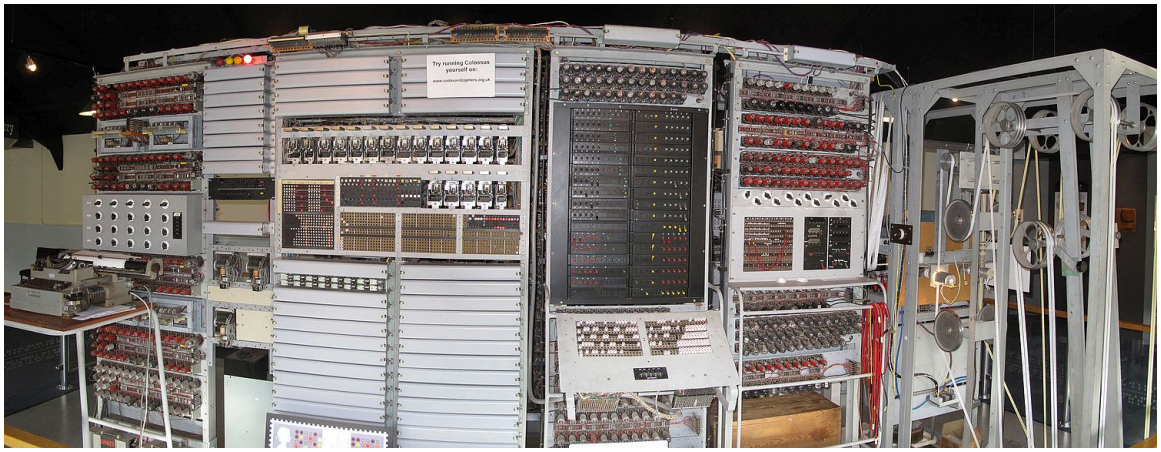
Project 4
Cracked

Time due: 11 PM Thursday, March 15

THIS SPEC IS NOT YET COMPLETE.
DETAILS OF FOUR CLASSES WILL SOON BE FLESHED OUT.
READ WHAT'S HERE NOW.



Before writing a single line of code, you MUST first read AND THEN RE-READ the *Requirements and Other Thoughts* section.

Introduction

Did you read the text in red on the previous page?

Before writing a single line of code, you **must** first read **and then re-read** the *Requirements and Other Thoughts* section. Print out that page, tape it to your wall or on the mattress above your bunk bed, etc. And read it over and over (it's nearly as exciting as Carey's novel, The Florentine Deception).

The NachenSmall Software Corporation has been contacted by the U.S. National Security Agency (NSA) and asked to create a software program that can crack encrypted messages sent by rogue groups around the world. These rogue groups have always encrypted their secret messages using advanced public-key cryptography programs, but recently have decided to go retro, ditching their complex encryption schemes (which may be vulnerable to quantum computers) for a Simple Substitution Cipher (which has been used for thousands of years).

If you've never heard of a Simple Substitution Cipher, it's pretty simple. To use it, you start by creating a mapping from each letter of the alphabet to a letter in the alphabet, like follows:

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

This table indicates that 'A' maps to 'C', 'B' maps to 'B', 'C' maps to 'Y', and so on. Given a message we want to encrypt, we can look up each letter of our unencrypted (*plaintext*) message on the top row and replace it with the associated letter on the bottom row. To decrypt a message, its recipient can look up each letter of the encrypted (*ciphertext*) message on the bottom row and replace it with its original plaintext letter from the top row.

In such a mapping, each plaintext letter on the top row **must** map to a unique letter on the bottom row, and each ciphertext letter on the bottom row **must** map to a unique letter on the top. If two different plaintext letters mapped to the same ciphertext letter, then the recipient would not know which plaintext letter that ciphertext letter should decrypt to. Similarly, if two different ciphertext letters were associated with the same plaintext letter, then with only 26 possible letters, some plaintext letter would not have a corresponding ciphertext letter. Notice that some letters, like B above, may map to themselves, so with the mapping above, the letter B in a plaintex message would be represented as B in the ciphertext message. Also notice that there is no symmetry requirement – for example, in the mapping above you can see that plaintext C on the top row maps to ciphertext Y on the bottom row, but in this particular mapping, plaintext Y on the top row does not map to ciphertext C on the bottom row.

Once you have created such a mapping, you then distribute this secret mapping to two people who want to secretly communicate: the sender (we'll call her Alice) and the

recipient (we'll call him Bob). Once both participants have a copy of the mapping, they can exchange secret messages without being spied upon.

To encrypt a message like "I LOVE UCLA CS", Alice simply looks up each letter of her message in the top row of the table, translates it to the corresponding letter in the bottom row, and writes this ciphertext letter. Thus, to encrypt the message "I LOVE UCLA CS" Alice would look up 'I' on the top row, and see that it maps to 'V' in the bottom row. She'd then look up 'L' in the top row, and see it maps to 'O' on the bottom row, and so on. The resulting encrypted message would be "V ONGF SYOC YX".

