COMP 333 — Week 6 Data Wrangling Example

Data Wrangling Example

In Week 6 (this lecture) and Week 7 we will cover Data Wrangling
which is the most time-consuming phase of Data Analytics.

Data Wrangling is the ETL process of data warehouses
applied more generally as part of Data Analytics.

It is very important to clean and organize your data.

Remember GIGO (Garbage-In, Garbage-Out)

Example video

The video example is *Data mining and integration with Python* by Isaac Vidas at PyTexas 2015. `https://www.youtube.com/watch?v=qWcas-OUE9I`

The video covers an example of Data Wrangling
to create an actual product on POI (Points-Of-Interest) for the travel industry.

Yes, it is called *Data mining and integration with Python*
but it actually is on Data Wrangling.

Workflow

He provides a workflow for Data Wrangling

- content acquisition
- enrichment, which is adding new features from related data
- entity resolution
- combine, or integrate data from different sources

Content Acquisition

Data is fetched from multiple sources to become the content.

The workflow for content acquisition is

- extract data
- clean data
- normalize data
- analyze data
- integrate data

Entity Resolution

Entity resolution associates a unique identifier with a description of an entity.

The key for use as a primary key or foreign key in database terms.

An entity often has many different names or descriptions.

Entity resolution is important for matching entities.

Some Practical Advice

He offers some practical advice for Data Wrangling:

- Iterate
- Start with a small sample of data from your data source.
- Test for bad content.
- Reports and dashboards are also useful to data analysts during Data Wrangling

  that is, use Descriptive Data Analysis during Data Wrangling

Summary

His summary is

- Handle new content with care
- Break down long processes
- Find bottlenecks
- Create well-defined methods
- Issues you meet in samples of data
  are probably systemic
  for all data in the data source

An Important Video

Go back to the video of the POI system development several times
to see each step of Data Wrangling in action.
And again after seeing Week 7 material.