COMP 333 — Week 7 OpenRefine Example Videos

OpenRefine Example Videos

OpenRefine is an open source spreadsheet-like tool
for data wrangling.

It used to be called GoogleRefine
because Google developed it.

The three videos show OpenRefine in use

- Video 1: Introduction to Data Cleaning
- Video 2: Transformation
- Video 3: Data Enrichment
  which they call *Data Augmentation*

WATCH the three videos.

Think about how you would do the OpenRefine steps using Python and pandas.

Video One

The video introduces data cleaning with OpenRefine

with government data about contracts

available as a csv file.

The text data is particularly messy

No doubt it was entered by different people at different times.

Note how the *facets* give a sorted list with counts.

The sorting puts similar text next to each other.

The counts let you know the most common usage.

Inconsistent terms are very evident this way.
they are nearby and have very low counts

Note the powerful editing/replacement commands

to make the text values consistent.

Later in the video, clustering is used for a similar purpose.

Note the recording of each step

the ability to undo/redo by selecting a step in the list

The second example is numeric data, the total cost of the contract.

Again, the data is very messy.

Here the facet is a histogram.

Note the benefit of using a log scale for the facet.

Note how values refer to several different amount “units”

dollars, millions of dollars, billions of dollars

Note that the value of zero may indicate a *missing value*
being flagged by whoever entered the data

Video Two

The second video shows data transformation using OpenRefine.
The example is Movie data from wikipedia
that is tabular text marked up for the wiki.

Text (or string) transformations are standard computations
especially useful in data cleaning.

Re-formatting, or re-structuring, of text is part of data wrangling.

You see here several examples of *feature engineering*
such as “adding a new column”.

Useful here for intermediate values to assist transformation
by making selection/filtering easy
and then removed later.

Note the recording of each step
the ability to encode the steps as JSON
and the capability to re-play the steps
by cut-and-paste of the JSON text

Video Three

The third video shows OpenRefine being used for web retrieval
and data enrichment
which they call *data augmentation*.

The example is Movie data.

The retrieval example refers to Freebase
"an open shared database of the world's knowledge"
that no longer exists.

There is an example of *entity resolution*,
which they call *reconciliation*
that is done using clustering.