

```
# install tabula python package
!pip install tabula.py
```

```
Collecting tabula.py
  Downloading tabula_py-2.9.0-py3-none-any.whl (12.0 MB)
    12.0/12.0 MB 42.1 MB/s eta 0:00:00
Requirement already satisfied: pandas>=0.25.3 in /usr/local/lib/python3.10/dist-packages (from tabula.py) (2.0.3)
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (from tabula.py) (1.25.2)
Requirement already satisfied: distro in /usr/lib/python3/dist-packages (from tabula.py) (1.7.0)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from pandas>=0.25.3->tabula.py)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=0.25.3->tabula.py) (2023.4)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=0.25.3->tabula.py) (2024.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2->pandas>=0.25.3->tabula.py)
Installing collected packages: tabula.py
Successfully installed tabula.py-2.9.0
```

```
!pip install tabulate
```

```
Requirement already satisfied: tabulate in /usr/local/lib/python3.10/dist-packages (0.9.0)
```

```
# import the necessary libraries
from tabula import read_pdf
from tabulate import tabulate
```

```
import warnings
```

```
# import all warnings
warnings.filterwarnings("ignore")
```

```
# filename variable of the pdf file which needs to be uploaded into the folder/environment
pdf_file = 'FoodList.pdf'
```

```
# extract data from page 1 of the pdf file
page_number = 1
```

```
# returns the extracted tables as pandas dataframes
tables_df = read_pdf(pdf_file, pages=page_number)
```

```
# print the tables from page 1 of the pdf
print(tables_df)
```

```
# ignore any warnings
```

```
WARNING:tabula.backend:Error importing jpytype dependencies. Fallback to subprocess.
WARNING:tabula.backend:No module named 'jpytype'
WARNING:tabula.backend:Got stderr: Apr 04, 2024 12:21:24 AM org.apache.pdfbox.pdmodel.font.FileSystemFontProvider loadDiskCache
WARNING: New fonts found, font cache will be re-built
Apr 04, 2024 12:21:24 AM org.apache.pdfbox.pdmodel.font.FileSystemFontProvider <init>
WARNING: Building on-disk font cache, this may take a while
Apr 04, 2024 12:21:25 AM org.apache.pdfbox.pdmodel.font.FileSystemFontProvider <init>
WARNING: Finished building on-disk font cache, found 17 fonts
```

```
[
    BREADS & CEREALS
    0 Bagel ( 1 average ) 140 cal (45g)
    1 Biscuit digestives 86 cal (per biscuit)
    2 Jaffa cake 48 cal (per biscuit)
    3 Bread white (thick slice) 96 cal (1 slice 40g)
    4 Bread wholemeal (thick) 88 cal (1 slice 40g)
    5 Chapatis 250 cal
    6 Cornflakes 130 cal (35g)
    7 Crackerbread 17 cal per slice
    8 Cream crackers 35 cal (per cracker)
    9 Crumpets 93 cal (per crumpet)
    10 Flapjacks basic fruit mix 320 cal
    11 Macaroni (boiled) 238 cal (250g)
    12 Muesli 195 cal (50g)
    13 Naan bread (normal) 300 cal (small plate size)
    14 Noodles (boiled) 175 cal (250g)
    15 Pasta ( normal boiled ) 330 cal (300g)
    16 Pasta (wholemeal boiled ) 315 cal (300g)
    17 Porridge oats (with water) 193 cal (350g)
    18 Potatoes** (boiled) 210 cal (300g)
    19 Potatoes** (roast) 420 cal (300g)
```

```

per 100 grams (3.5 oz)  Unnamed: 0 energy content
0      310 cal      NaN      Medium
1      480 cal      NaN      High
2      370 cal      NaN      Med-High
3      240 cal      NaN      Medium
4      220 cal      NaN      Low-med
5      300 cal      NaN      Medium
6      370 cal      NaN      Med-High
7      325 cal      NaN      Low Calorie
8      440 cal      NaN      Low / portion
9      198 cal      NaN      Low-Med
10     500 cal      NaN      High
11     95 cal      NaN      Low calorie
12     390 cal      NaN      Med-high
13     320 cal      NaN      Medium
14     70 cal      NaN      Low calorie
15     110 cal      NaN      Low calorie
16     105 cal      NaN      Low calorie
17     55 cal      NaN      Low calorie
18     70 cal      NaN      Low calorie
19     140 cal      NaN      Medium ]

