PAGE NO.
DATE : / /

✱ 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 matched with other similar sounding characters having same numerical code
- This results in retriving 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 same 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 occurence 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 seperated by vowels, have the same numeric value, retain only one of them Also two letters with the same number seperated by 'h' or 'w' are codded as a single number where as such a letters seperated by a vowel codded twice

→ Retain the first four characters, add with zero if there are less than four

Teacher's Signature.....

word	soundex code
create	C630
recreate	C630

create → CRATE

↓

CRT

↓

C63

↓

C630

recreate → 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

↓

ASCRFT

↓

A22613

↳ two or more adjacent letters not seperated by vowel have the same numeric value, retain one of them

↓

A2613

↓ Return the first four characters

A261

PAGE NO.
DATE : / /

* Tumczak
  ↓
  TmczK
  ↓
  T5222 → retain only one of them
  ↓
  T520

* Tjasnjm
  ↓
  TJSNM
  ↓
  T2255
  ↓
  T250