Alice then transmits this encrypted message to Bob, keeping her secret from the NSA.

To decrypt an encrypted message, Bob simply takes the same table and looks each ciphertext letter up in the bottom row to find its corresponding plaintext letter in the top row. So, first Bob looks up 'V' (the first letter of the encrypted message he received) in the bottom row, and sees it maps to I in the top row. Then he looks up O in the bottom row, and sees it translates to L in the top row, etc. He processes each such letter in the encrypted message in this way to arrive at the Alice's original, plaintext message.

Now if you think about it, there are 26! (about 4×10^{26}) possible mapping tables that Alice and Bob can pick from to transmit their secret messages! Why? Because there are 26! ways of shuffling all 26 letters on the bottom row of the mapping table.

For your final project of CS32, Winter '18, your job is to write a set of C++ classes capable of cracking an arbitrary encrypted message, like "V ONGF SYOC YX" and printing out the original message: "I LOVE UCLA CS". Since some encrypted messages have more than one English translation (since without the encryption key, there may be two or more possible translation tables that all produce English translations), your solution will be required to print out all possible English translations, sorted alphabetically.

If you're able to prove to NachenSmall's reclusive and bizarre CEO, Carey Nachenberg, that you have the programming skills to build the encryption-cracking tool described in this specification, he'll hire you and you'll be famous… at least among the people who work at the top secret National Security Agency.

How Do You Crack a Simple Substitution Cipher?

With 26! possible mapping tables, it would seem like it would take centuries to take an encrypted message like this

```
Vxgvab sovi jh pjhk cevc andi ngh iobnxdcjnh cn bdttook jb
pnio jpfnicvhc cevh vha nceoi nho cejhy.
```

and try all of the possible mappings to find the right one to decrypt it to its original form:

```
Always bear in mind that your own resolution to succeed is
more important than any other one thing.
```

(Abraham Lincoln wrote this.)

But, as it turns out, cracking a message encrypted with a simple substitution cipher is easier than you might think… especially if you have access to a dictionary of words and use some clever data structures and algorithms.

To make things even easier, in this assignment, we will guarantee that every word in the encrypted messages you have to crack is contained in a list of English words that we will give you. We'll also make things easier for you and ensure that every encrypted word is separated by a space, digit or limited set of punctuation marks so it is clear where one word ends and the next starts. Imagine if you had to crack this instead of the version above with spaces:

```
Vxgvabsovijhpjhkcevcandinghiobnxdcjnhcnbdttookjbpnio
jpfnicvhccevhvhanceoinhocejhy.
```

So how do you actually crack a simple substitution cipher? **Below is one possible algorithm that we require you to use in your project.** It relies upon a list of English words that we will provide you.

As you'll see, this algorithm is recursive and may require a helper function to implement.

string Crack(string ciphertext_message, vector<string>& output):

1. Start with an empty mapping (in this example, we'll show the ciphertext letters on the top row and the plaintext letters on the bottom for clarity – this is the opposite of the table shown in the introduction):

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
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?

   This is our initial input mapping – initially we know nothing about which plaintext letter each ciphertext letter maps to, which is why we show ? for each plaintext letter.

2. Break up ("tokenize") the ciphertext message into separate words and pick a ciphertext word w from the message that (a) has not yet been chosen and (b) has the most[1] ciphertext letters for which we don't have any translation. (When your algorithm first starts, this results in your picking the longest ciphertext word.)
3. Translate the chosen encrypted word using the current mapping table to get a partial decrypted translation. If there's no translation for a ciphertext letter on the

[1] This algorithm will work even if the chosen word does not have the *most* unknown letters, but choosing the one with the most increases the chance that an incorrect mapping will be rejected sooner.

top of your mapping table, then simply use a ? for the plaintext character. (So when the algorithm first starts, your translation of the word will be all ?s.)

