HPE TEXT SUMMARIZER

PDF DATA EXTRACTION

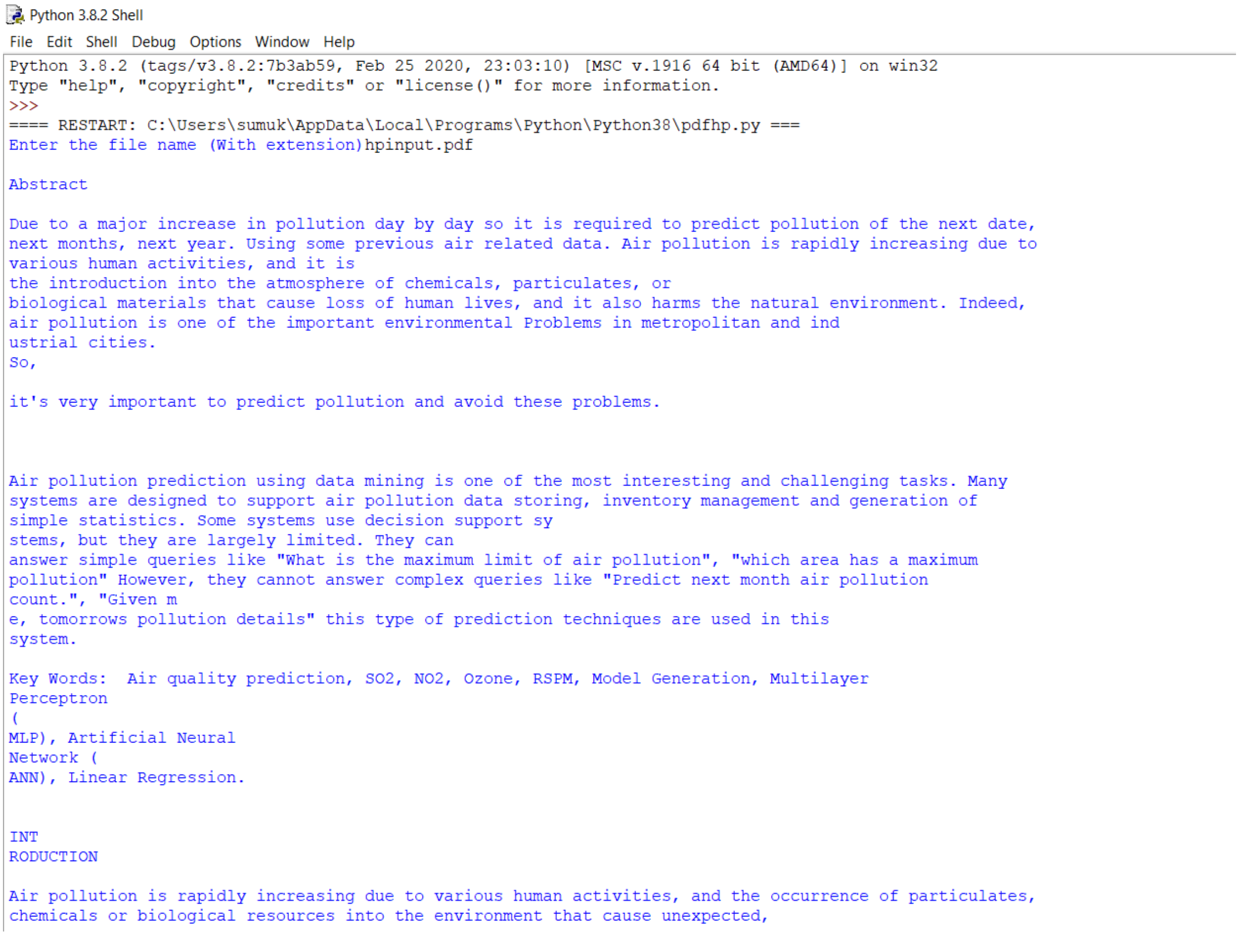
Explored few tools which could help in pdf data extraction.

1. PyPDF2

PyPDF2 is a pdf data extraction tool which we have used to extract data from pdf files. The text and the tabulated data will be extracted using this tool. Images which are present in pdf files are also extracted but the extraction is the image itself. The image itself will be stored in a .png format and the data from the image will not be extracted by this tool. Moreover, the tabulated data is extracted inline with the other text itself. The output breaks the words at places where it should not have done.

