Narges - Extracting PDF tabular data to Excel

December 17, 2020

[751]:
```
import glob
import tabula
import pandas as pd
import re
```

[872]:
```
Files = glob.glob("/Users/kabbas/Dropbox/CBS/Python - Wavin/*.PDF")
for i in range(len(Files)):
        df = tabula.read_pdf(Files[i], lattice=True, pages = [7, 8],␣
 ↪multiple_tables=True)
        df_32 = df[0]
        df_33 = df[1]
        df_34 = df[2]
        ##Renaming 3.2
        df_32.rename(columns = {
            "Abiotic\rdepletion\r(non-fossil)":"ADPM",
           "Abiotic\rdepletion\r(fossil fuels)": "ADPE",
            "Acidification": "AP",
            "Eutrophication" : "EP",
           "Global\rwarming" : "GWP",
            "Ozone layer\rdepletion" : "ODP",
           "Photochemical\roxidation" : "POCP"},
            inplace=True)
        ##removing the first column
        df_32.pop("Impact\rcategory")
        ##inserting new first column
        Parameter_column = ["Unit", "A1-A3", "A4-A5", "B1-B7", "C1-C4", "Total"]
        df_32.insert(0, "Parameter", Parameter_column)
        ##minor adjustments to make sure it comes out in the correct format
        df_32.set_index("Parameter", inplace=True)
        df_32 = df_32.transpose()



        #### We do the same proces for 3.3 and 3.3 with regards to their␣
 ↪idiosyncrasies ####

        ##3.3##
```

```
        df_33.rename(columns =
        {"Environmental\rparameter": "Environmental parameter",
         "Use␣
↪of\rrenewable\rprimary\renergy\rexcluding\rrenewable\rprimary\renergy\rresources\rused␣
↪as raw\rmaterials" : "RPEER",
         "Use of\rrenewable\rprimary\renergy\rresources\rused as␣
↪raw\rmaterials" : "RPEOR",
         "Total use of\rrenewable\rprimary\renergy\rresources\r(primary\renergy␣
↪and\rprimary\renergy\rresources\rused as raw\rmaterials)" : "TRPE",
         "Use of␣
↪non\rrenewable\rprimary\renergy\rexcluding\rnon\rrenewable\rprimary\renergy\rresources\ruse
↪as raw\rmaterials" : "NRPEER",
         "Use of non\rrenewable\rprimary\renergy\rresources\rused as␣
↪raw\rmaterials" : "NRPEOR",
         "Total use\rof␣
↪non\rrenewable\rprimary\renergy\rresources\r(primary\renergy␣
↪and\rprimary\renergy\rresources\rused as raw\rmaterials)" : "TNRPE",
         "Use of\rsecondary\rmaterial" : "SM",
         "Use of\rrenewable\rsecondary\rfuels" : "RSF",
         "Use of non\rrenewable\rsecondary\rfuels" : "NRSF",
         "Net use of\rfresh water" : "NFW"
        }, inplace = True)


        df_33.pop("Environmental parameter")
        df_33.insert(0, "Parameter", Parameter_column)

        df_33.set_index("Parameter", inplace=True)
        df_33 = df_33.transpose()


        ##3.4##
        df_34.rename(columns=
        {
        "Environmental\rparameter" : "Environmental parameter",
        "Hazardous waste" : "HW",
        "Non-hazardous waste" : "NHW",
        "Nuclear waste" : "NW"
        },
        inplace = True)

        df_34.pop("Environmental parameter")
        df_34.insert(0, "Parameter", Parameter_column)

        df_34.set_index("Parameter", inplace=True)
        df_34 = df_34.transpose()
```

```
        ####We collect all the dataframes and put them in a finished dataframe␣
↪that will be exported to excel####
        finished_df = [df_32, df_33, df_34]

        ## We determine file path##
        out_path = r"/Users/kabbas/Dropbox/CBS/Python - Wavin/Excel/
↪"+Files[i][41:]
        xlwriter = pd.ExcelWriter(out_path+".xlsx", engine = "xlsxwriter")

        ##We save to excel
        for k in range(len(finished_df)):
            finished_df[k].to_excel(xlwriter, sheet_name = str(round(3.2+k*0.1,␣
↪3)),
                                    index=True, header=True,␣
↪index_label="Parameter")
        xlwriter.close()
```