Pipeline ETL

September 13, 2022

[1]:
```
import petl as etl
```

[2]:
```
filename='drinks.json'
```

[3]:
```
dado1=etl.fromjson(filename)
```

[5]:
```
dado1.display(15)
```

[6]:
```
source=etl.sources.ZipSource('drinks.zip','drinks.json')
```

[8]:
```
etl.fromjson(source).display(15)
```

[9]:
```
dado1.nrows()
```

[9]:
```
193
```

[13]:
```
print(dado1[2][2])
```

```
4.9 Litres
```

[20]:
```
print(dado1['total_pure_alcohol'][3])
```

```
12.4 Litres
```

[18]:
```
fields = dado1.fieldnames()
print(fields)
```

```
('country', 'continent', 'total_pure_alcohol', 'servings')
```

[ ]:
```

```

[19]:
```
for f in fields:
    print(f, '\t', dado1.typecounter(f))
```

```
country          Counter({'str': 193})
continent        Counter({'str': 170, 'NoneType': 23})
total_pure_alcohol       Counter({'str': 193})
servings         Counter({'str': 193})
```

0.1 Transformação

[27]:
```
def para_float (texto):
    return float(texto[:-7])
```

[28]:
```
dado2 = dado1.convert('total_pure_alcohol', para_float)
```

[29]:
```
dado2
```

[29]:
```
+---------------+-----------+--------------------+-----------------------------
------+
| country       | continent | total_pure_alcohol | servings
|
+===============+===========+====================+=============================
======+
| 'Afghanistan' | 'AS'      |                0.0 | '[beer:0, spirit:0, wine:0]'
|
+---------------+-----------+--------------------+-----------------------------
------+
| 'Albania'     | 'EU'      |                4.9 | '[beer:89, spirit:132,
wine:54]'   |
+---------------+-----------+--------------------+-----------------------------
------+
| 'Algeria'     | 'AF'      |                0.7 | '[beer:25, spirit:0,
wine:14]'     |
+---------------+-----------+--------------------+-----------------------------
------+
| 'Andorra'     | 'EU'      |               12.4 | '[beer:245, spirit:138,
wine:312]' |
+---------------+-----------+--------------------+-----------------------------
------+
| 'Angola'      | 'AF'      |                5.9 | '[beer:217, spirit:57,
wine:45]'   |
+---------------+-----------+--------------------+-----------------------------
------+
…
```

[31]:
```
amostra = dado2['servings'][4]
print(amostra)
```

```
[beer:217, spirit:57, wine:45]
```

[32]:
```
print(type(amostra))
```

```
<class 'str'>
```

[35]:
```
def para_dict(texto):
    if texto[0]=='[' and texto[-1]==']': # expressão regular busca a informação␣
 ↪na string
```

```
        items = texto[1:-1].split(',') #separação da informação vetor de itens␣
↪separados
    else:
        return None # sai do if
    output = {} # armazenar os dicionários
    for i in items: #percorro o vetor de itens
        key, value = i.split(':') #separação de cada dicionario
        output[key]=int(value) #armazeno valor
    return output # retorno os dicionários
```

[36]:
```
print (para_dict(amostra))
```

```
{'beer': 217, ' spirit': 57, ' wine': 45}
```

[37]:
```
print(type(para_dict(amostra)))
```

```
<class 'dict'>
```

[38]:
```
dado3 = dado2.convert('servings', para_dict)
```

[39]:
```
dado3
```

[39]:
```
+---------------+-----------+--------------------+------------------------------
---------------+
| country       | continent | total_pure_alcohol | servings
|
+===============+===========+====================+==============================
===============+
| 'Afghanistan' | 'AS'      |                0.0 | {'beer': 0, ' spirit': 0, '
wine': 0}       |
+---------------+-----------+--------------------+------------------------------
---------------+
| 'Albania'     | 'EU'      |                4.9 | {'beer': 89, ' spirit': 132,
' wine': 54}    |
+---------------+-----------+--------------------+------------------------------
---------------+
| 'Algeria'     | 'AF'      |                0.7 | {'beer': 25, ' spirit': 0, '
wine': 14}      |
+---------------+-----------+--------------------+------------------------------
---------------+
| 'Andorra'     | 'EU'      |               12.4 | {'beer': 245, ' spirit': 138,
' wine': 312} |
+---------------+-----------+--------------------+------------------------------
---------------+
| 'Angola'      | 'AF'      |                5.9 | {'beer': 217, ' spirit': 57,
' wine': 45}    |
```

```
+---------------+-----------+--------------------+-----------------------------
---------------+
…
```

[40]:
```
dado4 = dado3.unpackdict('servings')
```

[50]:
```
dado4
```

[50]:
```
+---------------+-----------+--------------------+---------+-------+------+
| country       | continent | total_pure_alcohol |  spirit |  wine | beer |
+===============+===========+====================+=========+=======+======+
| 'Afghanistan' | 'AS'      |                0.0 |       0 |     0 |    0 |
+---------------+-----------+--------------------+---------+-------+------+
| 'Albania'     | 'EU'      |                4.9 |     132 |    54 |   89 |
+---------------+-----------+--------------------+---------+-------+------+
| 'Algeria'     | 'AF'      |                0.7 |       0 |    14 |   25 |
+---------------+-----------+--------------------+---------+-------+------+
| 'Andorra'     | 'EU'      |               12.4 |     138 |   312 |  245 |
+---------------+-----------+--------------------+---------+-------+------+
| 'Angola'      | 'AF'      |                5.9 |      57 |    45 |  217 |
+---------------+-----------+--------------------+---------+-------+------+
…
```

[54]:
```
def total_serving(row):
    return row['beer']+row[' spirit']+row[' wine']
```

[55]:
```
dado5=dado4.addfield('total_serving', total_serving)
```

[57]:
```
dado5.display(10)
```

[60]:
```
out_countries=dado5.
 ↪cut(['country','continent','total_pure_alcohol','total_serving'])
```

[61]:
```
out_countries
```

[61]:
```
+---------------+-----------+--------------------+---------------+
| country       | continent | total_pure_alcohol | total_serving |
+===============+===========+====================+===============+
| 'Afghanistan' | 'AS'      |                0.0 |             0 |
+---------------+-----------+--------------------+---------------+
| 'Albania'     | 'EU'      |                4.9 |           275 |
+---------------+-----------+--------------------+---------------+
| 'Algeria'     | 'AF'      |                0.7 |            39 |
+---------------+-----------+--------------------+---------------+
| 'Andorra'     | 'EU'      |               12.4 |           695 |
+---------------+-----------+--------------------+---------------+
| 'Angola'      | 'AF'      |                5.9 |           319 |
+---------------+-----------+--------------------+---------------+
```

...

[62]:
```
out_countries.tocsv('saida.csv')
```

[ ]:
```

```