4. Using our provided English word list and some clever data structures, create a collection C of all words in the word list that could possibly match the ciphertext word compatibly with the partially decrypted version of the word. For example:
    a. If the ciphertext word were *qbemme* and your current mapping table is empty, then C would contain the words *apollo, blotto, career, chilli, djinni, fusees, grotto, guerre, indeed, piazza, pierre, quagga, steppe, troppo.* All of these words have the same third and sixth letter (different from the other letters), the same fourth and fifth letter (different from the others), and first and second letters different from all others. Since we have no mappings in our mapping table, all of these choices are potential translations for *qbemme*.
    b. If the ciphertext word were *qbemme* and your current mapping table contains the two ciphertext-to-plaintext mappings e→o and b→r, then C would contain the words *grotto,* and *troppo.* Both words have the same pattern as above, and are consistent with our mappings from e→o and b→r. C does not contain *piazza*, since b→r, not b→i.
    c. If the ciphertext word were *qbemme* and your current mapping table has only the mapping e→i, then C would contain the words *chilli* and *djinni.* Both words have the pattern as above and are consistent with our mapping from e→i.
    d. If the ciphertext word were *qbemme* and your current mapping table has the mapping e→q, then C would be empty, since there are no words in the list that have a q in the third and sixth letter positions.
5. If there are no words in the list that could match the ciphertext word compatibly with the partially decrypted version of the word, so C is empty, then your current mapping table *must* be wrong, and your function can throw it away and return to the previous recursive call.
6. For each candidate plaintext word p in your collection C for ciphertext word w:
    a. Create a temporary mapping table by making a copy of the input mapping table and then updating the temporary mapping table by adding the mapping that would occur if your chosen encrypted word w actually did translate into the candidate English word p. Specifically:

       For each letter w[i] in your original ciphertext word w:
        i. Determine the letter p[i] in p that w[i] should map to.
        ii. If your current temporary mapping table does not yet contain a mapping between w[i] and p[i], then add one.
        iii. Otherwise, if the temporary mapping table already contains the same mapping between w[i] and p[i], then do nothing.
        iv. Otherwise, if the temporary mapping table already contains a conflicting mapping from w[i] to some letter that is not p[i], or from some letter that is not w[i] to p[i], then there is no way that the candidate word can work with the currently selected encoding,

so throw away your temporary mapping, and go back to step 6 to try another candidate p.

This now gives you a partial mapping (which may or may not be the right one) between the ciphertext alphabet (on the top row), and your proposed plaintext translations (on the bottom row).

b. Next use this partial proposed translation table to translate your entire ciphertext message into a proposed plaintext message.
c. Evaluate your just-decrypted message to see if all fully-translated words (those with no ?s in them) are in the word list.
    i. If at least one fully-translated word cannot be found in the word list, then the temporary mapping we chose is wrong. So the chosen word p can *not* be a valid candidate for the encrypted word w. Throw away your temporary mapping, and go back to step 6 to try another candidate p.
    ii. If all of the fully-translated words are found in our dictionary, but the message has *not* been completely translated (some words still have ?s in them that were not translated), then this partial solution is promising, as is the current proposed mapping table. Recursively call step 2, passing in our temporary mapping as the new input mapping (we will build upon this new mapping as we recurse deeper and deeper).
    iii. If every single word of the decrypted message was fully translated and in the dictionary, record this as a valid solution (potentially one of many) for eventual output to the user. Then discard the current temporary mapping and go back to step 6 to try other potential candidate translations for the chosen word w and their mappings.

In a nutshell, this algorithm works as follows: we pick a word from our ciphertext message, identify all of the candidate English translations for that word, then we update our mapping table based on each candidate and then apply that translation to the rest of the encrypted message. If a translation due to a given candidate results in all valid (or possibly valid) decrypted words, then we're on the right track, so we pick another word from our encrypted message, find all potential candidate English translations for it (that are consistent with our current mapping table), then update our mapping table based on each candidate (so now the mapping table reflects both the first candidate and this new one) and then apply that translation to the rest of the encrypted message. If this results in all valid (or possibly valid) decrypted words, then we're still on the right track, so we pick yet another encrypted word, identify all of the candidate translations for that word, then we update our mapping table based on that candidate (so now the mapping table reflects all three candidates) and then apply that translation to the rest of the encrypted message. Eventually we'll either reach a dead end because our translation will generate a word not in the word list (and we'll backtrack, trying other candidates), or we'll arrive at a completely translated sentence of valid words. In the latter case, we save the valid solution for later output, then continue hunting for other possibilities.

Let's trace through our algorithm for the following sentence and see how it might work:

```
y qook ra bdttook yqkook.
```

Each bulleted item is prefixed with its [step number] in the proposed algorithm above so you can see which step of the algorithm is doing what. Further, each step of the algorithm is listed as either *top-level*, *inside recursive call*, or *inside second recursive call* to help you understand where you are in the recursion.

- [Step 1, top level] Start with an empty mapping:

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
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?

- [Step 2, top level] Pick the word with the most untranslated letters. We haven't translated any words yet, so that would be *bdttook*, which has 7 untranslated letters.
- [Step 3, top level] Translate the chosen word, *bdttook*, to its decoded equivalent using the current mapping table. We get ?????? since we have no translations for b, d, t, o, or k in our mapping table.
- [Steps 4 and 5, top level] Locate all valid English words that might match *bdttook* and which are consistent with our current translation of ???????. For this example, let's assume there are only two possible matches: *balloon* and *succeed.*
- [Step 6a, top level] First try *balloon*, and assume that *bdttook* maps to *balloon*. Update our mapping table appropriately, mapping b to b, d to a, t to l, o to o, and k to n and leaving the rest of the letters untranslated:

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	?	A	?	?	?	?	?	?	N	?	?	?	O	?	?	?	?	L	?	?	?	?	?	?