Here are the output screenshots:

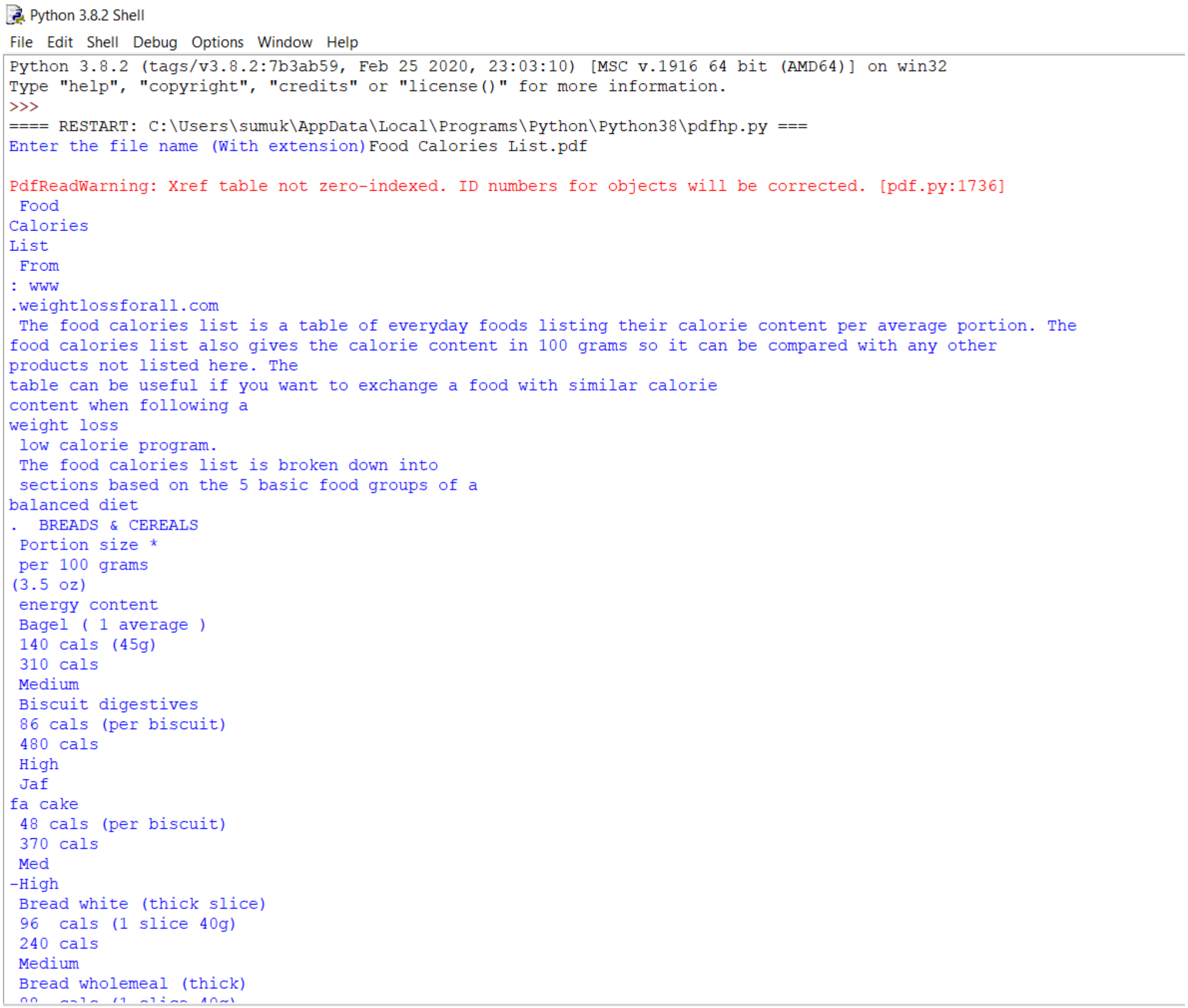
a. Output shown with input being just a text file.



```
Python 3.8.2 (tags/v3.8.2:7b3ab59, Feb 25 2020, 23:03:10) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
==== RESTART: C:\Users\sumuk\AppData\Local\Programs\Python\Python38\pdfhp.py ===
Enter the file name (With extension)hpinput.pdf

Abstract

Due to a major increase in pollution day by day so it is required to predict pollution of the next date,
next months, next year. Using some previous air related data. Air pollution is rapidly increasing due to
various human activities, and it is
the introduction into the atmosphere of chemicals, particulates, or
biological materials that cause loss of human lives, and it also harms the natural environment. Indeed,
air pollution is one of the important environmental Problems in metropolitan and ind
ustrial cities.
So,

it's very important to predict pollution and avoid these problems.



Air pollution prediction using data mining is one of the most interesting and challenging tasks. Many
systems are designed to support air pollution data storing, inventory management and generation of
simple statistics. Some systems use decision support sy
stems, but they are largely limited. They can
answer simple queries like "What is the maximum limit of air pollution", "which area has a maximum
pollution" However, they cannot answer complex queries like "Predict next month air pollution
count.", "Given m
e, tomorrows pollution details" this type of prediction techniques are used in this
system.

Key Words:  Air quality prediction, SO2, NO2, Ozone, RSPM, Model Generation, Multilayer
Perceptron
(
MLP), Artificial Neural
Network (
ANN), Linear Regression.


INT
RODUCTION

Air pollution is rapidly increasing due to various human activities, and the occurrence of particulates,
chemicals or biological resources into the environment that cause unexpected,
```

