hw3 Hbase with Thrift and Python

r06921105 葉津源

1. build docker image:

after creating a vm and installing docker ce (same procedure as hw2), run the following commands to setup the correct files for creating a docker image

```
git clone https://github.com/sdwangntu/hadoop-cluster.git
cd hadoop-cluster

git checkout 5ef6cd4fe3fadf4a990eba37318c61df9796a306
wget https://archive.apache.org/dist/hadoop/core/hadoop-3.1.2/
hadoop-3.1.2.tar.gz
wget https://archive.apache.org/dist/hbase/1.4.9/hbase-1.4.9-bin.tar.gz
wget https://github.com/cloudera/hue/archive/release-4.3.0.tar.gz
mv release-4.3.0.tar.gz hue-4.3.0.tar.gz
```

Then, modify Dockerfile and start-hadoop.sh according to TA
(add EXPOSE 9090 at line 99 of Dockerfile and add
$HBASE_HOME/bin/hbase-daemon.sh start thrift at line 28 of start-hadoop.sh)

build docker image by
docker build -t hbasecluster .

and wait for image to be built

2. set up docker container

docker run --hostname=hadoop-master -p 8088:8088 -p 9870:9870 -p 9864:9864 -p 19888:19888 -p 8042:8042 -p 8888:8888 -p 9090:9090 --name hadoop-master -d hbasecluster

and

docker run --hostname=hadoop-worker --name hadoop-worker -d hbasecluster

now we can connect from outside the docker with python thrift

3. Data files
we use air quality datafiles downloaded from https://taqm.epa.gov.tw/taqm/tw/YearlyDataDownload.aspx

I couldn't get libreoffice to work so I opened the files with excel, change chinese characters to english names, and saved the file as csv.

4. write to Hbase

the complete code is in src/run.py
the main part of the code is to read from csv and write to hbase row by row with client.MutateRow.

```
for i in range(3,26):
    qualifier = i-2
    value = frame[i]
    row = frame[0]      # date
    column = frame[2]   # attr
    mutate = Mutation(column=column+':'+str(qualifier),value=value)
    client.mutateRow(table_Name, frame[0], [mutate])
    InsertCounts += 1
```

After writing to hbase, it will create a table for each csv file listing air quality data from one station. It will save with dates as row keys and attributes as column keys.

5. read from Hbase and plot graph

the complete code is in src/plot.py
I choose to plot PM2.5
The main part of this script is to read from hbase with client.getRow(table, row)
row will be dates ranging from '2018/1/1' to '2018/12/31' in string format
table can be acquired by client.getTableNames().
I calculate the average of PM2.5 hourly values within day as one datapoint to be plotted.

not that try, except is needed because there are NA values in the data.

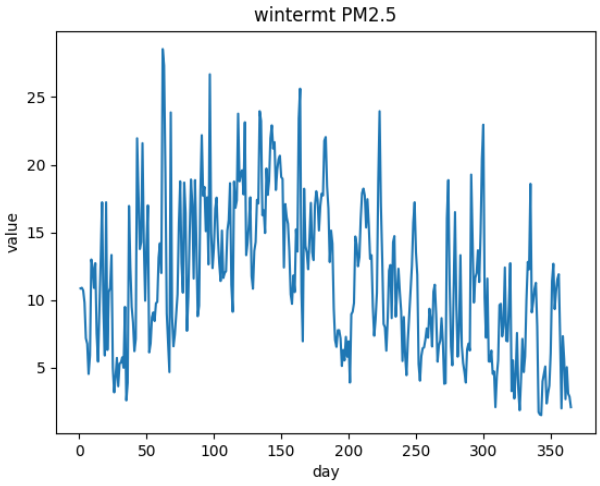
```
attr = ['PM2.5:%d'%i for i in range(1,23)]
while iterday <= lastday:
    row = iterday.strftime('%Y/%m/%d')
    DATA = client.getRow(table, row)
    columns = DATA[0].columns
    vals =[]
    for at in attr:
        try:
            val = columns[at].value
            vals.append(float(val))
        except:
            pass

    mean = np.mean(vals)
    X.append(i)
    y.append(mean)
    i+=1
    iterday = iterday+oneday
```

