Analysis of the PDF :

Harshvardhan Joshi

27 February 2025

1 Introduction

I analyzed PDF `Register_2024.pdf`, with data on roughly 700 workshops. The goal was to collect each workshop's services ("Auftragsarbeiten / Produkte"), website, and email address. I also added web scraping to check if each website has more updated services or a direct contact person.

2 Methodology

2.1 Scripts Used

2.1.1 `script_pagesplit.py`

Splits the PDF by lines such as "Seite - XX -" or "Baden-Württemberg(...)". It finds over 100 text chunks, some duplicates or partial segments. This is useful if you want every piece of text, even if repeated.

2.1.2 `script_regnr.py`

Splits the PDF by "(Reg.-Nr.)" and finds about 60–70 entries. Each entry usually belongs to one workshop (one registration number). This gives a cleaner list, with fewer duplicates or partial segments.

2.2 Data Fields

- **company_name**

- **offerings_raw** (services/products)
- **homepage**
- **email**
- **updated_services** (scraped info, if found)
- **contact_person** (scraped info, if found)

2.3 Tools

- Python 3.11
- PyPDF2 for reading the PDF
- Requests and BeautifulSoup for scraping each homepage
- CSV for writing out the data

3 Results

3.1 Pagesplit Script Results

- About 100–120 rows
- Includes some "Unknown" or repeated chunks
- Good if you want a full text extraction

3.2 RegNr Script Results

- Around 60–70 rows, each usually matching a single workshop
- Some "Unknown" names if the PDF chunk does not match our pattern
- More direct if you want one row per workshop registration

3.3 Web Scraping

For each valid website (starting with "http"), the scripts try to find extra service info and a contact name. Many websites do not have a uniform structure, so often no extra info is found.

4 Conclusion

I produced two CSV files:

- **companies_output_pagesplit.csv**: from `script_pagesplit.py`, shows ∼100+ rows.
- **companies_output_regnr.csv**: from `script_regnr.py`, shows 60–70 rows with a closer 1:1 mapping to each workshop.

Use **pagesplit** if you want every bit of text (including duplicates). Use **regnr** if you prefer fewer rows, each focusing on a single workshop.

Future Possibilities

- We can remove or skip "Unknown" rows by adding a check in the code.
- Gathering more data from each website if their HTML layout is consistent.
- We can analyze the CSV to see how many workshops offer specific services (e.g., "Elektromontage," "Verpackungsarbeiten").