1. In this question you will practice your code-breaking skills.

   (a) The following ciphertext was derived by applying a substitution cipher.

   ```
   NTVW STVD NVJV MEE GW STV YJMNGWRJPPB STV QUVFSGPWF NTGIT VEGHMCVST
   TMY MEJVMYD MFKVY NVJV PO IPUJFV JVZVMSVY CD STV PSTVJF MWY STVD FPPW
   OPUWY STMS XMWV TMY WP GWSVEEGRVWIV SP RGAV STV FMWRUGWV TPZV PO RPPY
   TPNVAVJ NTGIT STV CVWVAPEVWIV PO TVJ TVMJS FURRVFSVY TMY WPS DVS
   YVFVJSVY TVJ FTV FSGEE VLZVISVY STMS GS NPUEY MEE VWY NVEE MWY STMS
   VAVJD BPJWGWR NPUEY CJGWR FPBV EVSSVJ VGSTVJ OJPB EDYGM PJ TVJ OMSTVJ
   SP VLZEMGW STVGJ ZJPIVVYGWRF MWY ZVJTMZF MWWPUWIV STV BMJJGMRV
   ```

   It is a paragraph from *Pride and Prejudice*, whose main characters are `elizabeth`. Using this side information, decrypt the word `CVWVAPEVWIV`. (You may take advantage of the spaces.)

   Solution: `elizabeth` is nine letters long. There are three nine-letter words in the ciphertext: `QUVFSGPWF`, `VEGHMCVST`, and `FURRVFSVY`. The symbol `V` is by far the most frequent (66 occurrences, followed by `S` with 34), so it is likely to encrypt an `e`. We hypothesize that `VEGHMCVST` is an encryption of `elizabeth`. If that is so, the first line partially decrypts to

   ```
   NheW theD NeJe all iW the YJaNiWRJPPB the QUeFtiPWF NhiIh elizabeth
   ```

   We can now guess the matches `Nw`, `Wn`, `Dy`, `Ic`. With this `CVWVAPEVWIV` partially decrypts to `beneAPlence`. The word is `benevolence`.

   (b) Decrypt the short pad ciphertext from page 7 of the Lecture 1 notes using the method described there. You may use the information that the key is 4 symbols long.

   Solution: First, we split the ciphertext into four parts, consisting of those symbols in positions 1, 2, 3, and 4 modulo four:

   $c_1 =$ `EAKBBEXKEQLELOQBLBPPXY`$\cdots$
   $c_2 =$ `GDUVYOTGQJQTVINPOJGGIA`$\cdots$
   $c_3 =$ `YVFZZIUINKIVYVVWKFNEVK`$\cdots$
   $c_4 =$ `IMUUBOQEQEBMMVUZPCIOLP`$\cdots$.

   We then find the key symbols k_i that minimize the square distance for ciphertext c_i given by equation (1) in the notes. The minima are attained at $k_1 = 23, k_2 = 2, k_3 = 17, k_4 = 8$. The candidate key is therefore `XCRI`. The first three lines of the decrypted ciphertext are

   ```
   he had been some time with mr gardiner who with two or three other
   gentlemen from the house was engaged by the river and had left him
   only on learning that the ladies of the family intended a visit to
   ```

   (c) Complete the decryption of the short pad ciphertext from page 8 in the Lecture 1 notes.

   Solution: Rewriting the ciphertext in ten symbol chunks,

   ```
   CVWWRJBRVN CVWWRJBIFA MCSCFCBBCJ ZTCPYQUBMH FMJWRJBIFA MCSCFCBBTT
   MMZYLQUBMH FMJWRJBIFA MCSCFCBBCJ ZCSCFCBBCJ ZTCPYQUBM
   ```

   The first two chunks both start with `CVWWRJB`. For this alignment to happen, the first chunk likely encrypts two `buy`s and one `sell`. As the overlap consists of seven symbols, it is sensible to hypothesize that the plaintext starts with either `buysellbuy` or `sellbuybuy`. We derive two candidate keys `BBYENYQQBP` and `KRLLQPDQBP`. Only the first one yields the sensible plaintext

   ```
   buy sell buy buy sell sell buy sell buy sell sell sell sell sell buy
   sell sell buy sell sell sell sell buy sell buy buy sell buy sell sell
   ```

The same conclusion could have been derived by trying out all possibilities for the first ten plaintext symbols (`buybuybuyb`, `buybuysell`, and so on), deriving candidate keys, and filtering out the only one that is consistent with a sequence of `buy`s and `sell`s. This strategy would have worked even if we didn't know that the exact key length, but merely that it is bounded by ten.

2. These are the three most frequent digrams in ciphertexts obtained by encrypting the same 1715 character paragraph from *Pride and Prejudice* with (a) the shift cipher, (b) the substitution cipher, (c) the short pad with key length 3, and (d) the one-time pad. Which digram statistics arose from which cipher? Explain your reasoning.

1. digram	frequency	2. digram	frequency	3. digram	frequency	4. digram	frequency
`TV`	48	`LX`	22	`LW`	8	`GD`	48
`VJ`	43	`HH`	19	`LH`	8	`DQ`	43
`ST`	41	`AY`	18	`VT`	8	`SG`	41
`GW`	29	`XA`	18	`BG`	8	`HM`	29
`PW`	27	`YV`	17	`NS`	7	`NM`	27

Solution: In the shift cipher and substitution cipher, the same English digram always encrypts to the same ciphertext digram. We would therefore expect frequent digrams like `th`, `he`, and `er` to produce ciphertext digrams that are high on the list. In table 1, digrams `TV`, `VJ` and `ST` occur 40-50 times each, or with frequency about $2-3\%$. This is consistent with the fraction of occurrences of `th`, `he`, `er` in English. The same is true for digrams `GD`, `DQ`, `SG` in table 4?