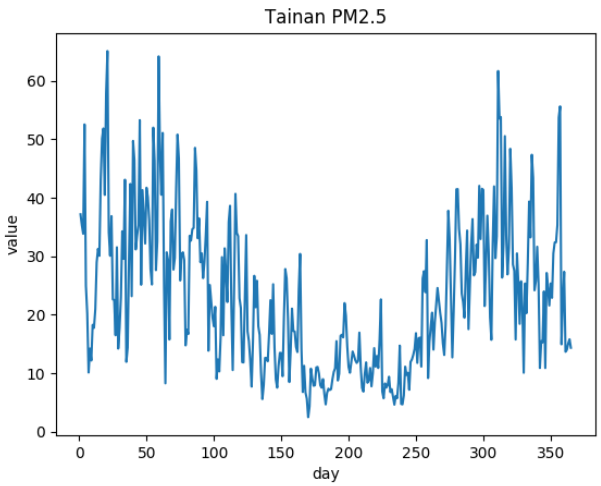
6. Results:

the resulting plots are shown below:

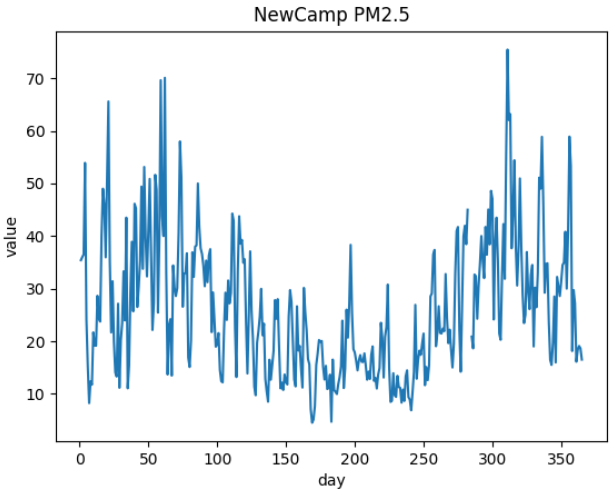
冬山測站



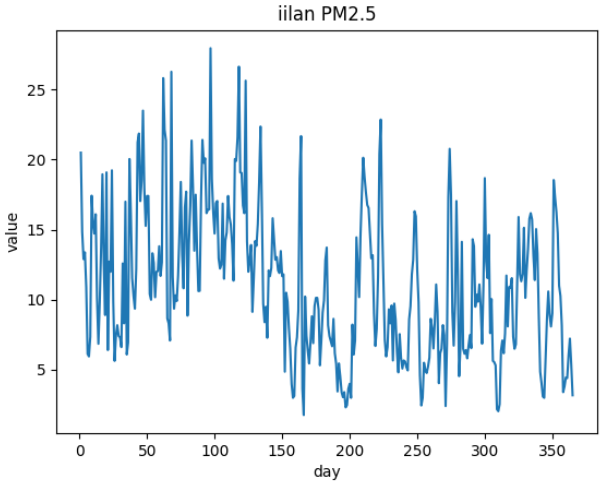
台南測站



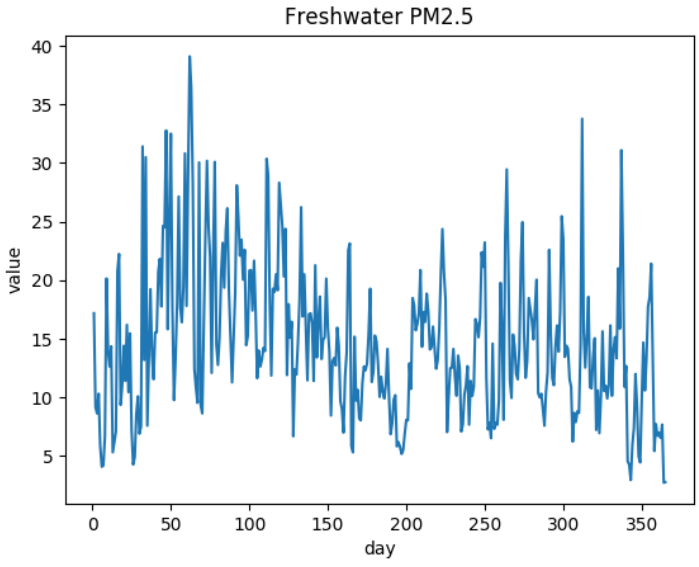
