```
# 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 19.1 MB/s e
Requirement already satisfied: pandas>=0.25.3 in /usr/local/lib/python
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist
Requirement already satisfied: distro in /usr/lib/python3/dist-package
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/li
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/d
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/d
```

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

```
import warnings

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

```
# filename variable of the paf file which needs to be uploaded into the fc
pdf_file = '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: Apr 01, 2024 4:58:47 PM org.apache.
WARNING: New fonts found, font cache will be re-built
Apr 01, 2024 4:58:47 PM org.apache.pdfbox.pdmodel.font.FileSystemFontP
WARNING: Building on-disk font cache, this may take a while
Apr 01, 2024 4:58:48 PM org.apache.pdfbox.pdmodel.font.FileSystemFontP
WARNING: Finished building on-disk font cache, found 17 fonts
```

[	BREADS & CEREALS	Portion size * \
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)

	per 100 grams (3.5 oz)	Unnamed: 0	energy content
0	310 cals	NaN	Medium
1	480 cals	NaN	High
2	370 cals	NaN	Med-High
3	240 cals	NaN	Medium
4	220 cals	NaN	Low-med
5	300 cals	NaN	Medium
6	370 cals	NaN	Med-High
7	325 cals	NaN	Low Calorie
8	440 cals	NaN	Low / portion
9	198 cals	NaN	Low-Med
10	500 cals	NaN	High
11	95 cals	NaN	Low calorie
12	390 cals	NaN	Med-high
13	320 cals	NaN	Medium

```
     14                70 cals         NaN    Low calorie
     15               110 cals         NaN    Low calorie
     16               105 cals         NaN    Low calorie
     17                55 cals         NaN    Low calorie
     18                70 cals         NaN    Low calorie
     19               140 cals         NaN         Medium  ]
```

```
# 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 val
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 grams (3.5 oz) energy content
0	Bagel (1 average)	140 cals (45g)	310 cals	Medium
1	Biscuit digestives	86 cals (per biscuit)	480 cals	High
2	Jaffa cake	48 cals (per biscuit)	370 cals	Med-High
3	Bread white (thick slice)	96 cals (1 slice 40g)	240 cals	Medium
4	Bread wholemeal (thick)	88 cals (1 slice 40g)	220 cals	Low-med
5	Chapatis	250 cals	300 cals	Medium
6	Cornflakes	130 cals (35g)	370 cals	Med-High
7	Crackerbread	17 cals per slice	325 cals	Low Calorie
8	Cream crackers	35 cals (per cracker)	440 cals	Low / portion
9	Crumpets	93 cals (per crumpet)	198 cals	Low-Med
10	Flapjacks basic fruit mix	320 cals	500 cals	High
11	Macaroni (boiled)	238 cals (250g)	95 cals	Low calorie
12	Muesli	195 cals (50g)	390 cals	Med-high
13	Naan bread (normal)	300 cals (small plate size)	320 cals	Medium
14	Noodles (boiled)	175 cals (250g)	70 cals	Low calorie
15	Pasta (normal boiled)	330 cals (300g)	110 cals	Low calorie
16	Pasta (wholemeal boiled)	315 cals (300g)	105 cals	Low calorie
17	Porridge oats (with water)	193 cals (350g)	55 cals	Low calorie
18	Potatoes** (boiled)	210 cals (300g)	70 cals	Low calorie
19	Potatoes** (roast)	420 cals (300g)	140 cals	Medium

```
# extract data from page 1 of the paf 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       Mediu
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)
```

```
[              BREADS & CEREALS              Portion size *  \
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)

   per 100 grams (3.5 oz)  Unnamed: 0 energy content
0               310 cals         NaN         Medium
1               480 cals         NaN           High
2               370 cals         NaN       Med-High
3               240 cals         NaN         Medium
```

4	220 cals	NaN	Low-med			
5	300 cals	NaN	Medium			
6	370 cals	NaN	Med-High			
7	325 cals	NaN	Low Calorie			
8	440 cals	NaN	Low / portion			
9	198 cals	NaN	Low-Med			
10	500 cals	NaN	High			
11	95 cals	NaN	Low calorie			
12	390 cals	NaN	Med-high			
13	320 cals	NaN	Medium			
14	70 cals	NaN	Low calorie			
15	110 cals	NaN	Low calorie			
16	105 cals	NaN	Low calorie			
17	55 cals	NaN	Low calorie			
18	70 cals	NaN	Low calorie			
19	140 cals	NaN	Medium	,	Rice (whit	
0	NaN	NaN	NaN	NaN		
1	Rice (egg-fried)	500 cals	200 cals	NaN		
2	Rice (Brown)	405 cals (300g)	135 cals	NaN		
3	Rice cakes	28 Cals = 1 slice	373 Cals	NaN		
4	Ryvita Multi grain	37 Cals per slice	331 Cals	NaN		
5	Ryvita + seed & Oats	180 Cals 4 slices	362 Cals	NaN		
6	Spaghetti (boiled)	303 cals (300g)	101 cals	NaN		

	Low calorie					
0	NaN					
1	High in portion					
2	Low calorie					
3	Medium					
4	Medium					
5	Medium					
6	Low calorie	,	Unnamed: 0		Unnamed: 1	

```
# 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,
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
```

```
Start coding or generate with AI.
```

```
Start coding or generate with AI.
```