

## A.4 Final Project - Creating Chatbot - Written Component

Hantao Lin

MSDS 453 Section 57: Natural Language Processing

November 27, 2023

## Introduction

In the rapidly evolving arena of international business, the comprehension of diverse laws and regulations becomes essential for business owners to circumvent potential penalties. However, not every business owner has the luxury of time or resources to delve into these complex legal frameworks, especially when considering the significant variations in laws across different countries. Although many enterprises might have specialized legal teams, a foundational understanding of pertinent laws and regulations remains invaluable for business owners overseeing their operations.

This research paper introduces 'Alex', an innovative world data policy chatbot, engineered to promptly equip stakeholders with pertinent global data regulation information. Alex is designed to narrow the informational divide and simplify legal data accessibility, thereby enhancing decision-making for business owners and stakeholders alike.

In our research, we leveraged the hierarchical structure of ontologies as our logical foundation for developing the chatbot. The dataset employed encompasses a comprehensive array of data regulation elements, including definitions, laws, and online privacy policies from 158 countries. Our research unfolds in several phases, each exploring different methodologies to ascertain the most relevant country, section, and content. We employed various models, including TF-IDF, Multi-QA (multi-qa-MiniLM-L6-cos-v1), and Distilled BERT (distilbert-base-uncased-distilled-squad), to enhance the chatbot's capabilities. These models were qualitatively evaluated based on their output relevance and accuracy. The findings suggest that the Multi-QA model is optimal for identifying relevant countries and content, while TF-IDF excels in pinpointing pertinent sections.

## Research Design and Modeling Method

### *Reading the PDF*

To process the data, we will first read the PDF file using the Fitz library and convert it into a text file. The outcome of this step is a DataFrame containing all countries, along with their associated sections and corresponding content.

### *Preprocessing*

In the preprocessing phase, I manually removed common terms that lack semantic significance, such as footnotes, website URLs, and line break symbols. This minimal preprocessing approach was deliberately chosen to preserve the integrity of the text for extracting law entities using Named Entity Recognition (NER). NER is integrated into the dataframe to establish an ontology, forming the logical foundation for the chatbot. Additionally, I defined several key functions: a greeting function, a function to find the most relevant country/section, a function to use laws for country location, a response generation function, and the main chatbot loop function.

### *Part I - Ontology*

I constructed the ontology to examine whether the hierarchical structure I built is suitable for the chatbot's logic. The ontology is designed such that each country, section, and law is represented as a node. The relationships are established from higher to lower levels, where each country encompasses multiple sections, and each section may contain zero or more laws.

### *Part II - Country*

In Part II of our study, we conducted experiments aimed at determining the optimal method for our chatbot to accurately recognize country names in user inputs. All three models were implemented and rigorously tested using the same set of five questions. The results were then

evaluated based on each algorithm's ability to correctly identify the country name from the user input.