新營測站

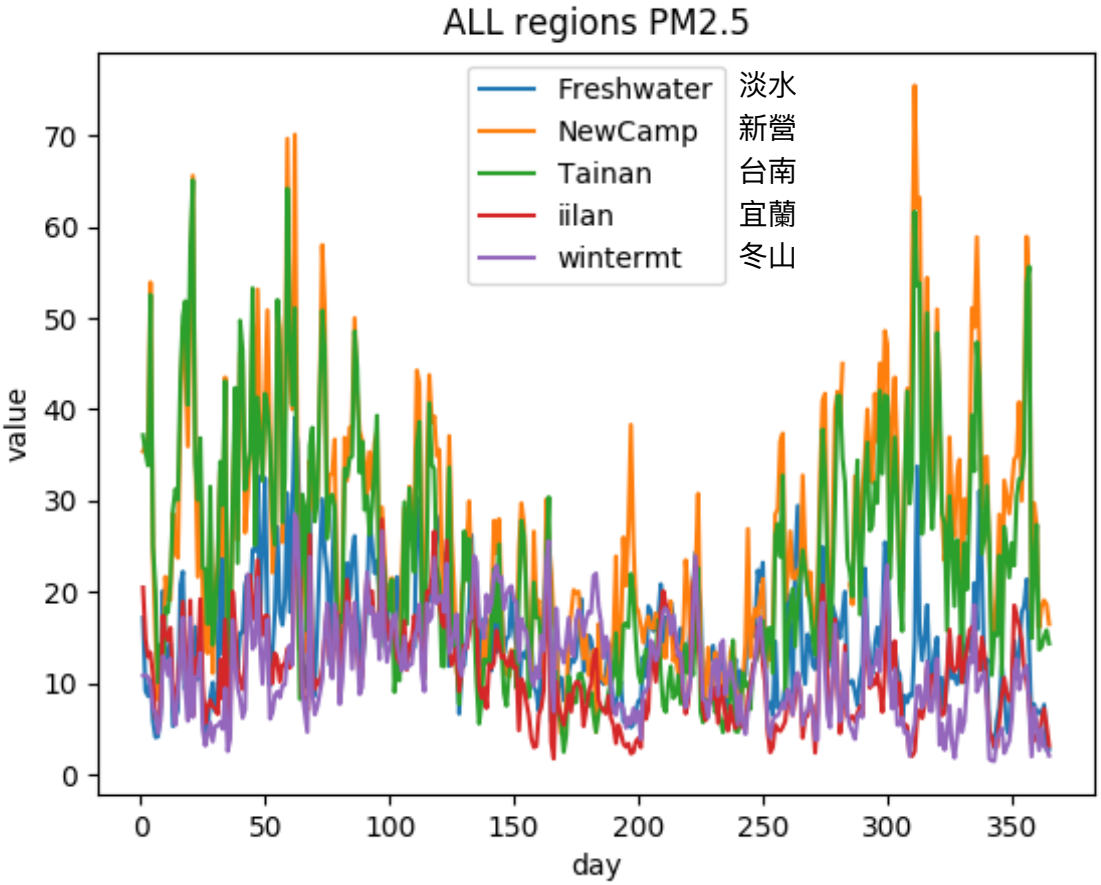


宜蘭測站



淡水測站





from the results we see that PM2.5 is higher at wintertimes

It is also higher at the southern part of Taiwan (台南新營）than at North or NorthEast part of Taiwan （淡水宜蘭冬山）