Read TOC from PDF

This code in this Notebook uses pymupdf to read the contents of the UFGS table of contents PDF. It then uses regex to find the lines starting with "DIVISION" and puts those lines into a list. The list is split into two indicies: div number and title. The list is written to a pandas dataframe, then exported to Excel.

Jupyter Notebook written by Ben Fisher on 26 November 2024

benjamin.s.fisher@usace.army.mil

Imports

In [1]:

```
import os, datetime
import pymupdf as pypdf
import pandas as pd
import re
```

Standard Characters

In [2]:

```
form_feed = 12
new_line = '\n'
```

Paths

In [3]:

```
root_path = os.getcwd()
toc_file = root_path + '\\TOC\\UFGS TOC.pdf'
```

Get All Division Info

In [4]:

```
sections_raw = []
pattern = '^DIVISION'

with pypdf.open(toc_file) as doc:
    all_lines = chr(form_feed).join([page.get_text() for page in doc]).split(new_line)
    for line in all_lines:
        if not re.search(pattern, line) == None:
            sections_raw.append(line.strip())

sections = []
if len(sections_raw) > 0:
    for line in sections_raw:
        split_line = re.split(' - ', line)
        sections.append([re.split(' - ', line)[0][9:].title(),re.split(' - ', line)[1]

sections
```

Out[4]:

```
[['00', 'Procurement And Contracting Requirements'],
 ['01', 'General Requirements'],
 ['02', 'Existing Conditions'],
 ['03', 'Concrete'],
 ['04', 'Masonry'],
 ['05', 'Metals'],
 ['06', 'Wood, Plastics, And Composites'],
 ['07', 'Thermal And Moisture Protection'],
 ['08', 'Openings'],
 ['09', 'Finishes'],
 ['10', 'Specialties'],
 ['11', 'Equipment'],
 ['12', 'Furnishings'],
 ['13', 'Special Construction'],
 ['14', 'Conveying Equipment'],
 ['21', 'Fire Suppression'],
 ['22', 'Plumbing'],
 ['23', 'Heating, Ventilating, And Air Conditioning (Hvac)'],
 ['25', 'Integrated Automation'],
 ['26', 'Electrical'],
 ['27', 'Communications'],
 ['28', 'Electronic Safety And Security'],
 ['31', 'Earthwork'],
 ['32', 'Exterior Improvements'],
 ['33', 'Utilities'],
 ['34', 'Transportation'],
 ['35', 'Waterway And Marine Construction'],
 ['40', 'Process Interconnections'],
 ['41', 'Material Processing And Handling Equipment'],
 ['43', 'Process Gas And Liquid Handling, Purification, And Storage'],
 ['46', 'Water And Wastewater Equipment'],
 ['48', 'Electrical Power Generation']]
```

In [5]:

```
if len(sections) > 0:
    sections_toc_path = root_path + '\\UFGS TOC {:%Y%m%d %H%M%S}'.format(datetime.date
    df = pd.DataFrame(sections, columns=['Division','Title'])
    df
    df.to_excel(sections_toc_path, index=None)
```