Data Science

Week 2

Getting Start with Data Manipulation

Outline

1. Data Collection
2. Data Loading and File Formats

Data from Public (Open) Dataset

- If you've ever worked on a personal data science project, you've probably spent a lot of time browsing the internet looking for interesting datasets to analyze.
- A dataset, or data set, is simply a collection of data.
- The simplest and most common format for datasets you'll find online is a spreadsheet or CSV format — a single file organized as a table of rows and columns.
- But some datasets will be stored in other formats, and they don't have to be just one file. Sometimes a dataset may be a zip file or folder containing multiple data tables with related data.
- When you use the open dataset, you must follow the requirements of using the dataset (copyright, citation, acknowledge, and so on).

Example of Open Dataset

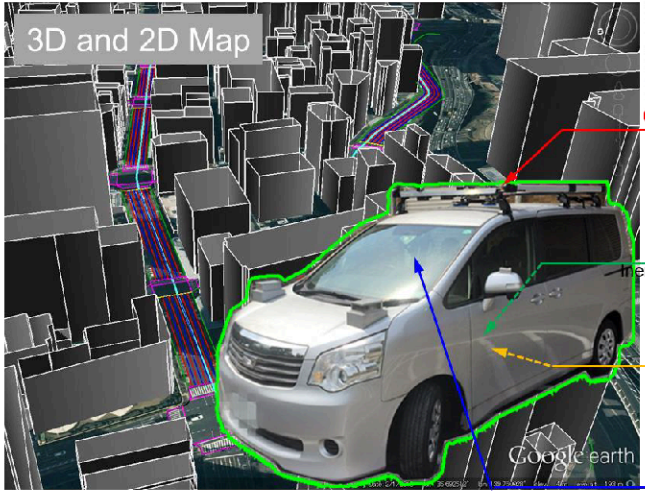
- Data science community:
  - E.g. Kaggle, https://www.kdnuggets.com/datasets/index.html
- Dataset platform
  - E.g. Google, https://cloud.google.com/bigquery/public-data/
- Dataset of Governmental institution
  - E.g. NASA, earth data: https://earthdata.nasa.gov/, data related space: https://pds.nasa.gov/datasearch/data-search/
- Specific website for benchmark
  - COCO dataset: https://cocodataset.org/#home
- And so on…

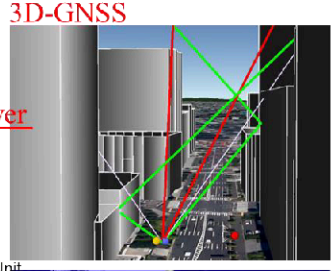
Building a Dataset from Scratch

- In my experience, we could not find the open dataset for most of projects, and we need to collect data by ourselves.
- Data collection is the process of gathering quantitative and qualitative information on specific variables. Good data collection requires a clear process to ensure the data you collect is clean, consistent, and reliable.
- Establishing that process involves taking stock of your objectives (limitations if any), identifying your data requirements, and finally organizing a data collection plan that synthesizes the most important aspects of your project.
- The following 5 examples (3 projects conducted by the instructor of this class) demonstrate how to perform the data collection for the projects with specific objectives.

Example 1: Multi-sensor based vehicle self-localization project

- Objective: using GNSS receiver, IMU (Inertial Measurement Unit) sensor, Speedometer, Camera, Map information for vehicle positioning





Example 1: Multi-sensor based vehicle self-localization project

- Limitation: this project needs the highly accurate map information for vehicle positioning, an accurate map at a specific place of Tokyo has been provided.
- This project has to collect other sensor data when vehicle drives at the place with map information.
- There is NO open dataset can be used for this research.

Example 1: Multi-sensor based vehicle self-localization project

- Data requirements
  - Synchronization of multi-sensor data
    - Use one Computer to control all sensors, and the local time of the computer is used for synchronization
  - Resolution of each data
    - It is necessary to consider the communication bandwidth between computer and each sensor
    - If it is possible, choose the highest resolution for each sensor data in the collection, and adjust the resolution in the later steps of the project
  - Do not forget to collect ground truth data
    - Evaluation needs ground truth (compare the estimated vehicle position and the ground truth position in the evaluation).
    - In this project, the vehicle lane-position (ground truth) is recorded by camera.

Example 2: Smartphone Lifelog project

Smartphone-based Lifelog automatically annotates the users' daily experience (position & activity) from multisensory streams on smartphones.

- Objective: using multisensory streams (GPS receiver, accelerometer, gyroscope, barometer, compass sensor, WiFi receiver) to calculate the position and activity of user.
- After survey, we found that there is NO open dataset can be used for this research.

