Python Technical Assignment

Requirements

“Given a single-page PDF, extract the table and output it into Excel format.”

- Each row in the table should correspond to one Excel row
- The excel file should contain the following columns:
  1. Value Date (Column `Value Date` in the PDF)
  2. Order Date (Column `Order Date` in the PDF)
  3. Description (Column `Description` in the PDF)
  4. Debit (Derived from `Amount` column in the PDF - Values in brackets are Debits)
  5. Credit (Derived from `Amount` column in the PDF - Values not in brackets are Credits)
  6. Currency (Derived from the section header in the PDF)

Example

Input PDF

Value date	Order date	Description	Amount
Deposits			
Deposits AUD		**Opening balance**	**32.29**
		Month end balance(31 Mar 2019)	**32.29**
Deposits HKD		**Opening balance**	**0.00**
01 Mar 2019	28 Feb 2019	INTEREST RECEIVED	2.39
04 Mar 2019	27 Feb 2019	INTEREST FOR BNP PARIBAS 4M CALLABLE FIXED COUPON NOTE 1810.HK 04MAR2019 8.01% ISIN:XS1881628199 #44368-0	20,025.00
04 Mar 2019	04 Mar 2019	EARLY REDEMPTION OF HKD 3,000,000.00 BNP PARIBAS 4M CALLABLE FIXED COUPON NOTE 1810.HK 04MAR2019 8.01% ISIN:XS1881628199 #12740-0	3,000,000.00
05 Mar 2019	05 Mar 2019	REPAY FIXED LOAN #31806-0	(5,443,428.95)
05 Mar 2019	05 Mar 2019	INTEREST CHARGE ON FIXED LOAN #31806-0	(1,834.73)
05 Mar 2019	05 Mar 2019	DRAWDOWN FIXED LOAN #38034-0	5,445,263.68
12 Mar 2019	12 Mar 2019	REPAY FIXED LOAN #38034-0	(5,445,263.68)
12 Mar 2019	12 Mar 2019	INTEREST CHARGE ON FIXED LOAN #38034-0	(1,998.34)
12 Mar 2019	12 Mar 2019	DRAWDOWN FIXED LOAN #43286-0	2,427,234.63
19 Mar 2019	19 Mar 2019	REPAY FIXED LOAN #43286-0	(2,427,234.63)
19 Mar 2019	19 Mar 2019	INTEREST CHARGE ON FIXED LOAN #43286-0	(849.20)
19 Mar 2019	19 Mar 2019	DRAWDOWN FIXED LOAN #49002-0	2,428,083.83
22 Mar 2019	20 Mar 2019	INTEREST FOR GOLDMAN SACHS INTERNATIONAL 4M CALLABLE FIXED COUPON NOTE 1398.HK 22MAR2019 8.01% ISIN:XS1901844396 #44870-0	20,025.00
22 Mar 2019	22 Mar 2019	EARLY REDEMPTION OF HKD 3,000,000.00 GOLDMAN SACHS INTERNATIONAL 4M CALLABLE FIXED COUPON NOTE 1398.HK 22MAR2019 8.01% ISIN:XS1901844396 #44868-0	3,000,000.00
26 Mar 2019	26 Mar 2019	REPAY FIXED LOAN #49002-0	(2,428,083.83)
26 Mar 2019	26 Mar 2019	INTEREST CHARGE ON FIXED LOAN #49002-0	(881.01)
		Month end balance(31 Mar 2019)	**591,060.16**
Deposits SGD		**Opening balance**	**302,404.96**
		Month end balance(31 Mar 2019)	**302,404.96**
Deposits USD		**Opening balance**	**64,338.48**
01 Mar 2019	28 Feb 2019	INTEREST RECEIVED	15.50
12 Mar 2019	12 Mar 2019	RECEIVE DIVIDEND FROM HOLDING OF VP GREATER CHINA HIGH YIELD INC P USD MCSH #54177-0	1,325.50
29 Mar 2019	29 Mar 2019	CONSENT FEE ON CFLD (CAYMAN) INVESTMENT LTD BOND 21-DEC-2020 6.50% (ISIN: XS1729851490)	400.00

