AIG150- Week 2

Working With Real Data

Reading Text:

Chap 04, 06: Pandas for everyone

Chap 06: Python for Data Science for Dummies

Agenda

- Importing and exporting data
  - Streaming, and Sampling Data
  - Accessing Data in Structured Flat-File Form
  - Sending Data in unstructured File Form
  - Managing Data from Relational Databases
  - Accessing Data from the Web
- Data Cleaning
- Data Assembly
  - Tidy data
  - Concatenation
  - Merging multiple datasets

Streaming, and Sampling Data

- Most efficient method to work with data is to load it directly in memory
- Reading the whole file at once and load it in memory using file. read(), not for larger files
- Streaming: Download individual pieces to avoid delays. Read observations one by one
- Sampling: Retrieve selected records

Accessing Data in Structured Flat-File Form

- Text file

- Csv file

- Excel file

Sending Data in Unstructured File Form

- No fixed structure like a csv or excel file where data is organized in rows and columns
- Contains a series of bits
- Need an interpretation algorithm to extract information
- File header contains hints on the type of data
- Examples are images, audio and video files

Managing Data from Databases

- Relation databases such as SQL, PostgresSQL, Oracle and so on
- To read data from a table, we use the read_sql_table() method
- create_engine() method is used to create an engine from a database URI
- NoSQL databases such as mongoDB
- Use pyMongo to work with MongoDB

Accessing Data from the Web

- Data collected from web services and microservices
- XML and JSON

What is Tidy Data ?

- Each row is an observation
- Each column is a variable
- Each type of observational unit forms a table

What is Tidy Data ?

- Each row is an observation
- Each column is a variable
- Each type of observational unit forms a table
- These 3 are interrelated rules, all must be satisfied for a tidy data

Which One Is Tidy ????

Table 1

Country	Year	Cases	Population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898

Table 2

Country	Year	type	count
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	745
Brazil	1999	cases	37737
Brazil	2000	population	174504898

Same data spread across two tables

Country	1999	2000
Afghanistan	745	266
Brazil	37737	80488

Country	1999	2000
Afghanistan	19987071	20595360
Brazil	172006362	174504898

How To Tidy Data Using Python

- **melt()** unpivots a DataFrame from wide to long format
- **Pivot()** reshapes DataFrame organized by given index/ column values
- See the sample code provided with the lecture

Merging Multiple Datasets

Identify:

- What needs to be combined ?
- Do we need to concatenate or join the data ?
- The appropriate function for merging data sets
- Assess if the merge was proper

- See the sample code provided with the lecture