- [Step 6b, top level] Use the new mapping table to decode the input message and see what we get:

  Ciphertext message: *y qook ra bdttook yqkook.*
  Partially-decrypted message: *? ?oon ?? balloon ??noon.*

- [Steps 6c, top level] Check all fully-translated words to make sure they're in the word list. Right now, only *balloon* has been fully translated, and it's in the word list, so we'll continue with the current mapping table and recursively call our function with this mapping table.
- [Step 2, inside recursive call] Pick the word with the most untranslated letters. That would be a tie between *ra* and *yqkook*. Let's assume algorithm picks *yqkook*, although either choice would work.

- [Step 3, inside recursive call] Translate the current word, *yqkook*, to its decrypted equivalent using the current mapping table. We get *??noon* since we have translations for k → n and o → o, but not for the other letters y and q.
- [Steps 4 and 5, inside recursive call] Locate all words in the word list that could match yqkook **and** that are compatible with our current translation of *yqkook*: *??noon*. Wait! As it turns out, the word list contains NO English words that are six letters long and end with the suffix "noon". So our current translation **must** be invalid. Return from our function, discarding our current translation map. We're now back to an empty translation table:

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
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?

- [Step 6a, top level] Ok, we just tried *balloon*, so now let's try our second candidate, *succeed*, and assume that *bdttook* maps to *succeed*. Update our mapping table appropriately:

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
?	S	?	U	?	?	?	?	?	?	D	?	?	?	E	?	?	?	?	C	?	?	?	?	?	?

- [Step 6b, top level] Use the new mapping table to translate the input message based on our proposed mapping table:

  Ciphertext message: *y qook ra bdttook yqkook.*
  Partially-decrypted message: *? ?eed ?? succeed ??deed.*

- [Steps 6c, top level] Check all fully-translated words to make sure they're in the word list. Right now, only *succeed* has been fully translated, and it's in the word list, so we'll continue with the current mapping table and recursively call our function with this mapping table.
- [Step 2, inside recursive call] Pick the word with the most unknown/untranslated letters. That would be a tie between *ra* and *yqkook*. Let's assume our algorithm picks *yqkook*, although either choice would work.
- [Step 3, inside recursive call] Translate the current word, *yqkook*, to its decrypted equivalent using the current mapping table. We get *??deed* since we have translations for k → d and o → e, but not for the other letters y and q.
- [Steps 4 and 5, inside recursive call] Locate all words in the word list that could match yqkook **and** that are compatible with our current translation of *yqkook*: *??deed*. Searching our word list results in only one word that matches this pattern: *indeed.*
- [Step 6a, inside recursive call] Update our mapping table appropriately, setting y to i and q to n:

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
?	S	?	U	?	?	?	?	?	?	D	?	?	?	E	?	N	?	?	C	?	?	?	?	I	?

- [Step 6b, inside recursive call] Use the new proposed mapping table to decode the input message (newly discovered letters shown in red):

  Ciphertext message: *y qook ra bdttook yqkook.*
  Partially-decrypted message: *i need ?? succeed indeed.*

- [Steps 6c, inside recursive call] Check all fully-translated words to make sure they're in the word list. We look up *i*, *need, succeed* and *indeed* in the dictionary and find all of them. This is promising! So we'll continue with the current mapping table and recursively call our function with this mapping table.
- [Step 2, inside second recursive call] Pick the word with the most untranslated letters. That would be the only untranslated word left: *ra*
- [Step 3, inside second recursive call] Translate the current word, *ra*, to its decrypted equivalent using the current mapping table. We get *??* since we have no translations for r and a.
- [Steps 4 and 5, inside second recursive call] Locate all words in the word list that could match ra **and** that are compatible with our current translation of *ra*: ??. That results in many words like *it*, *to, at, of, go,* etc.
- [Step 6a, inside second recursive call] For each such candidate word, create a new temporary mapping table and update it appropriately. Let's assume our algorithm starts with *ra* → *it*, so let's map r to i and a to t. Whoops! We have a problem – our table

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
?	S	?	U	?	?	?	?	?	?	D	?	?	?	E	?	N	?	?	C	?	?	?	?	I	?

  already has a mapping from y to i, so it's impossible for r to also map to i; having two encrypted characters (y and r) map to the same plaintext character (i) is incompatible with the way the substitution cipher works. Let's abandon this mapping (*ra* → *it*) and continue back with step 6a.
- [Step 6a, inside second recursive call] For the next candidate word, create a new temporary mapping table and update it appropriately. Let's assume our algorithm continues *with ra* → *to*, so let's map r to t and a to o. This yields the following updates to the mapping table.