### *Part III - Sections*

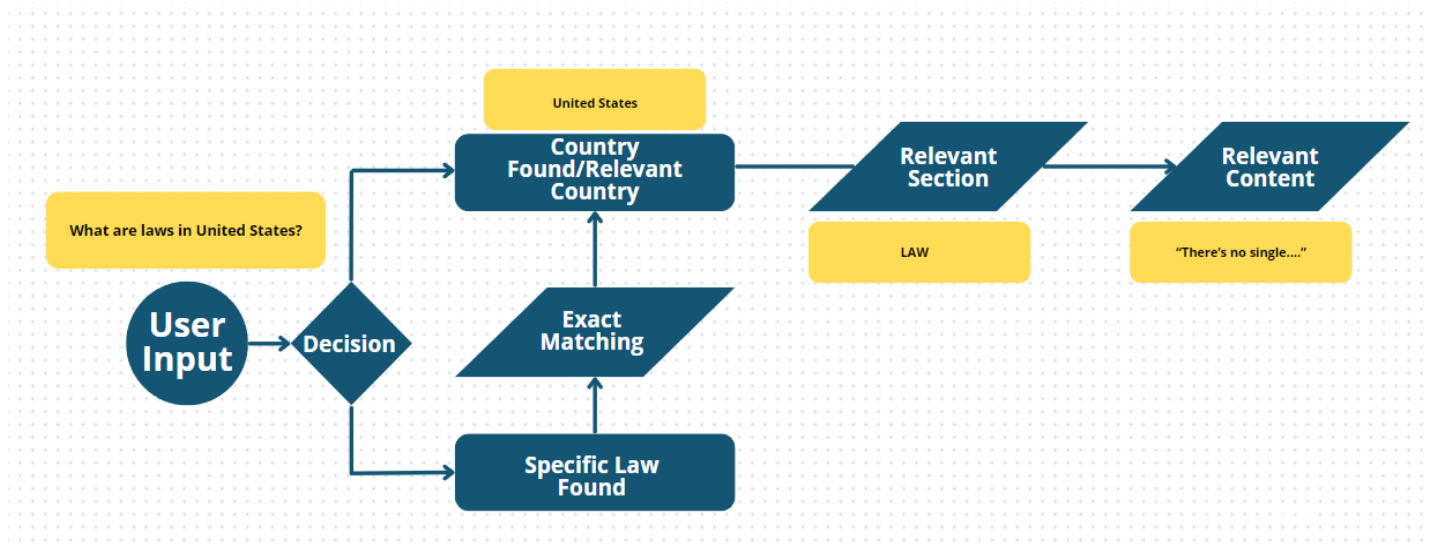
Similar to the methodology employed in Part II, Part III follows the same testing procedure.

However, the key difference lies in the structure of the questions used. In this phase, we are not only evaluating whether the algorithms can extract the correct sections, but also assessing their performance in scenarios where multiple sections appear in the user input. This includes determining if the algorithms can comprehend the question and accurately extract the relevant sections.

### *Part IV - Content*

In the final part of our study, we conducted experiments with all three models, integrating the optimal algorithms identified in Parts II and III. Furthermore, we incorporated additional logic to address situations where only a specific law is mentioned in the question. We qualitatively evaluated the output based on a set of five questions.

Here's an overview of the structure of Alex:



