Importing the required libraries

Tutorial To Scrape PDFs from a Given URL

This tutorial will help us to download all the PDFs in a given URL. In addition to downloading the PDF, this tutorial also helps us in reading a PDF and saving a table from the PDF to a conservative structured format like a CSV.

In [ ]:

```
import os
import requests
import urllib.request
import pandas as pd
from urllib.parse import urljoin
from bs4 import BeautifulSoup
```

In [ ]:

```
# Tabula scrapes tables from PDFs
!pip install tabula-py
import tabula
```

```
Collecting tabula-py
  Downloading tabula_py-2.3.0-py3-none-any.whl (12.0 MB)
     |████████████████████████████████| 12.0 MB 98 kB/s
Requirement already satisfied: pandas>=0.25.3 in /usr/local/lib/python3.7/dist-packages (
from tabula-py) (1.1.5)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from tabu
la-py) (1.19.5)
Collecting distro
  Downloading distro-1.6.0-py2.py3-none-any.whl (19 kB)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-pa
ckages (from pandas>=0.25.3->tabula-py) (2.8.2)
Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.7/dist-packages (fr
om pandas>=0.25.3->tabula-py) (2018.9)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from p
ython-dateutil>=2.7.3->pandas>=0.25.3->tabula-py) (1.15.0)
Installing collected packages: distro, tabula-py
Successfully installed distro-1.6.0 tabula-py-2.3.0
```

In [ ]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

The url input for dowloading pdfs and setting up output folder path

In [ ]:

```
# Save contents from url into folder_location
url = 'https://www.premierleague.com/publications'
folder_location = r'/content/drive/MyDrive/Colab Notebooks/premier_league'
if not os.path.exists(folder_location):
    os.mkdir(folder_location)
```

Actual Code

In [ ]:

```
response = requests.get(url)
soup = BeautifulSoup(response.text, "html.parser")
```

```
# Loop through all PDF links in the page
for link in soup.select("a[href$='.pdf']"):
    # Local lile name is the same as PDF file name in the URL (ignoring rest of the path)
    # https://premierleague-static-files.s3.amazonaws.com/premierleague/document/2016/07/02/e1648e96-4eeb-456e-8ce0-d937d2bc7649/2011-12-premier-league-season-review.pdf
    filename = os.path.join(folder_location, link['href'].split('/')[-1])
    with open(filename, 'wb') as f:
        f.write(requests.get(urljoin(url,link['href'])).content)
```

Reading a table from a PDF document and storing it in a csv file

In [ ]:

```
combined_pdf = folder_location + "/This-is-PL-Interactive-Combined.pdf"
tabula.read_pdf(combined_pdf,  pages='18')
```

Out[ ]:

```
[                                          Unnamed: 0  ... Total payment
 0                                                NaN  ...       £149.4m
 1                                                NaN  ...       £149.8m
 2                                                NaN  ...       £144.4m
 3                                                NaN  ...       £145.9m
 4                                                NaN  ...       £141.7m
 5                                                NaN  ...       £142.0m
 6                                                NaN  ...       £119.8m
 7                                                NaN  ...       £128.0m
 8                                                NaN  ...       £118.2m
 9                                               £100m  ...       £123.0m
 10                                               NaN  ...           NaN
 11                                               NaN  ...           NaN
 12                                  The amount the  ...       £114.3m
 13                                  Premier League  ...           NaN
 14                           invests per season in  ...       £111.2m
 15                              the development of  ...           NaN
 16                           community facilities,  ...       £116.1m
 17          sports participation, Central payments to  ...           NaN
 18                community and schools clubs 2017/18  ...       £106.3m
 19  programmes, and to The collective and central ...  ...           NaN
 20  support the well-being Premier League markets ...  ...       £107.7m
 21                               of players in lower  ...           NaN
 22  make using these central payments, leagues. Th...  ...       £102.4m
 23  League also supports all of them to not only i...  ...           NaN
 24  the English Football talented players and faci...  ...       £107.2m
 25  League with a further also their local communi...  ...           NaN
 26        £100m per season of wider football pyramid.  ...        £98.5m
 27   Solidarity payments The income generated by fans  ...           NaN
 28                           and ring-fenced Youth  ...        £98.9m
 29  Development grants. watching compelling matche...  ...           NaN
 30                the Premier League is what allows  ...        £94.7m

 [31 rows x 12 columns]]
```

In [ ]:

```
from tabula import convert_into

convert_into(combined_pdf, folder_location +"/table_output.csv", output_format="csv",pages = 18,area=[[275,504,640,900]])
pd.read_csv(folder_location+"/table_output.csv")
```

Out[ ]:

	Pos	Unnamed: 1	Club	W	D	L	GD	Pts	Total payment
0	1.0	NaN	Manchester City	32.0	4.0	2.0	79.0	100.0	£149.4m
1	2.0	NaN	Manchester United	25.0	6.0	7.0	40.0	81.0	£149.8m
2	3.0	NaN	Tottenham Hotspur	23.0	8.0	7.0	38.0	77.0	£144.4m
3	4.0	NaN	Liverpool	21.0	12.0	5.0	46.0	75.0	£145.9m

	Pos	Unnamed: 1	Club	W	D	L	GD	Pts	Total payment
4	5.0	NaN	Chelsea	21.0	7.0	10.0	24.0	70.0	£141.7m
5	6.0	NaN	Arsenal	19.0	6.0	13.0	23.0	63.0	£142.0m
6	7.0	NaN	Burnley	14.0	12.0	12.0	-3.0	54.0	£119.8m
7	8.0	NaN	Everton	13.0	10.0	15.0	-14.0	49.0	£128.0m
8	9.0	NaN	Leicester City	12.0	11.0	15.0	-4.0	47.0	£118.2m
9	10.0	NaN	Newcastle United	12.0	8.0	18.0	-8.0	44.0	£123.0m
10	NaN	125	NaN	NaN	NaN	NaN	NaN	NaN	NaN
11	NaN	E RY SA	NaN	NaN	NaN	NaN	NaN	NaN	NaN
12	11.0	NaN	Crystal Palace	11.0	11.0	16.0	-10.0	44.0	£114.3m
13	12.0	NaN	AFC Bournemouth	11.0	11.0	16.0	-16.0	44.0	£111.2m
14	13.0	NaN	West Ham United	10.0	12.0	16.0	-20.0	42.0	£116.1m
15	14.0	NaN	Watford	11.0	8.0	19.0	-20.0	41.0	£106.3m
16	15.0	NaN	Brighton & Hove Albion	9.0	13.0	16.0	-20.0	40.0	£107.7m
17	16.0	NaN	Huddersfield Town	9.0	10.0	19.0	-30.0	37.0	£102.4m
18	17.0	NaN	Southampton	7.0	15.0	16.0	-19.0	36.0	£107.2m
19	18.0	NaN	Swansea City	8.0	9.0	21.0	-28.0	33.0	£98.5m
20	19.0	NaN	Stoke City	7.0	12.0	19.0	-33.0	33.0	£98.9m
21	20.0	NaN	West Bromwich Albion	6.0	13.0	19.0	-25.0	31.0	£94.7m

In [ ]: