kvdb

January 25, 2022

[1]:
```
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)

src_data_dir = current_dir.parent.parent.parent.joinpath('data/external/
 ↪tidynomicon')

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')

print(current_dir)
print(results_dir)
print(kv_data_dir)
print(src_data_dir)
```

```
/home/jovyan/dsc650/dsc650/assignments/assignment02
/home/jovyan/dsc650/dsc650/assignments/assignment02/results
```

```
/home/jovyan/dsc650/dsc650/assignments/assignment02/results/kvdb
/home/jovyan/dsc650/data/external/tidynomicon
```

[2]:
```python
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)
```

[3]:
```python
def create_sites_kvdb():
    db = KVDB(sites_json)
    # df = read_cluster_csv('data/external/tidynomicon/site.csv')
    src_file_site = f"{src_data_dir}/site.csv"
    df =pd.read_csv(src_file_site,sep=",",header=0)
    for site_id, group_df in df.groupby('site_id'):
        db.set_value(site_id, group_df.to_dict(orient='records')[0])
    db.save()


def create_people_kvdb():
    db = KVDB(people_json)
    src_file_person = f"{src_data_dir}/person.csv"
    df = pd.read_csv(src_file_person, sep=",", header=0)
    for person_id, group_df in df.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)
    src_file_visited = f"{src_data_dir}/visited.csv"
    df = pd.read_csv(src_file_visited, sep=",", header=0)
```

```
    for key, group_df in df.groupby(['visit_id','site_id']):
        db.set_value(str(key),group_df.to_dict(orient='records')[0])
    db.save()


def create_measurements_kvdb():
    db = KVDB(measurements_json)
    src_file_measurements = f"{src_data_dir}/measurements.csv"
    df = pd.read_csv(src_file_measurements, sep=",", header=0)
    for key, group_df in df.groupby(['visit_id', 'person_id']):
        db.set_value(str(key), group_df.to_dict(orient='records')[0])
    db.save()
```

[4]:
```
create_sites_kvdb()
create_people_kvdb()
create_visits_kvdb()
create_measurements_kvdb()
```

[5]:
```
db = KVDB(visited_json)
src_file_visited = f"{src_data_dir}/visited.csv"
df = pd.read_csv(src_file_visited, sep=",", header=0)
for key,value in df.groupby(['visit_id','site_id']):
        print(key)
        print(value)
```

```
(619, 'DR-1')
   visit_id site_id  visit_date
0       619    DR-1  1927-02-08
(622, 'DR-1')
   visit_id site_id  visit_date
1       622    DR-1  1927-02-10
(734, 'DR-3')
   visit_id site_id  visit_date
2       734    DR-3  1930-01-07
(735, 'DR-3')
   visit_id site_id  visit_date
3       735    DR-3  1930-01-12
(751, 'DR-3')
   visit_id site_id  visit_date
4       751    DR-3  1930-02-26
(752, 'DR-3')
   visit_id site_id visit_date
5       752    DR-3        NaN
(837, 'MSK-4')
   visit_id site_id  visit_date
6       837   MSK-4  1932-01-14
(844, 'DR-1')
```

```
   visit_id site_id  visit_date
7       844    DR-1  1932-03-22
```

[ ]: