2022

Human Trafficking and Machine Learning: A Data Pipeline from Law Agencies to Research Groups

Nathaniel Hites
nhites@smu.edu

HUMAN TRAFFICKING AND MACHINE LEARNING:

A DATA PIPELINE FROM LAW AGENCIES

TO RESEARCH GROUPS

Approved by:

_____

Prof. Suku Nair

_____

Vanessa Bouché

_____

Frank Coyle

HUMAN TRAFFICKING AND MACHINE LEARNING:
A DATA PIPELINE FROM LAW AGENCIES
TO RESEARCH GROUPS

A Master's Thesis Presented to the Graduate Faculty of the

Lyle College of Engineering

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the Master's Thesis

with specialization in

Machine Learning

and

Artificial Intelligence

by

Nathan Hites

B.S., Computer Science, Southern Methodist University, Dallas

May 14, 2022

ACKNOWLEDGMENTS

I would like to dedicate this work to my mother and father and their everlasting belief in my ability to function, learn, grow, and ultimately pursue my dreams. Without such a strong belief in me, I do not believe I ever would have reached the stage in my life where I have become confident enough in myself and my abilities to create a work of this magnitude.

Hites, Nathan M.S. Computer Science, Southern Methodist University, Dallas

HUMAN TRAFFICKING AND MACHINE LEARNING:
A DATA PIPELINE FROM LAW AGENCIES
TO RESEARCH GROUPS

Advisor: Professor Suku Nair

Human trafficking is a form of modern-day slavery that, while highly illegal, is more dangerous with the advancements of modern technology (such as the Internet), which allows such a practice to spread more easily and quickly all over the world. While the number of victims of human trafficking is large (according to non-profit organization Safe House, there are estimated to be about 20.5 million human trafficking victims, worldwide ("Human Trafficking Statistics & Facts." Safe Horizon)- co-erced or manipulated by traffickers into either forced labor, or sexual exploitation and encounters), the number of heard cases is proportionally low- several thousand successful case prosecutions (Feehs K., p10-14). This disparaging fraction of unsettled human trafficking cases and trapped victims mandates that the system of fighting against human trafficking must be advanced.

This thesis presents an advancement of this field using a data pipeline that flows directly from law agencies and similar data-collecting groups to a web-based user-friendly interface that can be used for both research and analytical purposes and aims to allow legal-based efforts to proactively identify victims and traffickers as opposed to reacting to crimes after they happen. It displays data such as human trafficking case metadata (from title, to location, to verdict) and victim demographics (race, age, and sentence or conviction length, for example). This cleaned data is then stored and displayed through a Southern Methodist University-hosted infrastructure.

Currently, only one source of data is curated, used, and stored, but this groundwork pipeline is built for expansion for a wide variety of sources- one projected source being PACER, (Public Access to Court Electronic Records). This expansive and flexible quality adds to the pipeline's utility and projected future uses within the sphere of human trafficking discourse.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

What is human trafficking, and why is such a reprehensible practice so pervasive in the world, today? According to the United States' Department of Homeland Security, "Human trafficking involves the use of force, fraud, or coercion to obtain some type of labor or commercial sex act" (Homeland Security). Every single year, millions of men, women, and children are trafficked for one reason or another worldwide through these criminal operations and organizations. The repercussions of trafficking on a victim may very well scar them for life- whether that be physically, mentally, financially, or even all three at once.

> "[Traffickers] look for people who are susceptible for a variety of reasons, including psychological or emotional vulnerability, economic hardship, lack of a social safety net, natural disasters, or political instability. The trauma caused by the traffickers can be so great that many may not identify themselves as victims or ask for help, even in highly public settings" (Homeland Security).

While it is thus natural to want to persecute traffickers for the crimes they commit, it is not always as easy as pointing a finger at a human trafficker and sending them to jail. Many human trafficking rings are underground organizations that are difficult to reactively track and even harder to proactively uncover. At least six hundred of the reported cases in 2020 came directly from an investigation launched, not from curated evidence, but in response to information referral regarding an episode of human trafficking- typically by either the victim, a victim's family member, or another agency (such as a trafficking hotline) (Feehs, K, p64).

Moreover, the process of uncovering, arresting, and ultimately prosecuting human traffickers is drawn-out and time-consuming. Most successful human trafficking prosecutions stem from very lengthy and complex investigations, sometimes spanning up to multiple years in length- and at least 88% of all cases in 2020 involved the collaboration of multiple agencies, such as the FBI or the HSI (Feehs, K, p69). Typically, trafficking cases are prosecuted under US Code Chapter 77- a document that contains laws regarding peonage, slavery, and trafficking (Feehs, K, p74) and their appropriate sentences and fines for conviction. While prosecutors can charge *outside* of Chapter 77 (for instance, if the buyer's role in a trafficking case cannot be properly ascertained or is limited by ignorance, they cannot be charged under § 1584, as only "[w]hoever **knowingly** and **willfully** holds to involuntary servitude or sells into any condition of involuntary servitude, any other person for any term, or brings within the United States any person so held, shall be fined under this title or imprisoned not more than 20 years, or both"), charges outside of Chapter 77 are typically only in response to other determining factors in the case- such as the excessive use of force (which falls under § 1047.7) or exploitation of minors in a child-only case (which falls under § 2252) (Legal Information Institute). "Due to the often-complex nature of human trafficking prosecutions, it commonly takes federal courts several years to resolve charges against defendants in human trafficking cases, particularly when a defendant goes to trial." (Feehs, K, p108). In 2020, the average human trafficking prosecution lasted over three years to completely resolve all levied charges.

While much of this evidence required for these prosecutions comes from field work- for instance, discovering an abandoned warehouse or raiding a small business that covered the original base of operations- most US-based human trafficking takes place online (Portnoff, p1)- a sphere of criminal activity that requires very different types and amounts of time, effort, and

resources. Because multiple avenues of investigation are typically required to solve a case (not to mention the psychological damages of a human manually combing through scores of internet forums that contain nefarious imagery and activity), measures need to be taken to speed up and, ultimately, automate the ability to discover information about a case before it is even required to be legally used for prosecution.

As the current methods of information gathering are predominantly reactive, any solution's goal should be to allow prosecutions and investigations to be conducted proactively instead of reactively, with fewer resources and fewer legal organizations. Ideally, this solution would come in the form of "real time analysis" of any collected data to identify trends and act before traffickers are allowed to strike, as the current average human trafficking case turnabout time is two to three years (Feehs, K., p108), which allows plenty of time for undiscovered rings to relocate elsewhere to avoid punishment and further obscure their hapless victims from the law.

