Monday September 9th, 2013

Last time: Case Studies; Data Collection

Today: Text Similarity Metrics; Supervised Models

Next Time:

Announcements:

- Sam's OH Tues 11am, Porter 117
- Syria HRDAG/Linkage visit, Friday Sept 20th
- HW 2 due Monday 9/16

```
library(RecordLinkage)
data(RLdata500)
RLdata500
identity.RLdata500
```

How can/should we compare these records?

Text String Comparison Metrics

Properties of Similarity Metric:

How could we think of a similarity as a distance?

Exact Matching:

Partial Matching: What kinds of substrings might we be interested in? What would we look for?

Jaro-Winkler metric: taking into account typographical errors

Jaro (1972) - taking into account transpositions

Winkler update (1990) - taking into account transposition location

```
jarowinkler(str1, str2, W_1, W_2, W_3, r)
```

Levenshtein Edit Distance:
smallest number of single character insertions, deletions, and substitutions required to convert one string into another;
need dynamic algorithm to calculate

```
levenshteinSim(str1, str2)
levenshteinDist(str1, str2)
```

In practice, one similarity score per field;
let's say we have labels that tell us the unique identifiers?

What are some options for modeling matches/non-matches?

```
glm(Y~x1+x2+..., family=binomial())
```

More modeling:

```
library(tree)
tree(Y~x1+x2+...)
```