pdf_text_extract

April 11, 2023

1 Reading from PDF

Libraries that we need

[ ]:
```
from pypdf import PdfReader
import pandas as pd
```

[ ]:
```
# #we pass the page as input the pdfreader function from pypdf library
# reader = PdfReader('data/pdf/BatteryEnclosuresAcc.pdf')
# #using reader object to printing number of pages
# number_of_pages = len(reader.pages)
# print(number_of_pages)
# #now we can think pages as an array we can take a look at the first one
# page = reader.pages[0]
# # we can extract only text
# text = page.extract_text()
# # here is our first page text content
# text
```

the best format so far

[ ]:
```
category_names =␣
 ↪['CategoryName','Batteries','BatteryEnclosuresAcc','Chargers','Charging_Accessories','Contr
'DrivetrainComponents','EV_Conversion_Kits','Hose_Fittings','Instrumentation','Miscellaneous',
"Used_Components","Wiring_Parts"]
list_of_pdfs= [category_name +".pdf" for category_name in category_names]
list_of_csv= [category_name +".csv" for category_name in category_names]
```

[ ]:
```
def pdfReader(category,file_name_in,file_name_out):
    path_in = "data/pdf/"
    file_path_in = f"{path_in}{file_name_in}"
    reader = PdfReader(file_path_in)
    df = pd.DataFrame(columns=["category",'model', 'model_name',␣
 ↪'manufacturer', 'weight', 'price'])
    number_of_pages = len(reader.pages)
    data =[]
    category= category
    print("started")
```

```
    for page_num in range(number_of_pages):
        page = reader.pages[page_num]
        text = page.extract_text()
        chunks = text.split("Price:")
        for i,  chunk in  enumerate(chunks[1:], start=1): # Skip the first␣
↪chunk which is before the first "Price"
            #print("line", i)
            if "Show" in chunk:
                # If "Show" is in the chunk, break out of the loop
                break
            if "Model" in chunk:
                # Get the text between "Price" and "Model"
                model = chunk.split("Model")[0]
                # print("model Name")
                # print(model)
                # Get the text between "Model" and "Manufacturer"
                model_name = chunk.split("Model")[1].split("Manufacturer")[0]
                # print("model")
                # print(model_name)
                manufacturer = chunk.split("Manufacturer")[1].split("Weight")[0]
                # print("Manufacturer")
                # print(manufacturer)
                weight  = chunk.split("Weight")[1].split("\n")[0].strip()
                # print("weight")
                # print(weight)
                price  = chunk.split("Weight")[1].split("\n")[0].strip()
                # print("price")
                # print(price)
                # Add the data to the DataFrame as a new row
                df.loc[len(df.index)] = [category,model,model_name,␣
↪manufacturer,weight, price ]
            else:
                # print("Price:" + chunk)
                # print(" ")
                data.append(chunk)

    # # Create a pandas DataFrame from the data list
    # df = pd.DataFrame(data)
    # print("##########")
    # print("this is my df")
    # print(df)
    path_out = "data/ew_csv/"
    file_path_out = f"{path_out}{file_name_out}"
    df.to_csv(file_path_out, index=False)
    print(file_name_in, "done")
```

[ ]:
```
category_names =␣
 ↪['Batteries','BatteryEnclosuresAcc','Chargers','Charging_Accessories','Controller_Accessori
'DrivetrainComponents','EV_Conversion_Kits','Hose_Fittings','Instrumentation','Miscellaneous',
"Used_Components","Wiring_Parts"]
list_of_pdfs= [category_name +".pdf" for category_name in category_names]
list_of_csv= [category_name +".csv" for category_name in category_names]
# print(len(list_of_csv))
for item in range(len(category_names)):
    pdfReader(category=category_names[item],file_name_in=␣
 ↪list_of_pdfs[item],file_name_out=list_of_csv[item])
```

```
started
Batteries.pdf done
started
BatteryEnclosuresAcc.pdf done
started
Chargers.pdf done
started
Charging_Accessories.pdf done
started
Controller_Accessories.pdf done
started
Controllers.pdf done
started
DC_Converters.pdf done
started
DrivetrainComponents.pdf done
started
EV_Conversion_Kits.pdf done
started
Hose_Fittings.pdf done
started
Instrumentation.pdf done
started
Miscellaneous.pdf done
started
Motor_Accessories.pdf done
started
Motor_Adapters.pdf done
started
Motors.pdf done
started
Services.pdf done
started
Used_Components.pdf done
started
Wiring_Parts.pdf done
```

[ ]:
```
df = pd.DataFrame(columns=['category', 'model', 'model_name', 'manufacturer',␣
 ↪'weight', 'price'])
size =0
for item in range(len(list_of_csv)):
    path = "data/ew_csv/"
    file_path = f"{path}{list_of_csv[item]}"
    # print(file_path)
    df_new = pd.read_csv(file_path)
    size += len(df_new)
    df= pd.concat([df, df_new], axis=0)

print(len(df))
print(size)
df.to_csv("data/evWest.csv", index=False)
```

```
187
187
```

[ ]:
```
#testcode
# import pandas as pd
# # Create an empty list to store the data
# # Create an empty DataFrame to store the data
# df = pd.DataFrame(columns=["category",'model', 'model_name', 'manufacturer',␣
 ↪'weight', 'price'])
# category= "Battery Enclosures & Acc."
# reader = PdfReader('data/pdf/BatteryEnclosuresAcc.pdf')
# number_of_pages = len(reader.pages)
# data = []
# for page_num in range(number_of_pages):
#     page = reader.pages[page_num]
#     text = page.extract_text()
#     chunks = text.split("Price:")
#     for i,  chunk in  enumerate(chunks[1:], start=1): # Skip the first chunk␣
 ↪which is before the first "Price"
#         #print("line", i)
#         if "Show" in chunk:
#             # If "Show" is in the chunk, break out of the loop
#             break
#         if "Model" in chunk:
#             # Get the text between "Price" and "Model"
#             model = chunk.split("Model")[0]
#             print("model Name")
#             # print(model)
#             # Get the text between "Model" and "Manufacturer"
#             model_name = chunk.split("Model")[1].split("Manufacturer")[0]
#             # print("model")
#             # print(model_name)
```

```
#             manufacturer = chunk.split("Manufacturer")[1].split("Weight")[0]
#             # print("Manufacturer")
#             # print(manufacturer)
#             weight  = chunk.split("Weight")[1].split("\n")[0].strip()
#             # print("weight")
#             # print(weight)
#             price  = chunk.split("Weight")[1].split("\n")[0].strip()
#             # print("price")
#             # print(price)
#             # Add the data to the DataFrame as a new row
#             df.loc[len(df.index)] = [category,model,model_name,
 ↪manufacturer,weight, price ]
#         else:
#             # print("Price:" + chunk)
#             # print(" ")
#             data.append(chunk)

# # # Create a pandas DataFrame from the data list
# # df = pd.DataFrame(data)
# print("#########")
# print("this is my df")
# print(df)
```

```
#########
this is my df
                    category  \
0  Battery Enclosures & Acc.
1  Battery Enclosures & Acc.
2  Battery Enclosures & Acc.
3  Battery Enclosures & Acc.

                                                 model          model_name  \
0    MX150 Bulkhead and Connector -\nBattery Box …    : EVW -\nMX150\n
1    Rincon Power HVBD4AXR - 400 A\nContinuous - …         : RP-HVBD\n
2    VW Beetle Front Battery Box - '58-\n71\nAlum…       : EVW -FBB1\n
3    VW Beetle Rear Battery Box - '58-\n71\nAlumi…       : EVW -RBB1\n

         manufacturer   weight    price
0          : EV West\n   : 1.00   : 1.00
1   : Rincon\nPower\n   : 3.00   : 3.00
2          : EV West\n  : 20.00  : 20.00
3          : EV West\n  : 20.00  : 20.00
```