```

```

# use list comprehension to create a new list, loop through each dataframe drops any columns that contain NaN (missing) values
cleaned_tables = [table.dropna(axis='columns') for table in tables_df]

```

```

# loop through the table and print everything, should not have any NaN values
for idx, table in enumerate(cleaned_tables):
    print(f"Table {idx+1} after dropping the NaN values:")

```

```
print(table)
```

```

Table 1 after dropping the NaN values:
      BREADS & CEREALS      Portion size * \
0      Bagel ( 1 average )      140 cal (45g)
1      Biscuit digestives      86 cal (per biscuit)
2      Jaffa cake      48 cal (per biscuit)
3      Bread white (thick slice)      96 cal (1 slice 40g)
4      Bread wholemeal (thick)      88 cal (1 slice 40g)
5      Chapatis      250 cal
6      Cornflakes      130 cal (35g)
7      Crackerbread      17 cal per slice
8      Cream crackers      35 cal (per cracker)
9      Crumpets      93 cal (per crumpet)
10     Flapjacks basic fruit mix      320 cal
11     Macaroni (boiled)      238 cal (250g)
12     Muesli      195 cal (50g)
13     Naan bread (normal)      300 cal (small plate size)
14     Noodles (boiled)      175 cal (250g)
15     Pasta ( normal boiled )      330 cal (300g)
16     Pasta (wholemeal boiled )      315 cal (300g)
17     Porridge oats (with water)      193 cal (350g)
18     Potatoes** (boiled)      210 cal (300g)
19     Potatoes** (roast)      420 cal (300g)

```

```

per 100 grams (3.5 oz) energy content
0      310 cal      Medium
1      480 cal      High
2      370 cal      Med-High
3      240 cal      Medium
4      220 cal      Low-med
5      300 cal      Medium
6      370 cal      Med-High
7      325 cal      Low Calorie
8      440 cal      Low / portion
9      198 cal      Low-Med
10     500 cal      High
11     95 cal      Low calorie
12     390 cal      Med-high
13     320 cal      Medium
14     70 cal      Low calorie
15     110 cal      Low calorie
16     105 cal      Low calorie
17     55 cal      Low calorie
18     70 cal      Low calorie
19     140 cal      Medium

```

```

# extract data from page 1 of the pdf file
page_number = 3

```

```
# returns the extracted tables as pandas dataframes
tables_df = read_pdf(pdf_file, pages=page_number)
```

```
# print the tables from page 1 of the pdf
print(tables_df)
```

	Fish cake	90 cal	per cake	200 cal	Medium
0	Fish fingers	50 cal	per piece	220 cal	Medium
1	Gammon	320 cal		280 cal	Med-High
2	Haddock fresh	200 cal		110 cal	Low calorie
3	Halibut fresh	220 cal		125 cal	Low calorie
4	NaN	NaN		NaN	NaN
5	Ham	6 cal		240 cal	Medium
6	Herring fresh grilled	300 cal		200 cal	Medium
7	Kidney	200 cal		160 cal	Medium
8	Kipper	200 cal		120 cal	Low calorie
9	NaN	NaN		NaN	NaN
10	Liver	200 cal		150 cal	Medium
11	Liver pate	150 cal		300 cal	Medium
12	Lamb (roast)	300 cal		300 cal	Med-High
13	Lobster boiled	200 cal		100 cal	Low calorie
14	NaN	NaN		NaN	NaN
15	Luncheon meat	300 cal		400 cal	High
16	Mackerel	320 cal		300 cal	Medium
17	Mussels	90 cal		90 cal	Low-Med
18	Pheasant roast	200 cal		200 cal	Medium
19	Pilchards (tinned)	140 cal		140 cal	Medium
20	Prawns	180 cal		100 cal	Low-Med
21	Pork	320 cal		290 cal	Med-High
22	Pork pie	320 cal		450 cal	High
23	Rabbit	200 cal		180 cal	Medium
24	Salmon fresh	220 cal		180 cal	Medium
25	Sardines tinned in oil	220 cal		220 cal	Medium
26	Sardines in tomato sauce	180 cal		180 cal	Medium
27	Sausage pork fried	250 cal		320 cal	High
28	Sausage pork grilled	220 cal		280 cal	Med-High
29	Sausage roll	290 cal		480 cal	High
30	Scampi fried in oil	400 cal		340 cal	High
31	Steak & kidney pie	400 cal		350 cal	High

```
# use list comprehension to convert the dataframe into a JSON string
tables_json = [table.to_json() for table in tables_df]
```

```
# loop over each JSON string to print data from the table
```

```
for idx, table_json in enumerate(tables_json):
```

```
    print(f"Table {idx + 1}:")
```

```
    print(table_json)
```

```
    # add a space/newline between tables
```

```
    print()
```

