COMP 333 — Week 7 Unification

Data Unification

Unification ensures that your data has a consistent form.

A form that is standard, known, agreed-upon across the dataset
for each particular kinds of data

- unit conversions on numbers
- character code representations
- name unification
- time unification
- date unification
- financial unification

Unification makes the data comparable
both within a column
and across columns

Professor Skiena's Lecture 7 on Data Cleaning
`http://www3.cs.stonybrook.edu/~skiena/519/`
has a good presentation of unification.

String Unification

Unification of strings is often an important step that simplifies comparison of strings:

- a string *“ the Happiest day of My Life ”*
- to all lower case
- and without leading or trailing blanks
- and only one blank between words
- *“the happiest day of my life”*

Name Unification

There are two steps in cleaning name data.

The first is name unification
which produces a standard form as a string for a name
that is consistent about

full first names, such as “Stephen” rather than “Steve”

full middle names rather than initials

last names

and suffix such as “Jr” or “III”

as well as string unification issues such as whitespace.

The second is entity resolution:
which person is the name referring to?
and what do I use as an identifier for this entity?

The article
`https://www.kalzumeus.com/2010/06/17/falsehoods-programmers-believe-about-names/`
discusses some of the pitfalls common in computer systems’ treatment of names.

Entity resolution will be discussed later.

Date Unification

For example, date unification to unify the formats
05-11-2020, 11-05-2020, 2020-05-11 or 11-May-2020.

There is also the issue of different societies using different calendars

and that calendars have changed throughout history.

Time Unification

The common need to unify time is the 12-hour clock verus the 24-hout clock.

There is also the issue of time zones.

So you need a standard.

Use Coordinated Universal Time (UTC),
a modern standard subsuming GMT.

Financial Unification

There are issues

- different currencies worldwide
- the fluctuation in exchange rates over time
- the historical fluctuation in the value (*"purchasing power"*)
  of one unit of currency due to inflation
- splits and reverse splits of shares in a company
- company dividends and return-of-capital affecting *adjusted cost base*
- investment returns as percentages and not absolute price differences

Dealing with time series data in finance is complicated

because of weekdays vs weekends, public holidays

Scaling and Normalization

Normalization is about bringing data into a common form
so that values can be compared.

We need to normalize emphvalues in a column.

The first step is unit unification:
It requires data to be in the same units
for example, grams, kilograms, or pounds

You cannot be directly compared: 5.3 kg is not the same as 5.3 g
even if 5.3 equals 5.3

Second, you need to have numerical values on the same scale
for example, accounting use dollars, 1000 dollar, or million dollar
as the "unit"
This is illustrated in the OpenRefine videos.

To compare *values* from two columns
you normalize *values* in the columns to be comparable..
There are two common normalizations

- percentages
  of their range min..max
- Z-scores
  where you regard each column distribution as a normal distribution
  and then scale to $\mathcal{N}$(m,s)
  the normal distribution with mean m and sd s

Note how "scaling" using a log transformation can improve DDA
as is illustrated in the OpenRefine video No.1

You may want to normalize *observations*
so that each has the same *"weight"*
in future computations.

To do this we think of each observation as a vector
(x1, x2, ..., xn)
and normalize the vector, as in
`https://chrisalbon.com/machine_learning/preprocessing_structured_data/normalizing_observations/`

The article on data cleaning

The Ultimate Guide to Data Cleaning: When the data is spewing garbage, by Omar Elgabry

has sections on *Scaling/Transformation* and *Normalization*
and illustrates the difference between them.