IDEMIA Take Home Assignment - RAG

Chaojun(Vicky) Chen

June 16, 2024

Objective

Create a Python pipeline that demonstrates document processing capabilities, focusing on ingesting PDF documents containing tables and integrating them with a Large Language Model (LLM) for query-based information retrieval.

RAG Pipeline

1. Document Processing

- **PDF Document Ingestion**: Implement the ingestion of PDF documents, with a special focus on extracting tables.
- **Table Parsing**: Use the LLM to parse tables within the documents and extract structured data.
- **Superscript Relations**: Relate superscripts in table columns to their references.

File: `llm_parse_complex_table.py`
Description: Parses complex tables within PDF documents.
Output: Figure 1 is the input table while Figure 2 is the LLM output for parsing complex table. Note that this LLM approach is a significant improvement compared to traditional methods:

C-1 Field Edit Specifications for Type-2 Elements

Field Number	Identifier	Field Name	Character	Field Size (not including Character Separators)		Occurrences		Example	Comments/Special Characters
				Min	Max	Min[1]	Max		
2.001	LEN	LOGICAL RECORD LENGTH	N	2	7		1	2.001:909{GS}	
2.002	IDC	INFORMATION DESIGNATION CHARACTER	N	1	2		1	2.002:00{GS}	
2.003	FFN	FBI FILE NUMBER	N	10	10		1	2.003:2537597861{GS}	
2.005	RET	RETENTION CODE	A	1	1		1	2.005:Y{GS}	
2.006	ATN	ATTENTION INDICATOR	ANS	3	30		1	2.006:SA J Q DOE,RM 11867{GS}	Any printable 7-bit ASCII character with the exception of the period is allowed.
2.007	SCO	SEND COPY TO	ANS	9	19		9	2.007:NY030025P{GS}	
2.009	OCA	ORIGINATING AGENCY CASE NUMBER	ANS	1	20		1	2.009:Q880312465{GS}	Any printable 7-bit ASCII character with the exception of the period is allowed
2.010	CIN	CONTRIBUTOR CASE IDENTIFIER NUMBER	SET				5	2.010:INCIDENT NUMBER{US}1963BRT715{GS}	Any printable 7-bit ASCII character is allowed.
A	CIN_PRE	CONTRIBUTOR CASE PREFIX	ANS	1	24	1	1		
B	CIN_ID	CONTRIBUTOR CASE ID	ANS	1	24	1	1		
2.011	CIX	CONTRIBUTOR CASE IDENTIFIER EXTENSION	N	2	4		5	2.011:23{GS}	
2.012	LCN	FBI LATENT CASE NUMBER	ANS	11	11		1	2.012:MX-12345678{GS}	The only special character allowed is a hyphen
2.013	LCX	FBI LATENT CASE EXTENSION	N	5	5		1	2.013:98765{GS}	
2.014	FBI	FBI NUMBER/UCN	AN	1	9		1000	2.014:62760NY12{GS}	
2.015	SID	STATE IDENTIFICATION NUMBER	ANS	3	10		1000	2.015:NY12345678{GS}	OR and PA may use a hyphen in the last position
2.016	SOC	SOCIAL SECURITY ACCOUNT NUMBER	N	9	9		4	2.016:220565855{GS}	
2.017	MNU	MISCELLANEOUS IDENTIFICATION NUMBER	ANS	4	15		4	2.017:PP-12345678901P{GS}	A hyphen is allowed as a special character
2.018	NAM	NAME	AS	3	50		1	2.018:JONES,ANTHONY P{GS}	Commas, hyphens, and spaces are all allowed as special characters.

Figure 1: Complex Table

```
### Field Edit Specifications for Type-2 Elements
Guide on edit specifications, character, and field size for Type-2 elements.

| Field Number | Identifier | Field Name                  | Character | Field Size (Min) | Field Size (Max) | Occurrences (Min) | Occurrences (Max) | Example                | Comments/Special Characters
|--------------|------------|-----------------------------|-----------|------------------|------------------|-------------------|-------------------|------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------
| 2.001        | LEN        | LOGICAL RECORD LENGTH       | N         | 2                | 7                | 1                 | 1                 | 909{GS}                |
| 2.002        | IDC        | INFORMATION DESIGNATION CHARAC| N       | 1                | 2                | 1                 | 1                 | 00{GS}                 |
| 2.003        | FFN        | FBI FILE NUMBER             | N         | 10               | 10               | 1                 | 1                 | 2537597861{GS}         |
| 2.005        | RET        | RETENTION CODE              | A         | 1                | 1                | 1                 | 1                 | Y{GS}                  |
| 2.006        | ATN        | ATTENTION INDICATOR         | ANS       | 3                | 30               | 1                 | 1                 | SA J Q DOE,RM 11867{GS} | Any printable 7-bit ASCII character with the exception of the period is allowed.
| 2.007        | SCO        | SEND COPY TO                | ANS       | 9                | 199              | 1                 | 1                 | NY030025P{GS}          |
| 2.009        | OCA        | ORIGINATING AGENCY CASE NUMBER| ANS     | 12               | 12               | 1                 | 1                 | Q880312465{GS}         | Any printable 7-bit ASCII character with the exception of the period is allowed.
| 2.010        | CIN        | CONTRIBUTOR CASE IDENTIFIER NUMBER| ANS  | 1                | 5                | 1                 | 1                 | INCIDENT NUMBER{US}1963BRT715{GS} | Any printable 7-bit ASCII character is allowed.
| ACIN_PRE     |            | CONTRIBUTOR CASE PREFIX     | ANS       | 1                | 24               | 1                 | 1                 |                        |
| BCIN_ID      |            | CONTRIBUTOR CASE ID         | ANS       | 1                | 24               | 1                 | 1                 |                        |
| 2.011        | CIX        | CONTRIBUTOR CASE IDENTIFIER EXTENSION | N     | 2                | 4                | 1                 | 5                 | 23{GS}                 |
| 2.012        | LCN        | FBI LATENT CASE NUMBER      | ANS       | 1                | 11               | 1                 | 1                 | MX-12345678{GS}        | The only special character allowed is a hyphen.
| 2.013        | LCX        | FBI LATENT CASE EXTENSION   | N         | 5                | 5                | 1                 | 1                 | 98765{GS}              |
| 2.014        | FBI        | FBI NUMBER/UCN              | AN        | 1                | 9                | 1                 | 1000              | 62760NY12{GS}          |
| 2.015        | SID        | STATE IDENTIFICATION NUMBER | ANS       | 3                | 10               | 1                 | 1000              | NY12345678{GS}         | OR and PA may use a hyphen in the last position.
| 2.016        | SOC        | SOCIAL SECURITY ACCOUNT NUMBER| N       | 9                | 9                | 1                 | 4                 | 220565855{GS}          |
| 2.017        | MNU        | MISCELLANEOUS IDENTIFICATION NUMBER | ANS | 4              | 15               | 4                 | 4                 | PP-12345678901P{GS}    | A hyphen is allowed as a special character.
| 2.018        | NAM        | NAME                        | AS        | 3                | 50               | 1                 | 1                 | JONES,ANTHONY P{GS}    | Commas, hyphens, and spaces are allowed as special characters.
```

Figure 2: Extracting complex table using LLM

File: `llm_parse_pdf.py`

Description: Extracting information from tables in pdf (`./document/Type_II_table.pdf`) for generating knowledge graph.

Output:

```
["CAR", "subtype", "has subfield", "2.001 LEN", "subfield"]
["CAR", "subtype", "has subfield", "2.002 IDC", "subfield"]
["CAR", "subtype", "has subfield", "2.005 RET", "subfield"]
["CAR", "subtype", "has subfield", "2.006 ATN", "subfield"]
["CAR", "subtype", "has subfield", "2.007 SCO", "subfield"]
["CAR", "subtype", "has subfield", "2.009 OCA", "subfield"]
["CAR", "subtype", "has subfield", "2.014 FBI", "subfield"]
["CAR", "subtype", "has subfield", "2.015 SID", "subfield"]
["CAR", "subtype", "has subfield", "2.016 SOC", "subfield"]
["CAR", "subtype", "has subfield", "2.017 MNU", "subfield"]
["CAR", "subtype", "has subfield", "2.018 NAM", "subfield"]
["CAR", "subtype", "has subfield", "2.019 AKA", "subfield"]
["CAR", "subtype", "has subfield", "2.020 POB", "subfield"]
["CAR", "subtype", "has subfield", "2.021 CTZ", "subfield"]
["CAR", "subtype", "has subfield", "2.022 DOB", "subfield"]
["CAR", "subtype", "has subfield", "2.024 SEX", "subfield"]
["CAR", "subtype", "has subfield", "2.025 RAC", "subfield"]
["CAR", "subtype", "has subfield", "2.026 SMT", "subfield"]
["CAR", "subtype", "has subfield", "2.027 HGT", "subfield"]
```

Figure 3: Extracting information in PDF for knowledge graph visualization

2. LLM Inference

- **User Prompt Handling**: Handle user prompts.
- **RAG Model Integration**: Integrate with a Retrieval-Augmented Generation (RAG) model to generate responses based on user queries.
- **Information Retrieval**: Retrieve relevant information from the ingested documents to include in the responses.

File: `llm_multi_doc.py`

Description: Conducts real-time Q&A across multiple documents using RAG and LLM. Can be used to extract information and identify connections between different documents. It can also serve as a tool for onboarding to swiftly familiarize new employees with the company or for quick internal search of company documents.

Output: Conduct real-time Q&A across multiple documents using RAG and LLM

```
--------------------------------------------------------------------------------
Welcome! You are now ready to start interacting with your documents
--------------------------------------------------------------------------------
Prompt: Provide a summary of Appendix D
Answer: Appendix D of the FBI CJIS/NGI EBTS 11.0 document, dated April 16, 2021, is titled "Summary
Logical Record Layouts For Type-2 Identification and Verification Transactions." This appendix
includes detailed layouts and specifications for Type-2 logical records specifically used in the
context of identification and verification transactions. It provides a structured overview to ensure
that the data is formatted and utilized correctly within these transaction types.
Prompt: Is DOO portion of this field optional?
Answer: Yes, the DOO (Date of Offense) portion of the field is optional in Appendix D of the FBI
CJIS/NGI EBTS 11.0 document, but it should be provided if known.
Prompt: Are CDD and CCT mandatory fields for this TOT?
Answer: Yes, the CDD (Case Disposition Date) and CCT (Case Classification Type) fields are mandatory
for this Type of Transaction (TOT) in Appendix D of the FBI CJIS/NGI EBTS 11.0 document.
```

Figure 4: Real-time Q&A using RAG and LLM

3. Knowledge Graph Generation

- **Knowledge Graph**: Generate a knowledge graph from the retrieved document evidence.

File: `llm_main.py`

Description: Contains the entire process and generates the JSON file for Knowledge Graph visualization at the end.

Output: JSON file for knowledge graph visualization

```
[
    {
        "source": "\"AMN\"",
        "sourcetype": " \"subtype\"",
        "relation": " \"has subfield\"",
        "target": " \"2.001 LEN\"",
        "targettype": " \"subfield\""
    },
    {
        "source": "\"AMN\"",
        "sourcetype": " \"subtype\"",
        "relation": " \"has subfield\"",
        "target": " \"2.002 IDC\"",
        "targettype": " \"subfield\""
    },
    {
        "source": "\"AMN\"",
        "sourcetype": " \"subtype\"",
        "relation": " \"has subfield\"",
        "target": " \"2.005 RET\"",
        "targettype": " \"subfield\""
    },
    {
        "source": "\"AMN\"",
        "sourcetype": " \"subtype\"",
```


Figure 5: Knowledge graph JSON file

Usage

1. **Knowledge Graph JSON File Generation**:

   ```
   python llm_main.py
   ```

2. **LLM Inference**:

   ```
   python llm_multi_doc.py
   ```

3. **Knowledge Graph Visualization**: Run the following command to visualize the generated knowledge graph inside folder kg-demo.

   ```
   npm start
   ```

Generated Knowledge Graph

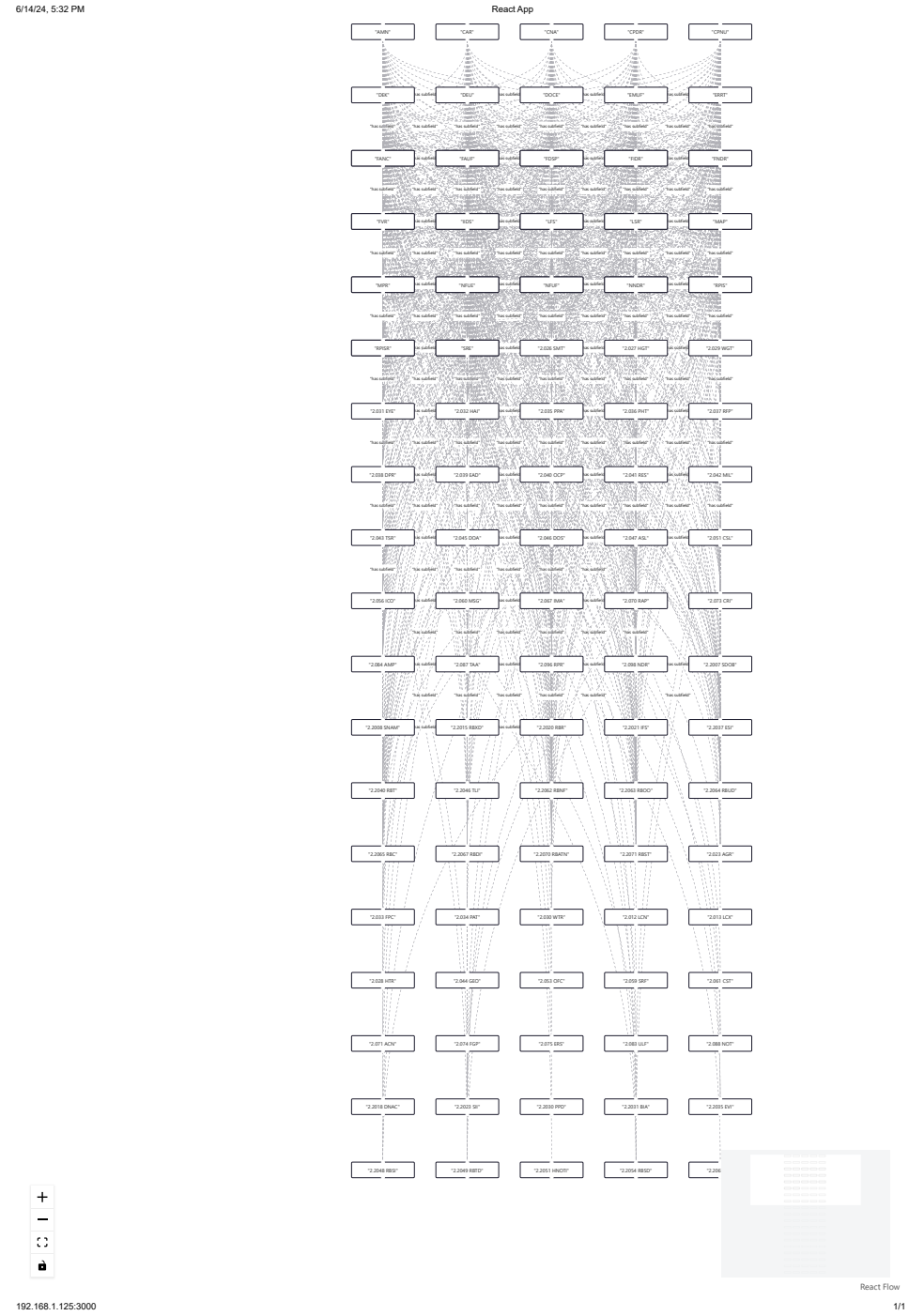

Figure 6: knowledge graph visualization

Appendix

The previous method proves inadequate for handling complex tables. I have also explored GitHub repositories such as `nlmatics/nlm-ingestor`, `nlmatics/llmsherpa`, and `Joshua-Yu/graph-rag` for PDF parsing and table extraction, as recommended. Currently, `llm_parse_table.py` demonstrates the best performance when it comes to handling complex tables.

A-1 NGI Maximum Transaction Response Times

Transaction	Priority	10 sec	20 sec	30 sec	2 min	5 min	10 min	15 min	30 min	1 hour	2 hours	4 hours	24 hours	48 hours	15 days
Criminal Fingerprint Identification Search	high (1)						•								
	routine (5)								•						
	low (7)												•		
	non-urgent (9)														•
Civil Fingerprint Identification Search	high (1)							•							
	routine (5)										•				
	low (7)												•		
	non-urgent (9)														•
Friction Ridge Investigation Search - Tenprint	high (1)			•											
	routine (5)				•										
	low (7)								•						
Friction Ridge Investigation Search - Latent	high (1)									•					
	routine (5)											•			
	low (7)												•		
Biometric/Biographic Maintenance								•							
Biometric Audit Trail Retrieval								•							
Biometric Image Retrieval, multiple UCN														•	
Biometric Image Retrieval, single UCN						•									
Biographic Search					•										
Cascaded Facial Recognition Search													•		
Cascaded Fingerprint Search													•		

Figure 7: Complex table

Type	Time Interval	Description
Cascaded Fingerprint Search	24 hours	·
	48 hours	·
	15 days	·
Cascaded Facial Recognition Search	24 hours	·
	48 hours	·
	15 days	·
Biographic Search	2 min	·
	5 min	·
	10 min	·
	15 min	·
	30 min	·
	1 hour	·
	2 hours	·
	4 hours	·
	24 hours	·
	48 hours	·
	15 days	·
Biometric Image Retrieval, single UCN	5 min	·
	10 min	·
	15 min	·
	30 min	·
	1 hour	·
	2 hours	·
	4 hours	·
	24 hours	·
	48 hours	·
	15 days	·
Biometric Image Retrieval, multiple UCN	48 hours	·
	15 days	·
Biometric Audit Trail Retrieval	15 min	·
	30 min	·
	1 hour	·
	2 hours	·
	4 hours	·
	24 hours	·
	48 hours	·
	15 days	·
Biometric/Biographic Maintenance	15 min	·
	30 min	·
	1 hour	·
	2 hours	·
	4 hours	·
	24 hours	·
	48 hours	·
	15 days	·
Friction Ridge Investigation Search - Latent	Priority	·

Figure 8: Extracting complex table using previous method