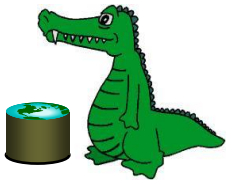
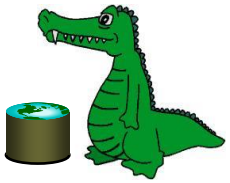


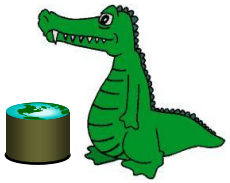
Logistics

- Lab 0 grading
- Lab 1 due Thursday 5pm
- Participation in NIST data science project – part of UF DSR team
  - Group
  - Register
  - Dataset
  - Results ensemble



Review

- Improve Data Quality Metrics
  - Use of Constraints
  - Data Cleaning is Cross-disciplinary
- Data Integration
  - Schema mapping
  - WebTables
  - Deduplication
    - Applications
    - Methods – Approximate matching, similarity measures over different data objects



Some Similarity Measures

Handle Typographical errors

- Equality on a boolean predicate
- Edit distance
  - Levenstein, Smith-Waterman, Affine
- Set similarity
  - Jaccard, Dice
- Vector Based
  - Cosine similarity, TFIDF

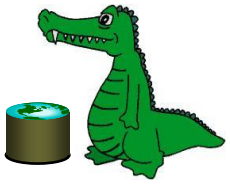
Good for Text like reviews/ tweets

Good for Names

- Alignment-based or Two-tiered
  - Jaro-Winkler, Soft-TFIDF, Monge-Elkan
- Phonetic Similarity
  - Soundex
- Translation-based
- Numeric distance between values
- Domain-specific

Useful for abbreviations, alternate names.

From: Getoor & Machanavajjhala: “Entity Resolution Tutorial”, VLDB 2012



Soundex Encoding

A phonetic algorithm that indexes names by their sounds when pronounced in english.

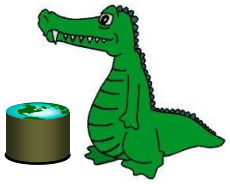
Consists of the first letter of the name followed by three numbers.
Numbers encode similar sounding consonants.

- Remove all W, H
- B, F, P, V encoded as 1, C,G,J,K,Q,S,X,Z as 2
- D,T as 3, L as 4, M,N as 5, R as 6, Remove vowels
- Concatenate first letter of string with first 3 numerals

Ex: great and grate become 6EA3 and 6A3E and then G63

More recent, metaphone, double metaphone etc.

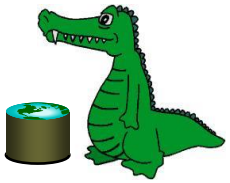
From: Koudas, Sarawagi, Strivastava, “Record Linkage: Similarity Measures and Algorithms”, VLDB 2006



Edit Distance

- Character Operations: I (insert), D (delete), R (Replace).
- Unit costs.
- Given two strings, s,t, edit(s,t):
  - Minimum cost sequence of operations to transform s to t.
  - Example: edit(Error,Eror) = 1, edit(great,grate) = 2
- Folklore dynamic programming algorithm to compute edit();
- Computation and decision problem: quadratic (on string length) in the worst case.
  - May be costly operation for large strings
  - Suitable for common typing mistakes
    - Comprehensive vs Comprenhensive
  - Problematic for specific domains
    - AT&T Corporation vs AT&T Corp
    - IBM Corporation vs AT&T Corporation

From: Koudas, Sarawagi, Strivastava, “Record Linkage: Similarity Measures and Algorithms”, VLDB 2006

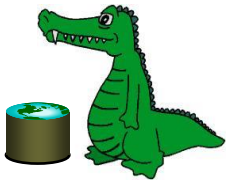


Overlap Metrics

- Given two sets of terms S, T
  - Jaccard coef.: Jaccard(S,T) = $|S \cap T| / |S \cup T|$
  - Variants
    - If scores (weights) available for each term (element in the set) compute Jaccard() only for terms with weight above a specific threshold.
- What constitutes a good choice of a term score?

- Terms can be words or "q-grams" (sequence of q characters in a field:
  - e.g., {'AT&', 'T&T', '&T ', 'T C', …} for AT&T Corp.

From: Koudas, Sarawagi, Strivastava, "Record Linkage: Similarity Measures and Algorithms", VLDB 2006



Some Similarity Measures

Handle Typographical errors

- Equality on a boolean predicate
- Edit distance
  - Levenstein, Smith-Waterman, Affine
- Set similarity
  - Jaccard, Dice
- Vector Based
  - Cosine similarity, TFIDF

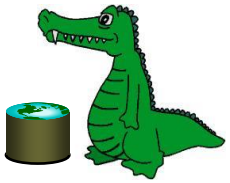
Good for Text like reviews/ tweets

Good for Names

- Alignment-based or Two-tiered
  - Jaro-Winkler, Soft-TFIDF, Monge-Elkan
- Phonetic Similarity
  - Soundex
- Translation-based
- Numeric distance between values
- Domain-specific

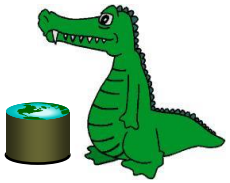
Useful for abbreviations, alternate names.

From: Getoor & Machanavajjhala: “Entity Resolution Tutorial”, VLDB 2012



More Sophisticated Techniques

- Evidence from multiple fields
  - Positive and Negative are possible
- Evidence from linkage pattern with other records
- Clustering-based approaches
- …



Summary

- Data Cleaning
  - Perspectives on “Dirty Data”
  - Perspectives on Data Quality
  - Some problems and solutions
- Data Integration
  - Schema Matching
  - Item Similarity for deduplication