ProjetL3_extraction_Table

December 12, 2018

1 Un exemple d'extraction de table d'une page html

```
In [16]: #!/usr/bin/env python3
         # -*- coding: utf-8 -*-
         """
         Created on Wed Dec 12 09:58:17 2018

         @author: moi
         """

         import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt

         # une page wikipedia avec des tableaux
         url = 'http://fr.wikipedia.org/wiki/Coupe_du_monde_de_rugby_%C3%A0_XV_2011'
```

On selectionne les tableaux de la page qui contiennent le champs "Joués"

```
In [17]: ts = pd.read_html(url, match='Joués', header=0)

In [18]: ts
```

```
Out[18]: [   No            Nation  Joués  V  N  D  EM  EE  BO  BD   PM   PE  Diff  Pts
          0   1  Nouvelle-Zélande      4  4  0  0  36   6   4   0  240   49   191   20
          1   2            France      4  2  0  2  13   9   2   1  124   96    28   11
          2   3             Tonga      4  2  0  2   7  13   0   1   80   98   -18    9
          3   4            Canada      4  1  1  2   9  20   0   0   82  168   -86    6
          4   5             Japon      4  0  1  3   8  25   0   0   69  184  -115    2,
             No      Nation  Joués  V  N  D  EM  EE  BO  BD   PM   PE  Diff  Pts
          0   1  Angleterre      4  4  0  0  18   1   2   0  137   34   103   18
          1   2   Argentine      4  3  0  1  10   3   1   1   90   40    50   14
          2   3      Écosse      4  2  0  2   4   4   1   2   73   59    14   11
          3   4     Géorgie      4  1  0  3   3   9   0   0   48   90   -42    4
          4   5    Roumanie      4  0  0  4   3  21   0   0   44  169  -125    0,
             No      Nation  Joués  V  N  D  EM  EE  BO  BD   PM   PE  Diff  Pts
          0   1     Irlande      4  4  0  0  15   3   1   0  135   34   101   17
```

```
1   2   Australie      4  3  0  1  25   4   3   0  173   48   125   15
2   3      Italie      4  2  0  2  13  11   2   0   92   95    -3   10
3   4  États-Unis      4  1  0  3   4  18   0   0   38  122   -84    4
4   5      Russie      4  0  0  4   8  29   0   1   57  196  -139    1,
   No          Nation  Joués  V  N  D  EM  EE  BO  BD   PM   PE  Diff  Pts
0   1  Afrique du Sud      4  4  0  0  21   2   2   0  166   24   142   18
1   2          Galles      4  3  0  1  23   4   2   1  180   34   146   15
2   3           Samoa      4  2  0  2  10   5   1   1   91   49    42   10
3   4           Fidji      4  1  0  3   7  19   1   0   59  167  -108    5
4   5         Namibie      4  0  0  4   5  36   0   0   44  266  -222    0]
```

On transforme en tableau numpy dont chaque ligne est (Pays, Points)

```
In [19]: rec = np.concatenate([df.filter(items=['Nation', 'PM']).to_records(index=False) for d
```

```
In [20]: rec
```

```
Out[20]: array([('Nouvelle-Zélande', 240), ('France', 124), ('Tonga',  80),
                ('Canada',  82), ('Japon',  69), ('Angleterre', 137),
                ('Argentine',  90), ('Écosse',  73), ('Géorgie',  48),
                ('Roumanie',  44), ('Irlande', 135), ('Australie', 173),
                ('Italie',  92), ('États-Unis',  38), ('Russie',  57),
                ('Afrique du Sud', 166), ('Galles', 180), ('Samoa',  91),
                ('Fidji',  59), ('Namibie',  44)],
               dtype=(numpy.record, [('Nation', 'O'), ('PM', '<i8')]))
```

On extrait le tableau 1d des pays et celui des points

```
In [21]: pays = [r[0] for r in tf]
         points = [r[1] for r in tf]
```

```
In [22]: pays
```

```
Out[22]: ['Nouvelle-Zélande',
          'France',
          'Tonga',
          'Canada',
          'Japon',
          'Angleterre',
          'Argentine',
          'Écosse',
          'Géorgie',
          'Roumanie',
          'Irlande',
          'Australie',
          'Italie',
          'États-Unis',
          'Russie',
          'Afrique du Sud',
```

```
          'Galles',
          'Samoa',
          'Fidji',
          'Namibie']
```

```
In [23]: points
```

```
Out[23]: [240,
          124,
          80,
          82,
          69,
          137,
          90,
          73,
          48,
          44,
          135,
          173,
          92,
          38,
          57,
          166,
          180,
          91,
          59,
          44]
```

On trace un camembert

```
In [24]: cam = plt.pie(points,labels=pays)
```