pdf_table_extraction

January 1, 2021

[55]:
```
import tabula
import camelot
from pymongo import MongoClient
import json
```

[15]:
```
conn = MongoClient('localhost', 27017)
db = conn.pdf_table
```

[71]:
```
def pdf_table_reader(file, collection):
    tables = tabula.read_pdf(file, pages='all', multiple_tables=True,␣
 ↪silent=True)
    if len(tables) == 0:
        tables = camelot.read_pdf(file, pages='all', multiple_tables=True,␣
 ↪silent=True)
        for elt in tables:
            records = json.loads(elt.df.T.to_json()).values()
            collection.insert_many(records)
    else:
        for elt in tables:
            records = json.loads(elt.T.to_json()).values()
            collection.insert_many(records)
```

[80]:
```
pdf_table_reader('./Rec_Task/Rec_Task/1c1edeee-a13e-4b2e-90be-eb1dd03c3384.
 ↪pdf', db.first)
collection = db.first
cursor = collection.find()
for record in cursor:
    print(record)
```

```
{'_id': ObjectId('5feb637a3751063f552791db'), 'Date': 'April 06, 2018',
'Particulars': 'SBICAP Securities Ltd', 'Type of Interactlon': 'One on One'}
```

[81]:
```
pdf_table_reader('./Rec_Task/Rec_Task/a6b29367-f3b7-4fb1-a2d0-077477eac1d9.
 ↪pdf', db.second)
collection = db.second
cursor = collection.find()
for record in cursor:
```

```
    print(record)
```

```
{'_id': ObjectId('5feb639f3751063f552791dc'), 'Day& Date': 'Friday.', 'Name
ofOrganisation': 'HDFCSecurities Limited', 'Type ofMeeting': 'One-on-one Call'}
{'_id': ObjectId('5feb639f3751063f552791dd'), 'Day& Date': 'April 06, 2018',
'Name ofOrganisation': None, 'Type ofMeeting': None}
```

[85]:
```
pdf_table_reader('./Rec_Task/Rec_Task/d9f8e6d9-660b-4505-86f9-952e45ca6da0.
 ↪pdf', db.third)
collection = db.third
cursor = collection.find()
for record in cursor:
    print(record)
```

```
PdfReadWarning: Invalid stream (index 16) within object 41 0: Stream has ended
unexpectedly [pdf.py:1573]
```

```
{'_id': ObjectId('5feb64073751063f55279200'), '0': 'Date', '1': 'Name of the
analyst/investor', '2': 'Type', '3': 'Location'}
{'_id': ObjectId('5feb64073751063f55279201'), '0': 'April 4, 2018  Motilal Oswal
Asset Management', '1': 'Company Limited', '2': 'One-on-One \nmeeting', '3':
'Mumbai'}
{'_id': ObjectId('5feb64073751063f55279202'), '0': '', '1': 'Credit Suisse \nI',
'2': 'Voice call', '3': '- M'}
{'_id': ObjectId('5feb64073751063f55279203'), '0': '', '1': 'April 5, 2018
Maybank Eng Securities India Private \nLimited', '2': 'One-on-One \nmeeting',
'3': 'umbai'}
```

[86]:
```
pdf_table_reader('./Rec_Task/Rec_Task/EICHERMOT.pdf', db.fourth)
collection = db.fourth
cursor = collection.find()
for record in cursor:
    print(record)
```

```
{'_id': ObjectId('5feb64093751063f55279204'), 'STATUTORY REPORTS': '(ii)
Recommending the amount of expenditure', 'Unnamed: 0': '7.1 Major Terms of
Reference'}
{'_id': ObjectId('5feb64093751063f55279205'), 'STATUTORY REPORTS': 'to be
incurred on the activities referred to in', 'Unnamed: 0': '(i)  To assist the
Board in formulating the Risk'}
{'_id': ObjectId('5feb64093751063f55279206'), 'STATUTORY REPORTS': 'CSR
policy.', 'Unnamed: 0': 'Management Plan and practices.'}
{'_id': ObjectId('5feb64093751063f55279207'), 'STATUTORY REPORTS': '(iii)
Monitoring the CSR Policy of the Company', 'Unnamed: 0': '(ii) To monitor and
review risk management plan'}
{'_id': ObjectId('5feb64093751063f55279208'), 'STATUTORY REPORTS': 'from time to
time.', 'Unnamed: 0': 'and practices of the Company as approved'}
{'_id': ObjectId('5feb64093751063f55279209'), 'STATUTORY REPORTS': None,
```