Explanation: This output as we have mentioned earlier creates blank spaces and cuts the words at wrong places. For example, in the 7th line of output, the word industrial is broken into two which makes it difficult to read. Similarly, at the end we can see the word INTRODUCTION also being divided into two.

b. Output shown with input being a text file with tables in it.



```
Python 3.8.2 (tags/v3.8.2:7b3ab59, Feb 25 2020, 23:03:10) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
==== RESTART: C:\Users\sumuk\AppData\Local\Programs\Python\Python38\pdfhp.py ===
Enter the file name (With extension)Food Calories List.pdf

PdfReadWarning: Xref table not zero-indexed. ID numbers for objects will be corrected. [pdf.py:1736]
 Food
Calories
List
 From
: www
.weightlossforall.com
 The food calories list is a table of everyday foods listing their calorie content per average portion. The
food calories list also gives the calorie content in 100 grams so it can be compared with any other
products not listed here. The
table can be useful if you want to exchange a food with similar calorie
content when following a
weight loss
 low calorie program.
 The food calories list is broken down into
 sections based on the 5 basic food groups of a
balanced diet
.  BREADS & CEREALS
 Portion size *
 per 100 grams
(3.5 oz)
 energy content
 Bagel ( 1 average )
 140 cals (45g)
 310 cals
 Medium
 Biscuit digestives
 86 cals (per biscuit)
 480 cals
 High
 Jaf
fa cake
 48 cals (per biscuit)
 370 cals
 Med
-High
 Bread white (thick slice)
 96  cals (1 slice 40g)
 240 cals
 Medium
 Bread wholemeal (thick)
 98  cals (1 slice 40g)
```

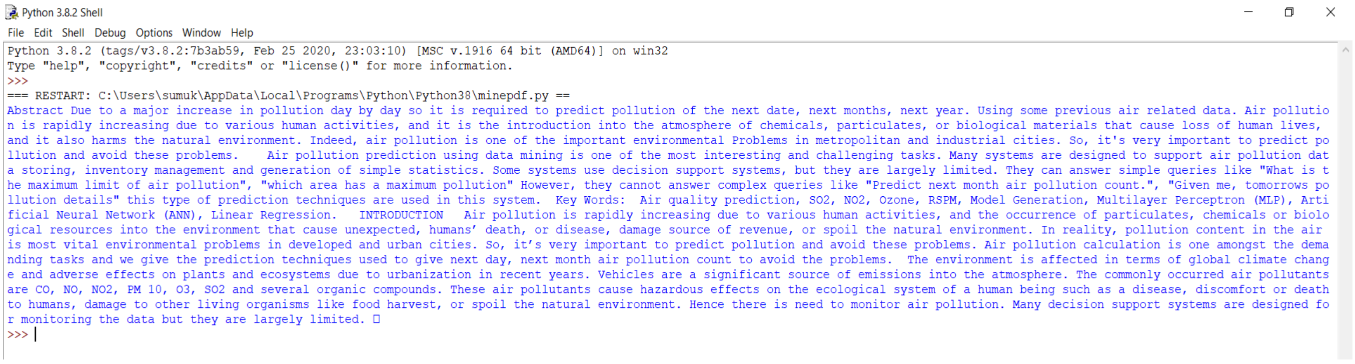
Explanation: This is the output with a file containing both text as well as tables in it. The first few lines are the text part and then the table data starts which is difficult to identify because the data extracted from tables is inline with the other text in the pdf file.

2. PDFMiner

PDFMiner is data extraction tool which is used to extract data from pdf files. The text and the tabulated data are extracted using this tool. PDFMiner extracts the text and displays it in a better fashion than the PyPdf2 tool which was stated above. The text extracted is similar to the input file and in the proper format too which was not the case with PyPdf2. However, the table data again is inline with the text in this tool too.