Hypothesizing these matches are correct, how can we tell apart the substitution cipher from the shift cipher? The mapping `tS`, `hT`, `eV`, `rJ` is not consistent with a shift by a fixed offset. In contrast, `tS`, `hG`, `eD`, `rQ` is consistent with every English letter being shifted by -1. Table 1 therefore arises from a substitution cipher, while table 4 arises from a shift cipher.

In the short pad with key length 3, common English digrams like `th` do not always map to the same ciphertext digram. However, they sometimes do. For example, if the word `the` occurs at positions 1 and 7 in the plaintext then the corresponding portions of the ciphertext will be identical. We can try to estimate the effect of this repetition. In a 1715 character text we might expect them to occur about 40-50 times each. (From tables 1 and 4 we can see that these statistics are correct.) Among these occurrences, there would be about 13-18 starting at the same position modulo 3. This is roughly consistent with the statistics in table 2.

In the one-time pad, all ciphertexts are equally likely. The statistics of the 1714 possible digrams among the $26^2 = 676$ possibilities roughly follows the occupancy numbers obtained by assigning 1714 balls randomly into 676 boxes. (This is not completely accurate because consecutive digrams are not independent.) On average, a digram will occur $1714/26^2 \approx 2.54$ and the digram occurrences should not deviate too much from that. The calculations that show this are too much for this course, but the table 3 statistics are the best fit for this model among the alternatives here.

3. Among the following four samplers, which produce indistinguishable outputs? Among those pairs that don't, what is the distinguisher with the highest advantage that you can come up with?

Sampler A: Output x, where x is a pure random number modulo 6.
Sampler B: Output $2x$ mod 6, where x is a pure random number modulo 3.
Sampler C: Output $2x$ mod 6, where x is a pure random number modulo 6.
Sampler D: Output x^2 mod 6, where x is a pure random number modulo 6.

Solution: The probability mass functions of the respective samplers are

x	0	1	2	3	4	5
$\mathrm{P}(A=x)$	1/6	1/6	1/6	1/6	1/6	1/6
$\mathrm{P}(B=x)$	1/3	0	1/3	0	1/3	0
$\mathrm{P}(C=x)$	1/3	0	1/3	0	1/3	0
$\mathrm{P}(D=x)$	1/6	1/3	0	1/6	1/3	0

The outputs of B and C are indistinguishable. After discarding C, the remaining pairs are distinguishable.

To distinguish B from A, Eve should accept even outcomes and reject odd ones. Her advantage is then

$$\mathrm{P}(B \text{ is even}) - \mathrm{P}(A \text{ is even}) = 1 - 1/2 = 1/2.$$

To distinguish D from A, Eve should accept outcomes 1 and 4 and reject outcomes 2 and 5. What she does on outcomes 0 and 3 is irrelevant as they are equiprobable under D and A. Suppose she rejects them. Here advantage is then

$$\mathrm{P}(D \text{ is 1 or 4}) - \mathrm{P}(A \text{ is 1 or 4}) = 2/3 - 1/3 = 1/3.$$

To distingiush D from B, Eve should accept outcomes $1, 3$ and reject outcomes $0, 2, 5$. What she does on outcome 4 is irrelevant. Her advantage is

$$\mathrm{P}(D \text{ is 1 or 3}) - \mathrm{P}(B \text{ is 1 or 3}) = 1/2 - 0 = 1/2.$$

4. This question is about message (in)distinguishability for the substitution cipher and the short pad cipher. In which of the following examples are encryptions of m_1 and m_2 indistinguishable? When they are not, what is the distinguisher of the highest advantage that you can think of?

   (a) $m_1 = \texttt{mama}$ and $m_2 = \texttt{papa}$ with the short pad cipher of length 4?

   Solution: As the message and key lengths are identical this cipher is effectively a one-time pad. Encryptions are indistinguishable. The advantage of any distinguisher is zero.

   (b) $m_1 = \texttt{mama}$ and $m_2 = \texttt{papa}$ with the substitution cipher?

   Solution: Encryptions of m_1 and m_2 are indistinguishable. Under the substitution cipher, both will encrypt to a ciphertext of the form `XYXY` where `X` and `Y` are random but distinct symbols. All such ciphertexts have the same probability $1/(26 \cdot 25)$ for either message.

   (c) $m_1 = \texttt{mamamama}$ and $m_2 = \texttt{mamapapa}$ with the short pad cipher of length 4?

   Solution: Encryptions are now distinguishable. The first and fifth symbol will always be the same in encryptions of m_1 and always different in encryptions of m_2. The distinguisher that accepts under this condition has advantage 1.

   (d) $m_1 = \texttt{mamamama}$ and $m_2 = \texttt{mamapapa}$ with the substitution cipher?

   Solution: Encryptions are distinguishable. The same distinguisher from part (c) has advantage 1 again.

   (e) $m_1 = \texttt{mamapapa}$ and $m_2 = \texttt{papamama}$ with the short pad cipher of length 4?

   Solution: They are distinguishable with advantage 1 again. The distinguisher accepts when symbols c_1 and c_5 of the ciphertext satisfy the identity $c_5 = c_1 + 3 \pmod{26}$. This distinguisher also gives advantage one in part (c).