Example 2: Smartphone Lifelog project

- Data requirements
  - Synchronization of multi-sensor data
    - Smartphone controls all sensors, and the local time of the smartphone is used for synchronization
  - Resolution of each data
    - If it is possible, choose the highest resolution for each sensor data in the collection, and adjust the resolution in the later steps of the project
  - Do not forget to collect ground truth data
    - Evaluation for positioning needs the position ground truth.
    - Evaluation for activity recognition needs to record the activity ground truth.
    - Learning process (activity recognition/classification task) need the activity ground truth.
    - We need to record the position and activity of user in the experiment (e.g. record synchronized user's voice-over which tells the ground truth).

Example 3: Stock performance prediction project

- Objective: Using historical data of stocks to predict the future of stock performance
- After survey, we found that there is NO open dataset (completed data) can be used for this research.
- We found that Thomson Reuters database (not free) can provide the data, and it is possible to use Datastream to access the financial data.
- We prepare a simple program to acquire the data automatically by batch processing.

Example 4: Data collection from internet-Wikipedia

- Wikipedia is one of modern humanity's most impressive creations. In just a few years, anonymous contributors working for free could create the greatest source of online knowledge the world.
- Not only is Wikipedia the best place to get information for writing your college papers, but it's also an extremely rich source of data that can fuel numerous data science projects from natural language processing to supervised machine learning.

Example 4: Data collection from internet-Wikipedia

- Using Wikipedia API and obtain the data from Wikipedia by python.

```
install wikipedia
```

```
import wikipedia

wikipedia.set_lang("en")
page = wikipedia.page('Ritsumeikan', auto_suggest=False)
print (page.content)
```

There are many useful functions wikipedia.*, please check the instruction of wikipedia lib when you use it.

Example 5: Data collection from internet-Website

- “Web scrapers” automatically collect information and data that's usually only accessible by visiting a website in a browser.
- By doing this autonomously, web scraping scripts open up a world of possibilities in data mining, data analysis, statistical analysis, and much more.

The copyright issue is existing when you use Web scrapers, please do NOT violate the **copyright laws** when you use Web scrapers

Example 5: Data collection from internet-Website

Collect data from website using python

install beautifulSoup4

```
from urllib import request

from bs4 import BeautifulSoup

# link of Yanlei Gu's Google Scholar
response = request.urlopen('https://scholar.google.co.jp/citations?user=2vLFO00AAAAJ&hl=en')
contents = BeautifulSoup(response,"html.parser")
print (contents)
```

Outline

1. Data Collection
2. Data Loading and File Formats

Getting Started with pandas

- Pandas adopts many coding idioms from NumPy, the biggest difference is that
  - Pandas is designed for working with tabular or heterogeneous (different types of) data.
  - NumPy, by contrast, is best suited for working with homogeneous numerical array data.

```
import pandas as pd
```

Data Loading function in pandas

Function	Description
`read_csv`	Load delimited data from a file, URL, or file-like object; use comma as default delimiter
`read_table`	Load delimited data from a file, URL, or file-like object; use tab (`'\t'`) as default delimiter
`read_fwf`	Read data in fixed-width column format (i.e., no delimiters)
`read_clipboard`	Version of `read_table` that reads data from the clipboard; useful for converting tables from web pages
`read_excel`	Read tabular data from an Excel XLS or XLSX file
`read_hdf`	Read HDF5 files written by pandas
`read_html`	Read all tables found in the given HTML document
`read_json`	Read data from a JSON (JavaScript Object Notation) string representation
`read_msgpack`	Read pandas data encoded using the MessagePack binary format
`read_pickle`	Read an arbitrary object stored in Python pickle format
`read_sas`	Read a SAS dataset stored in one of the SAS system's custom storage formats
`read_sql`	Read the results of a SQL query (using SQLAlchemy) as a pandas DataFrame
`read_stata`	Read a dataset from Stata file format
`read_feather`	Read the Feather binary file format

Comma-separated (CSV) text file

1) Use read_csv to read it into a DataFrame

2) A file will not always have a header row, we should use header=None to avoid losing the first row in this case