Here are the output screenshots:

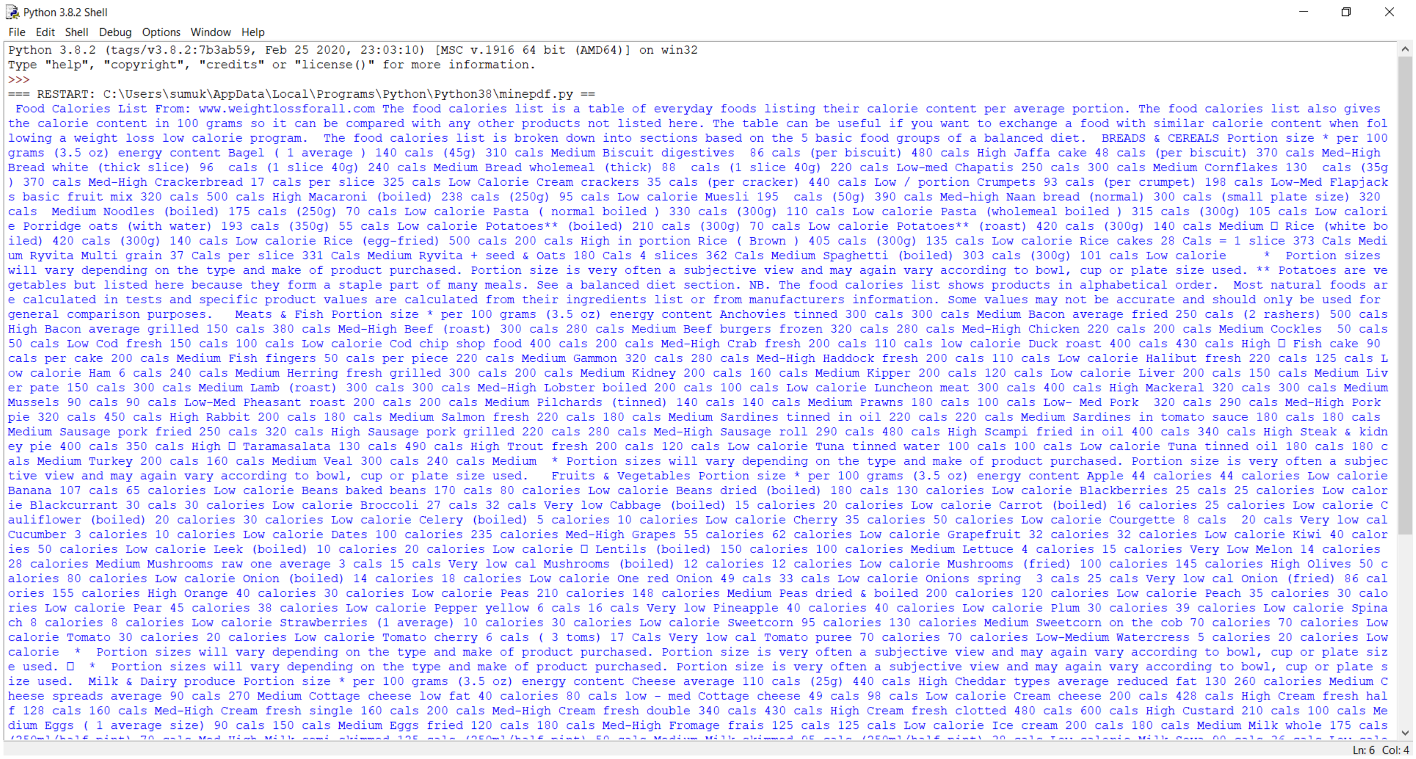
a. Output shown with input being just a text file.



```
Python 3.8.2 (tags/v3.8.2:7b3ab59, Feb 25 2020, 23:03:10) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> 
=== RESTART: C:\Users\sumuk\AppData\Local\Programs\Python\Python38\minepdf.py ==
Abstract Due to a major increase in pollution day by day so it is required to predict pollution of the next date, next months, next year. Using some previous air related data. Air pollutio
n is rapidly increasing due to various human activities, and it is the introduction into the atmosphere of chemicals, particulates, or biological materials that cause loss of human lives,
and it also harms the natural environment. Indeed, air pollution is one of the important environmental Problems in metropolitan and industrial cities. So, it's very important to predict po
llution and avoid these problems.    Air pollution prediction using data mining is one of the most interesting and challenging tasks. Many systems are designed to support air pollution dat
a storing, inventory management and generation of simple statistics. Some systems use decision support systems, but they are largely limited. They can answer simple queries like "What is t
he maximum limit of air pollution", "which area has a maximum pollution" However, they cannot answer complex queries like "Predict next month air pollution count.", "Given me, tomorrows po
llution details" this type of prediction techniques are used in this system.  Key Words:  Air quality prediction, SO2, NO2, Ozone, RSPM, Model Generation, Multilayer Perceptron (MLP), Arti
ficial Neural Network (ANN), Linear Regression.   INTRODUCTION   Air pollution is rapidly increasing due to various human activities, and the occurrence of particulates, chemicals or biolo
gical resources into the environment that cause unexpected, humans' death, or disease, damage source of revenue, or spoil the natural environment. In reality, pollution content in the air
is most vital environmental problems in developed and urban cities. So, it's very important to predict pollution and avoid these problems. Air pollution calculation is one amongst the dema
nding tasks and we give the prediction techniques used to give next day, next month air pollution count to avoid the problems.  The environment is affected in terms of global climate chang
e and adverse effects on plants and ecosystems due to urbanization in recent years. Vehicles are a significant source of emissions into the atmosphere. The commonly occurred air pollutants
are CO, NO, NO2, PM 10, O3, SO2 and several organic compounds. These air pollutants cause hazardous effects on the ecological system of a human being such as a disease, discomfort or death
to humans, damage to other living organisms like food harvest, or spoil the natural environment. Hence there is need to monitor air pollution. Many decision support systems are designed fo
r monitoring the data but they are largely limited. □
>>> |
```

Explanation: The output is shown properly as far as text alone is concerned with this tool. There is no breaking of words as in the case of PyPdf2.

b. Output shown with input being a text file with tables in it.



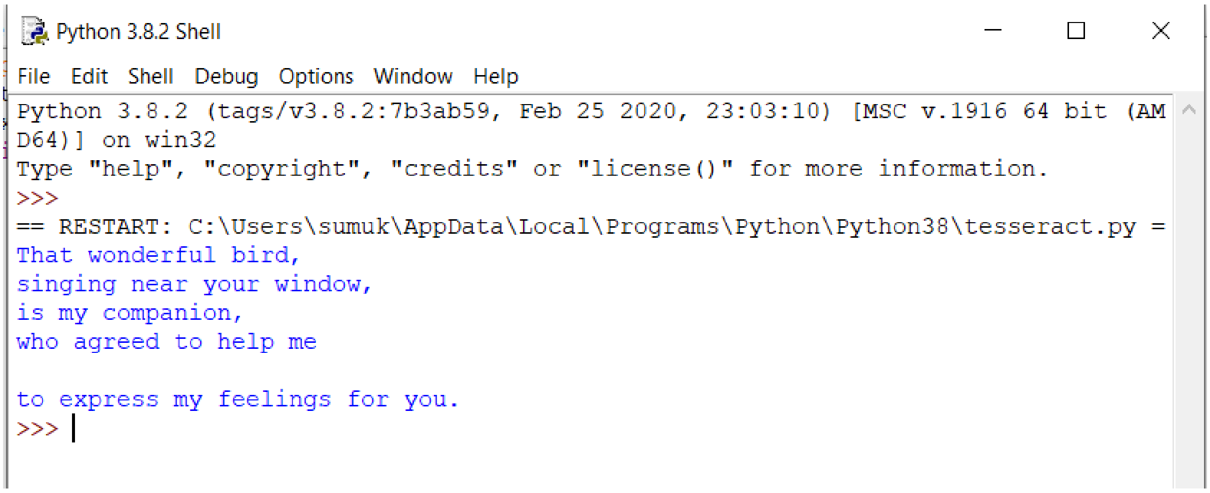
```
Python 3.8.2 (tags/v3.8.2:7b3ab59, Feb 25 2020, 23:03:10) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> 
=== RESTART: C:\Users\sumuk\AppData\Local\Programs\Python\Python38\minepdf.py ==
 Food Calories List From: www.weightlossforall.com The food calories list is a table of everyday foods listing their calorie content per average portion. The food calories list also gives
the calorie content in 100 grams so it can be compared with any other products not listed here. The table can be useful if you want to exchange a food with similar calorie content when fol
lowing a weight loss low calorie program.  The food calories list is broken down into sections based on the 5 basic food groups of a balanced diet.  BREADS & CEREALS Portion size * per 100
grams (3.5 oz) energy content Bagel ( 1 average ) 140 cals (45g) 310 cals Medium Biscuit digestives  86 cals (per biscuit) 480 cals High Jaffa cake 48 cals (per biscuit) 370 cals Med-High
Bread white (thick slice) 96  cals (1 slice 40g) 240 cals Medium Bread wholemeal (thick) 88  cals (1 slice 40g) 220 cals Low-med Chapatis 250 cals 300 cals Medium Cornflakes 130  cals (35g
) 370 cals Med-High Crackerbread 17 cals per slice 325 cals Low Calorie Cream crackers 35 cals (per cracker) 440 cals Low / portion Crumpets 93 cals (per crumpet) 198 cals Low-Med Flapjack
s basic fruit mix 320 cals 500 cals High Macaroni (boiled) 238 cals (250g) 95 cals Low calorie Muesli 195  cals (50g) 390 cals Med-high Naan bread (normal) 300 cals (small plate size) 320
cals  Medium Noodles (boiled) 175 cals (250g) 70 cals Low calorie Pasta ( normal boiled ) 330 cals (300g) 110 cals Low calorie Pasta (wholemeal boiled ) 315 cals (300g) 105 cals Low calori
e Porridge oats (with water) 193 cals (350g) 55 cals Low calorie Potatoes** (boiled) 210 cals (300g) 70 cals Low calorie Potatoes** (roast) 420 cals (300g) 140 cals Medium □ Rice (white bo
iled) 420 cals (300g) 140 cals Low calorie Rice (egg-fried) 500 cals 200 cals High in portion Rice ( Brown ) 405 cals (300g) 135 cals Low calorie Rice cakes 28 Cals = 1 slice 373 Cals Medi
um Ryvita Multi grain 37 Cals per slice 331 Cals Medium Ryvita + seed & Oats 180 Cals 4 slices 362 Cals Medium Spaghetti (boiled) 303 cals (300g) 101 cals Low calorie     *  Portion sizes
will vary depending on the type and make of product purchased. Portion size is very often a subjective view and may again vary according to bowl, cup or plate size used. ** Potatoes are ve
getables but listed here because they form a staple part of many meals. See a balanced diet section. NB. The food calories list shows products in alphabetical order.  Most natural foods ar
