

kvdb

March 28, 2021

1 Assignment 2

```
[1]: import pandas as pd
import s3fs

s3 = s3fs.S3FileSystem(
    anon=True,
    client_kwargs={
        'endpoint_url': 'https://storage.budsc.midwest-datascience.com'
    }
)

df1 = pd.read_csv(
    s3.open('data/external/tidynomicon/site.csv', mode='rb')
)

df1.head()
```

```
[1]:  site_id  latitude  longitude
0    DR-1    -49.85    -128.57
1    DR-3    -47.15    -126.72
2    MSK-4    -48.87    -123.40
```

```
[2]: df2 = pd.read_csv(
    s3.open('data/external/tidynomicon/measurements.csv', mode='rb')
)
df2.head()
```

```
[2]:  visit_id  person_id  quantity  reading
0        619        dyer        rad      9.82
1        619        dyer        sal      0.13
2        622        dyer        rad      7.80
3        622        dyer        sal      0.09
4        734         pb        rad      8.41
```

```
[3]: df3 = pd.read_csv(
    s3.open('data/external/tidynomicon/person.csv', mode='rb')
```

```
)  
df3.head()
```

```
[3]:  person_id personal_name family_name  
0      dyer      William      Dyer  
1        pb        Frank    Pabodie  
2      lake      Anderson      Lake  
3       roe    Valentina    Roerich  
4  danforth        Frank  Danforth
```

```
[4]: df4 = pd.read_csv(  
      s3.open('data/external/tidynomicon/visited.csv', mode='rb')  
      )  
df4.head()
```

```
[4]:  visit_id site_id  visit_date  
0        619    DR-1  1927-02-08  
1        622    DR-1  1927-02-10  
2        734    DR-3  1930-01-07  
3        735    DR-3  1930-01-12  
4        751    DR-3  1930-02-26
```

1.1 Assignment 2.1

```
[5]: import json  
from pathlib import Path  
import os  
  
import pandas as pd  
import s3fs  
  
def read_cluster_csv(file_path, endpoint_url='https://storage.budsc.  
↳midwest-datascience.com'):  
    s3 = s3fs.S3FileSystem(  
        anon=True,  
        client_kwargs={  
            'endpoint_url': endpoint_url  
        }  
    )  
    return pd.read_csv(s3.open(file_path, mode='rb'))  
  
current_dir = Path(os.getcwd()).absolute()  
results_dir = current_dir.joinpath('results')  
kv_data_dir = results_dir.joinpath('kvdb')  
kv_data_dir.mkdir(parents=True, exist_ok=True)
```

```

people_json = kv_data_dir.joinpath('people.json')
visited_json = kv_data_dir.joinpath('visited.json')
sites_json = kv_data_dir.joinpath('sites.json')
measurements_json = kv_data_dir.joinpath('measurements.json')

```

```

[6]: class KVDB(object):
    def __init__(self, db_path):
        self._db_path = Path(db_path)
        self._db = {}
        self._load_db()

    def _load_db(self):
        if self._db_path.exists():
            with open(self._db_path) as f:
                self._db = json.load(f)

    def get_value(self, key):
        return self._db.get(key)

    def set_value(self, key, value):
        self._db[key] = value

    def save(self):
        with open(self._db_path, 'w') as f:
            json.dump(self._db, f, indent=2)

```

```

[7]: def create_sites_kvdb():
    db = KVDB(sites_json)
    #df1 = read_cluster_csv('data/external/tidynomicon/site.csv')
    for site_id, group_df in df1.groupby('site_id'):
        db.set_value(site_id, group_df.to_dict(orient='records')[0])
    db.save()

    def create_measurements_kvdb():
        db = KVDB(measurements_json)
        # df2 = read_cluster_csv('data/external/tidynomicon/measurements.csv')
        for key, group_df in df2.groupby(['visit_id', 'person_id', 'quantity']):
            db.set_value(str(key), group_df.to_dict(orient='records')[0])

        db.save()

    def create_people_kvdb():
        db = KVDB(people_json)
        #df3 = read_cluster_csv('data/external/tidynomicon/person.csv')
        for person_id, group_df in df3.groupby('person_id'):
            db.set_value(person_id, group_df.to_dict(orient='records')[0])

