Hands on Data Cleaning of Messy Data

June 26, 2018

Motivation of The Lab

- to present messy data issues in real-world datasets.
- to explain what can be done to clean such data.

A Real-World Scenario: The Rodents dataset

- data on rodents during a survey.
- each row denotes the information collected on an individual rodent.
- useful for studying population dynamics and species interactions.
- data is provided by Ernest et al. (2018).
- meta-data is also available at `http://esapubs.org/archive/ecol/E090/118/Portal_rodent_metadata.htm`.

Data Wrangling Challenges in the Rodents Dataset

- NOTE: I WILL GIVE SOME EXAMPLES HERE.
- **missing data**.
- **format variabilities**: typos, abbreviations, leading and trailing whitespace.
- some issues are addressed in `http://www.datacarpentry.org/python-ecology-lesson/`.

A Data Wrangling Tool: OpenRefine

- a tool for working with messy datasets.
- see Verborgh and De Wilde (2013) for details.
- useful links:
  - the software at `http://openrefine.org`.
  - the documentation at `https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users`.

Installation

- **long answer**:
  - detailed installation instructions at `http://openrefine.org/download.html`.
- **short answer**:
  - download the file depending on the OS at `https://github.com/OpenRefine/OpenRefine/releases/tag/2.8`.
  - install OpenRefine as follows:
    - Linux: extract.
    - Mac: open, drag icon into the Applications folder.
    - Windows: unzip.

Running and Loading Data

- **run OpenRefine** depending on the operating system:
  - Linux: ./refine in your installation folder
  - Mac: OpenRefine in your Applications folder
  - Windows: .exe file in your installation folder
- **get the dataset**:
  - clone the git repository at `https://github.com/tahaceritli/acm-summer-school`.
  - use the file in datasets/Portal_rodents_19772002_scinameUUIDs.csv.
- **import the data**:
  - click "Create Project".
  - click "Choose Files".
  - select Portal_rodents_19772002_scinameUUIDs.csv.
  - click "Next".

Data Preview

- configuration page for importing.
- a subset of the data is shown.
- use the defaults.
- click "Create Project" on the top-right corner.

Checking for Unique Values

- click drop-down arrow in the “survey_id” column.
- select “Facet>Customized Facet>Duplicates Facet”.
- results in a binary facet of “true” or “false”.
- “true” facet denotes rows with unique values.

View Range of Values

- click the drop-down arrow in the “scientificName” column.
- select “Facet>Text Facet”.
- lists the values and their counts.
- any problems with the data?

Updating Cell Values

- notice the spelling errors, e.g. “Amphespiza bilineata” for “Amphispiza bilineata”.
- hover over the former and select “edit” to update its value.

Filtering Rows

- click the drop-down arrow in the “scientificName” column.
- select “Text Filter”.
- type “bai”, which lists matching rows.

Clustering

- data often contains inconsistencies due to data collection procedures.
- “Clustering” helps to find cells in a column, that refers to the same entity with different values.
- various methods to determine clusters.

Clustering

- click the drop-down arrow in the “scientificName” column.
- select “Edit cells>Cluster and edit...”.
- change the method to nearest neighbor.
- you can now check boxes and merge the clusters.

Trimming Whitespace

- click the drop-down arrow in the “scientificName” column.
- select “Edit cells>Common transforms>Trim leading and trailing whitespace”.

Reconciliation

- refers to the process of matching data to databases.
- details at https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation.
- OpenRefine uses a knowledge base named Wikidata https://www.wikidata.org/wiki/Wikidata:Main_Page.
- click the drop-down arrow in a column.
- select "Reconcile>Start reconciling...".
- select "Wikidata" and click "Start Reconciling".

Deliverables

1. **unique values**:
   - report whether every ID in the column "survey_id" is unique.
2. **filtering**:
   - report how many rows are left after filtering the "scientificName" column with the text "bai".
3. **missing data**:
   - report columns that have missing data.
   - report missing data encodings used for columns with missing entries.
   - check whether they are actually missing (hint: you may want to check meta-data!).
   - hint: commonly used encodings: "NA", "N/A", "Null", "-1", "-99", etc.

Deliverables

4. **format variabilities**:
   - report how many unique values exist in the “country” column of the original dataset.
   - is there any problems with the data? if so, explain how did you solve them?
   - report how many unique values you have after fixing the potential inconsistencies.
   - hint: “clustering” or “reconciliation” feature of OpenRefine could be useful for this task.

Submissions

- 1 page of either .txt or .pdf (not a .doc file!).
- answer the four questions mentioned earlier.
- send your report to `acm2018datacleaninglab@gmail.com` with the title ACMReport-NAME-SURNAME.

References I

Ernest, M., Brown, J., Valone, T., and White, E. P. (2018). Portal Project Teaching Database. `https://figshare.com/articles/Portal_Project_Teaching_Database/1314459`.

Verborgh, R. and De Wilde, M. (2013). *Using OpenRefine*. Packt Publishing Ltd.