In [1]:
```
import nltk
import pandas as pd
import re
import fitz
```

In [12]:
```
# read pdf file
document = fitz.open('Invoice_Sample03.pdf')
text = document.load_page(0).get_text()   # text extract

# find required details
invoice_date = re.search(r'Invoice Date: \s*(.*)', text).group(1).strip()
Vendor = re.search(r'Vendor:\s*(.*)', text).group(1).strip()
Address = re.search(r'Address:\s*(.*)', text).group(1).strip()
project = re.search(r'Project:\s*(.*)', text).group(1).strip()
Customer = re.search(r'Customer:\s*(.*)', text).group(1).strip()
Terms = re.search(r'Terms:\s*(.*)', text).group(1).strip()
# create datafrae
df=pd.DataFrame([invoice_date,Vendor,Address,project,Customer,Terms],
             index=["invoice_date","Vendor","Address","project","Customer","Terms
df
```

Out[12]:

	0
invoice_date	July 25, 2024
Vendor	GHI Building Supplies
Address	9101 Trade St, Buildville, ST 34567
project	Warehouse Expansion
Customer	JKL Construction Co.
Terms	Payment due by August 25, 2024.

In [ ]:
```
```