Exercise

Background: The Soundex algorithm (Odell and Russell, 1922; Knuth, 1973) is a method commonly used in libraries and older Census records for representing people’ s names. It has the advantage that versions of the names that are slightly misspelled or otherwise modified (common, for example, in hand-written census records) will still have the same representation as correctly-spelled names. (e.g., Jurafsky, Jarofsky, Jarovsky, and Jarovski all map to J612).

Exercise

Write a FST (a function) to implement the Soundex algorithm, which is explained as follows:

a. Keep the first letter of the name, and drop all occurrences of non-initial a, e, h, i, o, u, w, y

b. Replace the remaining letters with the following numbers:

b, f, p, v → 1
c, g, j, k, q, s, x, z → 2
d, t → 3
l → 4
m, n → 5
r → 6

c. Replace any sequences of identical numbers , only if they derive from two or more letters that were adjacent in the original name, with a single number (e.g., 666 → 6)

d. Convert to the form Letter Digit Digit Digit (e.g., J612) by dropping digits past the third or padding with trailing zeros if necessary

Tip: Use what we've learned so far to solve this problem.