SOUNDEX

CLASS OF HEURISTICS THAT EXPAND A QUERY INTO PHONETIC EQUIVALENTS.
THESE ALGS ARE LANGUAGE SPECIFIC AND MAINLY USED FOR NAMES
(CHEBYSHEV ⟶ TCHEBYCHEFF).

IN ORDER TO DO THIS, IT COMPUTES A REDUCED FORM OF 4 CHARS OF BOTH THE LEXICON AND THE QUERY TERM

↳ why?

EVEN IF THE queried name IS SPELLED WRONG (EX: Herman instead of Hermann), THE REDUCED FORM IS THE SAME

USE

THE SOUNDEX FAMILY IS NOT USED IN IR, BECAUSE THERE ARE BETTER ALTERNATIVES.

Typical alg.

(1) Take in input a word

EXAMPLE: Herman

(2) Retain the first letter of the word

EXAMPLE: H

(3) Change each occurence of A, E, I, O, U, H, W, Y with zero 0 and topped it to the first letter

EXAMPLE H0rm0n

(4) Change letters to digit as follow:

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

EXAMPLE H0650S

(5) Remove all pairs of consecutive equal digit

EXAMPLE: None, in this case

(6) Remove all zeroes

Example: H655

(7) PAD WITH TRAILING ZEROES AND RETURNS THE FIRST 4 positions as

<uppercase letter> <digit> <digit> <digit> <digit>

NOTE: Herman gives the same result

OSS: You might notice that the way the mapping is done in (4), the reduction is biased towards certain nationality names