This thesis will outline one such solution to this problem- a data pipeline that will serve as the framework to leverage the power of machine learning and artificial intelligence, combined with data gleaned from agencies such as the Public Access to Court Electronic Records (PACER) to proactively identify trends to discover human traffickers before their crimes can be brought to fruition. The second chapter will outline previous subject work- what has already been done within this sphere of computer science; the third chapter will outline the presented pipeline and what can be accomplished with its use; and the fourth chapter will outline the utility of such a pipeline and data it has gathered.

CHAPTER 2: PREVIOUS WORK

As the reach of human trafficking continues to spread around the world at an alarming rate, any work that can speed up the prosecution of human trafficking cases are invaluable assets against this ever-growing criminal empire. Machine learning (and associated exploratory data analysis) is a powerful tool that can be used to analyze and assess trends without the manual scraping and correlation of hundreds of thousands of pieces of data; and its uses are still being discovered and leveraged, today, to better combat the spread of human trafficking. The two most common applications for machine learning are the use of natural language processing and image identifications- both of which can be utilized in human trafficking prosecution.

2.1 Machine Learning, Image Identification, and Human Trafficking

Researchers at Microsoft are currently working to create a machine learning pipeline to identify student uniforms in the United Kingdom using computer vision, with the goal of finding from which school that specific uniform originated, based on nothing but a human trafficking image from the police. Ultimately, this type of analysis provides quick and somewhat conclusive information about the location where the child was abducted, which then allows authorities to mobilize and respond appropriately. This expediency is especially important when solving human trafficking cases, as current manual internet-scraping methods are very slow and increase the length of time of an initial investigation. According to Mukherjee, Sumit, et al, it only takes seventy-two hours for a human trafficking case involving younger students (who are typically less equipped to fight back than adults) to become unsolvable, and the culprit untraceable. While their model is still in development, it has shown effectiveness across smaller images, and they believe that, with more time devoted to fault tolerance and parameter tuning, their research will

substantially reduce the human verification time for uniforms and thus increase the speed at which police can respond to initial hits.

2.2 Natural Language Processing and Human Trafficking

When accuracy and response time are both important metrics to the success of a task, automation is almost always a consideration- whether that be through a pipeline, upgrading toolchains to prevent human oversight, or automating data acquisition and scraping. In 2019, during the European Intelligence and Security Informatics Conference (EISIC), researchers at MIT presented a model that would classify online ads by processing ad-specific grammatical structures (as opposed to simply a "bag of words" keyword classification), because traffickers commonly obfuscate text with coded keywords for their services- such as "#Prepago" (prepaid): "Current research has primarily relied on pre-determined keyword attributes when building human trafficking detection systems and have not had access to truth data" (Zhu, Jessica, p2). This lack of a ground truth is important when training machine learning algorithms, as without a ground truth, it becomes impossible to judge the "correctness" of a solution.

Testing this model on the Trafficking-10K dataset (a set of ten thousand adult service advertisements scraped and subset from Backpage.org, the second highest-traffic site for human traffickers besides Craigslist) yielded very strong results in non-obfuscated data instances (such as in a sample Oregon case, which spanned only local crime rings), but struggled in cases such as the infamous sex trafficking ring in Florida, where the original message is hidden behind another business or context such as a massage parlor or Air-BnB (Zhu, Jessica, p6).

Keywords are not the only natural language structures that can be processed and analyzed to discern the identity and actions of traffickers. One year earlier, as part of the 2018 IEEE International Conference on Big Data, several computer scientists at MIT (many of whom had

already worked on the previous EISIC model) published their work with an unsupervised-and-scalable natural language processor that utilized "template" human-trafficking online advertisements to quickly detect posts on social networks such as Twitter, Facebook, and Craigslist. Previously, most text-based machine learning assessments used some form of term-frequency-inverse-document-frequency (tf-idf) or linear discriminant analysis (lda), which is not easily scalable with respect to human trafficking advertisements- as such ads continue to evolve and change to better evade and confuse keyword-based filtering methods. Thus, this system instead uses 'word vector' distributional filters such as GloVe and FastText as its base instead of exclusively focusing on keywords. However, as these filters also rely on an assumption about the data- that "words [and phrases that] occur in the same context tend to have similar meanings" (Li, Lin, p3)- more methods must be used to properly cluster data.

The model employs density-based clustering with HDBSCAN, a common machine learning clustering algorithm, to help abate the inherent assumptions of GloVe and FastText to find template advertisements, as HDBSCAN does not require any assumption about the type of data and only clusters similar occurrences by contexts- regardless of the input data's form. This very quickly provides insight on current advertisement structure, as many human trafficking advertisements on social media tend to resemble other human trafficking advertisements in the same way that fraudulent tech support or banking scams are all similarly scripted. Within a sample space of one hundred thousand scraped ads (mainly from Backpage.org), this approach was able to identify over one hundred unique phone numbers of different organizations. This is extremely useful in situations where evidence is scarce- one or two discovered phone numbers during an investigation can be traced to like-grouped phone numbers to track down the rest of an associated crime ring. Because human trafficking does not typically stop until the root of an

organization is found (and any traffickers not found by an investigation are relocated elsewhere to continue operation within another sector of a larger crime ring), "[v]ictim-centered policing can [greatly] benefit from the ability of identifying the greater organization which in turns results in collaboration with federal or other jurisdiction law enforcement to target the breadth of the criminal enterprise and disrupt the root of the cycle of exploitation." (Li, Lin, p7).

2.3 Use of Blockchain to Analyze Human Trafficking

Traditionally, human traffickers used financial institutions such as banks and investment accounts to launder money to fund transactions. "As banks become better at identifying the proceeds of human trafficking, [however], criminals will find new ways to make it harder for investigators to understand and trace the movement of funds" (Ayden). Recently, many have turned to cryptocurrency and other online payment methods, as they are chiefly considered 'anonymous' and thus attractive to hide party identities. However, crypto transactions can more accurately termed 'pseudonymous': "[i]n most cases, the originating address, destination address, and amount of funds associated with a transaction are permanently and, generally, immutably recorded on the blockchain" (Ayden). Because several human trafficking websites and services use payment methods such as Bitcoin, analyzing the blockchain can help discover not only a trafficker's identity, but also any associates of that identity. The knowledge of common techniques, such as peeling chains (a technique to divide and send large amounts of money through a fleet of smaller transactions), or public addresses for sexual activity sites (such as Backpage.org, perhaps one of the largest crypto partners that offers such services) can also be leveraged to track down traffickers and all other wallets that interacted with them. Thus, if a trafficker is found, their accomplices may also be found, and in 2019, this same method was used to arrest the owner of Welcome to Video, the largest child sexual exploitation market by volume

(with over 250,000 child exploitation videos, downloaded by at least one million users), which commonly sold videos for Bitcoin, a form of currency used by most blockchain transactions. Following this chain of transactions also led to an arrest of 337 subjects all over the world (Ayden).

Manually scrutinizing blockchain strands is a useful tool for combatting the spread and reach of human trafficking, but, while largely successful, contemporary methods of analysis are slow at best. The investigation of Welcome to Video began in early 2018 and did not conclude until October 16, 2019, a turnaround time of roughly one and a half years. Thus, many researchers investigating the usefulness of blockchain as a criminal investigatory tool began work on automatic analyses of scraped internet advertisements, which mapped each ad to bitcoin wallets and searched for similarities between each mapping. Rebbecca Sorla Portnoff of the University of California published a dissertation in 2018 discussing this very problem: "Although these ad sites provide a significant source of potentially incriminating data for law enforcement, monitoring these sites is unfortunately a labor-intensive task", (Portnoff, p2) as new advertisements appear by the thousands each day. Further, understanding the nature of these advertisements (determining whether they *are* human trafficking advertisements, at all, and not simply bound to other sexual-related activities) requires a specialized subset of domain expertise that also limits how quickly analytics can be created, parsed, and understood.

Her work details a PCA-based stylometry classifier that maps different sexual advertisements by author with a ninety-percent true positive rate and a one-percent false positive rate. This classification can then be used to bridge Bitcoin wallets and transactions together, as the identity of one advertisement's author will match that of *all* transactions made from that same wallet, even across advertisements. This research will eventually lead to a ground truth that

can be used for further research, and she is currently working with several law organizations to continue to expand this tracking capability- as "finding more connections between previously unconnected ads – i.e., finding more owners and grouping those ads by owner – is key, [as, if] those ads include movement across multiple states/geographic locations, with multiple parties involved, it is highly likely that a trafficker or trafficking ring is responsible." (Portnoff, p76).

CHAPTER 3: THE DATA PIPELINE

3.1 Solution Objectives

Investigation and analysis of human trafficking cases poses two main challenges: a lack of time and a lack of data. Because most trafficking transactions (at least in the United States) currently take place over the Internet (Portnoff, Rebecca, p3), more attention should be given to data retrieval methods across this online medium that can easily scrape and process data or present it in a usable manner for quicker analysis and response times. Previous solutions to this problem allow for content to be gleaned and data to be analyzed, but these experiments needed to personally acquire and scrape their own data, which directly contributes to slower responses and extended time for analysis as data is required to be cleaned and prepared for analysis. This data pipeline, thus, aims to solve both issues. As a system that is lightweight, easily scalable, and easily accessible, it lays the foundation for a central database that law enforcement and researchers alike can access for both analytical and investigative purposes.

For heightened maintainability and expansion, this pipeline must also exist completely within the infrastructure of Southern Methodist University, as the predecessor to this project was unable to be quickly upgraded due to technology mismatches between that original project and its team. Thus, only languages and frameworks common to SMU's curricula will be used for development, allowing for more readily available maintenance and uptime to all parts of the pipeline by all members of any SMU-based team.

3.2 The Data Pipeline

This pipeline is built to ingest case-related data from sources such as PACER and manual data imputed from querying law agencies such as Bloomberg Law, West Law, and Lexus Nexus, and display it in a manner conducive to both end users and researchers. The data is protected behind servers internal to Southern Methodist University to prevent unauthorized or unapproved use and access, as such data is considered sensitive or (in PACER's case) private.

The data pipeline is split into three parts- the end users (and appropriate login servers), the internal SMU components (which houses both the frontend, the backend, and the middleware that connects them), and the APIs to source data provided by organizations like PACER. There are two main servers housed on SMU's infrastructure- a Windows 10 server that services the user-facing frontend accessed with the domain of Human Trafficking Data (abbreviated as HTD) and a MSSQL server that contains the database of all ingested case information.

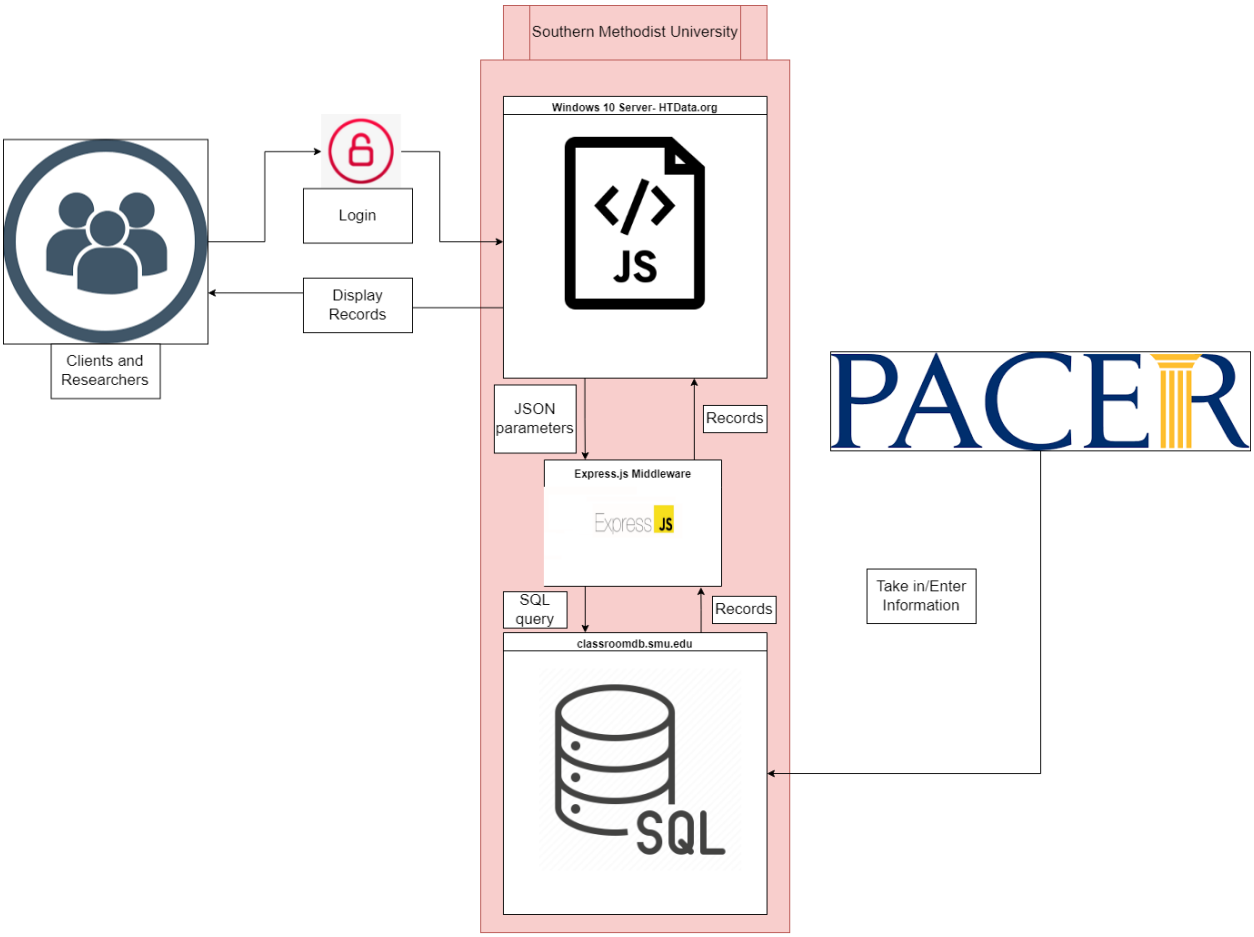

Figure 1: The Data Pipeline

Authentication is done with a user-based access control scheme before attempting to query any data to protect its integrity. Because new registered users can only be created internally, this allows SMU to control more easily who should (and should not) receive access to the data, based on the scope of projects that may concern such data. With a comparatively small number of researchers requiring access to very sensitive data, this level of heightened, controllable security, is more desirable than the broader scope and ease of role-based security (VanMSFT).

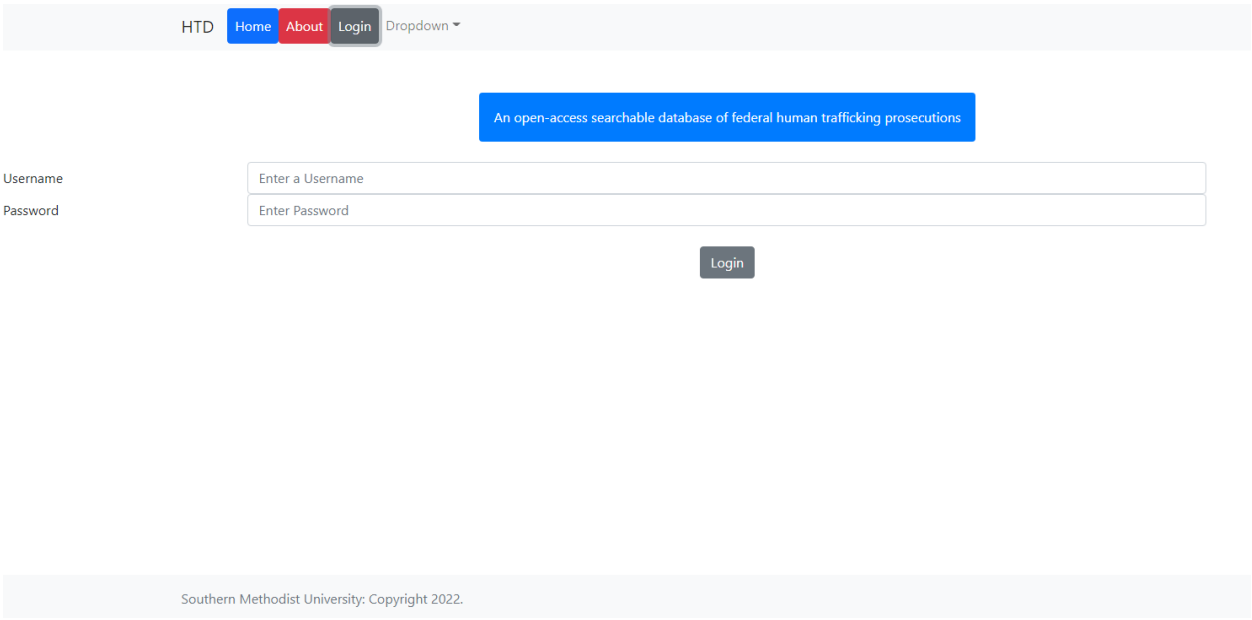

Figure 2: HTD Login Page

Once the user is authenticated, they can search for different cases by given keywords in their respective fields (for instance, by name of defendant, by date, or by state the case was prosecuted in). A list of all searchable keywords is provided in Appendix A1.

