Google Refine

A presentation for the Wolbach library

Hi and welcome to the presentation on Google Refine. In this 1-hour session we will demonstrate the power and utility of Google Refine to clean, transform, analyse and augment tabular data. We'll go through several exercises using some sample Worldcat library data. Time permitting, we'll finish with a couple of demonstrations of how NASA/ADS has made use of Google Refine.

Introduction

- Who we are, NASA/ADS, etc.
- What is Google Refine all about?
    - Free, Open Source
    - Runs on Windows, Mac & Linux
    - Desktop application that runs in your browser, therefore secure.
    - Brief history of the project.
- Installation (v2.5), download available from https://code.google.com/p/google-refine/

Data Normalization & Transformation

- Project creation, data import & column names
- Normalization via text faceting
- Normalization via clustering
- Numeric faceting
- Undo/Redo
- Dates & Timeline Faceting
- Text filtering and row removal
- Text transform using GREL (https://code.google.com/p/google-refine/wiki/GRELStringFunctions)
- Add column based on this column
- Splitting columns
- Exporting project/data

Google Refine @ ADS (time permitting)

- Extlinks - analyzing external links from XML full-text articles
- Affiliations Data
- Identifying undervalued pitchers using QERA (http://en.wikipedia.org/wiki/Nate_Silver#Other_tools)

The SAO/NASA Astrophysics Data System
Jay Luker <jluker@cfa.harvard.edu>
Benoit Thiell <bthiell@cfa.harvard.edu>

Google Refine Workshop

Exercises

Exercise 0: Creation of the project

- Download the CSV file at http://bit.ly/refinedemo.
- Open Google Refine at http://127.0.0.1:3333.
- Select *Create project*.
- Click *Choose Files* and select the file that you've downloaded. Click *Next*.
- The data is now being uploaded.
- Because the file is CSV (comma-separated values), select commas for the separator. You should now see a better preview.
- Change the name of the project to "My first project".
- Create the project.

Exercise 1: Basic functions

- How many rows are in the project?
- Change the number of visible rows from 10 to 25 rows.
- Move to the next page.
- Move to the last page.
- Move to the first page.

Exercise 2: Faceting

- Create a text facet on the column "Author" (click on the column's drop-down menu and invoke *Facet* > *Text facet*).
- How many different authors are there?
- Sort the facet by count.
- What is the most frequent author? How many times does it appear?

The SAO/NASA Astrophysics Data System
Jay Luker <iluker@cfa.harvard.edu>
Benoit Thiell <bthiell@cfa.harvard.edu>

- Edit the third author in the facet so that it matches the first author (remove the word "Project").
- How many different authors are there now?
- What is the most frequent author now? How many times does it appear?

Exercise 3: Clustering

- Cluster on the column "Author" (click on the column's drop-down menu and invoke *Edit cells* > *Cluster and edit...*).
- What is a cluster?
- How many clusters are there?
- For each cluster, select the best value by clicking on it. Alternatively, you can edit the value in the text field on the right of the cluster.
- Once you have selected a value for each cluster, click *Merge Selected & Close*.
- How many different authors are there now?
- Close the author facet.

Exercise 4: Numeric faceting

- Create a numeric facet on the column "Year" (click on the column's drop-down menu and invoke *Facet* > *Numeric facet*).
- How many rows don't have a year?
- Why are there non-numeric years?
- What is wrong with the numeric years for publications prior to 1100 AD? (Look at the Publication column).
- Correct the years for the 1978 and 1980 papers. Be careful to select type "number" in the edit menu.
- Delete the year for the 975 paper.
- Hit *Refresh* and see how the numeric facet changes.

Exercise 5: Undo & Redo

- Oops. You just discovered that the 975 year is correct. Using the *Undo / Redo* functionality, restore the cell to its previous value.
- Notice that the year facet goes back to its previous state.
- Close the year facet.

Exercise 6: Timeline facet

- Convert the value in the column "Added date" to a date format (click on the column's drop-down menu and invoke *Edit cells* > *Common transforms* > *To date*).
- Create a timeline facet on the column "Added date" (click on the column's drop-down menu and invoke *Facet* > *Timeline facet*).
- How long did it take between the initial addition to the last addition?

- Close the timeline facet.

Exercise 7: Text filtering and row removal

- We do not wish to make changes for the records that are articles, so we are going to remove them from the project.
- Open the filter menu for the column "Item type" (click on the column's drop-down menu and invoke *Text filter*).
- Select only the rows where the item type contains "art".
- How many rows are now selected?
- Delete these rows (click on the main drop-down menu and invoke *Edit rows* > *Remove all matching rows*).
- Why do we have now 0 matching rows?
- Remove the text filter.
- How many rows are there in the project now?

Exercise 8: Text transformation

- Create a text facet on the column "Publisher". Sort by count.
- How many different publishers are there?
- Cluster on the column "Publisher". Click *Select All* and then *Merge Selected & Close*.
- How many different publishers are there now?
- Using the cell transformation tool (click on the column's drop-down menu and invoke *Edit cells* > *Transform...*), replace "S. Burlington" with "South Burlington". The syntax for replacing is `value.replace("S. Burlington", "South Burlington")`.
- How many different publishers are there now?
- Close the publisher facet.

Exercise 9: Add column based on this column

- For each title, we are now going to create a short title which is the beginning (50 characters) of each title followed by an ellipsis.
- Open the column addition menu (click on the column's drop-down menu and invoke *Edit column* > *Add column based on this column*).
- Set the new column title to "Short title".
- Create the new column with the expression: `value.substring(0, 50) + "..."`

Exercise 10: Split into several columns

- Create a text facet on the column "Item type".
- How many different types of books do we have? What is the total number of books?
- Now we are going to create 2 columns for the item type, one for the main type and one for the secondary type. For example, for the type "book_digital", the main type is "book" and the secondary type is "digital".

- Open the column splitting menu for the column "Item type" (click on the column's drop-down menu and invoke *Edit column* > *Split into several columns*).
- Select the correct separator and split into two columns at most.
- Why is the facet now empty? Close it.
- Rename (click on the column's drop-down menu and invoke *Edit column* > *Rename column*) the two new columns to "Main type" and "Secondary type".
- Open a facet for both type columns.
- How many books do we have now? What are the secondary types of the book type?
- Close the facets.

Exercise 11: Exporting

- Using the export function, export your project to a CSV file.