```

```

db.save()

def create_visits_kvdb():
    db = KVDB(visited_json)
    #df4 = read_cluster_csv('data/external/tidynomicon/visited.csv')
    for key, group_df in df4.groupby(['visit_id', 'site_id']):
        db.set_value(str(key), group_df.to_dict(orient='records')[0])

    db.save()

```

```

[8]: create_sites_kvdb()
      create_measurements_kvdb()
      create_people_kvdb()
      create_visits_kvdb()

```

```

[9]: kvdb_path = 'sites.json'
      kvdb = KVDB(kvdb_path)
      key = ('DR-1')
      value = dict(
          site_id='DR-1',
          latitude=-49.85,
          longitude=-128.57
      )
      kvdb.set_value(key, value)
      retrieved_value = kvdb.get_value(key)
      retrieved_value

```

```

[9]: {'site_id': 'DR-1', 'latitude': -49.85, 'longitude': -128.57}

```

```

[10]: kvdb_path = 'measurements.json'
       kvdb = KVDB(kvdb_path)
       key = (619, 'dyer', 'rad')
       value = dict(
           visit_id=619,
           person_id='dyer',
           quantity='rad'
       )
       kvdb.set_value(key, value)
       retrieved_value = kvdb.get_value(key)
       retrieved_value

```

```

[10]: {'visit_id': 619, 'person_id': 'dyer', 'quantity': 'rad'}

```

```

[11]: kvdb_path = 'people.json'
       kvdb = KVDB(kvdb_path)

```

```

key = ('dyer')
value = dict(
    person_id='dyer',
    personal_name= 'William',
    family_name='Dyer'
)
kvdb.set_value(key, value)
retrieved_value = kvdb.get_value(key)
retrieved_value

```

```
[11]: {'person_id': 'dyer', 'personal_name': 'William', 'family_name': 'Dyer'}
```

```

[12]: kvdb_path = 'visits.json'
kvdb = KVDB(kvdb_path)
key = (619, 'DR-1')
value = dict(
    visit_id=619,
    site_id='DR-1',
    visit_date='1927-02-08'
)
kvdb.set_value(key, value)
retrieved_value = kvdb.get_value(key)
retrieved_value

```

```
[12]: {'visit_id': 619, 'site_id': 'DR-1', 'visit_date': '1927-02-08'}
```

```

[19]: with open('results/kvdb/sites.json','r') as file:
        sites_dict = json.load(file)
print(sites_dict)

```

```

{'DR-1': {'site_id': 'DR-1', 'latitude': -49.85, 'longitude': -128.57}, 'DR-3':
{'site_id': 'DR-3', 'latitude': -47.15, 'longitude': -126.72}, 'MSK-4':
{'site_id': 'MSK-4', 'latitude': -48.87, 'longitude': -123.4}}

```

```

[18]: with open('','r') as file:
        measurements_dict = json.load(file)
print(measurements_dict)

```

```

{"(619, 'dyer', 'rad')": {'visit_id': 619, 'person_id': 'dyer', 'quantity':
'rad', 'reading': 9.82}, "(619, 'dyer', 'sal')": {'visit_id': 619, 'person_id':
'dyer', 'quantity': 'sal', 'reading': 0.13}, "(622, 'dyer', 'rad')":
{'visit_id': 622, 'person_id': 'dyer', 'quantity': 'rad', 'reading': 7.8},
"(622, 'dyer', 'sal')": {'visit_id': 622, 'person_id': 'dyer', 'quantity':
'sal', 'reading': 0.09}, "(734, 'lake', 'sal')": {'visit_id': 734, 'person_id':
'lake', 'quantity': 'sal', 'reading': 0.05}, "(734, 'pb', 'rad')": {'visit_id':
734, 'person_id': 'pb', 'quantity': 'rad', 'reading': 8.41}, "(734, 'pb',
'temp')": {'visit_id': 734, 'person_id': 'pb', 'quantity': 'temp', 'reading':
-21.5}, "(735, 'pb', 'rad')": {'visit_id': 735, 'person_id': 'pb', 'quantity':

