Data Munging with Open Refine

Morning Session II

Jacob F. Koehler, PhD. [1]

[1] Data Carpentry

Rutgers Newark, February 11, 2018

Overview

What is Open Refine?

OPEN
Refine
A free, open source, powerful tool
for working with messy data

Figure: `openrefine.org`

- All actions tracked and reversible
- Raw data kept raw
- Repeatable across multiple files
- Simple Clustering

Creating a Project

1. click Create Project and select Get data from This Computer.
2. Click Choose Files and select the file Portal_rodents_19772002_scinameUUIDs.csv. Click Open or double-click on the filename.
3. Click Next under the browse button to upload the data into OpenRefine.
4. OpenRefine gives you a preview - a chance to show you it understood the file. If, for example, your file was really tab-delimited, the preview might look strange, you would choose the correct separator in the box shown and click Update Preview (bottom left). If this is the wrong file, click Start Over (upper left).
5. If all looks well, click Create Project (upper right).

Faceting

We will facet the **scientificName** column. In OpenRefine, faceting:

- seeing a big picture of your data, and
- filtering down to just the subset of rows that you want to change in bulk.

Challenge

1. Using faceting, find out how many years are represented in the census.
2. Is the column formatted as Number, Date, or Text? How does changing the format change the faceting display?
3. Which years have the most and least observations?

Clustering

"finding groups of different values that might be alternative representations of the same thing"

1. cluster **scientificName**
2. **key collision** and **metaphone3**

Split

1. Let us suppose we want to split the scientificName column into separate colums for genus and for species.
2. Click the down arrow at the top of the scientificName column. Choose Edit Column > Split into several columns...
3. In the pop-up, in the Separator box, replace the comma with a space.
4. Uncheck the box that says Remove this column.
5. Click OK. You'll get some new columns called scientificName 1, scientificName 2, and so on.
6. Notice that in some cases scientificName 1 and scientificName 2 are empty. Why is this? What do you think we can do to fix this?

Split

1. Let us suppose we want to split the scientificName column into separate colums for genus and for species.
2. Click the down arrow at the top of the scientificName column. Choose Edit Column > Split into several columns...
3. In the pop-up, in the Separator box, replace the comma with a space.
4. Uncheck the box that says Remove this column.
5. Click OK. You'll get some new columns called scientificName 1, scientificName 2, and so on.
6. Notice that in some cases scientificName 1 and scientificName 2 are empty. Why is this? What do you think we can do to fix this?

Try changing second new column to "species".

Undo/Redo

1. Click where it says Undo / Redo on the left side of the screen. All the changes you have made so far are listed here.
2. Click on the step that you want to go back to, in this case the previous step. The added columns will disappear.
3. Notice that you can still click on the last step and make the columns reappear, and toggle back and forth between these states.
4. Leave the dataset in the state in which the scientificNames were clustered, but not yet split.

Trim Leading and Trailing Whitespace

1. In the header for the column scientificName, choose Edit cells> Common transforms > Trim leading and trailing whitespace.
2. Notice that the Split step has now disappeared from the Undo / Redo pane on the left and is replaced with a Text transform on 3 cells
3. Perform the same Split operation on scientificName that you undid earlier. This time you should only get two new columns. Why?

Filtering and Sorting with OpenRefine

1. Click the down arrow next to scientificName > Text filter. A scientificName facet will appear on the left margin.
2. Type in bai and press return. There are 48 matching rows of the original 35549 rows (and these rows are selected for the subsequent steps).
3. At the top, change the view to Show 50 rows. This way you will see all the matching rows.

Exercise

- What scientific names (genus and species) are selected by this procedure?
- How would you restrict this to one of the species selected?

Excluding Entries

Filtering rather than faceting: choose observations instead of listing all

- Use **include/exclude** to select only entries from one of these two species.

Sort

- Sort by month. How can you ensure that months are in order?

Sort

- Sort by month. How can you ensure that months are in order?
- Sort the data by plot. What year(s) were observations recorded for plot 1 in this filtered dataset.

Sorting by Multiple Columns

You might like to look for trends in your data by month of collection across years.

- How do you sort your data by month?
- How would you do this differently if you were instead trying to see all of your entries in chronological order?

Sorting by Multiple Columns

You might like to look for trends in your data by month of collection across years.

- How do you sort your data by month?
- How would you do this differently if you were instead trying to see all of your entries in chronological order?
- Sort by year, month and day in some order. Be creative: try sorting as numbers or text, and in reverse order (largest to smallest or z to a).
  Use > Sort > Remove sort to remove the sort on the second of three columns. Notice how that changes the order.