```
# 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 1.6 MB/s eta
    Requirement already satisfied: pandas>=0.25.3 in /usr/local/lib/python3
    Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-
    Requirement already satisfied: distro in /usr/lib/python3/dist-packages
    Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib
    Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.1
    Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/di
    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/di
```

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

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

```
# filename variable of the pdf file which needs to be uploaded into the folde
pdf_file = '/content/FoodList.pdf'
# extract data from page 1 of the pd 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 jpype dependencies. Fallback to
WARNING:tabula.backend:No module named 'jpype'
WARNING:tabula.backend:Got stderr: Mar 29, 2024 1:43:10 AM org.apache.p
WARNING: New fonts found, font cache will be re-built
Mar 29, 2024 1:43:10 AM org.apache.pdfbox.pdmodel.font.FileSystemFontPr
WARNING: Building on-disk font cache, this may take a while
Mar 29, 2024 1:43:10 AM org.apache.pdfbox.pdmodel.font.FileSystemFontPr
WARNING: Finished building on-disk font cache, found 17 fonts

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

[20 rows x 5 columns]]
```

```
# use list comprehension to create a new list, loop through each dataframe,
cleaned_tables = [table.dropna(axis='columns') for table in tables_df]
# loop through the table and print everything, should not have any NaN valu
for idx, table in enumerate(cleaned_tables):
  print(f"Table {idx+1} after dropping NaN values;")
  print(table)
```

```
Table 1 after dropping NaN values;
              BREADS & CEREALS               Portion size * per 100 gra
0          Bagel ( 1 average )               140 cals (45g)
1           Biscuit digestives        86 cals (per biscuit)
2                   Jaffa cake        48 cals (per biscuit)
3    Bread white (thick slice)       96  cals (1 slice 40g)
4      Bread wholemeal (thick)       88  cals (1 slice 40g)
5                     Chapatis                     250 cals
6                   Cornflakes              130  cals (35g)
7                 Crackerbread            17 cals per slice
8               Cream crackers        35 cals (per cracker)
9                     Crumpets        93 cals (per crumpet)
10   Flapjacks basic fruit mix                     320 cals
11           Macaroni (boiled)              238 cals (250g)
12                      Muesli              195  cals (50g)
13         Naan bread (normal)  300 cals (small plate size)
14            Noodles (boiled)              175 cals (250g)
15     Pasta ( normal boiled )              330 cals (300g)
16   Pasta (wholemeal boiled )              315 cals (300g)
17  Porridge oats (with water)              193 cals (350g)
18         Potatoes** (boiled)              210 cals (300g)
19          Potatoes** (roast)              420 cals (300g)
```

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

```
# extract tables from all pages
tables = read_pdf(pdf_file, pages='all', multiple_tables=True)
# print the tables extracted from each page
print(tables)
```

```
9                              Jam             38 cals  ...        NaN
10                            Lard            225 cals  ...        NaN
11                  Low fat spread             50 cals  ...        NaN
12                       Margarine             50 cals  ...        NaN
13                        Mars bar            240 cals  ...        NaN
14                    Mint sweets     10 cals per piece  ...        NaN
15  Oils -corn, sunflower, olive  135 cals (1 Tbspoon)  ...        NaN
16                Popcorn average            150 cals  ...        NaN
17         Sugar white table sugar   20 cals (1 tspoon)  ...        NaN
18                Sweets (boiled)            100 cals  ...        NaN
19                           Syrup             15 cals  ...        NaN
20                          Toffee            100 cals  ...        NaN

[21 rows x 5 columns],                              Fruit Calories per
0                Apple (1 average)       44 calories          10.5
1                    Apple cooking       35 calories             9
2                          Apricot       30 calories           6.7
3                          Avocado      150 calories             2
4                           Banana      107 calories            26
5                Blackberries each         1 calorie           0.2
6                Blackcurrant each       1.1 calorie          0.25
7          Blueberries  (new) 100g   49 Cals ( 100g )          15 g
8                      Cherry each       2.4 calories           0.6
9                       Clementine           24 cals             5
10                        Currants        5 calories           1.4
11                          Damson       28 calories           7.2
12             One average date 5g            5 cals           1.2
13  Dates with inverted sugar 100g      250 calories            63
14                            Figs       10 calories           2.4
```

```
15                 Gooseberries    2.6 calories        0.65
16          Grapes 100g Seedless        50 cals          15
17          one average Grape 6g      3 calories         0.9
18              Grapefruit whole    100 calories          23
19                         Guava     24 calories         4.4
20                          Kiwi     34 calories           8
21                         Lemon     20 calories         3.4
22                       Lychees      3 calories         0.7
23                         Mango     40 calories         9.5
24         Melon Honeydew (130g)     36 calories           9
25       Melon Canteloupe (130g)        25 cals           6
26                    Nectarines     42 calories           9
27                        Olives    6.8 calories       trace
0               Orange large 350g        100 Cals  22g  75 %
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 cals  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 cals  2.2  93 %
```

```
# set flag to process information page by page, performance optimizer
stream_option = True
# extract contents from page 4
page_number = 4
# extract tables in a rectangular area defined by coordinates (top, left, b
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, are
# 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 cals   65 calories    Low calorie
2      Beans baked beans       170 cals   80 calories    Low calorie
3   Beans dried (boiled)       180 cals  130 calories    Low calorie
4           Blackberries        25 cals   25 calories    Low calorie
5           Blackcurrant        30 cals   30 calories    Low calorie
6               Broccoli        27 cals       32 cals       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 cals       20 cals   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
```