Introduction to Data Cleaning with OpenRefine

Alexander Botzki and Herwig Van Marck

2015-11

VIB

BITS VIB
BioInformatics Training and Services

What is OpenRefine?

A free, open source, power tool for working with messy data

Formarly known as Google Refine

Being rebranded as OpenRefine (http://openrefine.org)

VIB-Bits plugins to facilitate usage (http://www.bits.vib.be/software)

How does OpenRefine compare with other tools?

- Compared to spreadsheets
  - Basic unit of interaction is column (versus cells)
  - Pro: easier to import data, explore, manipulate and export again
- Compared to scripting
  - Pro: you see your data, while it is being transformed
  - Con: for medium size data sets
- Compared to databases
  - Pro: you see your data, while it is being queried
  - Con: for simple data structures

What are typical use cases for OpenRefine?

- Explore unknown/new data files
- Manipulate/clean data to prepare for other tools
  - E.g. GraphPad Prism
- Get data from web services
- Use as a workflow tool
- Create dashboards

How to install OpenRefine

Installation instructions can be found on

https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions

(version 2.6 beta 1 is preferable)

Platforms:

- Windows
- Mac OSX
- Linux

How to run OpenRefine

- Windows

  Run the .exe file in the installation folder

- Mac OSX

  Double click the OpenRefine app in the Applications folder

- Linux

  Start ./refine in the installation folder

How to shut down OpenRefine

- Windows

  Press Ctrl-C in the OpenRefine Command windows

- Mac OSX

  Invoke the Quit command on the OpenRefine app

- Linux

  Press Ctrl-C in the shell

User Interface overview

Home screen

- Create Project
- Import Project
- Open Project
  - Delete projects
  - Rename projects

User Interface overview

- Project
  - Data table
  - Column menu
  - Side bar
    - Facet / Filter
    - Undo / Redo
  - Home button

Input data demo

Configuration screen

- Preview area
- Options area
  - Parsing format
  - Parsing options
    - Ignore first lines
    - Parse as column header

Exercise 1: importing files

a) Import the file qPcr results.txt

Exercise 1: importing files

a) Import the file qPcr results.txt

Hints:

- Use Ignore first ... line(s) at beginning of file
- Check if correct separator is used

Explore data using facets demo

- Text facet
  - Sorting by name/count
  - Facet counts in tab separated format
  - Querying
    - Select choice
    - Inverted query
    - Resetting facet
    - Facet counts reflect query

Explore data using facets demo

Numeric facet

- Query using slider

Timeline facet

- Histogram
- Query using slider
- Query is reflected in histogram

Explore data using facets demo

Custom text facet

- Create selectable items using expression
  - e.g. value.toLowercase().contains(“good”)
- Query by selecting item

Explore data using facets demo

Data and history is always saved, but query is not!

Current query can be saved

- Using the Permalink link and bookmarking the page
- Facet box size can not be saved

Exercise 2: using facets

Use the file syst-nocallsCG69.bed to determine

a) the number of no-call regions that are larger than 1040 bases long in chromosome 21

b) the length of the longest region in chromosome 1

Exercise 2: using facets

Use the file syst-nocallsCG69.bed to determine

a) the number of no-call regions that are larger than 1040 bases long in chromosome 21

b) the length of the longest region in chromosome 1

Hints:

- The 5th column contains the length of a region
- Use a custom facet with ‘>‘ in the expression
- Use sort to determine the longest region

Cleaning data demo

Cluster facet choices

- Try different keying functions
- Merge clusters of similar values
- Browse this cluster link opens cluster in a new window

Exercise 3a: cleaning data

Clean the file qPcr results.txt you loaded earlier.

Exercise 3a: cleaning data

Clean the file qPcr results.txt you loaded earlier.

Hints:

- Check columns using facets
- Use the clustering tool in facets

Manipulate data demo

Use ‘Line based text files’ for complex data files

Edit column>Split into several coumns...

- Regular expressions for separator

  E.g. ‘ +’

- Split into ... columns at most

Edit cells>Blank down on index column (1st)

Use custom text facet to check if separator is used

E.g. value.contains(“|”)

Expression window details

Preview tab

Help tab

- Expressions can also use chained form
  - e.g. value.contains(“|”) instead of contains(value,”|”)

History tab

- Reuse recently used expressions

Starred tab

- Expressions that were starred in the history tab

Exercise 3b: cleaning data

Clean the numerical column in file qPcr results.txt you loaded earlier.

Exercise 3b: cleaning data

Clean the numerical column in file qPcr results.txt you loaded earlier.

Hints:

- Use numeric facet to explore column
- Use the replaceChars(...) command and the toNumber() command in transform column

Extended course material

For extended course material

- Go to www.bits.vib.be
- Click on Training
- Click on Previous trainings
- Click on Data Manipulation with OpenRefine

and also

- Click on Custom trainings
- See OpenRefine