Lab #1: Data Wrangling and Exploration with OpenRefine

Everything Data
CompSci 290.01 Spring 2014
Announcements (Tue. Jan. 14)

• Have you been receiving emails from the class mailing list?
• Please sign up for this class on Piazza and use it for Q&A!
• Permission #’s will be emailed tonight
• Crazy coincidence: OpenRefine class by Duke Library at 3pm today!
  – http://library.duke.edu/events/data/event.do?id=6517&occur=14499
Seat assignment

Front of D106

A       B       C

D       E       F

G       H       I

J       K       L

M       N       O

Back of D106
Format of this lab (tentative)

• Catch up, or head start (15 min.)
• Discuss last homework (20 min.)
  – Go over solutions with your team
  – Share your solutions with class
• Work on new exercises (30 min.)
  – Win challenge to get extra credits!
• Summarize lessons learned (10 min.)
Catch up or head start (15 min.)

If you were stuck somewhere in the last homework, catch up
• For help, ask us or your teammates

Otherwise, get a head start on the challenge
• Lab 1, Part 2 (we will come back to Part 1 later together)
Discuss last homework (20 min.)

Homework 1, Part 3

• Go over solutions with your teammates (10 min.)

• Share your solution with class (10 min.)
  – Ring me on Google Hangout if you need to show your screen
  • You may need to install a Google Hangout plugin/app on your computer
  • My gmail id is junyang
New exercises (30 min.)

Lab 1, Part 1 (~10 min.)
• Fun with facets
• Feel free to start Part 2 once you are done
• Share your solution with class

Lab 1, Part 2 (~20 min.)
• Challenge!
Lesson learned: reality check

• Data is messy
• Reality is messy
• Garbage in, garbage out

☞ One reason why you need data wrangling

Lesson learned: UI or not UI

• Exploratory analysis goes a long way
  – You can get fair amount of insight without a single line of code

• Interactivity is nice

• Easy to learn, difficult to master
  – Tool-specific recipes vs. universal primitives

Image credit: http://www.24tee.com/movie_related_shirts/lunch_is_for_wimps_tee_shirt
Lesson learned: “abstraction”

- More structure and semantics ➞ more powerful analysis, and
- Different data models ➞ different questions and processing methods

☞ Another reason you need data wrangling

- Examples in this lab
  - Character strings: regular expressions
  - Dates: more meaningful range conditions
  - Tables: filtering, grouping, counting, sorting…
Finally

• Remember to **submit by midnight tonight**
• Sample solutions (to Homework #1 and Lab #1) will be posted tomorrow
• Feel free to give us feedback (anonymously if you prefer) on Piazza

• On Thursday Ashwin will tell you all about **text analytics**!