## Result & Analysis

### Part I - Knowledge Graph

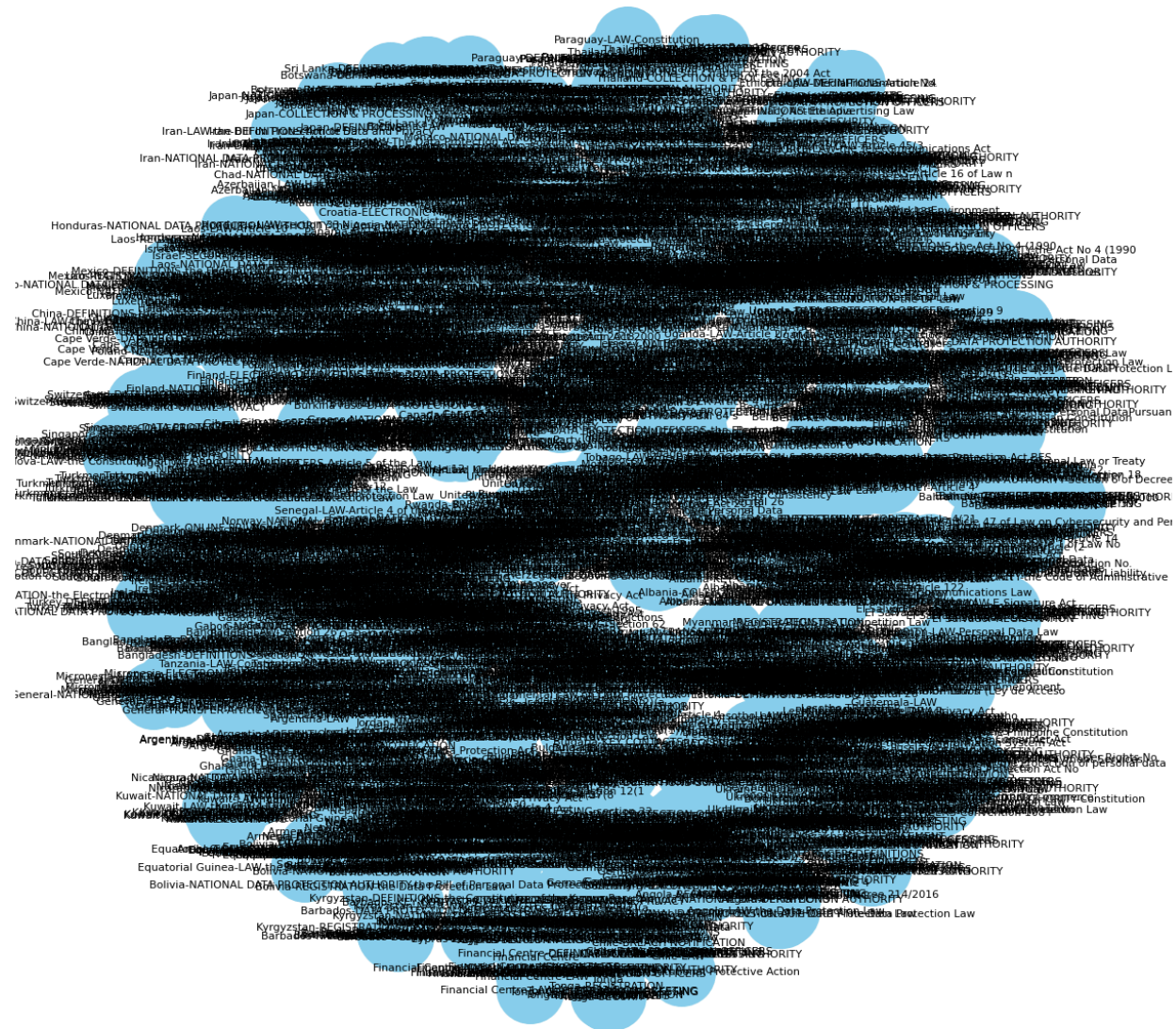


Figure 1. Knowledge Graph on 158 Country's data policy

As depicted in Figure 1, due to the size of the dataset, making it challenging to extract meaningful information. Consequently, we made the decision to concentrate our attention on specific pairs of entity relationships.



Figure 2. Knowledge Graph on United Kingdom

In Figure 2, we analyze the knowledge graph of the United Kingdom. As depicted, the United Kingdom exhibits a well-structured hierarchical system, wherein each section is associated with exactly one law entity. Upon closely examining the laws, it becomes apparent that some are truncated, and most do not specify the full name of the law. However, the method used (NER -

Named Entity Recognition) did yield some valuable insights. This ontology can be instrumental in addressing the challenge posed when only laws are mentioned in user queries.

