COMP 333 — Week 7 Entity Resolution

Preamble: Entity Resolution

Entity resolution, also called entity recognition or entity reconciliation,
is an important step in data cleaning
as it removes *duplicate* descriptions of an entity.

Entity resolution is a major step in text mining
and NLP (natural language processing)
because we talk about real-world entities.

However, the methods are advanced methods
involving similarity of strings and/or machine learning
that are beyond the scope of this course.

You need to know what entity resolution is,
and you get that in these notes.

You need to see entity resolution in action,
and you do in Lab 7.

The supplementary material for this lecture segment
are **beyond the scope of this course**.

Slides The slides are very advanced dealing with the algorithms
from the machine learning, AI, and Big Data community
for entity resolution.

This is much more than you need to know for the course!

Entity Resolution: Tutorial, VLDB 2012, by Lise Getoor and Ashwin Machanavajjhalar
`http://users.umiacs.umd.edu/~getoor/Tutorials/ER_VLDB2012.pdf`

Article The article is a good introduction to entity resolution
but it is limited in coverage of the available Python tools,
focussing on Dedupe.

You should find it easy to read, a good example of Python,
and an introduction to Dedupe and its algorithms.

Basics of Entity Resolution with Python and Dedupe, by Kyle Rossetti and Rebecca Bilbro
`https://www.districtdatalabs.com/basics-of-entity-resolution`

Video The Data Wrangling video has a good discussion of entity resolution.

Entity Resolution

Definition *Entity Resolution* is the task of disambiguating manifestations of real world entities in various records or mentions by linking and grouping.

Definition *Entity Resolution* is about determining when references to real-world entities are equivalent (refer to the same entity) or not equivalent (refer to different entities).

For example, there could be different ways of
addressing the same person in text,
different addresses for businesses, or
photos of a particular object.

The three primary tasks involved in entity resolution are

Deduplication : eliminating duplicate (exact) copies of repeated data.

Record linkage : identifying records that reference the same entity across different sources.

Canonicalization : converting data with more than one possible representation into a standard form.

Deduplication may create a *reference table* of entities
listing a unique description, or identifier, for each entity.

Linking is appending a common identifier (the one in the reference table)
to records / descriptions to denote the decision that they are equivalent.

Canonicalization also called Reference Matching
where we match noisy records to clean ones in a deduplicated reference table.

This is normalization.

The OpenRefine video 1 shows *canonicalization*
when applying text facets to eliminate the various descriptions
of terms like “Fixed Firm Price”.