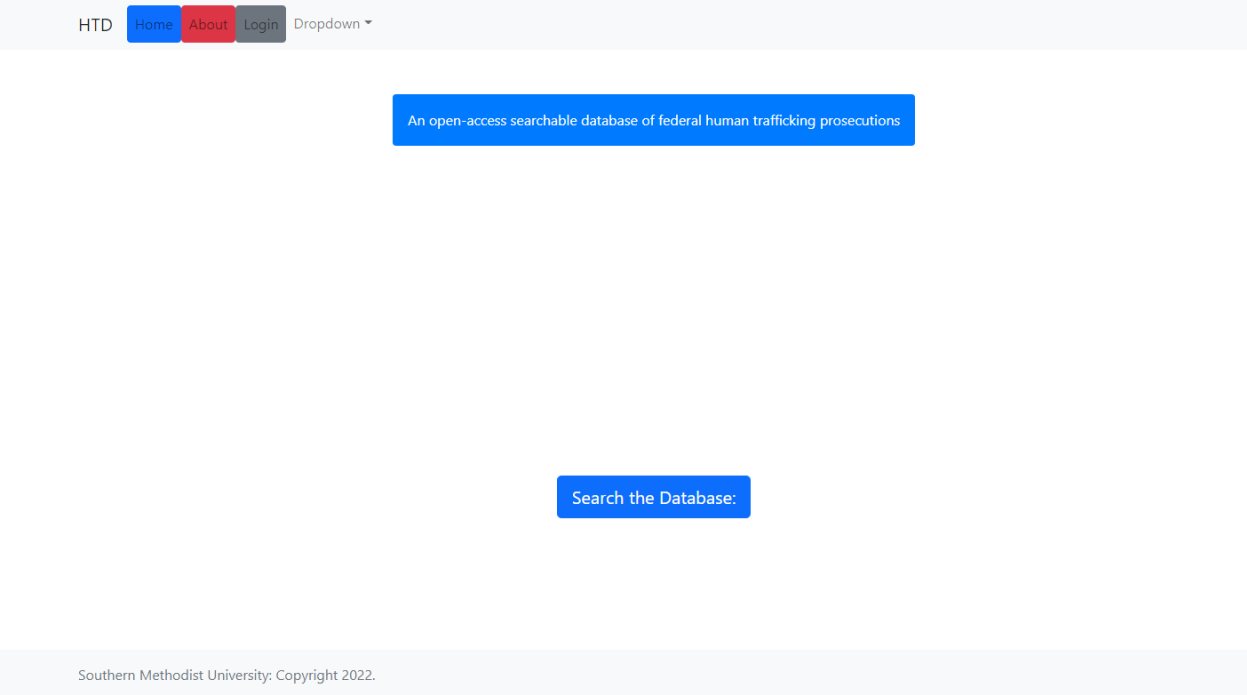

Figure 3: HTD Landing Page (unauthenticated user)

All results will be displayed in internal database order, unless a range of dates is specified, in which case each result will be displayed in descending order (from most recent case to least recent case).

3.3 The Frontend User Interface

The pipeline's frontend is built of three main pages- an initial landing page, a login form, and a displayable form that doubles as a search results page. This landing page presents introductory information to the user- information regarding HTD and its purposes- and gives them an option to search the database. This searching is done through an off-canvas bar form that can be easily dismissed or refreshed. Searching can be done using any or all the given criteria- a query will be dynamically created based on the content of the search form upon submission.

Cases are shown using the following four fields: the case's name, the starting date for prosecution of the case, the state in which the case was prosecuted, and a link to the public case file for further information.

