W2

2016年8月8日 17:40

- W2L1
  - Approximate String Search and Matching
    - Hard
      - Because need to define the closest or best match
    - 2 main app
      - Spelling correction
        - Need the notion of a dict
          - Entry
          - Substring
          - Item
          - Given some item of interest — which does not appear in our dictionary — which entry from the dictionary was truly intended?
      - Computational Genomics
        - ->Given a substring, find whether the sequence occurs within a larger string, possibly with “errors”
        - (but much larger)
    - Some
      - Name matching
      - Query repair
      - Phonetic matching
      - Data cleaning
      - ...
  - Common Applications
  - Methods:
    - Neighbourhood Search
      - Insert Delete Replace(Substitute) transpose
      - Efficiency
        - alphabet size is Σ, length of string is|w|, k edits:
        - O(Σ^k ·|w|^k) neighbours
        - O(|w|^k logD) string comparison
    - Edit Distance
      - Not really a "distance"
      - operations
        - Insert delete replace match
      - Global
        - Needleman–Wunsch algorithm
      - Local
        - Substring (particularly suitable for diff len)
        - Smith–Waterman
        - Match must have different +/—sign to Insert/Delete/Replace
      - Efficiency
        - Given string f, entry string t
        - 1 turn
          - O(|f||t|) in space and time
        - Each t in D
          - $\mathcal{O}(|f| \sum_{t \in D} |t|)$
    - N-Gram Distance
    - Phonetic methods
  - Evaluation

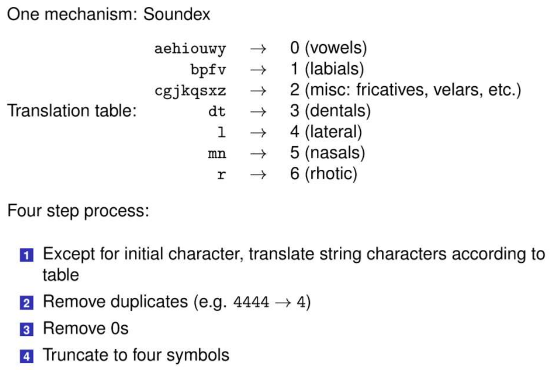
- W2L2
  - Ssh @dimefox.eng.unimelb.edu.au
  - Agrep -1 "^ther$"

  - n-gram distance
    - Same goal as Edit Distance
      - Compare two strings to determine "best" match
    - n-gram
      - Sub-string of length n
      - a true "distance"
    - n-gram dist
      - N-Gram Distance between n-grams of string s ($G_n(s)$) and t ($G_n(t)$):
        $|G_n(s)| + |G_n(t)| - 2 \times |G_n(s) \cap G_n(t)|$
      - The smaller the better (more common parts)
    - Efficiency
      - Much simpler
      - Occasionally useful as a simpler variant of Edit Distance
      - More sensitive to long substring matches, less sensitive to relative ordering of strings (matches can be anywhere!)
        - Quite useless for very long strings and/or very small alphabets (Why?)
          - Why
            - For example, Computational Genomics

- Quite useless for very long strings and/or very small alphabets (Why?)
    - Why
        - For example, Computational Genomics
            - Neighbourhood search
                - Too many possible neighbours (string too long)
            - Global Edit Distance
                - Too many insertions (string too long)
            - Local Edit Distance
                - cannot fit table into mem(string too long)
            - N-Gram Distance
                - With huge n (e.g. 80% of length of shorter string) can (almost) work!
                - Though tends to prefer shorter chromosomes like Global Edit Distance
- But not faster: Despite its simplicity, takes roughly the same time to compare entire dictionar

Orthography(spelling) and phonetics(sound)

- soundex



One mechanism: Soundex

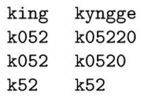
Translation table:

aehiouwy	→	0 (vowels)
bpfv	→	1 (labials)
cgjkqsxz	→	2 (misc: fricatives, velars, etc.)
dt	→	3 (dentals)
l	→	4 (lateral)
mn	→	5 (nasals)
r	→	6 (rhotic)

Four step process:

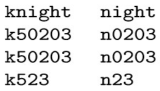
1. Except for initial character, translate string characters according to table
2. Remove duplicates (e.g. 4444 → 4)
3. Remove 0s
4. Truncate to four symbols

example

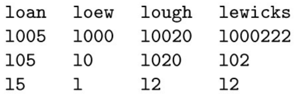


king	kyngge
k052	k05220
k052	k0520
k52	k52

Not good enough



knight	night
k50203	n0203
k50203	n0203
k523	n23



loan	loew	lough	lewicks
1005	1000	10020	1000222
105	10	1020	102
15	1	12	12

Evaluation

- whether the system is effective at solving the user’s problem
    - for a misspelled word, does the system identify the correct word?
        - Need
            - A number of cases of misspelled words
            - The intended (correct) word for each case
            - An evaluation metric
        - To compare sys
            - Accuracy
            - Precision
            - Recall