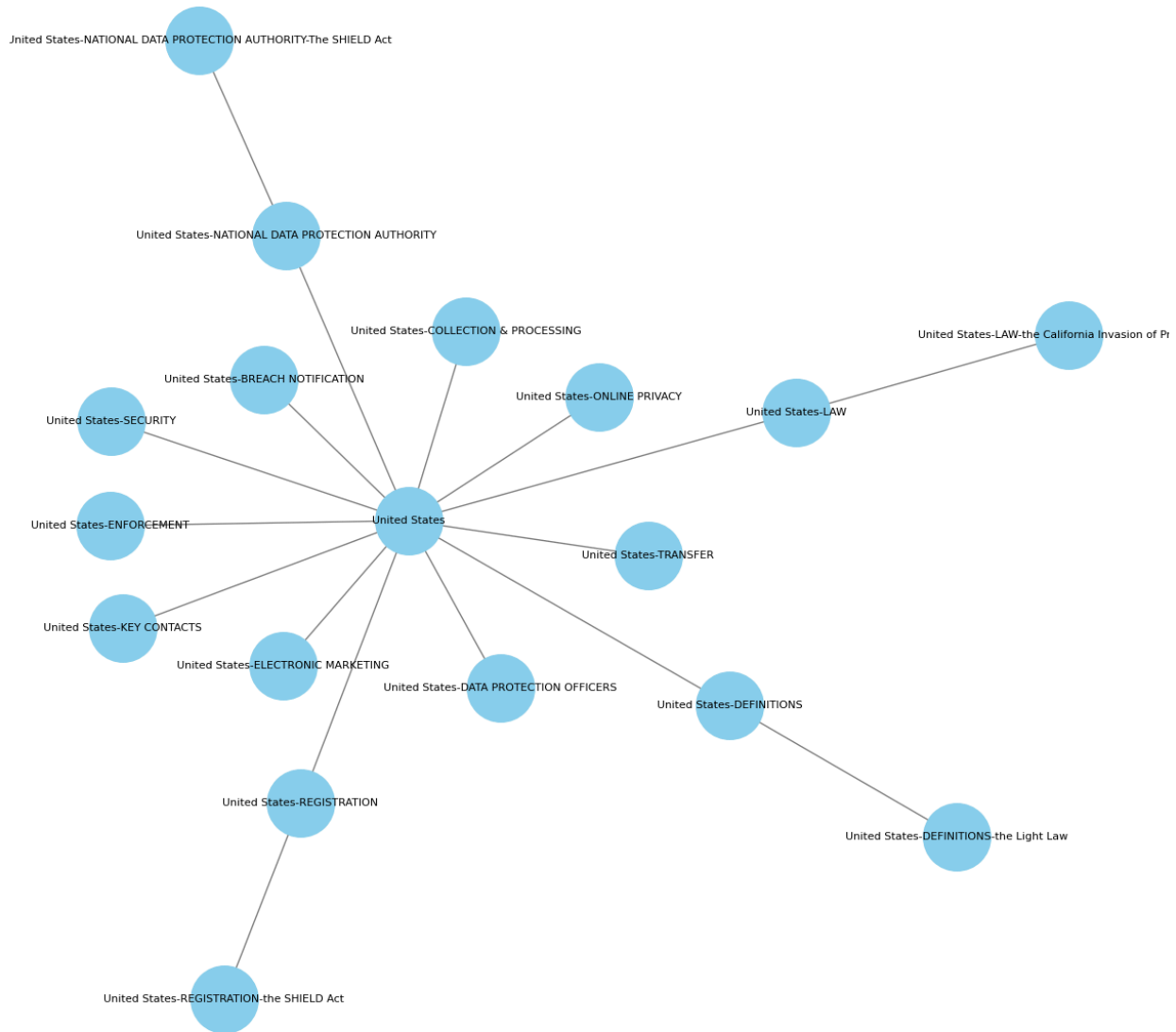


Figure 3. Knowledge Graph on United States

We also conducted an examination of the ontology for the United States. The analysis reveals that not all sections within this ontology have an associated law entity. Additionally, it is

observed that different sections may reference the same law. This insight is valuable and can be utilized to further refine and improve our chatbot.