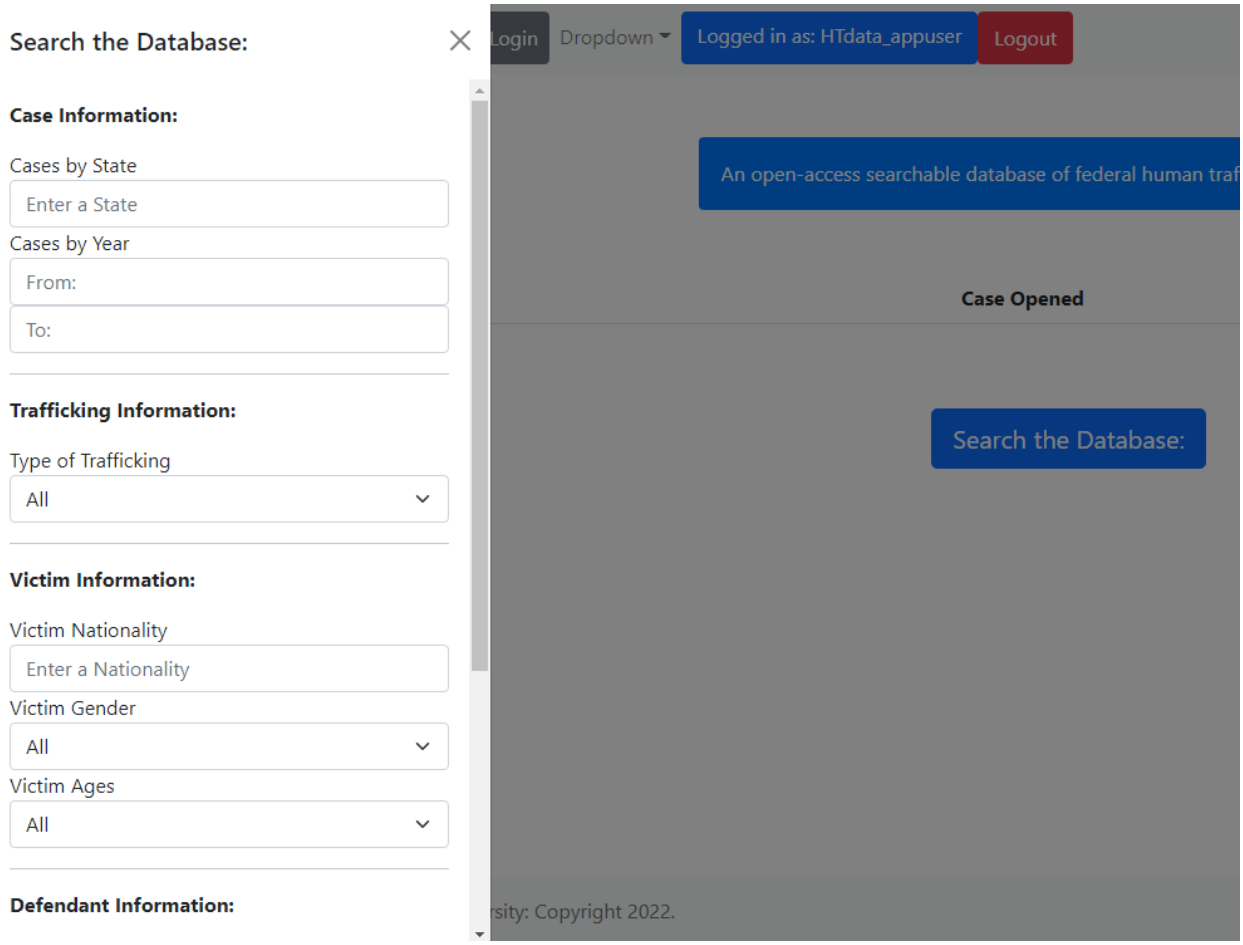

Figure 4: HTD Database Search Form

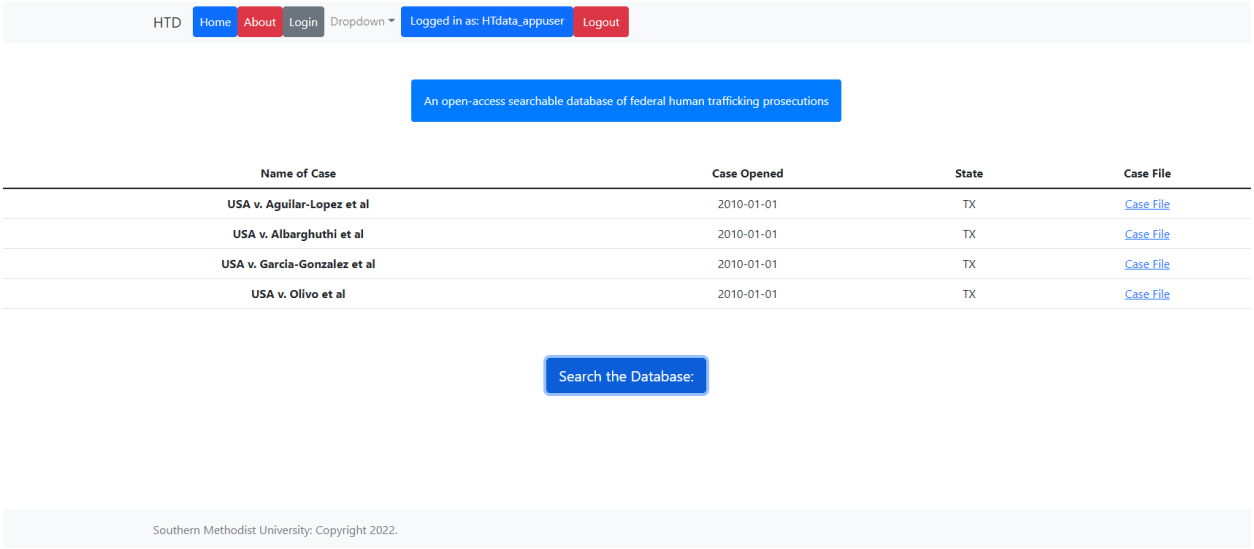

Figure 5: HTD Search Results Page

No information is permanently stored on the frontend, neither login information nor database query results, to prevent unwanted data retrieval. Inputs are sanitized using standard JSX practices to prevent injection attacks and further secured by Southern Methodist University's VPN as a primary data egress measure.

3.4 The Backend Server

The backend server was created from the work of Ms. Vanessa Bouche, featuring information that categorizes each human trafficking case based on a predefined protocol provided by Ms. Bouche's work- from the metadata that defines it (such as the case's name, the state in which the case was heard, a summary of the proceedings) to demographics of each victim, defendant, and presiding judge. It is divided into nine different tables- HTAppUsers, HTCrimeLocations, HTEntryPoints, HTJudges, HTCases, HTDefendants, HTVictimCountries, HTCriminalMethods, and BaseLocations.

HTAppUsers is only accessed and used when first connecting to the database through the frontend. It contains secured user data and is queried when attempting to login, which is used to validate authentication against stored, protected credentials. This is common practice as a stand-in for native back-end authentication and is used in most ASP.net and React applications.

HTCrimeLocations contains geospatial information about where the predominant crime of a case occurred. There can be more than one crime location for a case, so this table is commonly grouped with HTEntryPorts, HTCriminal Methods, and BaseLocations- other tables that contain data regarding crime locale and all associated travel of traffickers- to better chart and understand all the criminal activity of a specific case. HTEntryPorts denotes ports where traffickers entered the country (either by sea or by air) by their address, city, state, and country. HTCriminalMethods denotes the types of transportation and documents- legal and misused or illegal and faked- used to facilitate the trafficking in question. BaseLocations contains information on all known bases of operations for human trafficking- whether that be a cover-up business, an abandoned building, or other hub for a given case's criminal activity.

HTJudges contains professional information about the judge or judges that sentenced a given case, including the court they were appointed to (and by whom) and the number of human trafficking cases that they have presided over during their career.

HTDefendants holds all the necessary demographic information (such as name, age, and race) about a defendant in a case, as well as any sentences, penalties, fines, or probations accrued as part of the case they were tried for- including the specific statutes cited for each punishment (how many counts of specific charges under what U.S. law statutes were given, and how those counts were then dealt with- guilty by plea, guilty by trial, or otherwise).

HTVictimCountries is an auxiliary table that contains the country of origin for any victim or victims of a case. This is most useful when compared with HTEntryPorts to discover patterns and uncover rings by analyzing their common activities- where a trafficking operation starts (in what country, found in HTVictimCountries) and where an operation continues or ends (found in HTEntryPorts).

Finally, HTCases contains all the "case-level information"- data and metadata values that do not change based on the defendants, location, or other factors, in a case. This table serves as the bridge to the other tables in the database- as the "case_id" of HTCrimeLocations or HTCriminalMethods or HTDefendants will reference the same "case_id" of each case in HTCases.