3) Give the column name (header) manually

4) Use one column to be the index of the returned DataFrame

Files used for programming

ex1.csv

	A	B	C	D	E
1	a	b	c	d	message
2	1	2	3	4	hello
3	5	6	7	8	world
4	9	10	11	12	foo

ex2.csv

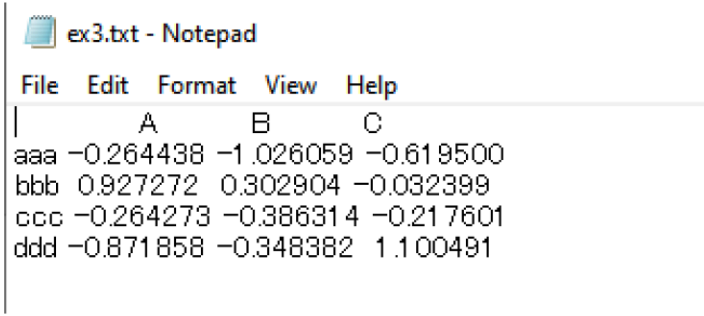
	A	B	C	D	E
1	1	2	3	4	hello
2	5	6	7	8	world
3	9	10	11	12	foo

Check the code by yourself

```
# (1) read standard CSV
# Code Example 1
print ("--------------------")
df = pd.read_csv('ch06/ex1.csv')
print (df)
print ("--------------------")
# (2 )A file will not always have a header row. Consider this file:
df2 = pd.read_csv('ch06/ex2.csv')
print (df2)
print ("--------------------")
# should use header=None to avoid lossing the first row
df3 = pd.read_csv('ch06/ex2.csv', header=None)
print (df3)
print ("--------------------")
# (3) give the column name manually
df4 = pd.read_csv('ch06/ex2.csv', names=['a', 'b', 'c', 'd', 'message'])
print (df4)
print ("--------------------")
# (4) use message column to be the index of the returned DataFrame
names = ['a', 'b', 'c', 'd', 'message']
df5 = pd.read_csv('ch06/ex2.csv', names=names, index_col='message')
print (df5)
print ("--------------------")
```

Whitespace-separated text file

- If the data are separated by the whitespace, we can pass a regular expression as a delimiter for `read_csv`. This can be expressed by the regular expression `\s+`:



```
# read other type Text
# Code Example 2
#In some cases, a table might not have a fixed delimiter, using
whitespace or some
#other pattern to separate fields. Consider a text file that looks
like this:
list(open('ch06/ex3.txt'))
result = pd.read_csv('ch06/ex3.txt')
print ("--------------------")
print (result)
print ("--------------------")
print (result.iloc[0,[0]])

result1 = pd.read_csv('ch06/ex3.txt', sep='\s+')
print ("--------------------")
print (result1)
print ("--------------------")
print (result1.iloc[0,[0]])
print ("--------------------")
print (result1.loc['aaa',['A']])
print ("--------------------")
```

File with missing data

ex5.csv

	A	B	C	D	E	F
1	somethin	a	b	c	d	message
2	one	1	2	3	4	all
3	two	5	6		8	world
4	three	9	10	11	12	foo

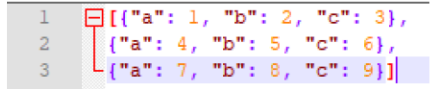
```
# read a file with missing data
# Code Example 3
result = pd.read_csv('ch06/ex5.csv')
print ("--------------------")
print (result)
print ("--------------------")
```

JSON Data

- Short for JavaScript Object Notation
- JSON has become one of the standard formats for sending data by HTTP request between web browsers and other applications.
- It is a much more free-form data format than a tabular text form like CSV.

Load JSON Data

JSON_example.json



- The pandas.read_json can automatically convert JSON datasets in specific arrangements into a Series or DataFrame.
- The default options for pandas.read_json assume that each object in the JSON array is a row in the table

Check the code by yourself

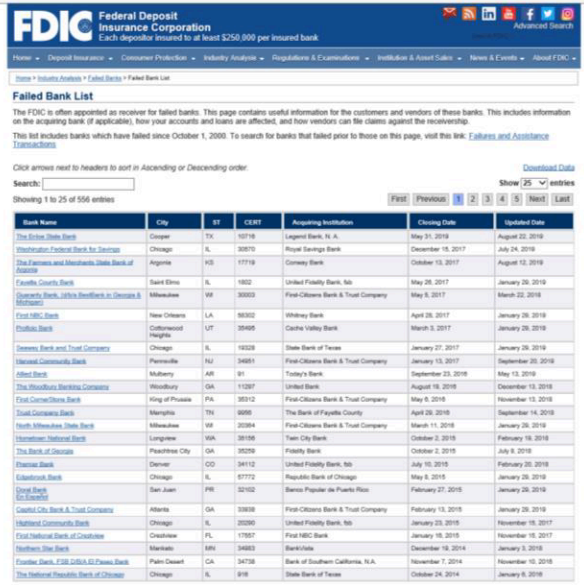
```
# read a json file
data = pd.read_json('ch06/JSON_example.json')
print ("--------------------")
print (data)
print ("--------------------")
print (data.to_json())
print ("--------------------")open
```

