

Resources / Labs (/COMP9321/22T1/resources/72107) / Week 2 (/COMP9321/22T1/resources/72108)
/ Data Access

Data Access

Prerequisites:

Data scientist are required to access various kinds of data from various sources. In this Lab, you will practice how to read textual material from most common sources: CSV (https://en.wikipedia.org/wiki/Comma-separated_values) , Relational Databases (https://en.wikipedia.org/wiki/Relational_database) , NoSQL databases (<https://en.wikipedia.org/wiki/NoSQL>) , and RESTful APIs (https://en.wikipedia.org/wiki/Representational_state_transfer) . (There are numerous Relational and non-relational databases; however, this lab will specifically focus on SQLite (<https://en.wikipedia.org/wiki/SQLite>) and MongoDB (<https://en.wikipedia.org/wiki/MongoDB>) . It is assumed that you will take a look at the following packages in python before heading to activities:

- pandas (<https://pandas.pydata.org/>)
- sqlite3 (https://sebastianraschka.com/Articles/2014_sqlite_in_python_tutorial.html)
- pymongo (<https://docs.mongodb.com/drivers/pymongo/>)
- requests (<http://docs.python-requests.org/en/master/>)
- json (<https://realpython.com/python-json/>)

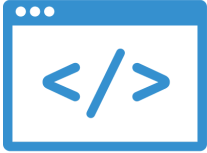
All of the following activities will be based on the *Demographic Statistics By Zip Code dataset* (Download the csv file (https://raw.githubusercontent.com/mysilver/COMP9321-Data-Services/master/Week2_DataAccess/Demographic_Statistics_By_Zip_Code.csv)). This dataset includes 236 zip code along with their demographic statistics for New York City. For consistency, in all of the following activities, you will load the data from a specific source (and format), and convert it into panda dataframe.

Activity-1: CSV Files

Description : You will learn how to into CSV files pandas' dataframe and how to create a CSV file out of a dataframe

Steps :

1. Download the dataset (https://raw.githubusercontent.com/mysilver/COMP9321-Data-Services/master/Week2_DataAccess/Demographic_Statistics_By_Zip_Code.csv)
2. Read the CSV file and put the samples into a pandas' dataframe.
(<https://towardsdatascience.com/how-to-read-csv-file-using-pandas-ab1f5e7e7b58>)
3. Programmatically print the columns of the dataframe (<https://code-examples.net/en/q/129495a>)
4. Programmatically print the rows of the dataframe (<https://stackoverflow.com/questions/16476924/how-to-iterate-over-rows-in-a-dataframe-in-pandas>)
5. Save the dataframe as a CSV file
(https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.to_csv.html)



(<https://github.com/mysilver/COMP9321-Data->

[Services/blob/master/Week2_DataAccess/activity_1.py](#)) Sample Code

Activity-2: Relational DBs (SQLite)

Description : You will learn how to store a dataframe in SQLite and how query the database and create a dataframe

Steps :

1. Load the CSV file using the method you created in the previous activity
2. Store the dataframe in sqlite database (https://pandas.pydata.org/pandas-docs/version/0.22/generated/pandas.DataFrame.to_sql.html)
3. Query the database and load the data into a new dataframe again (<https://pythonspot.com/sqlite-database-with-pandas/>)



(<https://github.com/mysilver/COMP9321-Data->

[Services/blob/master/Week2_DataAccess/activity_2.py](#)) Sample Code

Activity-3: NoSQL DBs (MongoDB)

Description : You will learn how to store a dataframe in MongoDB and how query the database and create a dataframe

Steps :

1. Load the CSV file using the method you created in the first activity
2. Install mongodb (<https://docs.mongodb.com/manual/installation/>) on your local machine (install pymongo version 3.5.1, or you need to edit the sample code)
3. Create a database named **comp9321** with a collection named **Demographic_Statistics**
4. Write the dataframe in mongodb (<https://stackoverflow.com/questions/20167194/insert-a-pandas-dataframe-into-mongodb-using-pymongo>)
5. Query the database and load the data into a new dataframe again (<https://stackoverflow.com/questions/16249736/how-to-import-data-from-mongodb-to-pandas>)



(<https://github.com/mysilver/COMP9321-Data->

[Services/blob/master/Week2_DataAccess/activity_3.py](#)) Sample Code

Activity-4: RESTful

Description : You will learn how to create a RESTful request, and process JSON objects into dataframes

Steps :

1. Fortunately, the dataset is also available as a simple RESTful API:
GET <https://data.cityofnewyork.us/api/views/kku6-nxdu/rows.json>
(<https://data.cityofnewyork.us/api/views/kku6-nxdu/rows.json>)
2. Using the above url, create a get request and fetch the JSON programmatically
(<https://stackoverflow.com/questions/6386308/http-requests-and-json-parsing-in-python>) .
3. Create a dataframe out of the JSON object created in the previous step
(https://chrisalbon.com/python/data_wrangling/load_json_file_into_pandas/)




(<https://github.com/mysilver/COMP9321-Data->

[Services/blob/master/Week2_DataAccess/activity_4.py](#)) Sample Code

Resource created [3 months ago \(Monday 07 February 2022, 01:46:42 AM\)](#), last modified [2 months ago \(Monday 21 February 2022, 12:35:07 PM\)](#).