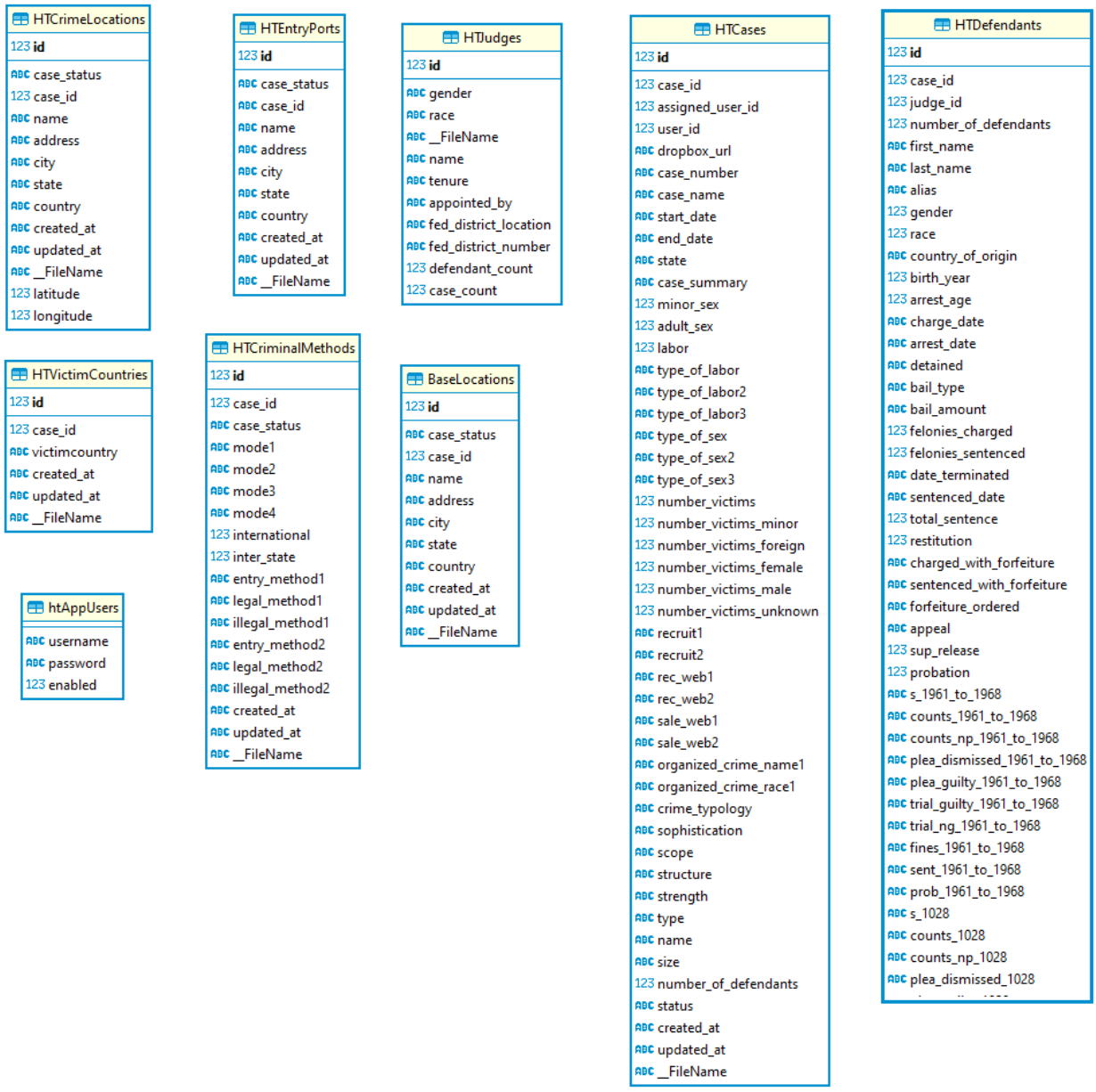

Figure 6: HTD Backend Tables

A brief explanation of each table's key constraints, and other implementation details are contained in Chapter 3.5.

3.5 Implementation

As the pipeline will be maintained and curated through Southern Methodist University, each piece of the pipeline's infrastructure is implemented in a language or framework that is either known by or familiar to most students and faculty that would be staffed to help maintain it.

The frontend is built with a React framework- a component-based declarative programming platform that is used in SMU's CS 3345, Graphical User Interfaces, to create the user interface for the class's joint final project (with the correlated CS 3330, Database Concepts) by combining Javascript and HTML to create scalable web applications. Released in 2003, React is a stateful, input driven framework that is both conceptually easy to understand and still widely used to this day, even in the professional world, due to its rules mandating that an application be broken down into its components, or "building blocks". The decision to render components to construct entire pages both adheres to SMU's programming paradigms and allows for each piece of the application to be edited and contained separately.

The backend is written in standard ANSI SQL and hosted using a 2007 MSSQL server, bridging the knowledge of SMU's Computer Science curriculum (CS 3330, Data Concepts; and CS 5/7330, File Organization and Database Structures) and the technology currently used for research and analytics on campus- classroomdb.smu.edu. According to the Office of Information Technology's Lane Duncan, "classroomdb.smu.edu is a MSSQL server dedicated to academic and research use- which allows any student, faculty, or organization on the campus [(or through the use of a SMU-approved VPN)] to access it without having to worry about firewall rules or middleware," (Bouche, Vanessa) provided they possess valid user credentials for the database. Because classroomdb.smu.edu is used for the curation and housing of many different servers,

each different application is hosted through its own port. This application uses port 55433 for its MSSQL traffic.