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
O	S	?	U	?	?	?	?	?	?	D	?	?	?	E	?	N	T	?	C	?	?	?	?	I	?

- [Step 6b, inside second recursive call] Use the new mapping table to translate the input message based on our proposed mapping table (newly discovered letters shown in red):

  Ciphertext message: *y qook ra bdttook yqkook.*
  Partially-decrypted message: *i need to succeed indeed.*

- [Step 6c, inside second recursive call] Check all fully-translated words to make sure they're in the word list. We look up *i*, *need, to, succeed* and *indeed* in the word list and find all of them. Every single word has been translated, so we have a full translation! Produce the current translation (*i need to succeed indeed*) as a valid solution, discard our current translation table, and then proceed to the next candidate word.
- [Step 6a, inside second recursive call] For each such candidate word, create a new temporary mapping table and update it appropriately. We'll continue with *ra* → *at*, so let's map r to a and a to t:

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
T	S	?	U	?	?	?	?	?	?	D	?	?	?	E	?	N	A	?	C	?	?	?	?	I	?

- [Step 6b, inside second recursive call] Use the new mapping table to translate the ciphertext message (newly discovered letters shown in red):

  Ciphertext message: *y qook ra bdttook yqkook.*
  Partially-decrypted message: *i need at succeed indeed.*

- [Step 6c, inside second recursive call] Check all fully-translated words to make sure they're in the word list. We look up *i*, *need, at, succeed* and *indeed* in the word list and find all of them. Every single word has been translated, so we have a full translation! Produce the current translation as a valid solution. Note that this translation may not make much sense, but it *is* a valid translation as far as our program is concerned – every translated word is in the dictionary. Now proceed to the next candidate word.
- [Step 6a, inside second recursive call] For each such candidate word, create a new temporary mapping table and update it appropriately. We'll continue with *ra* → *of*, so let's map r to o and a to f:
- And so on….

As you can see, by using the above algorithm, your program can quickly home in on the proper mapping. Our program might output something that looks like this:

i need ab succeed indeed.
i need ah succeed indeed.
… *items skipped*
i need go succeed indeed.
… *items skipped*
i need to succeed indeed.
… *items skipped*

There are many other approaches that can be used to crack substitution ciphers such as letter and bigram (e.g., qu, th, er, …) frequency analysis, and simulated annealing. Feel free to look them up. But for this problem, you **must** use the algorithm described above.

What Do You Need to Do?

Question: So, at a high level, what do you need to build to complete Project #4?

Answer: You'll be building five complete classes that together can be used to crack a simple substitution cipher. These classes are detailed below:

Class #1: You need to build a map class template named *MyHash* that implements a templated, resizable, open hash-table-based map:

You need to create a new map class template, based on a resizable *open hash table*. By resizable, we mean if the hash table's load factor gets too high (because you've inserted too many items) the hash table will automatically resize itself (that is, allocate a new, bigger dynamic array and move all of the entries over to the new table, then delete the old dynamic array). Since the hash table is templated, it can be used to map any type of data to any other type of data. You **must not** use any STL container classes to implement your *MyHash* class.

Class #2: You need to build a class named *TokenizerImpl*:

You need to build a "tokenizer" class that can be used to chop up a string (e.g., a sentence) into a vector of tokens (e.g., words in the sentence) - that's basically what tokenizing means. Typically when you tokenize a string, you must specify what characters separate tokens (e.g., spaces, periods, commas, quotes, etc.). Then the tokenizer cuts up the string by gathering only the tokens, excluding the separators.

Class #3: You need to build a class named *WordListImpl*:

You need to build a word list class which is used to load up the contents of a text file consisting of English words and allow the user to efficiently search for words in that word list (e.g., "Is the word "onomatopoeia" in the word list", or "Give me all the words in the list that follow the letter pattern 122134, like 'GOOGLE' or 'TOOTHY'"). Your *WordListImpl* class **must** use your *MyHash* class in the implementation of its data structures and **must not** use any STL containers in its data members.

Class #4: You need to build a class named *TranslatorImpl*:

You need to build a class that maintains a partially or completely filled mapping table, like this:

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
?	S	?	U	?	?	?	?	?	?	D	?	?	?	E	?	?	?	?	C	?	?	?	?	?	?

which can be used to translate a ciphertext message such as:

y qook ra bdttook yqkook.

to a (possibly partially) decrypted message like this:

? ?eed ?? succeed ??deed.

The *TranslatorImpl* class allows the user to specify one or more new mappings, e.g., that Q should map to N, which will update the table:

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
?	S	?	U	?	?	?	?	?	?	D	?	?	?	E	?	N	?	?	C	?	?	?	?	?	?