e calculated in tests and specific product values are calculated from their ingredients list or from manufacturers information. Some values may not be accurate and should only be used for
general comparison purposes.   Meats & Fish Portion size * per 100 grams (3.5 oz) energy content Anchovies tinned 300 cals 300 cals Medium Bacon average fried 250 cals (2 rashers) 500 cals
High Bacon average grilled 150 cals 380 cals Med-High Beef (roast) 300 cals 280 cals Medium Beef burgers frozen 320 cals 280 cals Med-High Chicken 220 cals 200 cals Medium Cockles  50 cals
50 cals Low Cod fresh 150 cals 100 cals Low calorie Cod chip shop food 400 cals 200 cals Med-High Crab fresh 200 cals 110 cals low calorie Duck roast 400 cals 430 cals High □ Fish cake 90
cals per cake 200 cals Medium Fish fingers 50 cals per piece 220 cals Medium Gammon 320 cals 280 cals Med-High Haddock fresh 200 cals 110 cals Low calorie Halibut fresh 220 cals 125 cals L
ow calorie Ham 6 cals 240 cals Medium Herring fresh grilled 300 cals 200 cals Medium Kidney 200 cals 160 cals Medium Kipper 200 cals 120 cals Low calorie Liver 200 cals 150 cals Medium Liv
er pate 150 cals 300 cals Medium Lamb (roast) 300 cals 300 cals Med-High Lobster boiled 200 cals 100 cals Low calorie Luncheon meat 300 cals 400 cals High Mackeral 320 cals 300 cals Medium
Mussels 90 cals 90 cals Low-Med Pheasant roast 200 cals 200 cals Medium Pilchards (tinned) 140 cals 140 cals Medium Prawns 180 cals 100 cals Low- Med Pork  320 cals 290 cals Med-High Pork
pie 320 cals 450 cals High Rabbit 200 cals 180 cals Medium Salmon fresh 220 cals 180 cals Medium Sardines tinned in oil 220 cals 220 cals Medium Sardines in tomato sauce 180 cals 180 cals
Medium Sausage pork fried 250 cals 320 cals High Sausage pork grilled 220 cals 280 cals Med-High Sausage roll 290 cals 480 cals High Scampi fried in oil 400 cals 340 cals High Steak & kidn
ey pie 400 cals 350 cals High □ Taramasalata 130 cals 490 cals High Trout fresh 200 cals 120 cals Low calorie Tuna tinned water 100 cals 100 cals Low calorie Tuna tinned oil 180 cals 180 c
als Medium Turkey 200 cals 160 cals Medium Veal 300 cals 240 cals Medium  * Portion sizes will vary depending on the type and make of product purchased. Portion size is very often a subjec
tive view and may again vary according to bowl, cup or plate size used.   Fruits & Vegetables Portion size * per 100 grams (3.5 oz) energy content Apple 44 calories 44 calories Low calorie
Banana 107 cals 65 calories Low calorie Beans baked beans 170 cals 80 calories Low calorie Beans dried (boiled) 180 cals 130 calories Low calorie Blackberries 25 cals 25 calories Low calor
ie Blackcurrant 30 cals 30 calories Low calorie Broccoli 27 cals 32 cals Very low Cabbage (boiled) 15 calories 20 calories Low calorie Carrot (boiled) 16 calories 25 calories Low calorie C
