Problem 20: Soundex Encoding

Difficulty: Medium

Originally Published: Code Quest 2015

Problem Background

Soundex is a phonetic hashing algorithm that groups together names that sound similar yet have minor differences in spelling. It can be useful for genealogical studies by identifying variations for a given surname.

Problem Description

The hashing part of the algorithm uses the following character groups:

- Group 1: b, f, p, v
- Group 2: c, g, j, k, q, s, x, z
- Group 3: d, t
- Group 4: l
- Group 5: m, n
- Group 6: r
- Wild: h, w
- Vowels: a, e, i, o, u, y

An American Soundex code for a name consists of a letter followed by a three digit number. The code can be determined by the following steps:

- Find the first letter in the name that belongs to one of the numbered groups. Starting from that letter and working left to right, if two or more letters from the same numbered group are adjacent to one another, remove all but the first letter. Note that h and w are "wild" meaning that they will match letters from any group 1-6.
- Retain the first letter of the name and remove all vowels and wild letters.
- Retain the first letter of the name and replace all other letters with their group number.
- If there are less than three numbers, add zeroes until there are three. If there are more than three numbers, just keep the first three. Make sure the letter at the beginning is capitalized.

Examples:
Ashcroft --> Asroft --> Asrft --> A2613 --> A261
Pfister --> Pister --> Pstr --> P236
Williams --> Wiliams --> Wlms --> W452

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 positive number N representing the number of names to follow.
- N lines, containing one name per line.

```
2
3
Williams
Ashcroft
Pfister
10
Gary
Clare
Jane
Gore
Geier
June
Claire
George
John
Jenny
```

Sample Output

Your program should print out the list of Soundex codes generated by the list of names (ordered alphabetically) and the number of times that code was generated separated by a space for each test case. The first line of each test case's output should be the word OUTPUT.

```
OUTPUT
A261 1
P236 1
W452 1
OUTPUT
C460 2
G600 3
G620 1
J500 4
```