Table 1:

```
{"Fish cake":{"0":"Fish fingers","1":"Gammon","2":"Haddock fresh","3":"Halibut fresh","4":null,"5":"Ham","6":"Herring fresh grille"}}
```

```
# extract tables from all pages
```

```
tables = read_pdf(pdf_file, pages='all', multiple_tables=True)
```

```
# print the tables extracted from each page
```

```
print(tables)
```



```

5      77 %
6      81 %
7      83 %
8      66 %
9      16 %
10     70 %
11     14 %
12     12 %
13     24 %
14     80 %
15     82 %
16     82 %
17     65 %
18     85 %
19     75 %
20     85 %
21     80 %
22     80 %
23     90 %
24     93 %
25     80 %
26     63 % ,
27
0      Orange large 350g      Orange average      35 calories 8.5 73 %
1      Papaya Diced (small handful) 67 Cals (20g) 17g -
2      Passion Fruit      30 calories 3 50 %
3      Paw Paw      28 calories 6 70 %
4      Peach      35 calories 7 80 %
5      Pear      45 calories 12 77 %
6      Pineapple      50 calories 12 85 %
7      Plum      25 calories 6 79 %
8      Prunes      9 calories 2.2 37 %
9      Raisins      5 calories 1.4 13 %
10     Raspberries each 1.1 calories 0.2 87 %
11     Rhubarb      8 calories 0.8 95 %
12     Satsuma one average 112g      29 calcs 6.5 88 %
13     Satsumas 100g      35 calories 8.5 88 %
14     Strawberries (1 average) 2.7 calories 0.6 90 %
15     Sultanas      5 calories 1.4 16 %
16     Tangerine      26 calories 6 60 %
17     Tomatoes (1 average size)      9 calcs 2.2 93 %
18     Tomatoes Cherrv (1 average size) 2 calories 0.5 90 %

```

```

# set flag to process informationpage by page, performance optimier
stream_option = True

```

```

# extract contents from page 4
page_number = 4

```

```

# extarct tables in a rectangular area defined by cordinates (top,left,bottom,rihgt)
area = (270,13,790,900)

```

```

# extract from the specified area using the stream option
tables_df = read_pdf(pdf_file, pages=page_number, stream=stream_option, area=area)

```

```

# loop over the table, print the information
for idx, table in enumerate(tables_df):
    print(f"Table {idx + 1}:")
    print(table)

```

```

Table 1:
Fruits & Vegetables Portion size *      oz) energy content
0      Apple      44 calories 44 calories Low calorie
1      Banana      107 calcs 65 calories Low calorie
2      Beans baked beans      170 calcs 80 calories Low calorie
3      Beans dried (boiled)      180 calcs 130 calories Low calorie
4      Blackberries      25 calcs 25 calories Low calorie
5      Blackcurrant      30 calcs 30 calories Low calorie
6      Broccoli      27 calcs 32 calcs Very low
7      Cabbage (boiled) 15 calories 20 calories Low calorie
8      Carrot (boiled) 16 calories 25 calories Low calorie
9      Cauliflower (boiled) 20 calories 30 calories Low calorie
10     Celery (boiled) 5 calories 10 calories Low calorie
11     Cherry      35 calories 50 calories Low calorie
12     Courgette      8 calcs 20 calcs Very low cal
13     Cucumber      3 calories 10 calories Low calorie
14     Dates      100 calories 235 calories Med-High
15     Grapes      55 calories 62 calories Low calorie
16     Grapefruit      32 calories 32 calories Low calorie

```

17	Kiwi	40 calories	50 calories	Low calorie
18	Leek (boiled)	10 calories	20 calories	Low calorie