**Decryption Assignment**

To decrypt literally means “to unhide”. **Decryption** is the art of breaking a code, usually in the form of a message that someone has altered so that it cannot be read. **Encryption** is the reverse of decryption, it is the act of coding a message. Encryption and decryption is big business in the computer science industry. Mathematicians have created encryption schemes that are virtually unbreakable for even the most advanced computers on Earth. This assignment will have you working with a partner to decrypt an almost unbreakable code using your combined programming skills and smarts.

**The Caesar Shift**

The earliest and simplest form of encryption is the Caesar shift. This encryption scheme involves taking every letter in the alphabet and shifting it some set amount. This set amount is called the “key” to the code. Once you know the key, you can unlock the code. Here is an example of a Caesar shifted word:

bnlotsdq

Although it looks like gibberish, it is the code for the word “computer”. The key in this case is -1, since every letter was shifted back 1 in the alphabet. To “decrypt” the coded message, you would have to shift each letter forward 1:

bnlotsdq 🡪 computer

**The Substitution Cypher**

The code you will break in this assignment is called a “substitution cypher”. The substitution cypher involves substituting each letter with any other letter in the alphabet. For example, the key could be A becomes R, B becomes X, C becomes M and so on. Since each letter is randomly switched, the number of possible keys is:

26x25x24x23x22x... x 3x2x1 = 403291461126605635584000000 combinations.

For this reason, the substitution cypher is almost impossible to break... but not totally impossible. Your assignment is to find the key and decode a text that has been encrypted with a substitution cypher. Obviously, randomly guessing the key is not going to work!

**How to break the code**

When encrypted, a substitution cypher text looks like pure gibberish

QWMIPOJVZPNOBCJDBCIKLPNCLDNCDSLNLDQWNCRTCUSIILIJFLIJSLIJFIJ etc.

We know it is not English. How do we know this? Because ... it doesn’t look like English.

What should it look like then?

For one thing, the cypher text doesn’t have many vowels. In fact, it has no letter E, which should be the most popular letter in the alphabet. This is the key that will let us open the door.

**Frequency Analysis**

Program 1:

This program opens a text file and counts the number of As, Bs, Cs, and so on. You will put these numbers into a list called a frequency list. The program will then sort the list from most popular to least popular. This is called ***frequency analysis***. If you do it right, E, T, A should be at the top of your list, and Z, X or Q should be at the bottom. Steal a medium sized bit of text from the internet and save it as a .txt file to use as a sample. It must be English, and preferrably not about Zinedine Zidane also known as ZouZou.

Program 2:

This program takes an encrypted text file, switches letters according to the cypher key and prints out the result. Which letter gets switched with which letter is not important right now. You can do this easily with 26 if statements, but you should come up with something a little more elegant than that for full marks.

When you are both ready, you apply program 1 to your English sample text and get a sorted list of most frequent to least frequent letters in the text. Then you apply the same program to the cypher text. Again, this will give you a sorted list of the most frequent letters in the cypher text.

Once this is done, you run program 2, getting the program to switch the most popular letter in your cypher back to the most popular letter in English, the second to the second, third to the third, and so on. You run your code and hope it works.

The odds that you get it right on the first try are about 1 in a billion billion. But some letters will be correct. You will have to look for patterns and see if you can figure out the rest. This is not easy, but that’s why it’s good to have a partner.

Read the following helpful hints when you are ready to decode:

**Helpful Hints**

1. **Look at the beginning and end of lines**

Looking along the edges is a good way to spot a potential word. It’s easy to get lost in the jungle in the middle.

1. **Look for the words “THE” and “AND”.**

Especially “THE” because it is in a lot of words (THEN, THERE etc) and it contains two very popular letters – T- E.

1. **Look for repeated “words”.**

If you see “PILSEIJERJ” once, it’s meaningless. If you see it six times, you know it must be an important word. Crack that word and you’re golden.

1. **Only switch letters that are close together on your list**

If I see this:

TOMPUCERSTIENTE

I might realise this could be

COMPUTER SCIENCE

so I switch the T with C. Keep in mind, if this is right then T and C should be close to each other on the list (of similar frequency). If they are more than 4 or 5 letters apart, it’s an iffy move. For example, in English the letter “E” is similar in frequency to the letter “T” but not to “Q”.

1. **Look for double letters**

In English, common double letters are T, E, S, L, M, P, R, N. If you see “QQ” or “UU” you know there needs to be a switch with a letter close by on your list.

1. **Two samples is better than one**

If you are stuck, I recommend that you run the code independently using a different sample English file. That way you can compare your results and look for patterns.