XML and HTML files

Extensible Markup Language and
Hypertext Markup Language

- Python has many libraries for reading and writing data in HTML and XML formats.
- Examples include lxml, Beautiful Soup, and html5lib.
- pandas has a built-in function, read_html, which uses libraries like lxml and Beautiful Soup to automatically parse tables out of HTML files as DataFrame objects.
- First, you must install some additional libraries used by read_html

Downloaded an HTML file from the United States FDIC government agency showing bank failures.



FDIC Federal Deposit Insurance Corporation
Each depositor insured to at least $250,000 per insured bank

Advanced Search

Home | Deposit Insurance | Consumer Protection | Industry Analysis | Regulations & Examinations | Institution & Asset Sales | News & Events | About FDIC

Home > Industry Analysis > Failed Banks > Failed Bank List

Failed Bank List

The FDIC is often appointed as receiver for failed banks. This page contains useful information for the customers and vendors of these banks. This includes information on the acquiring bank (if applicable), how your accounts and loans are affected, and how vendors can file claims against the receivership.

This list includes banks which have failed since October 1, 2000. To search for banks that failed prior to those on this page, visit this link: Failures and Assistance Transactions

Click arrows next to headers to sort in Ascending or Descending order.

Download Data

Search:

Show 25 entries

Showing 1 to 25 of 556 entries

First Previous 1 2 3 4 5 Next Last

Bank Name	City	ST	CERT	Acquiring Institution	Closing Date	Updated Date
The Enloe State Bank	Cooper	TX	10716	Legend Bank, N. A.	May 31, 2019	August 22, 2019
Washington Federal Bank for Savings	Chicago	IL	30870	Royal Savings Bank	December 15, 2017	July 24, 2019
The Farmers and Merchants State Bank of Argonia	Argonia	KS	17719	Conway Bank	October 13, 2017	August 12, 2019
Fayette County Bank	Saint Elmo	IL	1802	United Fidelity Bank, fsb	May 26, 2017	January 29, 2019
Guaranty Bank, (d/b/a BestBank in Georgia & Michigan)	Milwaukee	WI	30003	First-Citizens Bank & Trust Company	May 5, 2017	March 22, 2018
First NBC Bank	New Orleans	LA	58302	Whitney Bank	April 28, 2017	January 29, 2019
Proficio Bank	Cottonwood Heights	UT	35495	Cache Valley Bank	March 3, 2017	January 29, 2019
Seaway Bank and Trust Company	Chicago	IL	19328	State Bank of Texas	January 27, 2017	January 29, 2019
Harvest Community Bank	Pennsville	NJ	34951	First-Citizens Bank & Trust Company	January 13, 2017	September 20, 2018
Allied Bank	Mulberry	AR	91	Today's Bank	September 23, 2016	May 13, 2019
The Woodbury Banking Company	Woodbury	GA	11297	United Bank	August 19, 2016	December 13, 2018
First Cornerstone Bank	King of Prussia	PA	35312	First-Citizens Bank & Trust Company	May 6, 2016	November 13, 2018
Trust Company Bank	Memphis	TN	9956	The Bank of Fayette County	April 29, 2016	September 14, 2018
North Milwaukee State Bank	Milwaukee	WI	20364	First-Citizens Bank & Trust Company	March 11, 2016	January 29, 2019
Hometown National Bank	Longview	WA	35156	Twin City Bank	October 2, 2015	February 19, 2018
The Bank of Georgia	Peachtree City	GA	35259	Fidelity Bank	October 2, 2015	July 9, 2018
Premier Bank	Denver	CO	34112	United Fidelity Bank, fsb	July 10, 2015	February 20, 2018
Edgebrook Bank	Chicago	IL	57772	Republic Bank of Chicago	May 8, 2015	January 29, 2019
Doral Bank En Español	San Juan	PR	32102	Banco Popular de Puerto Rico	February 27, 2015	January 29, 2019
Capitol City Bank & Trust Company	Atlanta	GA	33938	First-Citizens Bank & Trust Company	February 13, 2015	January 29, 2019
Highland Community Bank	Chicago	IL	20290	United Fidelity Bank, fsb	January 23, 2015	November 16, 2017
First National Bank of Crestview	Crestview	FL	17557	First NBC Bank	January 16, 2015	November 16, 2017
Northern Star Bank	Mankato	MN	34983	BankVista	December 19, 2014	January 3, 2018
Frontier Bank, FSB D/B/A El Paseo Bank	Palm Desert	CA	34738	Bank of Southern California, N.A.	November 7, 2014	November 10, 2015
The National Republic Bank of Chicago	Chicago	IL	916	State Bank of Texas	October 24, 2014	January 8, 2016