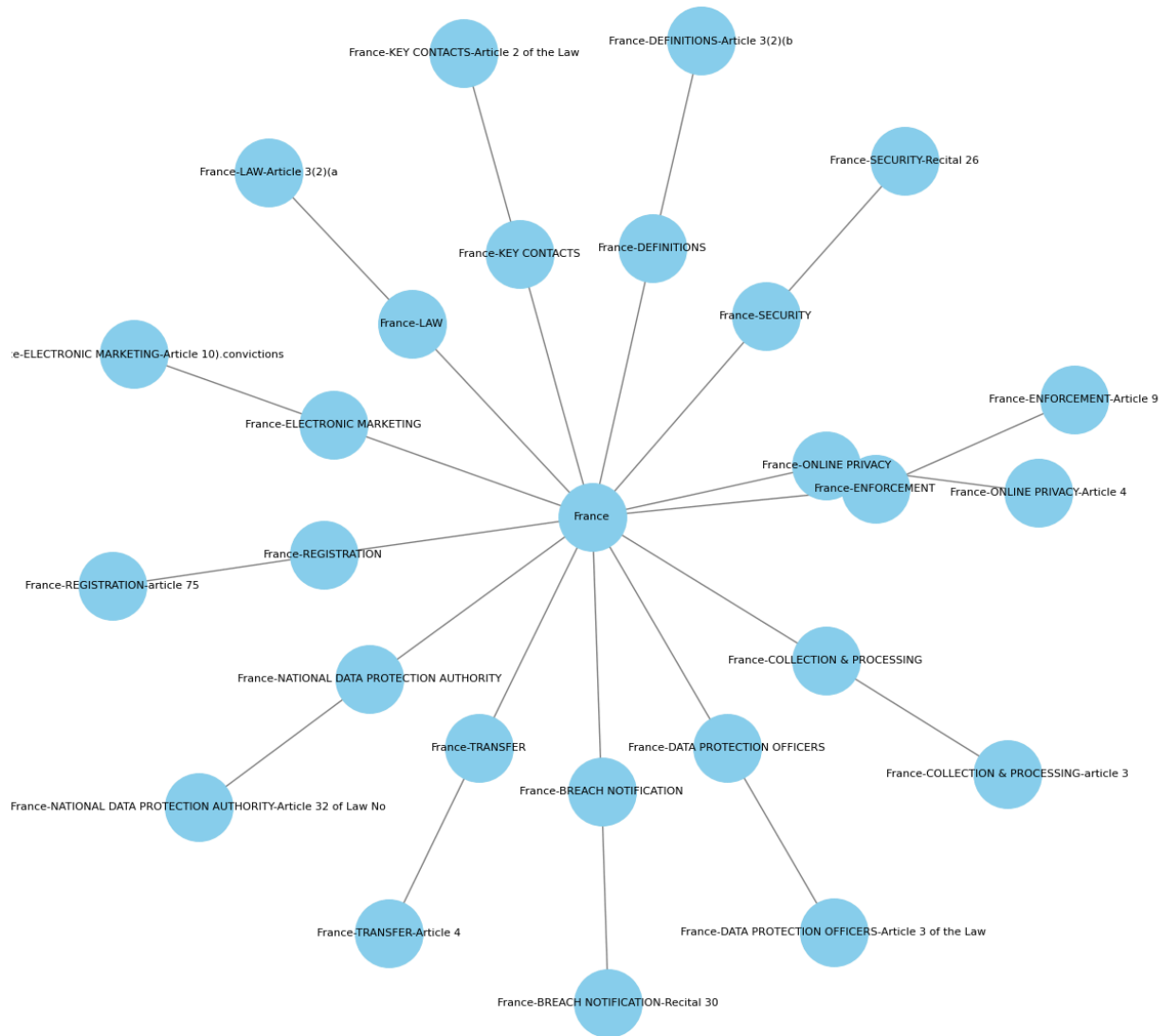


Figure 4. Knowledge Graph on France



The final ontology we examined pertains to France. An interesting insight from this ontology is that it demonstrates the possibility of multiple law entities being categorized under a single section. Additionally, we observed that some laws are truncated following the term 'No', an issue that was subsequently addressed in the response function of our system.

## *Part II - Country*

```
Welcome to the World Data Protection Chatbot. My name is Alex. To end session please type exit
You: what are the laws in United States?
Alex: United States
You: who are the data authority in Albania?
Alex: Albania
You: Wht are the regulation for online privacy in United Kingdom?
Alex: United Kingdom
You: What's the defintion of personal identify information in France?
Alex: France
You: Who are the key contacts for data regulations in Greece?
Alex: Greece
You: exit
Thank you for using the World Data Protection Chatbot. Good bye.
```