Expected Output

Value Date	Order Date	Description	Currency	Debit	Credit
2019-03-01	2019-02-28	INTEREST RECEIVED	HKD		2.39
2019-03-01	2019-03-01	INTEREST RECEIVED	USD		15.50
2019-03-04	2019-02-27	INTEREST FOR BNP PARIBAS 4M CALLABLE FIXED COUPON NOTE 1810.HK 04MAR2019 8.01% ISIN:XS1881628199 #44368-0	HKD		20025.00
2019-03-04	2019-03-04	EARLY REDEMPTION OF HKD 3,000,000.00 BNP PARIBAS 4M CALLABLE FIXED COUPON NOTE 1810.HK 04MAR2019 8.01% ISIN:XS1881628199 #12740-0	HKD		3000000.00
2019-03-05	2019-03-05	REPAY FIXED LOAN #31806-0	HKD	5443428.95	
2019-03-05	2019-03-05	INTEREST CHARGE ON FIXED LOAN #31806-0	HKD	1834.73	
2019-03-05	2019-03-05	DRAWDOWN FIXED LOAN #38034-0	HKD		5445263.68
2019-03-12	2019-03-12	RECEIVE DIVIDEND FROM HOLDING OF VP GREATER CHINA HIGH YIELD INC P USD MCSH #54177-0	USD		1325.50
2019-03-12	2019-03-12	REPAY FIXED LOAN #38034-0	HKD	5445263.68	
2019-03-12	2019-03-12	INTEREST CHARGE ON FIXED LOAN #38034-0	HKD	1998.34	
2019-03-12	2019-03-12	DRAWDOWN FIXED LOAN #43286-0	HKD		2427234.63
2019-03-19	2019-03-19	REPAY FIXED LOAN #43286-0	HKD	2427234.63	
2019-03-19	2019-03-19	INTEREST CHARGE ON FIXED LOAN #43286-0	HKD	849.20	
2019-03-19	2019-03-19	DRAWDOWN FIXED LOAN #49002-0	HKD		2428083.83
2019-03-22	2019-03-20	INTEREST FOR GOLDMAN SACHS INTERNATIONAL 4M CALLABLE FIXED COUPON NOTE 1398.HK 22MAR2019 8.01% ISIN:XS1901844396 #44870-0	HKD		20025.00
2019-03-22	2019-03-22	EARLY REDEMPTION OF HKD 3,000,000.00 GOLDMAN SACHS INTERNATIONAL 4M CALLABLE FIXED COUPON NOTE 1398.HK 22MAR2019 8.01% ISIN:XS1901844396 #44868-0	HKD		3000000.00
2019-03-26	2019-03-26	REPAY FIXED LOAN #49002-0	HKD	2428083.83	
2019-03-26	2019-03-26	INTEREST CHARGE ON FIXED LOAN #49002-0	HKD	881.01	
2019-03-29	2019-03-29	CONSENT FEE ON CFLD (CAYMAN) INVESTMENT LTD BOND 21-DEC-2020 6.50% (ISIN: XS1729851490)	USD		400.00

Tips

How you decide to identify the table is entirely up to you.

Some ideas include but are not limited to:

- A custom config file indicating where each table is (string-based matching, column distinction, etc.)
- Using OpenCV for table detection or row separation
- A third-party library such as Camelot-Py or Tabula-Py
- Hardcoded strings

To streamline the process, it is recommended to use Poppler to extract text from the PDF. For instance, the following command line invocation of the Poppler pdftotext utility will output a file with each word's location on the page, removing any text watermarks in the process:

```
pdftotext -f 1 -l 1 -r 300 -q -nodiag -bbox-layout "test_input.pdf" "test_ pdftotext_ou
```

> Poppler can also be used in a Conda environment (especially relevant for Windows developers)

Bonuses

> Note that the below asks are only bonuses. Ultimately, preference will be given to clean, modularized code.

- If any columns contain dates, then standardize all entries in `YYYY-MM-DD` format
- If any columns contain numbers, then standardize all entries in `###.##` format (removing commas; see above Excel screenshot for an example)