Comments

 /COMP9321/22T1/forums/search?forum_choice=resource/72116

 </COMP9321/22T1/forums/resource/72116>

 Add a comment



Arjun Sharma (</users/z5228961>) [2 months ago \(Wed Mar 02 2022 00:07:41 GMT+0800 \(China Standard Time\)\)](#)

Hi, I'm a bit late in asking this but when I run activity 3, this is the output from my first line:

```
621e19ea278d12a23c531caa,10001.0,44.0,22.0,0.5,22.0,0.5,0.0,0.0,44.0,100.0,0.0,0.0,16.0
,0.36,0.0,0.0,3.0,0.07,1.0,0.02,21.0,0.48,3.0,0.07,0.0,0.0,44.0,100.0,2.0,0.05,42.0,0.95,0.0,0.0,0.0,0.0,44.0,100.0,20.0,0.45,24.0,0.55,0.0,0.0,44.0,100.0
```

is the first part normal at all? if not, how can it be fixed?

Reply



Gordon Chen (</users/z5161163>) [2 months ago \(Fri Mar 04 2022 10:42:04 GMT+0800 \(China Standard Time\)\)](#)

What was the command to produce that output?

Reply



Arjun Sharma (</users/z5228961>) [2 months ago \(Sun Mar 06 2022 23:54:30 GMT+0800 \(China Standard Time\)\)](#)

No command, I just used the sample code in activity 3. Only change I made was using `c.insert_many(records)` due to my pymongo version.

Reply

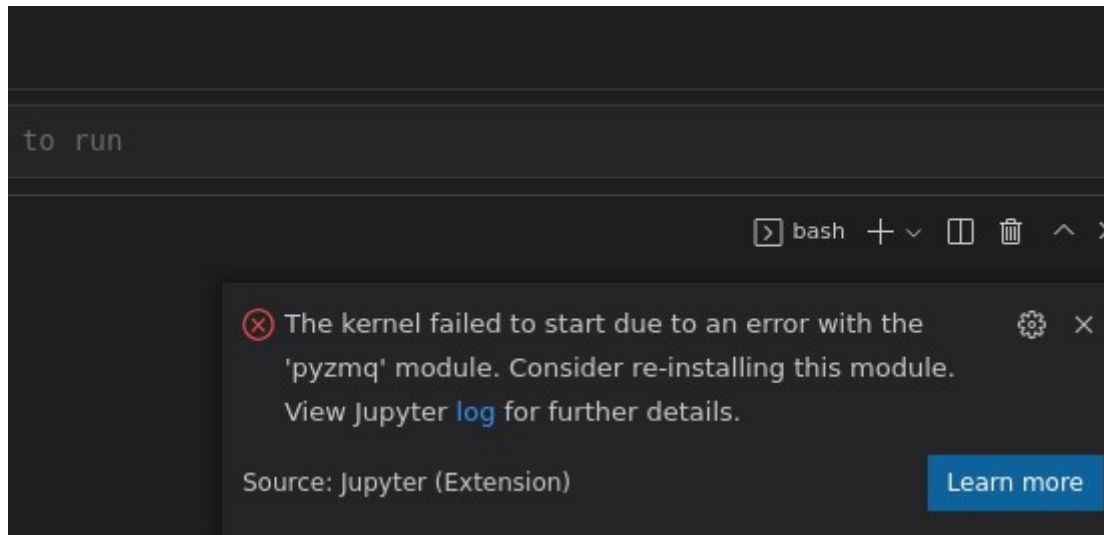


Maher Mesto (/users/z5272300) 2 months ago (Sat Feb 26 2022 22:41:13 GMT+0800 (China Standard Time))

Activity3 Error.

Hi,

When I tried to run activity3 code on cse machine I got the below error:



can you try to solve this issue please?

Thx

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) 2 months ago (Sun Feb 27 2022 11:15:45 GMT+0800 (China Standard Time))

Hi,

MongoDB is not installed in CSE machine. Please use your personal laptop and install the latest version of the libraries unless it is stated in the spec

Reply



Wisesa Resosudarmo (/users/z5256103) 2 months ago (Sat Feb 26 2022 17:24:55 GMT+0800 (China Standard Time))

Relating to question 4. The json format that the REST response spits out seems pretty convoluted. It was very tedious to get through. I doubt that most others URLs would output their responses like that exactly.

In the real world, how would we go about automating processing data from REST requests. Or will we have to manually go through it one by one or just hope that the website host has a more convenient API?

