Data Schools

Lecture 2: Introduction to OpenRefine

Presented by Dr. Bianca Peterson

Motivations for the OpenRefine Lesson

- Data is often very messy and it has to be refined before it is useful. OpenRefine provides a set of tools to allow you to identify and amend the messy data.
- It is important to know what you did to your data (i.e., the steps you took to clean the data).
- OpenRefine does not modify your original dataset. You must export your modified dataset to a new file.
- All actions are easily reversed (undone) in OpenRefine.
- OpenRefine saves as you go. You can return to the project at any time to pick up where you left off, or export your data to a new file.
- Data cleaning steps often need repeating with multiple files. OpenRefine keeps track of all your actions and allows them to be applied to multiple datasets.
- Some concepts such as clustering algorithms are quite complex, but OpenRefine makes it easy to introduce them, use them, and show their power.
- Open source.
- Works with large-ish datasets (100,000 rows). Can adjust memory allocation to accommodate larger datasets.
- A large growing community, from novice to expert, ready to help.

What is OpenRefine?

- OpenRefine is a free and open source Java program that runs on your machine (not in the cloud) - it is a desktop application that uses your web browser as a graphical interface.
- No internet connection is needed, and none of the data or commands you enter in OpenRefine are sent to a remote server. Sensitive data thus never have to leave your computer to get cleaned up.

Basics of OpenRefine

- You can find out a lot more about OpenRefine at http://openrefine.org and check out some great introductory videos.
- There is a Google Group that can answer a lot of beginner questions and problems.
- The OpenRefine GitHub wiki page has a reference of the General Refine Expression Language (GREL).

Key Points

- OpenRefine is a powerful, free and open source tool that can be used for data cleaning.
- OpenRefine will automatically track any steps you take in working with your data.