auliflower (boiled) 20 calories 30 calories Low calorie Celery (boiled) 5 calories 10 calories Low calorie Cherry 35 calories 50 calories Low calorie Courgette 8 cals  20 cals Very low cal
Cucumber 3 calories 10 calories Low calorie Dates 100 calories 235 calories Med-High Grapes 55 calories 62 calories Low calorie Grapefruit 32 calories 32 calories Low calorie Kiwi 40 calor
ies 50 calories Low calorie Leek (boiled) 10 calories 20 calories Low calorie □ Lentils (boiled) 150 calories 100 calories Medium Lettuce 4 calories 15 calories Very Low Melon 14 calories
28 calories Medium Mushrooms raw one average 3 cals 15 cals Very low cal Mushrooms (boiled) 12 calories 12 calories Low calorie Mushrooms (fried) 100 calories 145 calories High Olives 50 c
alories 80 calories Low calorie Onion (boiled) 14 calories 18 calories Low calorie One red Onion 49 cals 33 cals Low calorie Onions spring  3 cals 25 cals Very low cal Onion (fried) 86 cal
ories 155 calories High Orange 40 calories 30 calories Low calorie Peas 210 calories 148 calories Medium Peas dried & boiled 200 calories 120 calories Low calorie Peach 35 calories 30 calo
ries Low calorie Pear 45 calories 38 calories Low calorie Pepper yellow 6 cals 16 cals Very low Pineapple 40 calories 40 calories Low calorie Plum 30 calories 39 calories Low calorie Spina
ch 8 calories 8 calories Low calorie Strawberries (1 average) 10 calories 30 calories Low calorie Sweetcorn 95 calories 130 calories Medium Sweetcorn on the cob 70 calories 70 calories Low
calorie Tomato 30 calories 20 calories Low calorie Tomato cherry 6 cals ( 3 toms) 17 Cals Very low cal Tomato puree 70 calories 70 calories Low-Medium Watercress 5 calories 20 calories Low
calorie  *  Portion sizes will vary depending on the type and make of product purchased. Portion size is very often a subjective view and may again vary according to bowl, cup or plate siz
e used. □  *  Portion sizes will vary depending on the type and make of product purchased. Portion size is very often a subjective view and may again vary according to bowl, cup or plate s
ize used.  Milk & Dairy produce Portion size * per 100 grams (3.5 oz) energy content Cheese average 110 cals (25g) 440 cals High Cheddar types average reduced fat 130 260 calories Medium C
heese spreads average 90 cals 270 Medium Cottage cheese low fat 40 calories 80 cals low - med Cottage cheese 49 cals 98 cals Low calorie Cream cheese 200 cals 428 cals High Cream fresh hal
f 128 cals 160 cals Med-High Cream fresh single 160 cals 200 cals Med-High Cream fresh double 340 cals 430 cals High Cream fresh clotted 480 cals 600 cals High Custard 210 cals 100 cals Me
dium Eggs ( 1 average size) 90 cals 150 cals Medium Eggs fried 120 cals 180 cals Med-High Fromage frais 125 cals 125 cals Low calorie Ice cream 200 cals 180 cals Medium Milk whole 175 cals
```