```

```
'rad', 'reading': 7.22}, "(735, 'pb', 'sal')": {'visit_id': 735, 'person_id': 'pb', 'quantity': 'sal', 'reading': 0.06}, "(735, 'pb', 'temp')": {'visit_id': 735, 'person_id': 'pb', 'quantity': 'temp', 'reading': -26.0}, "(751, 'pb', 'rad')": {'visit_id': 751, 'person_id': 'pb', 'quantity': 'rad', 'reading': 4.35}, "(751, 'pb', 'temp')": {'visit_id': 751, 'person_id': 'pb', 'quantity': 'temp', 'reading': -18.5}, "(752, 'lake', 'rad')": {'visit_id': 752, 'person_id': 'lake', 'quantity': 'rad', 'reading': 2.19}, "(752, 'lake', 'sal')": {'visit_id': 752, 'person_id': 'lake', 'quantity': 'sal', 'reading': 0.09}, "(752, 'lake', 'temp')": {'visit_id': 752, 'person_id': 'lake', 'quantity': 'temp', 'reading': -16.0}, "(752, 'roe', 'sal')": {'visit_id': 752, 'person_id': 'roe', 'quantity': 'sal', 'reading': 41.6}, "(837, 'lake', 'rad')": {'visit_id': 837, 'person_id': 'lake', 'quantity': 'rad', 'reading': 1.46}, "(837, 'lake', 'sal')": {'visit_id': 837, 'person_id': 'lake', 'quantity': 'sal', 'reading': 0.21}, "(837, 'roe', 'sal')": {'visit_id': 837, 'person_id': 'roe', 'quantity': 'sal', 'reading': 22.5}, "(844, 'roe', 'rad')": {'visit_id': 844, 'person_id': 'roe', 'quantity': 'rad', 'reading': 11.25}}
```

```
[20]: with open('results/kvdb/people.json','r') as file:
        people_dict = json.load(file)
        print(people_dict)
```

```
{'danforth': {'person_id': 'danforth', 'personal_name': 'Frank', 'family_name': 'Danforth'}, 'dyer': {'person_id': 'dyer', 'personal_name': 'William', 'family_name': 'Dyer'}, 'lake': {'person_id': 'lake', 'personal_name': 'Anderson', 'family_name': 'Lake'}, 'pb': {'person_id': 'pb', 'personal_name': 'Frank', 'family_name': 'Pabodie'}, 'roe': {'person_id': 'roe', 'personal_name': 'Valentina', 'family_name': 'Roerich'}}
```

```
[21]: with open('results/kvdb/visited.json','r') as file:
        visited_dict = json.load(file)
        print(visited_dict)
```

```
{"(619, 'DR-1')": {'visit_id': 619, 'site_id': 'DR-1', 'visit_date': '1927-02-08'}, "(622, 'DR-1')": {'visit_id': 622, 'site_id': 'DR-1', 'visit_date': '1927-02-10'}, "(734, 'DR-3')": {'visit_id': 734, 'site_id': 'DR-3', 'visit_date': '1930-01-07'}, "(735, 'DR-3')": {'visit_id': 735, 'site_id': 'DR-3', 'visit_date': '1930-01-12'}, "(751, 'DR-3')": {'visit_id': 751, 'site_id': 'DR-3', 'visit_date': '1930-02-26'}, "(752, 'DR-3')": {'visit_id': 752, 'site_id': 'DR-3', 'visit_date': nan}, "(837, 'MSK-4')": {'visit_id': 837, 'site_id': 'MSK-4', 'visit_date': '1932-01-14'}, "(844, 'DR-1')": {'visit_id': 844, 'site_id': 'DR-1', 'visit_date': '1932-03-22'}}
```

```
[ ]:
```