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) 2 months ago (Sat Feb 26 2022 19:11:25 GMT+0800 (China Standard Time))

Usually there are documentation about the apis, and you can get to know the attributes by reading the documentation. However, not all services provide good documentation , and manual inspections are often required

Reply



Richard Liu (/users/z5165455) 2 months ago (Thu Feb 24 2022 12:58:57 GMT+0800 (China Standard Time)), last modified 2 months ago (Thu Feb 24 2022 12:59:11 GMT+0800 (China Standard Time))

For activity 1 & 2, I was just wondering, why do you not use 'JURISDICTION NAME' as an index column (as it looks like it could be a primary key) or turn off indexing when writing the csv file, since I feel like you wouldn't want to store extra unnecessary data - being the index we introduce.

Unless the whole point of adding an additional index is because we don't want to unreasonable determine a primary key/index and so we introduce our own one.

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) 2 months ago (Thu Feb 24 2022 13:09:22 GMT+0800 (China Standard Time))

Hi Richard,

Yes, it makes sense; the sample code is just a simple solution, of course, there are better ones.

Having a good index always makes sense

Reply



Xiu Cheng Zhang (/users/z3188003) 2 months ago (Thu Feb 24 2022 07:16:24 GMT+0800 (China Standard Time))

Using the provided sample code for activity 3 I get this error when trying to write into mongodb.

```
ServerSelectionTimeoutError: localhost:27017: [Errno 61] Connection refused
```

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) 2 months ago (Thu Feb 24 2022 09:04:20 GMT+0800 (China Standard Time))

Make sure, you are running MongoDB in your local machine and check the port; the port in the sample code is the default port number for MongoDB if you have changed it when you were installing it, you should also change it in the code

Reply



Zheng Fu (/users/z5285691) 2 months ago (Wed Feb 23 2022 17:29:16 GMT+0800 (China Standard Time))

mongodb is still not supported on M1 Mac, Anybody having the solutions?

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) 2 months ago (Wed Feb 23 2022 18:11:36 GMT+0800 (China Standard Time))

You can use mongodb as a service:
<https://mlab.com/> (<https://mlab.com/>)

But this will require you modify the sample code a bit

Reply



Jonathan Wong (/users/z3375329) 2 months ago (Tue Feb 22 2022 12:22:33 GMT+0800 (China Standard Time))

Hi, anyone know how to get a .db file viewer to work on the CSE computers?

Thanks!

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) 2 months ago (Tue Feb 22 2022 12:32:15 GMT+0800 (China Standard Time))

This might be helpful but not in CSE machines.
<https://chrome.google.com/webstore/detail/sqlite-v...>
(<https://chrome.google.com/webstore/detail/sqlite-viewer/golagekponhmgfoofmlepfbdmhpajia?hl=en>)

Reply



Gordon Chen (/users/z5161163) 2 months ago (Mon Feb 21 2022 23:02:46 GMT+0800 (China Standard Time))

For those who needs help on MongoDB -
<https://www.prisma.io/dataguide/mongodb/setting-up...>
(<https://www.prisma.io/dataguide/mongodb/setting-up-a-local-mongodb-database>)
(<https://www.prisma.io/dataguide/mongodb/setting-up-a-local-mongodb-database#setting-up-mongodb-on-windows>)

Reply



Morty Al-Banna (/users/z3445371) 2 months ago (Tue Feb 22 2022 12:40:29 GMT+0800 (China Standard Time))

thanks for Sharing

Reply



Solomon Rachamim (/users/z5375417) 2 months ago (Mon Feb 21 2022 13:20:55 GMT+0800 (China Standard Time))

Purpose of lab consultations:

Hi guys,

super excited about the course, looking really practical <3

I just wanted to check the purpose of the lab consultation, if I can complete all the tasks you have requested on my own, do I need to attend the lab consults, or rather they are there if you need help with something?

Wishing you a great week ahead!

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) [2 months ago \(Tue Feb 22 2022 09:49:56 GMT+0800 \(China Standard Time\)\)](#)

Hi,
Labs are for consultation. If you need help you can make use of consultation labs and tutors will be more than happy to help

Reply



Sung Go (/users/z5310199) [2 months ago \(Mon Feb 21 2022 09:11:44 GMT+0800 \(China Standard Time\)\)](#), last modified [2 months ago \(Mon Feb 21 2022 09:27:51 GMT+0800 \(China Standard Time\)\)](#)

Hello ~~

How exactly are we supposed to do the quiz and submit it? Do we share our github repo or submit on webcms or something else? Also the sample codes seems like the answers -> can't everyone just copy it? Do you know where the detailed lab instructions are including how to submit?

Edit: just realised this is the lab -> so are labs not assessed and quizzes are assessed? if so, when will the quizzes be released?

Reply



Morty Al-Banna (/users/z3445371) [2 months ago \(Mon Feb 21 2022 11:34:09 GMT+0800 \(China Standard Time\)\)](#)

Hi Sung,

once we release the quiz on WebCMS will will announce that and you will see the activity. the first quiz will be released after the next lecture (just making sure we cover more material) :)

Reply



Sung Go (/users/z5310199) [2 months ago \(Mon Feb 21 2022 11:43:22 GMT+0800 \(China Standard Time\)\)](#)

Ahh right makes much more sense. Thanks for the info!

Reply

Load More Comments