Class #5: You need to build a class named *DecrypterImpl*:

Finally, you need to build a decrypter class that uses your other classes to process a ciphertext message and produce all possible valid decryptions of that ciphertext to the user. It **must** present these decryptions in alphabetical order. Your decrypter class **must** use the algorithm described above to produce these decrypted messages. This class **must not** use the STL map or unordered_map classes.

What Will We Provide?

We'll provide a simple *main.cpp* file that lets you test your overall decoder implementation. You can compile this *main.cpp* file (and our other provided header file) with your own source files to build a complete working test program. You can then run this program from the Windows command line or in a macOS Terminal window. You can read about our provided *main.cpp* in the *Test Harness* Section of this document.

We'll also provide you with a text file listing English words as well as some abbreviations (e.g., CA for California), one per line.

We'll also provide you with a header file named *provided.h* that will declare and implement the four classes *Tokenizer*, *WordList*, *Translator*, and *Decrypter*. Stop and think about that sentence. Is it saying that we're writing most of this project's code for you? Well, no. The code in each of these classes simply delegates work to a corresponding class that you will write. Let's see how will work, and then explain why we're doing it this way.

Part of the *provided.h* file will be (edited to make this example simpler):

```
class WordListImpl;

class WordList
{
public:
    WordList();
    ~WordList();
```

```
    bool loadWordList(std::string dictFilename);
    bool contains(std::string word) const;
private:
    WordListImpl* m_impl;
};
```

This class contains only one data member, a pointer to a *WordListImpl* object (which you can define however you want in *WordList.cpp*). As shown in the skeleton *WordList.cpp* that you will complete and turn in, the member functions of *WordList* simpy delegate their work to functions in *WordListImpl*[2]:

```
#include "provided.h"

    // You may add additional headers, using statements, helper functions or
    // classes, etc.

class WordListImpl
{
    // You may implement this however you want, provided that WordList objects
    // behave as required by the spec.
};

  // Implementation of WordList that delegates everything to WordListImpl.
  // You will not want to change anything below this line.

WordList::WordList()
{
    m_impl = new WordListImpl;  // create a new implementation object
};

WordList::~WordList()
{
    delete m_impl;  // destroy the implementation object
}

bool WordList::loadWordList(string dictFilename)
{
    return m_impl->loadWordList(dictFilename);  // delegate work to the
                                                // implementation object
}

bool WordList::contains(string word) const
{
    return m_impl->contains(word);  // delegate work to the implementation object
}
```

There's an important restriction we will place on you: Other than *WordList.cpp* itself, no source file that you turn in may contain the name *WordListImpl* or that of any new helper functions or classes you might introduce in *WordList*. So some other file whose code wants to use a word list **must not** have

[2] This is an example of what is called the pimpl idiom (for "**p**ointer-to-**impl**ementation").

```
void f()
{
    WordListImpl wl;  // No! Other files must not use the name WordListImpl
    ...
}
```

but instead should have

```
void f()
{
    WordList wl;
    if ( ! wl.loadFile("wordlist.txt"))
        ...
}
```

When this code constructs a *WordList*, the skeleton code creates a *WordListImpl* object and has code you write construct it. When this code calls *WordList::loadFile*, the skeleton code calls the *WordListImpl::loadFile* you write, and returns the value that you have *WordListImpl::loadFile* return.

Why are we having you do this instead of just implementing *WordList* directly? For two reasons: one that restricts you and one that helps you.

The restriction: We want to make sure you don't introduce a public function into one of the required classes, e.g. *WordList*, that you then call in another, e.g. *Decrypter*; we want you to use the interface we provided. By having none of your implementation go in *WordList*, we can put the *WordList* class in a file *provided.h* that you will not turn in. We will build our tester using the *provided.h* we gave you, so it's useless for you to modify it, since we'll never use or even see those modifications. All of your implementation related to *WordList* will go into *WordList.cpp*. By forbidding you to mention *WordListImpl* outside of *WordList.cpp* (which is easy for us to detect), code in other files can't use a *WordListImpl* (whose interface you might have extended) instead of a *WordList*.

The help: Suppose you cannot figure out how to write a correct *WordListImpl*. If your *Decrypter* code would work correctly if it used a correctly implemented *WordList*, but fails with your incorrectly implemented one, would you lose points on *Decrypter*, too? No, because to test your implementation of one class, such as *Decrypter*, we will build *your* implementation (*DecrypterImpl* in this example) together with *our* correct implementations of the other classes. This is much easier for us when a class's interface and implementaion are separated the way we're doing it.

The *MyHash* class template isn't subject to this separation, since template implementations aren't put in a separate .cpp file; they go in the header file. Your other classes will use the *MyHash* class template directly; there's no need for a separate implementation class.

Details: The Classes You MUST Write

You **must** write correct versions of the following classes to obtain full credit on this project. Your classes **must** work correctly with our provided code, and you **must not** modify our provided *main.cpp* or *provided.h* to make them work with your code. Doing so will result in a **zero score** on that part of the project.

