Problem 176: Codebreaker Returned

Difficulty: Medium

Author: Brett Reynolds, Annapolis Junction, Maryland, United States

Originally Published: Code Quest 2022

Problem Background

In 2021, we presented a problem in which the National Security Agency had asked Lockheed Martin to help develop a frequency analysis database for cracking substitution ciphers. The Agency was so impressed with your work, they've asked you to continue your efforts!

As a reminder, substitution ciphers encrypt messages by replacing letters in the original message with different letters, in an effort to hide the content of the message. A simple substitution cipher might replace each letter with the next one in the alphabet; for example, "cat" would become "dbu." These are generally regarded as weak ciphers, however, because in each language, certain letters appear more commonly than others. Vowels are particularly common because of how important they are to the language. These common letters serve as guideposts to anyone trying to break such a cipher; if you know the original message was in English, you can safely assume that the most common letter is likely to be 'E'. This method of breaking ciphers is called "frequency analysis."

More advanced ciphers take this into account and go to greater lengths to obscure the text; however, more advanced frequency analysis can circumvent this. Rather than looking at individual letters, looking at pairs or triplets of letters (known as digraphs and trigraphs) can provide even greater insight and help with the identification of individual letters.

Problem Description

As mentioned above, the National Security Agency has asked Lockheed Martin to expand the frequency analysis database created in 2021 to include an analysis of digraphs and trigraphs. A digraph is a pair of letters that appear next to each other in the same word; a trigraph is a group of three letters in the same word. As with letters, certain digraphs and trigraphs appear much more commonly in certain languages, and certain combinations never appear at all. For example, you'll never see the letters 'qxz' appear next to each other in any English word, but the trigraph 'the' is the most common - largely because 'the' is the most common word in the English language.

When performing your analysis, keep in mind that digraphs and trigraphs do not cross word boundaries, and all punctuation and numbers should be removed prior to starting your analysis. For example, the phrase 'code quest' includes the digraphs 'co', 'od', 'de', 'qu', 'ue', 'es', and 'st'. It does not contain the digraph 'eq'; even though those letters appear next to each other, they are separated by a space. The same applies to trigraphs; that phrase includes 'ode' and 'que', but not 'deq' or 'equ'.

As with the 2021 problem, your team will be analyzing a large amount of text in order to identify the relative frequencies of the digraphs and trigraphs you find within that text. The relative frequency of a term can be calculated using this formula:

$$Relative\ frequency = \left(\frac{Occurrences\ of\ term}{Total\ number\ of\ terms\ of\ the\ same\ type}\right) \times 100\%$$

As noted, keep the count of all digraphs and the count of all trigraphs separate when performing this calculation.

Sample Input

The first line of your program's input, received from the standard input channel, will contain a positive integer representing the number of test cases. Each test case will include:

- A line containing a positive integer, X, indicating the number of lines of text to be provided.
- X lines, containing text to be analyzed. Lines may contain up to 2000 characters each, and can contain any US ASCII character.

```
1
3
The quick red fox jumps over the lazy brown dog.
The above sentence contains every letter in the English language.
Don’t forget to ignore punctuation and numbers; they’re not relevant!
```

Sample Output

For each test case, your program must print the results of your analysis, first printing a single line for each digraph identified in the text, then printing a single line for each identified trigraph. Each list should be presented in alphabetical order. Each line should contain the following information:

- The digraph or trigraph, in uppercase
- A colon (:)
- A space
- The relative frequency of the given term within the text provided in the test case, relative to other terms of the same type, rounded to three decimal places and including any trailing zeroes.
- A percent sign (%)

The sample output is too large to print in full here; instead, a small selection of lines from the sample output is presented in order to demonstrate the format. For the full sample output, please download the file from the contest website.

```
AB: 0.840%
AG: 0.840%
```

[...additional lines for AI through WN...]

```
YR: 0.840%
ZY: 0.840%
ABO: 1.124%
AGE: 1.124%
```

[...additional lines for AIN through VAN...]

```
VER: 2.247%
YRE: 1.124%
```