Load <table> data of HTML

- The pandas.read_html function has a number of options, but by default it searches for and attempts to parse all tabular data contained within <table> tags.
- The result is a list of DataFrame objects.

```
# read a HTML file
tables = pd.read_html('ch06/FDICFailedBankList.html')
print (tables)
```

Check the code by yourself

HDF5 Format

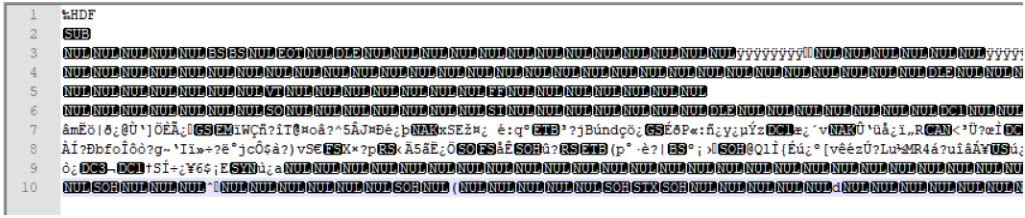
- The “HDF” in HDF5 stands for hierarchical data format.
- HDF5 is a well-regarded file format intended for storing large quantities of scientific array data.
- Each HDF5 file can store multiple datasets and supporting metadata.
- Compared with simpler formats, HDF5 supports compression modes, enabling data with repeated patterns to be stored more efficiently.
- HDF5 can be a good choice for working with very large datasets that don’t fit into memory, as you can efficiently read and write small sections of much larger arrays.

Load HDF5 File

Check the code by yourself

```
# read a HDF file
frame = pd.DataFrame({'a': np.random.randn(100)})
frame.to_hdf('ch06/mydata.h5', key='frame', mode='w')
frame2 = pd.read_hdf('ch06/mydata.h5')
```

mydata.h5



Microsoft Excel Files

- pandas also supports reading tabular data stored in Excel 2003 (and higher) files using pandas.read_excel function.
- Internally these tools use the add-on packages xlrd and openpyxl to read XLS and XLSX files, respectively.
- You may need to install these manually with pip or conda.

Interacting with Web APIs

- Many websites have public APIs providing data feeds via JSON or some other format.
- There are a number of ways to access these APIs from Python.
- One easy-to-use method is the requests package.

Example for Web APIs

- To find the last 30 GitHub issues for pandas on GitHub, we can make a GET HTTP request using the add-on requests library:

Check the code by yourself

```
# find the last 30 GitHub issues for pandas on GitHub
url = 'https://api.github.com/repos/pandas-dev/pandas/issues'
resp = requests.get(url)
print (resp)
data = resp.json()

print (data[0]['title'])
issues = pd.DataFrame(data, columns=['number', 'title'])
print (issues)
```

Example for Web APIs

- The Response object's json method will return a dictionary containing JSON parsed into native Python objects.

- Each element in data is a dictionary containing all of the data found on a GitHub issue page (except for the comments). We can pass data directly to DataFrame and extract fields of interest

Image file and image processing lib

- Image Processing is a field of knowledge that falls in Computer Vision.
- The premises of machine learning were first laid down by computer vision theory, applying a whole set of techniques to process and analyze imagery data to extract valuable information that computers and machines may use for a wide range of applications.
- Some of the commonly used image processing techniques leveraging a very popular Computer Vision library, OpenCV. We can use the functions of OpenCV lib load image files.

Reference

- Devin Pickell, Structured vs Unstructured Data – What's the Difference?, https://learn.g2.com/structured-vs-unstructured-data.
- 21 Places to Find Free Datasets for Data Science Projects, https://www.dataquest.io/blog/free-datasets-for-projects/
- Will Koehrsen, Wikipedia Data Science: Working with the World's Largest Encyclopedia, https://towardsdatascience.com/wikipedia-data-science-working-with-the-worlds-largest-encyclopedia-c08efbac5f5c
- McKinney, Wes. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.", 2012.
- Provost, Foster, and Tom Fawcett. Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc., 2013.

Data Science

Week 3

Statistics for Data Science