Problem Statement

Business Contract Validation -To Classify Content within the Contract Clauses and Determine Deviations from Templates and highlight them.

Description:

The first task is to parse these documents. Determine the key details within the contract document. Every contract has clauses and sub-clauses. Use NER to determine the important feature from the document. Then Use Knowledge-Graph to see the important feature. Using Bert Model analyse the semantics and classify the contents of the parsed documents to these clauses. A contract has an associated template to it, and it is important to determine the deviations from that template and highlight them and analyse the semantics to generate corresponding suggestions.

Unique Idea Brief (Solution)

The Advanced Contract Analysis and Deviation Detection System is an innovative solution designed to streamline and automate the meticulous process of business contract analysis. This system integrates multiple advanced technologies, including machine learning, natural language processing (NLP), and network analysis, to provide a comprehensive tool for contract management. Key features include the generation of synthetic contracts using the Faker library to create realistic datasets for model training, utilizing BERT (Bidirectional Encoder Representations from Transformers) for precise sequence classification to identify and categorize contract clauses, and extracting text from PDF contracts with PyMuPDF (Fitz) for accurate information retrieval. The system employs spaCy for Named Entity Recognition (NER) to pinpoint relevant entities within the contract text and checks for deviations against a predefined template, ensuring each clause meets specified standards. In cases where deviations are detected, the system provides context-specific suggestions for corrections, thereby enhancing contract accuracy and compliance. Additionally, it constructs a knowledge graph from the extracted entities using NetworkX, offering a visual representation of the relationships within the contract data, which is further illustrated with Matplotlib. The system also highlights deviations directly within the contract and saves these annotated contracts as PDFs for easy reference and review. This comprehensive approach not only improves efficiency and accuracy in contract analysis but also offers valuable insights and corrective measures, making it an indispensable tool for businesses handling complex contractual agreements.

Features Offered

1. Synthetic Contract Generation: Utilizing the Faker library to create realistic contract datasets.

2. PDF Text Extraction: Extracting text from PDF contracts using PyMuPDF (Fitz).

3. Clause Categorization: BERT-based sequence classification to identify and categorize contract clauses.

4. Named Entity Recognition (NER): Employing spaCy to identify relevant entities within the contract text.

5. Deviation Detection and Suggestion: Checking for deviations against a predefined template and providing context-specific suggestions for corrections.

6. Knowledge Graph Construction: Using NetworkX to create and visualize the relationships within the contract data.

7. Highlighting Deviations: Annotating deviations directly within the contract and saving them as PDFs for easy reference and review.

8. Clause Validity Checks: Ensuring the correctness of company names, dates, amounts, notice periods, and other critical contract details.

9. Suggestion: Generate suggestion based on the deviation.

Process flow

[Start]

|

[Generate Synthetic Contracts]

|

[Save Contracts to CSV]

|

[Create PDF from Contracts]

|

[Load Dataset and Preprocess]

|

[Tokenize and Label Mapping]

|

[Define Dataset Class]
|
[Initialize and Train BERT Model]
|
[Extract Text from PDF]
|
[Perform NER on Text]
|
[Extract Clauses from Text]
|
[Detect Deviations]
|
[Generate Suggestions]
|

[Build Knowledge Graph]

|

[Visualize Knowledge Graph]

|

[Highlight Deviations in PDF]

|

[Save Highlighted PDF]

|

[End]

Technologies used

- **Python**: The primary programming language used for the entire workflow.
- **pandas**: For data manipulation and analysis.
- **csv**: For reading from and writing to CSV files.
- **pdfkit**: For converting HTML content to PDF files.
- **reportlab**: For generating PDF files.
- **fitz (PyMuPDF)**: For extracting text from PDF files.
- **spacy**: For natural language processing, particularly Named Entity Recognition (NER).
- **networkx**: For creating and manipulating complex networks (graphs).
- **matplotlib**: For plotting and visualizing graphs.
- **datetime**: For handling dates and times.
- **transformers**: From Hugging Face, for leveraging pre-trained models like BERT and DistilBERT for sequence classification.
- **torch**: For working with PyTorch, a deep learning library used for training models.

- **Faker**: For generating synthetic data.
- **random**: For generating random values.
- **sklearn**: For machine learning tasks (specifically mentioned, but not used in the provided code).
- **BERT and DistilBERT models**: For sequence classification tasks, using pre-trained models from Hugging Face.

Conclusion

This solution offers a comprehensive system for analyzing PDF contracts. It uses machine learning to classify clauses, performs NLP tasks like entity recognition, and detects deviations from a standard template. By highlighting these deviations and suggesting corrections, the system can significantly improve the efficiency and accuracy of contract review processes.