

## ★ Soundex algorithm

- soundex is considered as a phonetic algorithm used primarily in NLP for indexing names by sound
- Algorithm used to group similar sounding letters together and assign each group a numerical number.
- main goal of this technique is to use homophones for encoding text with numerical indexing
  - As a result the numerical representation can easily be matched with other similar sounding characters having same numerical code
  - This results in retrieving a list of words that are pronounced similarly with very little variation in their homophones
- Used as a simple phonetic based spell checker
- code for a word consists of its first letter followed by three numbers that encode the remaining consonants
- Those consonants that generate same sound have same numbers
- Uses code to check for closest word



map to <sup>same</sup> letters digit

substitute with integers

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

### Algorithm

- Remove all punctuation marks & capitalize letters in the word
- Retain the first letter of the word
- Remove any occurrence of the letters - A, E, I, O, U, H, W, Y apart from very first letter
- Replace the letters by numbers shown in table
- If two or more adjacent letters, not separated by vowels, have the same numeric value, retain only one of them. Also two letters with the same number separated by 'h' or 'w' are coded as a single number where as such a letters separated by a vowel coded twice
- Return the first four characters, pad with zero if there are less than four

Teacher's Signature.....



word

soundex code

create

C630

creat

C630

create

→ CRATE

↓

CRT

↓

C63

↓

C630

creat

→ CREAT

↓

CRT

↓

C630

TORN → TORN

↓

TRN

↓

T65

↓

T650

WORN → W650

HORN → H650

used to find good possible candidate for correction to be effected for a misspelled word

✱

ASHCRAFT

↓

ASHCRAFT

↓

ASCRAFT

↓

A22613

↳

two or more adjacent letters not separated by vowel have the same numeric value, return one of them

↓

A2613

↓

Return the first four characters

A261



★ Tumczak  
 ↓  
 TmCzk  
 ↓  
 T5222 → TS222  
 ↓  
 T520

retain one from there 'z'

~~TS222~~ ~~TS222~~ ~~TS222~~

★ Tjcsnim  
 ↓  
 TJSNM  
 ↓  
~~TJSNM~~  
 ↓  
~~TJSNM~~

TS2(2) → This 'z' will be retained since z & k are separated by vowel.

T2255  
 ↳ Both will be retained

T2255  
 ↳ Both 's' will be retained

T225 → Only first 4 characters will be returned.