Simple Geohash Index

April 10, 2021

1 3.2

1.1 3.2.a Simple Geohash Index

[1]:

```
import os
import sys
import gzip
import json
from pathlib import Path
import csv

import pandas as pd
import s3fs
import pyarrow as pa
from pyarrow.json import read_json
import pyarrow.parquet as pq
import fastavro
import pygeohash as pgh
import snappy
import jsonschema
from jsonschema.exceptions import ValidationError


endpoint_url='https://storage.budsc.midwest-datascience.com'

current_dir = Path(os.getcwd()).absolute()
schema_dir = current_dir.joinpath('schemas')
schema_dir.mkdir(parents=True, exist_ok=True)
results_dir = current_dir.joinpath('results')
results_dir.mkdir(parents=True, exist_ok=True)

def read_jsonl_data():
    s3 = s3fs.S3FileSystem(
        anon=True,
        client_kwargs={
            'endpoint_url': endpoint_url
        }
    )
    src_data_path = 'data/processed/openflights/routes.jsonl.gz'
    with s3.open(src_data_path, 'rb') as f_gz:
        with gzip.open(f_gz, 'rb') as f:
            records = [json.loads(line) for line in f.readlines()]


    return records

records = read_jsonl_data()
```

[2]:

```
df = pd.json_normalize(records)

df = df.rename({'dst_airport.latitude': 'dst_airport_latitude', 'dst_airport.
↪longitude': 'dst_airport_longitude'}, axis=1)  # new method
df.head()
```

[2]:

```
   codeshare equipment  airline.airline_id airline.name  \
0      False     [CR2]                 410   Aerocondor
1      False     [CR2]                 410   Aerocondor
2      False     [CR2]                 410   Aerocondor
3      False     [CR2]                 410   Aerocondor
4      False     [CR2]                 410   Aerocondor

           airline.alias airline.iata airline.icao airline.callsign  \
0  ANA All Nippon Airways           2B          ARD       AEROCONDOR
1  ANA All Nippon Airways           2B          ARD       AEROCONDOR
2  ANA All Nippon Airways           2B          ARD       AEROCONDOR
3  ANA All Nippon Airways           2B          ARD       AEROCONDOR
4  ANA All Nippon Airways           2B          ARD       AEROCONDOR

  airline.country  airline.active  …  dst_airport_latitude  \
0        Portugal            True  …             55.606201
1        Portugal            True  …             55.606201
2        Portugal            True  …             44.225101
3        Portugal            True  …             55.606201
4        Portugal            True  …             55.012600

  dst_airport_longitude dst_airport.altitude dst_airport.timezone  \
0             49.278702                411.0                  3.0
1             49.278702                411.0                  3.0
2             43.081902               1054.0                  3.0
3             49.278702                411.0                  3.0
4             82.650703                365.0                  7.0

  dst_airport.dst dst_airport.tz_id  dst_airport.type  dst_airport.source  \
0               N     Europe/Moscow           airport         OurAirports
1               N     Europe/Moscow           airport         OurAirports
```

```
2               N     Europe/Moscow           airport         OurAirports
3               N     Europe/Moscow           airport         OurAirports
4               N  Asia/Krasnoyarsk           airport         OurAirports

   dst_airport  src_airport
0          NaN          NaN
1          NaN          NaN
2          NaN          NaN
3          NaN          NaN
4          NaN          NaN

[5 rows x 40 columns]
```

[10]:

```
dst_airport_latitude = df['dst_airport_latitude']
dst_airport_longitude = df['dst_airport_longitude']

df['geohash'] = df.apply(lambda x: pgh.encode(x.dst_airport_latitude,x.
 ↪dst_airport_longitude,precision=5), axis=1)
df.head(5)
```

[10]:

```
   codeshare equipment  airline.airline_id airline.name  \
0      False     [CR2]                 410   Aerocondor
1      False     [CR2]                 410   Aerocondor
2      False     [CR2]                 410   Aerocondor
3      False     [CR2]                 410   Aerocondor
4      False     [CR2]                 410   Aerocondor

            airline.alias airline.iata airline.icao airline.callsign  \
0  ANA All Nippon Airways           2B          ARD       AEROCONDOR
1  ANA All Nippon Airways           2B          ARD       AEROCONDOR
2  ANA All Nippon Airways           2B          ARD       AEROCONDOR
3  ANA All Nippon Airways           2B          ARD       AEROCONDOR
4  ANA All Nippon Airways           2B          ARD       AEROCONDOR

  airline.country  airline.active  …  dst_airport_longitude  \
0        Portugal            True  …              49.278702
1        Portugal            True  …              49.278702
2        Portugal            True  …              43.081902
3        Portugal            True  …              49.278702
4        Portugal            True  …              82.650703

  dst_airport.altitude dst_airport.timezone dst_airport.dst dst_airport.tz_id  \
0                411.0                  3.0               N     Europe/Moscow
1                411.0                  3.0               N     Europe/Moscow
2               1054.0                  3.0               N     Europe/Moscow
3                411.0                  3.0               N     Europe/Moscow
4                365.0                  7.0               N  Asia/Krasnoyarsk
```

```
  dst_airport.type  dst_airport.source  dst_airport  src_airport  geohash
0          airport          OurAirports          NaN          NaN    v1gh3
1          airport          OurAirports          NaN          NaN    v1gh3
2          airport          OurAirports          NaN          NaN    szyes
3          airport          OurAirports          NaN          NaN    v1gh3
4          airport          OurAirports          NaN          NaN    vcfbb

[5 rows x 41 columns]
```

```
[11]: df['geohash']
```

```
[11]: 0        v1gh3
      1        v1gh3
      2        szyes
      3        v1gh3
      4        vcfbb
               …
      67658    r1f90
      67659    txsuy
      67660    ucfgn
      67661    tx5z0
      67662    txsuy
      Name: geohash, Length: 67663, dtype: object
```

```
[4]: df.to_json(r'/home/jovyan/dsc650/schemas/results/geoindex\geoindex.json')
```