding program used was micanalysisq4.java.**

24, 7, 2, 5, 0, 7, 11, 4, 2 and 23.

None of the keywords produced look recognizable. Here are a couple that maybe look like the resemble something:

GENILGNRKID
TRAVYTAEXVQ

Just for kicks, let's try the first few letters of the ciphertext with these two key words:

M – G = 12 – 6 = 6 G
Y – E = 24 – 4 = 20 U
E – N = 4 – 13 = -9 = 17 R
K – I = 10 – 8 = C

M – T = 12 – 19 = -7 = 19 T
Y – R = 24 – 17 = 7 H
E – A = 4 – 0 = E
K – V = 10 – 21 = -11 = 15 P

The latter message “THE P” seems more promising. Let’s use this as a basis and then see which letters in this keyword have to be fixed.

So, let’s decrypt with this and see what we get, maybe we’ll be able to make out something:

thepsroobyc**umus**hfenrho**sana**aeshwcv**isan**onwgfaaofrajeotodwemanrionswosani
bsprichortocundsportmebtxuhhsisnohnawhoicfintaytwfmougohophsebginesren
utkoatrwuiabdzookudykukizl**seeovarmlorge**pvopourophofviofocsbtewsejtvege
twoqlqegiomcontizebtmouwizlbiuufeoutkhkihigsolobgwsmoicandscnydthhisms
soaueondifmoqgcthhisforphsnhhepagssofd**aight**vehpmoi

If we think we have the first four letters right, we can try to highlight these and see if we can guess a word at all in between the correct letters.

If we see the first underlined phrase, it seems like we can make out the word “see” in the beginning and “large” at the end. If we can figure out the letters in between, we’ll have the keyword! (Any 11 letters in a row will designate the keyword.) One guess would be “SEE OVER A LARGE”. Let’s quickly give that a shot, calculating the corresponding keyword. This would be “TRAVYTWEJVD”. Here is what we get when we try this keyword:

thepsrsopyoumushfinfhasanaaewhkchisanonagtamofrajestcdiemanrisngwasanibstr
wctortocurdgpartmebtbuvheisnohnewvoucfintactkfyougohothgenginesrinitwoatrw
umapdlookudyouyillseeoveralargepvotoiraphofvisfccebtewsentjesetwoqlueuiamc
ontideptyouwizlfiiureoutkhoivissolobgasaoucandscryrtthismsssaieandifmougqt
thisforthgnthepagswotdmightvelpaou

This seems to be quite a bit closer, but it’s definitely not correct. But, if you use this version, it seems like we can do a better job of guessing the first few words of the message: “THE PERSON YOU MUST FIND”. Just using the first three words, we can work out the corresponding keyword by subtracting this proposed plaintext from the ciphertext. This gives us a keyword of “TRAVMTWELVE”. Now, let’s put this keyword in and see what we get:

The person you must find has a name which is an anagram of ravestadiem and is new as an instructor to our department but he is not new to ucf. In fact, if you go to the engineering two atrium and look up you will see a very large photograph of his face. Between these two clues, I am confident you will figure out who it is. So long as you can decrypt this message and if you got this far, then the password might help you.

The password is Travis Meade’s email address, spelled out. This is who had the cash prize! His picture is in the ENG2-202 atrium banner from when he was a programming team member. That banner he is on was to commemorate the 2013-4 UCF Programming Team, which finished tied for 19th at the ACM ICPC World Finals.