Information Retrieval
Phonetic Correction
Lecture-16

Prepared By

Dr. Rasmita Dash & Dr. Rasmita Rautray
Associate Professor
Dept. of CSE

Content

- Phonetic Correction
- Soundex Algorithm

Phonetic Correction

- **phonetic correction**: misspellings that arise because the user types a query that sounds like the target term.
- Such algorithms are especially applicable to searches on the names of people.
- The main idea here is to generate, for each term, a ``phonetic hash" so that similar-sounding terms hash to the same value.

Contd..

- Algorithms for such phonetic hashing are commonly collectively known as **soundex** algorithms. However, there is an original soundex algorithm, with various variants, built on the following scheme:
    - Turn every term to be indexed into a 4-character reduced form. Build an inverted index from these reduced forms to the original terms; call this the soundex index.
    - Do the same with query terms.
    - When the query calls for a soundex match, search this soundex index.

Soundex Algorithm

The variations in different soundex algorithms have to do with the conversion of terms to 4-character forms. A commonly used conversion results in a 4-character code, with the first character being a letter of the alphabet and the other three being digits between 0 and 9.

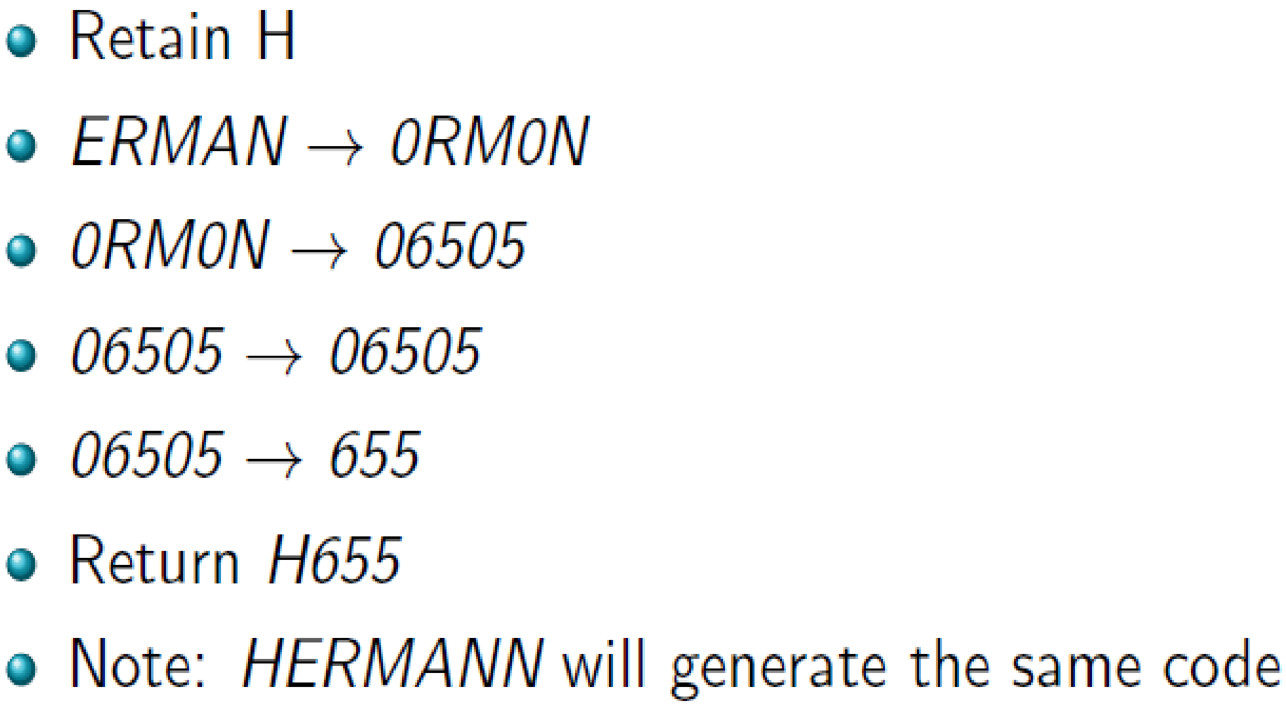
- Retain the first letter of the term.
- Change all occurrences of the following letters to '0' (zero): 'A', E', 'I', 'O', 'U', 'H', 'W', 'Y'.
- Change letters to digits as follows: B, F, P, V to 1. C, G, J, K, Q, S, X, Z to 2. D,T to 3. L to 4. M, N to 5. R to 6.
- Repeatedly remove one out of each pair of consecutive identical digits.

Contd..

- Remove all zeros from the resulting string. Pad the resulting string with trailing zeros and return the first four positions, which will consist of a letter followed by three digits.
- Change letters to digits as follows: B, F, P, V to 1. C, G, J, K, Q, S, X, Z to 2. D,T to 3. L to 4. M, N to 5. R to 6.
- Repeatedly remove one out of each pair of consecutive identical digits.
- Remove all zeros from the resulting string. Pad the resulting string with trailing zeros and return the first four positions, which will consist of a letter followed by three digits.

Example

Soundex of HERMAN with HERMANN



- Retain H
- $ERMAN \rightarrow 0RM0N$
- $0RM0N \rightarrow 06505$
- $06505 \rightarrow 06505$
- $06505 \rightarrow 655$
- Return $H655$
- Note: $HERMANN$ will generate the same code

Thank You