MyHash Class Template

You **must** write a class template named *MyHash* that implements an open hash table that resizes its internal hash table size (the number of buckets) when its load factor gets too high.

Your *MyHash* class template **must** have the following public interface:

```
template <class KeyType, class ValueType>
class MyHash
{
public:
    MyHash(double maxLoadFactor = 0.5);
    ~MyHash();
    void reset();
    void associate(const KeyType& key, const ValueType& value);
    ValueType* find(const KeyType& key) const;
    int getNumItems() const;
    double getLoadFactor() const;
};
```

You **must not** add any additional public methods or data members to this class. You may add as many private methods or data members as you like.

MyHash(double maxLoadFactor) and ~MyHash()

You **must** implement a constructor and destructor for your *MyHash* class:

The constructor **must** initialize your hash table, setting the size of its initial dynamic array to 100 buckets. The user specifies the maximum load factor of the hash table by passing it into the constructor, and the default maximum load factor must be 0.5 if no parameter is passed in. This method **must** run in O(B) time where B is the number of buckets in the table. If the user passes in a zero or negative value for the maximum load factor, you must use a maximum load factor value of 0.5. If the user passes in a load factor greater than 2, then you must use 2.0 for the maximum load factor.

The destructor **must** free all memory associated with hash table. This method **must** run in O(B) time where B is the number of buckets in the table.

void reset()

Your hash table's *reset()* method **must** free all of the memory associated with the current hash table, then allocate a new empty hash table of the default size of 100 buckets. The maximum load factor is unchanged. This method **must** run in O(B) time where B is the number of buckets in the table to be reset.

void associate(const KeyType &key, const ValueType &value)

The *associate()* method adds a new association to the hash table associating the specified *key*, e.g. "122134", with the associated *value*, e.g., "google". If you pass in a key which is already in the hash table, then this function will update the association in the hash table, discarding the old value that the key mapped to and replacing it with the new value. So a given key in the hash table may map to only one value.

If inserting a new key → value association into the hash table would cause the hash table's load factor to exceed its maximum load factor (as specified during construction), then your function **must** allocate a new dynamic array that has double the number of buckets the current dynamic array has, and must move all of the items from the old array to the new array. It **must** delete the contents of the old array.

Be careful: Items that hashed to bucket X in the old array may not hash to the same bucket in the new array because its number of buckets is different! You'll have to re-compute the proper location for every key → value association in the new array.

Assuming items in your hash table are roughly uniformly distributed across all the buckets, your *associate()* function **must** run in O(1) time if the number of buckets does not change. (In a pathological case, such as when most keys hash to just a few buckets, this function may run in O(X) time where X is the number of items in the hash table.) In those cases where the *associate()* call results in the number of buckets changing, your *associate()* function **must** run in O(B) time, where B is the number of buckets in the hash table.

ValueType* find(const KeyType& key) const

The *find()* method attempts to locate the item with a key equal to *key* in the hash table, and if successful, returns a pointer to the value associated with that key. If the key can not be found within the hash table then this function returns *nullptr*.

Assuming items in your hash table are roughly uniformly distributed across all the buckets, your *find()* function **must** run in O(1) time. (In a pathological case, such as when most keys hash to just a few buckets, this function may run in O(X) time where X is the number of items in the hash table.)

int getNumItems() const

The *getNumItems()* method returns the number of unique key → value associations that the hash table holds. Note, that overwriting an old association k1→v1 with a new association k1→v2 does not increase the number of items in the hash table. It simply replaces the old association with the new one. Your *getNumItems()* function **must** run in O(1) time. Note: This function does not return the number of buckets in the table, but the actual number of stored associations.

double getLoadFactor() const

The *getLoadFactor()* method **must** return the actual current load factor of the hash table based on its current number of buckets and the number of associations it holds. Your *getLoad()* function **must** run in O(1) time.

Hash Functions

We are providing you with two hash functions (for common types like *string*, and *unsigned int*) which you may call within your *MyHash* class to obtain an unsigned int value between 0 and roughly 4 billion:

```
inline unsigned int hash(std::string s)
{
	return std::hash<std::string>()(s);
}

inline unsigned int hash(unsigned int u)
{
	return std::hash<unsigned int>()(u);
}
```

These hash functions leverage the STL's *hash* class which can be found in the <algorithm> header file. You'll need to transform the values provided by our hash functions to obtain the proper bucket number in your hash table.

If you'd like to define your own hash function, you **must** define it in one of your .cpp files, **not** in *MyHash.h*. Each such hash function **must** be named *hash()* and **must** have the signature:

```
	unsigned int hash(const YourTypeHere& k);
```

However, you shouldn't need to create additional hash functions to complete this project.

