Announcements (Mon. Jan. 22)

• Office Hours: https://sites.duke.edu/compsci216s2019/help/

• Use Piazza!

• Data+ Information Fair
  – 3:00-4:00 today, 3rd Floor of Gross Hall
  – Data+, Code+, and Computer Science projects
  – Applications due Feb 25
Data+, Code+, & CompSci

Data+ 2019 Information Fair

- 10-week summer research experience
- Gain broad exposure to the modern world of data science
- Explore new data-driven approaches to interdisciplinary challenges
- Small project teams in a communal environment!

UNDERGRADS:
Learn more about our Interdisciplinary Project Teams!

Grad & Ph.D Students:
Gain valuable project management experience with Data+

January 17, 2019 3:00 – 4:00 p.m. Ahmadieh Atrium (Gross Hall 3rd Floor)

For more information, contact Paul Bendich (bendich@math.duke.edu) or Ariel Dawn (ariel.dawn@duke.edu)
Summer Undergrad Projects in Computer Science

• Automated Agenda Management  
  – Dr. Kamesh Munagala
• Scaling Up “Pop-Up” Fact Checking  
  – Dr. Jun Yang
• Breadcrumbs: Analyzing Classroom Data  
  – Dr. Kristin Stephens-Martinez and Dr. Jeff Forbes
• Duke Human/ML Decision Making  
  – Dr. Cynthia Rudin
• Protecting Individual Privacy Using Differential Privacy  
  – Dr. Ashwin Machanavajjhala
Working with the Boston Uber data

- Counting rows
- Grouping by one or more values

1. csv DictReader
2. Python Data Analysis Library (pandas)
3. datascience
Sources of data …

- Paper records
- Sensors
- Web pages
- Activity Logs
- …
… are noisy

- Data entry errors
- Measurement errors
- Extraction errors

Mamma mia! Why there are 20,000 pregnant MEN in Britain

- Meanwhile more than 3,000 children and teenagers have apparently accessed geriatric services

By CLAIRE BATES
PUBLISHED: 11:18 EST, 9 April 2012 | UPDATED: 03:46 EST, 10 April 2012
... are in a variety of formats

- Tables
- Text
- Tweets
- HTML Files
- PDFs
- XML Files
- RDF
- Custom Formats
- ...
- ...
... and sometimes plain wrong

Typical data quality issues

1) parsing text into fields (separator issues)
2) Naming conventions: NYC vs New York
3) Missing required field (e.g. no SSN)
4) Different representations (2 vs Two)
5) Fields too long (get truncated)
6) Primary key violation (two people with the same social security number)
7) Redundant Records (exact match or other)
8) Formatting issues – especially dates
9) Licensing issues/Privacy/ keep you from using the data as you would like

Adapted from Berkeley CS 194 Lecture notes
Domain knowledge is critical

Ages of employees (US)

- Median: 37
- Mean: 58.52632
- Variance: 9252.041

Adapted from Berkeley CS 194 Lecture notes
Domain knowledge is critical

• Same data values in a different domain may not have errors ...

   12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 68 450

• Number of friends in a social network
  – A few people in Facebook have millions of friends, while the average number of friends is on the order of hundreds
Data Cleaning Makes Everything Okay?

The appearance of a hole in the earth's ozone layer over Antarctica, first detected in 1976, was so unexpected that scientists didn't pay attention to what their instruments were telling them; they thought their instruments were malfunctioning.

In fact, the data were rejected as unreasonable by data quality control algorithms.

National Center for Atmospheric Research

Adapted from Berkeley CS 194 Lecture notes
Regular Expressions
Regular expressions

• A formal language to specify sets of strings.
  – Useful for extracting data fields from text
  – Useful for searching through text
  – (think “Find” or “grep”)
Disjunctions

• Find all occurrences of letter ‘n’ in the following sentence.

“I am nobody.
Nobody is perfect.
Therefore I am perfect.”

• /[Nn]/
Disjunctions

<table>
<thead>
<tr>
<th>Expression</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>/[Nn]obody/</td>
<td>Nobody or nobody</td>
</tr>
<tr>
<td>/[A-Z]/</td>
<td>All upper case letters</td>
</tr>
<tr>
<td>/[0-9]/</td>
<td>All numbers</td>
</tr>
<tr>
<td>/[A-Za-z]/</td>
<td>All letters</td>
</tr>
<tr>
<td>/[]/</td>
<td>All white spaces</td>
</tr>
</tbody>
</table>

- **How do you search for ‘nobody’ or ‘perfect’?**

`nobody | perfect`
Negations

• *Find all letters that are not ‘n’ in the following sentence.*

“*I am nobody. Nobody is perfect. Therefore I am perfect.*”

• `/[^Nn] `/
## Negations

<table>
<thead>
<tr>
<th>Expression</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>/[^ ]/</td>
<td>Non space characters</td>
</tr>
<tr>
<td>/[A-Z]/</td>
<td>Not capital letters</td>
</tr>
<tr>
<td>/[A-Za-z]/</td>
<td>Not letters</td>
</tr>
</tbody>
</table>

- *How do you search for ‘^’?*

/\^[\^]/
Regular operators: * + ? .

<table>
<thead>
<tr>
<th>Expression</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>/no?body/</td>
<td>nobody, nbody</td>
</tr>
<tr>
<td>/no*body/</td>
<td>nbody, nobody, noobody, nooobody, ....</td>
</tr>
<tr>
<td>/no+body/</td>
<td>nobody, noobody, nooobody ...</td>
</tr>
<tr>
<td>/n.body/</td>
<td>n body, nobody, n3body, npbody ...</td>
</tr>
</tbody>
</table>

• ? : 0 or 1 occurrence
• * : 0 or more occurrences
• + : 1 or more occurrences
• . : any one character
Anchors and Flags

• Start and End
  – ^ : start of string
  – $ : end of string

• Flags
  – g – global don’t return after first match
  – i – insensitive case insensitive match
() operator

• (no)?body matches body and nobody

• String matched by the expression within () can be captured
  – /(no)body/ when applied to the string “nobody” returns an array of length 1 with the string “no”.
Exercise

• Find all occurrences of the word ‘am’ in the sentence:

“Am I amply ashamed of the sham that I am?”

Ans: the 2\textsuperscript{nd} group in
\[
/(^\text|[^A-Za-z])([Aa]m)([^A-Za-z]|$)\]/

Matches start of line or non letter

Matches end of line or non letter
RegEx practice

• Go to https://regex101.com

• Complete the first six (6) tasks in the Quiz
Summary

• Cleaning is an important step before making sense of data.

• Diverse set of techniques are usually employed to fix many types of errors.

• Regular expressions are a useful tool to parse the data into the right format.