Explanation: This is the table file output using PDFMiner tool. Like in the PyPdf2 this also displays text extracted from tables in line itself which makes it difficult to read and understand.

3. Tesseract

This is tool mainly used to extract data from images. This takes a image file as input and extracts data from it, if any.

Here is the output screenshot:

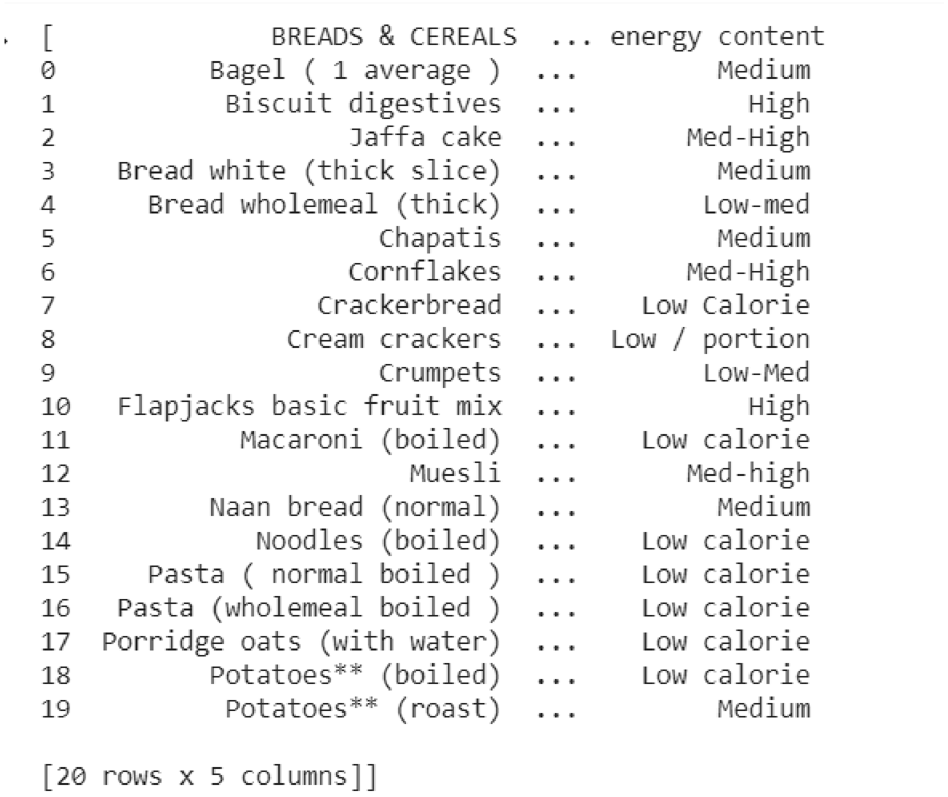


```
Python 3.8.2 (tags/v3.8.2:7b3ab59, Feb 25 2020, 23:03:10) [MSC v.1916 64 bit (AM
D64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
== RESTART: C:\Users\sumuk\AppData\Local\Programs\Python\Python38\tesseract.py =
That wonderful bird,
singing near your window,
is my companion,
who agreed to help me

to express my feelings for you.
>>> 
```

4. Tabula

This tool is mainly used to extract data from tables itself. The data extracted will be in a tabulated format itself which makes it easy to read and observe.

Here is the output screenshot:



```
[                 BREADS & CEREALS  ... energy content
0              Bagel ( 1 average )  ...         Medium
1               Biscuit digestives  ...           High
2                       Jaffa cake  ...       Med-High
3      Bread white (thick slice)    ...         Medium
4         Bread wholemeal (thick)   ...        Low-med
5                         Chapatis  ...         Medium
6                       Cornflakes  ...       Med-High
7                     Crackerbread  ...    Low Calorie
8                   Cream crackers  ...  Low / portion
9                         Crumpets  ...        Low-Med
10     Flapjacks basic fruit mix    ...           High
11               Macaroni (boiled)  ...    Low calorie
12                          Muesli  ...       Med-high
13            Naan bread (normal)   ...         Medium
14                 Noodles (boiled) ...    Low calorie
15        Pasta ( normal boiled )   ...    Low calorie
16      Pasta (wholemeal boiled )   ...    Low calorie
17   Porridge oats (with water)     ...    Low calorie
18           Potatoes** (boiled)    ...    Low calorie
19            Potatoes** (roast)    ...         Medium

[20 rows x 5 columns]]
```

Queries:

1. By examining the above tools, we would be using pdfminer as data extraction tool for extracting data from pdf files.
2. Tabula and Tesseract are working fine and giving us the desired results but the problem is that they need to be incorporated with pdfminer in order to extract data from them well.
3. Tesseract takes image file as input and is not working with pdf files as input. Moreover, the thing is that the tool should judge an image and extract data from it by itself from the file, but when we combine tesseract with pdfminer, the output is not working well.
4. The problem with tabula is also the same. It needs to be incorporated with pdfminer. The problem is when it is incorporated with pdfminer, pdfminer itself will extract data from the tables and tabula part will not do its work.
5. Any ideas as to how to go about will help us further.