Homework 11

The Soundex algorithm is used to transform any word (and especially names) in English into a code that consists of a letter and three digits. There are variations and elaborations of the code. This version does **not** conform to the rules applied for the US Census (with the additional benefit that students submitting a program glanced from the web will submit one that is not in compliance with these specifications). The algorithm tries to give the same code to similar sounding words. It proceeds in a number of steps.

1. Replace the bigrams in Table 1 with the specified letter.
2. Change to string to capitalized letters.
3. The first letter of the string is maintained.
4. All other occurrences of ”A”, ”E”, ”I”, ”O”, ”U”, “H”, “W”, and ”Y” are changed to ”0”.
5. All letters with exception of the first one are changed into digits according to the following translation table.
6. Repeatedly, one of all adjacent pairs of the same integer is removed.
7. All digits ”0” are removed.
8. Smaller codes are padded with zeroes to four letters total and larger codes are shortened to four by removing the last digits.

Table 1: Bigram substitution

DG	G
GN	N
KN	N
PH, PF	F
PS	S
TCH	CH

Table 2: Letter substitution

B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M,N	5
R	6

Notice that step 3 is important because it allows for the same digit in two locations.

Examples for Soundex:

“Dadaism” —> DADAISM —> D0D00SM —> D030025—>D325

“Trump” —> TRUMP —> TR0MP —> T6051 —> T651
“Pence” —> PENCE —> P0520 —> P520 —> P52 —> P520
“Tymczak” —> TYMCZAK —> T0MCZ0K —> T052202 —> T05202 —> T522
“Ashcraft” —> ASHCRAFT —> AS0CR0FT —> A2026013 —> A22613 —> A226
“Schwarz” —> SCHWARZ —> SC0000RZ —> S262