'Unnamed: 0': 'by the Board.'}
{'_id': ObjectId('5feb64093751063f5527920a'), 'STATUTORY REPORTS': '6.2 Meetings and Attendance', 'Unnamed: 0': None}
{'_id': ObjectId('5feb64093751063f5527920b'), 'STATUTORY REPORTS': None, 'Unnamed: 0': '7.2 Members of the Committee'}
{'_id': ObjectId('5feb64093751063f5527920c'), 'STATUTORY REPORTS': 'T wo Meetings of the Corporate Social', 'Unnamed: 0': None}
{'_id': ObjectId('5feb64093751063f5527920d'), 'STATUTORY REPORTS': 'Responsibility Committee of the Company were', 'Unnamed: 0': None}
{'_id': ObjectId('5feb64093751063f5527920e'), 'STATUTORY REPORTS': 'held during the Financial Year 2016-17 on May', 'Unnamed: 0': 'Sl. Name Chairman/'}
{'_id': ObjectId('5feb64093751063f5527920f'), 'STATUTORY REPORTS': '5, 2016 and October 28, 2016. The names of the', 'Unnamed: 0': 'No. Member'}
{'_id': ObjectId('5feb64093751063f55279210'), 'STATUTORY REPORTS': 'members, chairperson and attendance details are', 'Unnamed: 0': '1. Mr Siddhartha Lal (Managing Director & Chairman'}
{'_id': ObjectId('5feb64093751063f55279211'), 'STATUTORY REPORTS': 'as under:', 'Unnamed: 0': 'Chief Executive Officer)'}
{'_id': ObjectId('5feb64093751063f55279212'), 'STATUTORY REPORTS': None, 'Unnamed: 0': '2. Mr S. Sandilya (Chairman and Non-Executive Member'}
{'_id': ObjectId('5feb64093751063f55279213'), 'STATUTORY REPORTS': None, 'Unnamed: 0': 'Independent Director)'}
{'_id': ObjectId('5feb64093751063f55279214'), 'STATUTORY REPORTS': 'Sl. Name Chairman/ No. of No. of', 'Unnamed: 0': '3. Mr Lalit Malik (Chief Financial Officer) Member'}
{'_id': ObjectId('5feb64093751063f55279215'), 'STATUTORY REPORTS': 'No. Member meetings meetings', 'Unnamed: 0': None}
{'_id': ObjectId('5feb64093751063f55279216'), 'STATUTORY REPORTS': 'held attended', 'Unnamed: 0': '8. SHARES COMMITTEE'}
{'_id': ObjectId('5feb64093751063f55279217'), 'STATUTORY REPORTS': '1. Mr S. Sandilya Chairman 2 2', 'Unnamed: 0': 'The Shares Committee of the Company consists of'}
{'_id': ObjectId('5feb64093751063f55279218'), 'STATUTORY REPORTS': '2. Mr Siddhartha Lal Member 2 2', 'Unnamed: 0': 'three members i.e. Mr Siddhartha Lal - Managing'}
{'_id': ObjectId('5feb64093751063f55279219'), 'STATUTORY REPORTS': None, 'Unnamed: 0': 'Director & Chief Executive Officer, Mr Lalit Malik - Chief'}
{'_id': ObjectId('5feb64093751063f5527921a'), 'STATUTORY REPORTS': '3. Mr Prateek Jalan Member 2 2', 'Unnamed: 0': None}
{'_id': ObjectId('5feb64093751063f5527921b'), 'STATUTORY REPORTS': None, 'Unnamed: 0': 'Financial Officer and Mr Manhar Kapoor - General'}
{'_id': ObjectId('5feb64093751063f5527921c'), 'STATUTORY REPORTS': 'M r Manhar Kapoor, General Counsel & Company', 'Unnamed: 0': 'Counsel & Company Secretary to look after and'}
{'_id': ObjectId('5feb64093751063f5527921d'), 'STATUTORY REPORTS': 'Secretary acts as the Secretary to the Corporate', 'Unnamed: 0': 'approve transfer/transmission of equity shares,'}
{'_id': ObjectId('5feb64093751063f5527921e'), 'STATUTORY REPORTS': 'Social

```
Responsibility Committee.', 'Unnamed: 0': 'dematerialisation, issue of duplicate
certificates, etc.'}
{'_id': ObjectId('5feb64093751063f5527921f'), 'STATUTORY REPORTS': None,
'Unnamed: 0': 'All valid requests for transfer & transmission of shares in'}
{'_id': ObjectId('5feb64093751063f55279220'), 'STATUTORY REPORTS': '7. RISk
MANAGEMENT COMMITTEE', 'Unnamed: 0': 'physical form, duplicate issue of share
certificate were'}
{'_id': ObjectId('5feb64093751063f55279221'), 'STATUTORY REPORTS': 'In terms of
Regulation 21 of the SEBI (LODR)', 'Unnamed: 0': 'processed in time.'}
```

[ ]: