Project Overview

Automated Legal Document Compliance Checker

Group Members:
Gnyani Enugandula – genugand@syr.edu
Khushi Shah – kbshah@syr.edu
Mansi Jagdale – mjagdale@syr.edu
Neha Shirodkar – nmshirod@syr.edu
Shaurya Jain – sjain42@syr.edu

1. Data Sets for Use:

For an automated legal document compliance checker, the datasets will consist of legal documents and standards relevant to specific industries or compliance regulations. The following are potential data sources:

• Public Legal Documents: Datasets of publicly available legal documents such as contracts, terms of service, and privacy policies can be obtained from:
- Common Crawl: A vast repository of openly available web data, including legal texts.
- Government Regulatory Data: Regulatory compliance guidelines such as GDPR, HIPAA, and CCPA available from official websites or open-source repositories (e.g., U.S. Government websites, European Union portals).
- Industry Standards Documents: Regulatory standards for industries (e.g., ISO standards for IT security, financial regulations), either downloaded or scraped.
- Annotated Legal Datasets: Publicly available datasets that have been annotated for compliance, such as the "Contracts Noun Phrase Dataset" or "LEXGLUE" for legal-specific NLP tasks.

2. Methods of Data Acquisition and Analysis

Method 1: Web Scraping

• Acquisition: Python libraries like BeautifulSoup, Scrapy, or Selenium will be used to scrape legal documents, regulations, and compliance rules from regulatory websites and government portals.
• Analysis: Natural Language Processing (NLP) techniques powered by Hugging Face transformers and libraries such as SpaCy or NLTK will preprocess and analyze the scraped documents. Key compliance terms and clauses can be extracted using techniques like Named Entity Recognition (NER) and Topic Modeling.
• Pros: Real-time data acquisition from diverse sources.
• Cons: Web scraping rules need to be followed, and regular updates may be required.

Method 2: API Access

• Acquisition: Several governments and regulatory bodies provide APIs that allow access to legal regulations and compliance data (e.g., European Union Open Data Portal, U.S. Federal Register API).
• Analysis: API data will be integrated into the system to pull real-time compliance rules and documents, which will then be analyzed using NLP frameworks like Hugging Face. Preprocessing will convert text into structured formats to feed into rule-checking algorithms.
• Pros: Real-time data availability and easier maintenance.
• Cons: Limited to available APIs and may not cover all legal domains.

Method 3: Machine Learning for Document Classification

• Acquisition: Use publicly available labeled datasets of legal documents (e.g., contracts or terms of service) that indicate whether they are compliant or non-compliant.
• Analysis: Supervised machine learning methods will be employed to classify documents as compliant or non-compliant. Using tools like Hugging Face's pretrained models, along with supervised algorithms (e.g., Logistic Regression, Support Vector Machines, Random Forest), document compliance can be determined based on regulatory standards.
• Pros: Scalability to large datasets and automation of compliance checks.

• Cons: Requires labeled data and is dependent on data quality for performance.

3. Data Storage

For document storage, we will utilize AWS S3 (Simple Storage Service) to handle legal document storage dynamically. AWS S3 offers highly scalable and secure storage, allowing us to manage large volumes of compliance documents and contracts efficiently. This service will support:

• Storing legal document files (such as PDFs, text files).
• Metadata management to facilitate quick retrieval and organization of documents.

With AWS S3, the scalability and durability of the storage are ideal for this project, ensuring easy access and management of documents as the system expands.

4. Potential Development Tasks

Task 1: Data Acquisition and Preprocessing

• Details: Data gathering will occur through web scraping or API integration. Legal documents will be cleaned and preprocessed using Hugging Face and NLP libraries (SpaCy/NTLK). This will include removing boilerplate text, normalizing the formatting, and structuring the data.
• Guidance Needed: Moderate. Guidance may be needed in selecting the right data sources, handling large volumes of legal text, and managing API connections.

Task 2: Building the NLP Compliance Checker

• Details: The project will use Hugging Face transformers and NLP techniques to parse legal documents, identify key clauses, and match them against compliance rules. Named Entity Recognition (NER) will be employed to detect key terms such as "data controller" or "consent."
• Guidance Needed: High. Support may be required for defining compliance rules and setting up accurate entity recognition systems for legal documents.

Task 3: Machine Learning Model for Document Classification

• Details: A machine learning model will be trained using

Hugging Face-based document transformers on labeled legal datasets. This model will classify documents as compliant or non-compliant, using feature extraction, model training, and evaluation techniques.
• Guidance Needed: Moderate. Assistance with feature engineering and interpreting the model outputs to ensure legal validity might be necessary.

5. Benefits of the Compliance Checker

This project will provide significant benefits, particularly for individuals unfamiliar with legal rights and obligations, such as new residents in a country. When faced with lengthy legal documents, they may overlook key clauses or hidden fees, especially in cases like rental agreements. Our tool will help users identify compliance issues quickly, preventing potential legal pitfalls.
Additionally, users can store their contracts securely in AWS S3, and as a future extension, we could develop a chatbot-based interface where users can inquire about their legal rights and document compliance in real-time, using our NLP-powered compliance checker.