Information Visualization lab. Cleaning Data

Pere-Pau Vázquez

Dept. Computer Science – UPC

Data cleaning

- The problem: Garbage in – garbage out
  - When the input is not correct, the output can be unexpected

- Major problem in Computer Science

Data cleaning

- The quality of the data determines the quality of the outcome. Some criteria:
  - Valid
  - Accurate
  - Complete
  - Consistent
  - Uniform
  - Traceable
  - Updated

Data cleaning

- Valid: the data must meet certain constraints
  - Some columns must not be empty
  - Values must be of a certain data type
  - Values must fall within certain ranges
  - Sometimes values reference to other values in other columns/databases… (foreign keys)
  - Fields or values must be unique

Data cleaning

- Accuracy:
  - The data must conform to a true value
  - Valid values does not imply correctness
    - We can have a valid street address that does not exist (e.g. Av. Diagonal, 42000042)
  - Accuracy is different from precision

Data cleaning

- Completeness:
  - How complete the data and related measures are known
  - Missing data is going to happen
    - Should be mitigated as much as possible
- Consistency:
  - Measures are equivalent within the same dataset and across multiple datasets
    - Values contradicting each other depict inconsistency
    - A valid age, e.g. 2, is inconsistent with the marital status of *divorced*

Data cleaning

- Uniformity: The degree to which data is specified using the same units across systems
  - E.g. distances should be encoded either in meters or miles, dates should be stored either in the USA or European format, currency should always be the same
- Traceable: The source of the data is known
- Timeliness: Is the data updated properly?

Data cleaning

- Common data cleaning tasks
  - Formatting values
    - For example: changing date formats
    - Converting between units
  - Dealing with Anomalies
    - Detecting and dealing with outliers
    - Detecting and removing duplicate values

Data cleaning

- Common data cleaning tasks (ii)
  - Standardizing values
    - Making sure all values are formatted consistently for each variable
    - Detecting and standardizing values which have the same meaning but are formatted differently
      - E. g. Values like “Charles St” and “Charles Street” in an address.
  - Data Augmentation or Extension
    - Adding new columns to your dataset
    - Combining your dataset with another dataset

Data cleaning

- Data cleaning process
  - Data parsing
  - Data profiling

Data cleaning

- Data parsing
  - Take each observation and ensure it matches the schema
    - Schema: High level representation of the data observations
  - Check types (e.g. expected numbers are not strings…)
  - Ensure all the fields are filled
  - Identify duplicates
  - …
  - Once the data is syntactically correct, we can proceed to the semantic analysis

Data cleaning

- Things to look for:
  - Irrelevant data
  - Duplicates
  - Types
  - Typos
  - Syntax errors
  - Pad strings (strings padded with spaces or other characters)

Data cleaning

- Data profiling
  - Analyzes the data to ensure consistency
  - Check for outliers
  - Check for the correct order of the data (e.g. in time-dependent datasets)
  - …

Data cleaning

- Tools
  - Libraries: pandas, dora, datacleaner…
  - Applications: Excel, Tableau, Open Refine, Wrangler…

Data cleaning

- Challenges & features
  - Ability to work with large datasets
  - Task automation
  - Easy data exploration

- Extra features
  - Data augmentation
  - Bulk task executions

Open Refine

- Free open source tool for cleaning data
  - http://openrefine.org/
- Originally Google Refine
- Works in the browser
  - But locally
- Requires Java

Open Refine

- Procedure
  - Open a file
  - Work with it
    - Modify fields (all at once) to get consistent data
    - Add/remove columns
    - Transform data
  - Save
  - Optionally apply the same editions to other files

Open Refine

- Now take the tutorial (DataCleainingWithOpenRefine) and follow the examples there.

- Analyze and clean the file “inca_od_20191115.csv”.
- Analyze and clean the file “airQualityDailySummary.csv”.

Information Visualization lab. Cleaning Data

Pere-Pau Vázquez

Dept. Computer Science – UPC