INFORMATION RETRIEVAL

Faculty: Venkatesh Vinayaka Rao

Term: Aug – Sep, 2020

Chennai Mathematical Institute

Guest Speaker: Vinoothna Sai K

QUERY UNDERSTANDING

Phonetic Correction

Understanding the true need of Phonetic Correction

What is a Phonetic??

/fəˈnɛtɪk/

Describes the sounds of words in a language using the symbols of the International Phonetic Alphabet (IPA)

Phonetic Correction

- Misspellings that arise because the user types a query that sounds like the target term.

- The main idea here is to generate, for each term, a "phonetic hash" so that similar-sounding terms hash to the same value.

- Algorithms for such phonetic hashing are commonly collectively known as Soundex algorithms.

Standard Soundex Algorithm

1. Retain the first character
2. Convert each character to digit using the rules in the table.
3. Repeatedly remove one out of each pair of consecutive identical digits.
4. Remove all the zeros.
5. Add trailing zeros, and return the first four positions.

Alphabets to be replaced	Digit
A, E, I, O, U, H, W, Y	0
B, F, P, V (Labial)	1
C, G, J, K, Q, S, X, Z (Gutterals and sibilants)	2
D, T (Dental)	3
L (Long liquid)	4
M, N (Nasal)	5
R (Short liquid)	6

Any characters not included in the above table are just ignored from the term

Standard Soundex Algorithm

Alphabets to be replaced	Digit
A, E, I, O, U, H, W, Y (Gym, Gim, Candy, Deny, yellow, yeah, sigh)	0
B, F, P, V (pfister, obvious)	1
C, G, J, K, Q, S, X, Z (Example, Egsample, eksample, eczampl, gibberish, jibberish, clique, click)	2
D, T (Midterms, goldtone)	3
L	4
M, N (Solemn, damnation, damn, autumn)	5
R	6

Implementation using an example

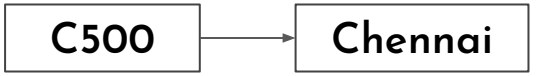
Let the term be "CHENNAI".

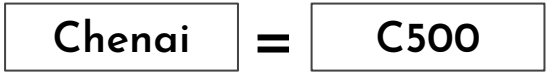
Step No	Step	Changes in the term
1	Retain the first character & Convert each character to digit using the rules of the table	C 0 0 5 5 0 0
2	Repeatedly remove one out of each pair of consecutive identical digits	C 0 5 0
3	Remove all the zeros	C 5
4	If number of integers is less than 3, add trailing zeros	C 5 0 0 0
5	Return the first four positions	C 5 0 0

Alphabets to be replaced	Digit
A, E, I, O, U, H, W, Y	0
B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

Scheme of a soundex algorithm

1. Turn every term to be indexed into a 4-character reduced form.
2. Build an inverted index from these reduced forms to the original terms; call this the soundex index.
3. Do the same with query terms.
4. When the query calls for a soundex match, search this soundex index.





Test your understanding

- Find two differently spelled proper nouns whose soundex codes are the same.
- Find two phonetically similar proper nouns whose soundex codes are different.

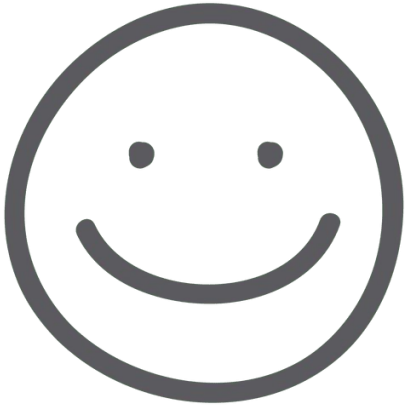
LINKS TO EXPLORE

How to use Soundex Search in MYSQL

```
SELECT * FROM `people` where SOUNDEX(`last_name`) =
SOUNDEX('Wlliams')
```

So, what did we learn? 🤔

- Phonetic Correction
- Standard Soundex Algorithm
- Scheme of Soundex Algorithm



THANK YOU !

Vinoothna Sai K

Batch 2020, IIIT Sri City

vinoothna.kinnera@gmail.com

Link to the YouTube Video for the same lecture