Tokenizer Class

[This part coming soon.]

WordList Class

[This part coming soon.]

Translator Class

[This part coming soon.]

Decrypter Class

[This part coming soon.]

Test Harness

[This part coming soon.]

Encrypted Messages to Test With

Here are some interesting encrypted messages that you can test your program with.

```
Trcy oyc koon oz rweelycbb vmobcb, wyogrcn oecyb; hjg ozgcy tc
moox bo moya wg grc vmobck koon grwg tc ko yog bcc grc oyc trlvr
rwb hccy oecyck zon jb. -Rcmcy Xcmmcn

Jxwpjq qwrla glcu pcx qcn xkvv dw uclw ekarbbckpjwe dq jzw jzkpta
jzrj qcn ekep'j ec jzrp dq jzw cpwa qcn eke ec. -Urls Jxrkp

Xjzwq gjz cuvq xz huri arwqvudiy fuk ufjrqoq svquxiy. -Lzjk
Nqkkqcy

Axevfvu lvnelvp bxqp mvpprjv rgl bvoop Grnxvgkvuj dqupb jvbp
buvrbvl be lqggvu.
```

Requirements and Other Thoughts

Make sure to read this entire section before beginning your project!

1. You should backup your code to an online repository like dropbox or google drive frequently (e.g., creating a new function successfully). If you come to us and complain that your computer crashed and you lost all of your work, we'll ask you where your backups are.
2. In Visual C++, make sure to change your project from UNICODE to Multi Byte Character set, by going to Project → Properties → Configuration Properties → General → Character Set
3. No matter what you do, and how much you finish, make sure your project compiles and at least runs (even if it crashes after a while). WHATEVER YOU DO, don't turn in code that doesn't compile.
4. Whatever you do, DO NOT MODIFY OUR PROVIDED HEADER FILES IN ANY WAY! YOU WILL NOT TURN THEM IN, SO ANY MODIFICATIONS WILL NOT BE GRADED!
5. The entire project can be completed in less than 550 lines of C++ code beyond what we've already written for you, so if your program is getting much larger than this, talk to a TA – you're probably doing something wrong.
6. Be sure to make copious use of the C++ STL where we permit it – it can make things much easier for you!
7. If you need to define your own inline comparison operators, feel free to do so! However you MUST place these functions within your own header file(s) that you will submit as part of the project. You MUST NOT modify our provided header files!!
8. Before you write a line of code for a class, think through what data structures and algorithms you'll need to solve the problem. Plan before you program!
9. Don't make your program overly complex – use the simplest data structures possible that meet the requirements.
10. You MUST NOT modify any of the code in the files we provide you that you will not turn in; since you're not turning them in, we will not see those changes. We will incorporate the required files that you turn in into a project with special test versions of the other files.
11. Make sure to implement and test each class independently of the others that depend on it. Once you get the simplest class coded, get it to compile and test it with a number of different unit tests. Only once you have your first class working should you advance to the next class.
12. Try your best to meet our big-O requirements for each method in this spec. If you can't figure out how, then solve the problem in a simpler, less efficient way, and move on. Then come back and improve the efficiency of your implementation later if you have time.

If you don't think you'll be able to finish this project, then take some shortcuts. For example, implement the *MyHash* class first by using the STL's map or unordered_map

class to do all of the hard work. Then use it to get your decoder working. Once you get your decoder working, then go back and implement your full *MyHash* class.

You can still get a good amount of partial credit if you implement most of the project. Why? Because if you fail to complete a class (e.g., *MyHash*), we will provide a correct version of that class and test it with the rest of your program. If you implemented the rest of the program properly, it should work perfectly with our version of the class you couldn't get working, and we can give you credit for those parts of the project you completed.

But whatever you do, make sure that ALL CODE THAT YOU TURN IN BUILDS without errors with both g32 and either Visual C++ or clang++!

What to Turn In

You will turn in **six** files:

MyHash.h	Contains your resizable hash table map class template implementation
Tokenizer.cpp	Contains your tokenizer implementation
WordList.cpp	Contains your word list implementation
Translator.cpp	Contains your translator implementation
Decrypter.cpp	Contains your decrypterer implementation
report.docx, report.doc, or report.txt	Contains your report

You **must** submit a report that describes:

1. Whether any of your classes have known bugs or other problems that we should know about. For example, if you didn't finish the *DecoderImpl::crack()* method or it has bugs, tell us.
2. A high-level description of what data structures and algorithms you chose for each of your classes' non-trivial methods and data structures. Brief means a paragraph or two description or half a page pseudocode per class.
3. Whether or not each method satisfies our big-O requirements, and if not, what you did instead and what the big-O is for your version.

Grading

- 95% of your grade will be assigned based on the correctness of your solution.
- 5% of your grade will be based on your report.

Good luck!