Simple Search Feature

April 10, 2021

1 3.2.b Simple Search Feature

[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
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)
```

[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_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]
```

1.1 the airports which is in a Radius of 300 Km from Bellevue University

[18]:
```
from math import radians, cos, sin, asin, sqrt
def haversine(lon1, lat1, lon2, lat2):
  # convert decimal degrees to radians
    lon1, lat1, lon2, lat2 = map(radians, [lon1, lat1, lon2, lat2])

  # haversine formula
    dlon = lon2 - lon1
    dlat = lat2 - lat1
    a = sin(dlat/2)**2 + cos(lat1) * cos(lat2) * sin(dlon/2)**2
    c = 2 * asin(sqrt(a))
    r = 6371 # Radius of earth in kilometers. Use 3956 for miles
    return c * r
```

[37]:
```
jersey_city_long_lat=(95.9182,41.1506)
def row_hsign(row):
    return␣
 ↪haversine(*jersey_city_long_lat,row['dst_airport_longitude'],row['dst_airport.
 ↪altitude'])

df['distance']=df.apply(row_hsign,axis=1)
df[df['distance']<=600]
```

[37]:
```
       codeshare        equipment  airline.airline_id  \
443        False            [SF3]               20710
1141       False            [320]               17885
1755       False            [DHT]                1581
1756       False            [DHT]                1581
1759       False            [DHT]                1581
…            …                …                   …
63957      False            [73W]                4547
64074      False            [73W]                4547
64129      False  [73C, 73W, 733]                4547
64301      False            [73W]                4547
64547      False            [73W]                4547
```

```
                  airline.name         airline.alias airline.iata  \
443          Silver Airways (3M)                   nan           3M
1141   Interjet (ABC Aerolineas)                   nan           4O
1755         CAL Cargo Air Lines  SN Brussels Airlines           5C
1756         CAL Cargo Air Lines  SN Brussels Airlines           5C
1759         CAL Cargo Air Lines  SN Brussels Airlines           5C
…                            …                   …          …
63957         Southwest Airlines              SkyWork           WN
64074         Southwest Airlines              SkyWork           WN
64129         Southwest Airlines              SkyWork           WN
64301         Southwest Airlines              SkyWork           WN
64547         Southwest Airlines              SkyWork           WN

      airline.icao airline.callsign airline.country  airline.active  …  \
443            DAK     Silver Wings   United States            True  …
1141           IBS          INTERJET          Mexico            True  …
1755           ICL               CAL          Israel            True  …
1756           ICL               CAL          Israel            True  …
1759           ICL               CAL          Israel            True  …
…              …                 …               …               …  …
63957          SWA         SOUTHWEST   United States            True  …
64074          SWA         SOUTHWEST   United States            True  …
64129          SWA         SOUTHWEST   United States            True  …
64301          SWA         SOUTHWEST   United States            True  …
64547          SWA         SOUTHWEST   United States            True  …

       dst_airport.altitude dst_airport.timezone dst_airport.dst  \
443                  2302.0                 -5.0               U
1141                 3021.0                 -6.0               U
1755                 3021.0                 -6.0               U
1756                 3021.0                 -6.0               U
1759                 3021.0                 -6.0               U
…                     …                   …              …
63957                 501.0                 -5.0               A
64074                 501.0                 -5.0               A
64129                 501.0                 -5.0               A
64301                 501.0                 -5.0               A
64547                 501.0                 -5.0               A

        dst_airport.tz_id dst_airport.type dst_airport.source  dst_airport  \
443       America/New_York          airport        OurAirports          NaN
1141   America/Costa_Rica          airport        OurAirports          NaN
1755   America/Costa_Rica          airport        OurAirports          NaN
1756   America/Costa_Rica          airport        OurAirports          NaN
1759   America/Costa_Rica          airport        OurAirports          NaN
…                     …                …                  …          …
```

```
63957    America/New_York          airport        OurAirports          NaN
64074    America/New_York          airport        OurAirports          NaN
64129    America/New_York          airport        OurAirports          NaN
64301    America/New_York          airport        OurAirports          NaN
64547    America/New_York          airport        OurAirports          NaN

      src_airport  geohash    distance
443           NaN    dnwz6  471.429287
1141          NaN    d1u0g  239.379747
1755          NaN    d1u0g  239.379747
1756          NaN    d1u0g  239.379747
1759          NaN    d1u0g  239.379747
…             …      …          …
63957         NaN    dng11  277.468141
64074         NaN    dng11  277.468141
64129         NaN    dng11  277.468141
64301         NaN    dng11  277.468141
64547         NaN    dng11  277.468141

[243 rows x 42 columns]
```