Extracting information from invoice from Amazon

Code:

```
import pdfplumber
import re
from word2number import w2n
import csv

pdf_files = ['invoice.pdf', 'invoice1.pdf', 'invoice2.pdf', 'invoice3.pdf']
csv_file = "invoice_data.csv"
fieldnames = [
    "Invoice Number", "Invoice Details", "Invoice Date",
    "Order Number", "Order Date", "PAN No",
    "GST Registration Number", "Amount in Words", "Total Amount"
]

with open(csv_file, mode='w', newline='') as file:
    writer = csv.DictWriter(file, fieldnames=fieldnames)
    writer.writeheader()

    for loc in pdf_files:
        with pdfplumber.open(loc) as pdf:
            page = pdf.pages[0]
            text = page.extract_text()

        invoice_number = re.search(r'Invoice Number :\s*([A-Z0-9-]+)', text)
        invoice_details = re.search(r'Invoice Details :\s*(.*?)\s+Invoice Date', text)
        invoice_date = re.search(r'Invoice Date :\s*(\b\d{2}\.\d{2}\.\d{4}\b)', text)
        order_number = re.search(r'Order Number:\s*([A-Z0-9-]+)', text)
        order_date = re.search(r'Order Date:\s*(\b\d{2}\.\d{2}\.\d{4}\b)', text)
        PAN_No = re.search(r'PAN No:\s*([A-Z0-9-]+)', text)
        gst_registration_number = re.search(r'GST Registration No:\s*([A-Z0-9-]+)', text)
        amount_in_words = re.search(r'Amount in Words:\s*(.+?)\s+only', text,
re.IGNORECASE)

        total_amount = re.search(r'Total Amount:\s*([A-Z0-9\s,.-]+)', text)
        if total_amount:
            try:
                amount_number = float(total_amount.group(1).replace(',', '').strip())
            except:
                amount_number = "Could not convert"
        elif amount_in_words:
            try:
                amount_number = w2n.word_to_num(amount_in_words.group(1))
            except:
                amount_number = "Conversion failed"
        else:
```

```
            amount_number = "Not found"

        data = {
            "Invoice Number": invoice_number.group(1) if invoice_number else "Not found",
            "Invoice Details": invoice_details.group(1) if invoice_details else "Not
found",
            "Invoice Date": invoice_date.group(1) if invoice_date else "Not found",
            "Order Number": order_number.group(1) if order_number else "Not found",
            "Order Date": order_date.group(1) if order_date else "Not found",
            "PAN No": PAN_No.group(1) if PAN_No else "Not found",
            "GST Registration Number": gst_registration_number.group(1) if
gst_registration_number else "Not found",
            "Amount in Words": amount_in_words.group(1) if amount_in_words else "Not
found",
            "Total Amount": amount_number
        }

        print(f"Processed {loc}: {data}")
        writer.writerow(data)

print(f"Data from all invoices saved to {csv_file}")
```

Output:

One instance of all the output for clear understanding:

```
Processed invoice.pdf: {'Invoice Number': 'CCX1-3271', 'Invoice Details': 'WB-CCX1-161657241-2526', 'Invoice Date': '22.05.2025', 'Order Number': '408-5871091-1105120', 'Order
Date': '22.05.2025', 'PAN No': 'AYFPC6294H', 'GST Registration Number': '19AYFPC6294H1Z6', 'Amount in Words': 'Four Hundred Seventy-nine', 'Total Amount': 479}
```

The saved CSV file:

	A	B	C	D	E	F	G	H	I
1	Invoice Number	Invoice Details	Invoice Date	Order Number	Order Date	PAN No	GST Registration Num	Amount in Words	Total Amount
2	CCX1-3271	WB-CCX1-161657241-2526	22.05.2025	408-5871091-1105120	22.05.2025	AYFPC6294H	19AYFPC6294H1Z6	Four Hundred Seventy-nin	479
3	GAX1-837020	AS-GAX1-1044-2526	19.05.2025	408-7390656-7547550	18.05.2025	AAICA3918J	29AAICA3918J1ZE	Seven	7
4	BOM5-5134	MH-BOM5-166755631-2526	03.05.2025	408-6719700-1634717	03.05.2025	AABCF8391H	27AABCF8391H1ZC	Four Hundred Sixty-seven	467
5	CJB1-261202	TN-CJB1-1293787125-2526	12.04.2025	171-0350028-7774745	12.04.2025	AAECR0564M	33AAECR0564M1ZA	Five Hundred Forty-eight F	548.05
6									
7									
8									
9									
10									
11									