Figure 5. Country output from Multi-QA

As observed in Figure 5, the output of the Multi-QA model accurately extracts the country from user input

```
Welcome to the World Data Protection Chatbot. My name is Alex. To end session please type exit
You: what are the laws in United States?
Alex: Sint Maarten
You: who are the data authority in Albania?
Alex: Sint Maarten
You: Wht are the regulation for online privacy in United Kingdom?
Alex: Abu Dhabi Global Market Free Zone
You: What's the defintion of personal identify information in France?
Alex: Abu Dhabi Global Market Free Zone
You: Who are the key contacts for data regulations in Greece?
Alex: Sint Maarten
You: exit
Thank you for using the World Data Protection Chatbot. Good bye.
```

Figure 6. Country output from Distilled Bert Based Model

When analyzing the output presented in Figure 6, it becomes evident that the Distilled BERT-Based Model is not an optimal solution for extracting countries from user input. Not only

did it fail to accurately identify any countries, but it also appeared to be stuck in a loop, consistently outputting a specific country.

```
Welcome to the World Data Protection Chatbot. My name is Alex. To end session please type exit
You: what are the laws in United States?
Alex: United States
You: who are the data authority in Albania?
Alex: Albania
You: What are the regulation for online privacy in United Kingdom?
Alex: United Kingdom
You: What's the definition of personal identify information in France?
Alex: France
You: Who are the key contacts for data regulations in Greece?
Alex: Greece
You: exit
Thank you for using the World Data Protection Chatbot. Good bye.
```

Figure 7. Country output from TF - IDF

Based on the output shown in Figure 7, TF-IDF has proven to be a competent algorithm for identifying countries, as it correctly answered all queries related to countries.

In conclusion, both TF-IDF and the Multi-QA model are excellent choices for accurately determining the country based on user input. I have decided to choose the Multi-QA model as the foundation for the Chatbot's country matching algorithm.

### *Part III - Section*

```
Welcome to the World Data Protection Chatbot. My name is Alex. To end the session, please type exit.
You: What are the laws in United States?
Alex [United States - LAW]
You: Who are the data uauthorith in Albania?
Alex [Albania - NATIONAL DATA PROTECTION AUTHORITY]
You: Provide me the contacts of data protection officers in United States
Alex [United States - DATA PROTECTION OFFICERS]
You: Should I have a breach notification for online privacy data leakage in France?
Alex [France - BREACH NOTIFICATION]
You: who are the key contacts in United States regarding data privacy law
Alex [United States - NATIONAL DATA PROTECTION AUTHORITY]
You: exit
Thank you for using the World Data Protection Chatbot. Goodbye.
```

Figure 8. Section output from Multi-QA

From the results presented in Figure 8, it is apparent that the Multi-QA model is competent at extracting sections when only one section is mentioned in the user input but falls short when multiple sections are presented. Additionally, we observed that the algorithm is self-corrective when dealing with abbreviated section names.

```
Welcome to the World Data Protection Chatbot. My name is Alex. To end the session, please type exit.
You: What are the laws in United States?
Alex [United States - NATIONAL DATA PROTECTION AUTHORITY]
You: Who are the data uathorith in Albania?
Alex [Albania - NATIONAL DATA PROTECTION AUTHORITY]
You: Provide me the contacts of data protection officers in United States
Alex [United States - DATA PROTECTION OFFICERS]
You: Should I have a breach notification for online privacy data leakage in France?
Alex [France - NATIONAL DATA PROTECTION AUTHORITY]
You: who are the key contacts in United States regarding data privacy law
Alex [United States - NATIONAL DATA PROTECTION AUTHORITY]
You: exit
Thank you for using the World Data Protection Chatbot. Goodbye.
```

Figure 9. Section output from Distilled Bert Based Model

In Figure 9, the Distilled BERT-Based Model correctly answered only one question out of the five asked, making it an undesirable algorithm for this task.

```
Welcome to the World Data Protection Chatbot. My name is Alex. To end the session, please type exit.
You: What are the laws in United States?
Alex [United States - LAW]
You: Who are the data uathorith in Albania?
Alex [Albania - DATA PROTECTION OFFICERS]
You: Provide me the contacts of data protection officers in United States
Alex [United States - DATA PROTECTION OFFICERS]
You: Should I have a breach notification for online privacy data leakage in France?
Alex [France - BREACH NOTIFICATION]
You: who are the key contacts in United States regarding data privacy law
Alex [United States - KEY CONTACTS]
You: exit
Thank you for using the World Data Protection Chatbot. Goodbye.
```

Figure 10. Section output from TF - IDF

Surprisingly, with TF-IDF, the algorithm not only correctly identified questions with a single section mentioned but also outperformed the Multi-QA model when addressing the fifth question, where multiple sections were presented.

In conclusion, based on our evaluation of the five questions asked, it is evident that both TF-IDF and Multi-QA are viable choices for identifying sections within user input, with TF-IDF performing slightly better. Additionally, the Distilled BERT-Based model is not a desirable option for the task presented.

#### *Part IV - Content*

**You: what are the laws in United States?**

Alex [United States - LAW]: US states have also passed privacy and data security laws and regulationsthat apply across sectors and go beyond federal law—such as data security laws, secure destruction, Social Security numberprivacy, online privacy, biometric information privacy, and data breach notification laws

**You: what is GDPR**

Alex [Guatemala - LAW]: Guatemala does not have a personal data protection law, however the Law on Access to Public Information (Ley de Acceso a la Información Pública – Decree 57-2008 of the Congress of the Republic), even if it pertains to information in public files and records, does address the matter in certain provisions which can be applicable to private parties.

**You: what are the laws in Albania**

Alex [Albania - LAW]: As of December 2022 this draft law has yet to be approved by the Albanian Parliament. Earlier in the year, on 28 January 2022, Albania signed the Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, which was later ratified by Law No. 49/2022, dated 12 May 2022 "On the". Ratification of the Protocol Amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Dat

**You: who are the key contacts in United States regarding data privacy law**

Alex [United States - KEY CONTACTS]: DATA PRIVACY TOOL You may also be interested in our to assess your organization's level of data protection maturity. Data Privacy Scorebox Kate Lucente Partner and Co-Editor, Data Protection Laws of the World T +1 813 222 5927 kate.lucente@dlapiper.com Andrew Serwin Partner, Global Co-Chair Data Protection, Privacy and Security Group T +1 858 677 1418 andrew.serwin@dlapiper.com Jennifer Kashatus Partner T +1 202 799 4448 jennifer.kashatus@dlapiper.com

**You: Who are the data authority in United Kingdom**

Alex [United Kingdom - NATIONAL DATA PROTECTION AUTHORITY]: The Information Commissioner (whose functions are discharged through the Information Commissioner's Office ) is the ("ICO") supervisory authority for the UK for the purposes of Article 51 of the UK GDPR

Figure 11. Content output from Multi-QA

When examining the results produced by Multi-QA in Figure 11, we observe that the algorithm accurately extracts information relevant to the given question. Question 2, however, is an example of exact law matching that led the model to locate an irrelevant country and section.



Figure 12. Content output from Distilled Bert-Based Model

Considering the answers produced by the Distilled BERT-Based Model, the model generates reasonable responses. It is challenging to determine whether the answers produced by Multi-QA

or Distilled BERT-Based are superior for questions 1 to 4, but it is quite evident that Multi-QA yields far more relevant information for question 5.

**You: what are the laws in United States?**  
 Alex [United States - LAW]: There is no comprehensive national privacy law in the United States

**You: what is GDPR**  
 Alex [Guatemala - LAW]: Guatemala does not have a personal data protection law, however the Law on Access to Public Information (Ley de Acceso a la Información Pública – Decree 57-2008 of the Congress of the Republic), even if it pertains to information in public files and records, does address the matter in certain provisions which can be applicable to private parties.

