"GARBAGE IN, GARBAGE OUT": WATCHING OUT FOR BAD DATA

Polonsky/Brine Summer Internships - 2018

Prof. Deena Engel

First, study your data structure and values:

- For example, in a CSV file, how many values are there per line? And is the number of fields consistent from one record or row to the next?
- Is the delimiter character consistent?
- For fixed width columns … are the columns set up consistently?
- *For class discussion: What are some of the approaches you might use to correct these problems?*

Field Validation

- Are all of the values of a given field cast in the same data type?
  - For example, does a field for numbers consist of only numeric data?
- Is each field limited to one value?

  - *For class discussion: how might you check for these errors? Which corrections would require human intervention (i.e. could not be handled programmatically)?*

Value Validation

- Do all of the values for a given field have the same meaning? (For example, in a field called “height”, are all of the values related to the height of each entity and not its weight?)
- Are all of the values within a range that is reasonable for that field? (For example, 10 feet is not a reasonable value for the height of a person.)
- Do all of the values use the same unit of measure (pounds, meters, years, etc)?
- Do all of the values use an appropriate unit of measure (dollars for prices, meters for length, kilograms for weights, etc.)?
- *For class discussion: which of the above can be checked programmatically? And which require human intervention?*

Use Simple Statistics to evaluate your data

- For numerical data, what are some of the statistics that could help to inform you if the values are meaningful?
  - For example, taking minimum and maximum values could highlight “outliers” or errors.
  - Checking both the mean and the median will help to clarify an overview of the data.
    - *What does it mean if the mean (average) is wildly different from the median (middle value)?*
  - Use visualization (e.g. graphing your data) can give one a quick overview.

Data that is intended for human readers … not machines!

In some cases, data are widely spread out or organized in a logical and visual way for readers:

- The Federal CPI index is one of many examples: http://www.bls.gov/cpi/cpid1407.pdf

  - *For class discussion:* How would you use these data?

Another problem: Bad data in the unstructured textual data

- Sometimes textual data includes characters that are specific to its presentation (e.g. markup) and this must be removed or ignored programmatically before running analysis.
- What happens if you spell-check your text?
- Tokenizing text can create problems if not done carefully enough.
  - For example, the word “don’t” might be cut into the tokens “don” and “t” which are both meaningless for the purposes of textual analysis.

Liars!?

- Just because it is on the web … does that make the data valid? !!
- What are some of the ways that you might determine whether the provider of your data is lying to you? Or simply posting sloppy data?
- Many of the current techniques for evaluating textual data are outside of the scope of this meeting (e.g. sentiment classification, polarized language and other problems) and numeric data (manipulated results, selectively posted results) but I would be happy suggest further resources if you are interested or if this impacts your project.

Other possible sources for errors

- It is important to evaluate your data in light of the context for common errors that can misshape your results:
  - For example, stock splits in a database of financial transactions
  - Political data that might be skewed by the provider
  - Survey data that are limited to specific socio-economic subgroups in society
- These considerations are outside of the scope of this workshop but we are happy to discuss this with you.
- Subject Librarians at the Bobst to assist you: http://library.nyu.edu/research/lib_arc.html.

Good Practices

- □ **Ensure data traceability**: Keep careful notes on your data sources (e.g. URLs) as well as sample data files for future reference.
- □ **Ensure reproducibility**: Keep copies and document any changes that you make as you work with your data so that you can reproduce the results with the same or more current datasets.

In conclusion …

- □ The four standards for evaluating and maintaining good data*:

1. Complete (Do you have all of the data that you need or if not, is this a representative sample?)
2. Coherent (Do the data and the data structure “make sense”?)
3. Correct (Are the values correct?)
4. aCcountability (Can you trace the data and reproduce any modifications?)

**Adapted from The Bad Data Handbook, edited by Q. Ethan McCallum, published by O’Reilly*