University of California, Berkeley
Master of Information and Data Science (MIDS)
W205 – Fundamentals of Data Engineering

Week 6 – Data Wrangling,
Part I: Common File Formats I/O

Agenda for Today’s Class

- Attendance and Participation
- Announcements
- Schedule and Due Dates
- Work / Life / School Balance
- Asynch High Level Review in a Nutshell
- Project 2
- Breakouts
- Summary

Attendance and Participation

Please record your attendance and participation for today’s class:

GitHub => ucb_mids_w205_repo => README.md => Attendance and Participation

Announcements

- Upcoming holidays and/or breaks
- Makeup classes for holidays
- Upcoming events
- Student evaluations
- Etc.

Schedule and Due Dates

Take a quick look at the next couple of weeks’ due dates:

GitHub => ucb_mids_w205_repo => README.md => Schedule and Due Dates

Work / Life / School Balance

Open Discussion

Student feedback

- About 5 minutes
- How are things going related to work / life / school balance?
- How is w205 going? Difficulty? Time?
- Impact of any natural and/or man-made disasters
- Etc.

Asynch High Level Review in a Nutshell

Each week we will spend about 15 minutes reviewing the most important high level concepts from the asynch

Data Wrangling

- Data engineers prepare data for data science
- Analogous to cattle wrangling from the wild into pens
    - Cattle out on the range
    - Find them in several groups
    - Herd them into one group
    - Move them into pens
    - Clean up, vet, etc.
- Munging – alternative term, originally just a destructive irrevocable change to data

Data Encoding

- Bit = 0 or 1
- Byte = 8 bits
- Nibble = 4 bits
- Hex
  - Base 16
  - 0..9 then A=10… F=15
  - 1 hex digits = nibble, 2 hex digits = byte

Storage Units

- Kilobyte = KiB = 1024
- Megabyte = MiB = 1024^2
- Gigabyte = GiB = 1024^3
- Terabyte = TiB = 1024^4
- Petabyte = PiB = 1024^5
- Exabyte = EiB = 1024^6
- Zettabyte = ZiB = 1024^7
- Yottabyte = YiB = 1024^8

- Tips:
  - Always use the official units
  - It’s more professional
  - No room for any confusion
  - Unofficial units will mark you as an amateur

Encoding Natural Languages

- EBCDIC – English, IBM mainframe
- ASCII – English, standard, 7 bits
- Unicode – all languages, 2 or 3 bytes per character, wastes space for English
- UTF-8 – Unicode without wasted space for English, uses the “extra bit” to signal the number of bytes
- Emojis – part of Unicode and handled by UTF-8
- Uuencoding – allows binary data to pass through networks, MIME, Base64

CSV: Comma-Separated Values

- No official standard, MS Excel format is de facto
- Mimics structure of relational database table
  - First line – typically list of fields (columns)
  - Remaining lines – records (rows)
  - Easy to load into database table
  - Easy to extract from database table
  - Some products can treat a CSV file like a database table
- Exceptions – common in field, double quotes, etc.

JSON: JavaScript Object Notation

- XML – original format
- JSON – lighter weight version of XML
  - Similar to Python dictionary and lists
  - Key / Value pairs
  - Value can be a scalar, a dictionary, or a list
  - Nest multiple levels deep

JSON Types and Issue

- Flat JSON – easy to convert to CSV and load
- Nested JSON – each nesting would be a separate CSV file and separate database table
  - Advantage – hold multiple tables
  - Disadvantage – extraction is more difficult
- Issue
  - Holes in data, XML did not allow, JSON does!

MS Excel

- Widely used on business side at most companies
- A lot of knowledge we want to tap sitting in Excel on desktops
- User seem to want everything in Excel
- Workbook has a collection of Worksheets (tabs)
- Simple Excel, column & rows, mimics database table
- Complicated Excel needs complicated custom code

Project 2

GitHub => ucb_mids_w205_repo => projects => project_2

- Videos going over project 2 are provided, so we won’t spend class time going over it today
- Breakouts today and next week will be related to project 2

Breakouts

GitHub => ucb_mids_w205_repo => breakouts

(time permitting, we may not get to all of them)

Summary

Instructor will give a brief (about 2 minute) summary of today’s class.