**You: what are the laws in Albania**  
 Alex [Albania - LAW]: GDPR The complete harmonization of the current Albanian legislation in force on data protection with the GDPR has been one of the main objectives of the Office of Information and Data Protection Commissioner since 2018, however this objective has yet to be achieved (due in part to the Covid-19 pandemic). In June 2022 the Ministry of Justice of the Republic of Albania launched a public consultation process on a draft law "On Personal Data Protection" which is approximated with the GDPR

**You: who are the key contacts in United States regarding data privacy law**  
 Alex [United States - KEY CONTACTS]: DATA PRIVACY TOOL You may also be interested in our to assess your organization's level of data protection maturity. Data Privacy Scorebox Kate Lucente Partner and Co-Editor, Data Protection Laws of the World T +1 813 222 5927 kate.lucente@dlapiper.com Andrew Serwin Partner, Global Co-Chair Data Protection, Privacy and Security Group T +1 858 677 1418 andrew.serwin@dlapiper.com Jennifer Kashatus Partner T +1 202 799 4448 jennifer.kashatus@dlapiper.com

**You: Who are the data authority in United Kingdom**  
 Alex [United Kingdom - NATIONAL DATA PROTECTION AUTHORITY]: The Information Commissioner (whose functions are discharged through the Information Commissioner's Office ) is the ("ICO") supervisory authority for the UK for the purposes of Article 51 of the UK GDPR

Figure 13. Content output from TF - IDF

Comparing the answers produced by TF-IDF, it is also difficult to determine whether the previous two models significantly outperformed this one, as all models produced reasonable answers with the exception of the fifth question. In conclusion, I will use Multi-QA as my final algorithm for extracting relevant content because it not only outputs more content but also appears slightly more relevant than the other two models tested.



## Conclusion & Recommendation

Throughout our research, which involved visualizing ontologies for 158 countries' data policies and identifying the best models for country, section, and context relevancy, we have identified key findings and recommendations for each part:

### *Part I - Knowledge Graphs:*

Our ontology visualization process provided valuable insights, particularly in potential refinements such as customizing sentence breakpoints. It also highlighted the effectiveness of NER (Named Entity Recognition) in identifying law entities. While not flawless, NER was reasonably effective in identifying most laws present in the text. Further enhancement of the ontology is recommended through additional data cleaning, especially since some laws were incompletely extracted or truncated. Additionally, incorporating OSCR (Ontology-based Semantic Composition Regularization) could significantly improve our model's ability to recognize entity relationships, resulting in a cleaner and more effective model, as opposed to relying solely on hard coded logic.

### *Part II - Country:*

The experiments conducted demonstrated that both TF-IDF and the Multi-QA model are excellent choices for accurately identifying countries from user input. Future analyses involving more complex questions are suggested to further assess their capabilities.

### *Part III - Section:*

Our evaluation of five specific questions revealed that both TF-IDF and Multi-QA are effective for identifying sections within user input, with TF-IDF having a slight performance edge. However, the Distilled BERT-Based model was not a suitable option for this task. Introducing

more complex questions in future studies could help differentiate which model performs best. It is hypothesized that TF-IDF might struggle when the target section name is shorter in word count than non-target section names. Therefore, exploring additional refinements or alternative models is necessary to address these challenges.

#### *Part IV - Content:*

Qualitatively evaluating the results produced by all three models was challenging. Both Multi-QA and TF-IDF slightly outperformed the Distilled BERT-Based Model, particularly on question 5, where Multi-QA provided more detailed and subjectively more accurate information. A common issue with all models was the unclean formatting of output, primarily due to the initial conversion of PDFs into dataframes, where words were not properly segmented. Further data cleaning is required for clearer results. To enhance accuracy and relevance, the use of ChatGPT, an NLU (Natural Language Understanding) level model, is proposed. Furthermore, implementing OSCAR could notably improve semantic understanding.

In summary, this research venture offers a profound exploration into the integration of Knowledge Graphs and advanced natural language processing techniques within a question-answering chatbot framework. The findings not only underscore the efficacy of these technologies but also illuminate their transformative impact in the realm of information retrieval. By leveraging these sophisticated tools, stakeholders can now access pertinent data with unprecedented speed and precision, thereby revolutionizing the way information is consumed and utilized in decision-making processes. This study paves the way for future innovations in the field, heralding a new era of intelligent information systems that are more intuitive, responsive, and user-centric.