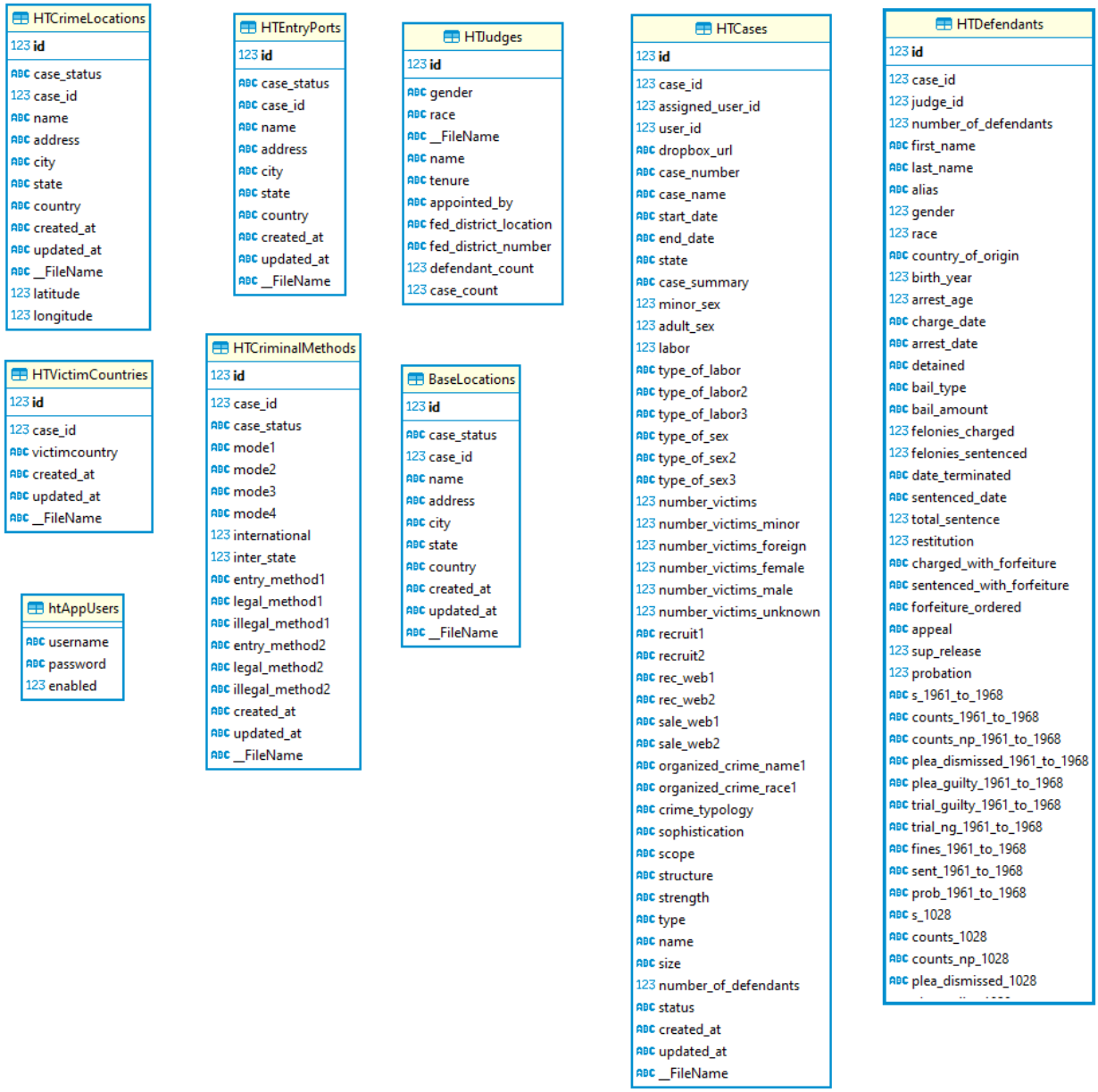

Figure 7: HTD Backend Key Constraints

The original database schema was generated by a Ruby on Rails application that hosted Vanessa's previous work, following an Active Association convention that connects tables and records by named variables rather than foreign keys. This allows for express-automated creation and testing, as one can create records and query them without satisfying strict foreign key constraints. This named unit variable for human trafficking cases is the "case_id" value, contained in the HTCases table, that then links to every other table in the database. As an example, to find a crime's location, the "case_id" of a particular case (as given by HTCases), indexes into HTCrimeLocations's "case_id", which holds the matching address, country, and state fields. All other information across tables can be found with similar procedures- by linking that one "case_id" to all matching "case_id's" in other tables.

As a single instance, with respect to this dataset, is a case, and any further information regarding other entities within that case (such as crime locations, defendants, and victims) can only be found through these smaller tables (who all also index by that same "case_id"), this organizational scheme represents data in the commonly accepted standard for databases- third normal form.

Because React cannot communicate directly to a backend database, Node.js (within the Express.js framework) services database queries and serves as the middleman between the frontend and the backend. The Express.js server's functionality mirrors the types of Node.js servers used connect CS 3330 and CS 3345 projects, and a list of all back-end endpoints is contained in Appendix A2. The frontend interacts with this server using standard Axios HTML request-and-response conventions that are handled internally by Express.js' JSON parser. The server runs on port 5000, which is commonly known for TCP and UDP, but can also be utilized for these types of internal data transfers.

To allow for more controlled data moderation, each query that is received is internally displayed and printed to a log file for bookkeeping purposes. The format of this output is shown in Figure 8.

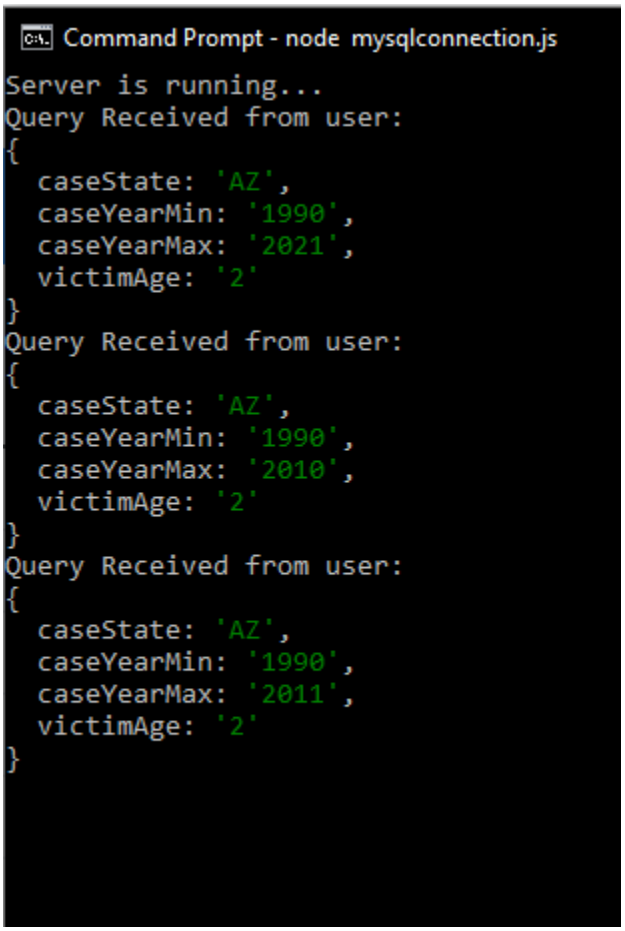

Figure 8: HTD Express.js Server Output

Both frontend and backend are hosted on independent 2012 Windows Servers (the R2 standard) with 4GB of RAM and 2 CPUs dedicated to processing, maintained as virtual machines in VMware, with the middleware existing on the same server as the backend to interface with it more easily. This is technology common to most student-built servers and

applications of this size hosted within SMU and was deemed appropriate for the current size of the pipeline. These specifications can be expanded without issue when the project grows beyond its current scope of a single-sourced frontend.

CHAPTER 4: BENCHMARKS AND ANALYTICS

The most common method to assess the quality of a user-facing frontend is the speed and simplicity of the user experience- namely, its adherence to KISS (namely, "Keep it Simple, Stupid") web design principles. The purpose of the pipeline is to deploy and maintain a searchable frontend for human trafficking case data- and as such, it must offer that functionality quickly, easily, and painlessly. For the pipeline to be considered usable, it should not be challenging to perform any task associated with querying the database- from first connection and authentication to querying and data retrieval.

4.1 User Experience Guidelines

Software development and usability consultant Steve Krug presents several tenants in his book, *Don't Make Me Think*, many of which are presented within SMU's user interface curricula. Chiefly among these tenants is his definition of usability: "Usability is making sure something works well, and that a person of average ability or experience can use it for its intended purpose without getting hopelessly frustrated" (Lawrimore, John, p3). Ideally, all features that require user interaction, such as forms and links, should be easily accessible, intuitive, and, by their design, construct and create the complete scope of the website- in this case, a central, interactive human trafficking repository that can be queried for research and analytical purposes. "If [any part of this process] cannot be self-evident, [the website should] make it self-explanatory" (Lawrimore, John, p9).

Krug's first principle establishes the need for a clear visual hierarchy between all content within any page. For this pipeline, queried results are the target information for any end user and

are featured prominently in the center of the webpage in a bold-faced table. No other information is presented while database queries are present to allow all focus to remain on these results.

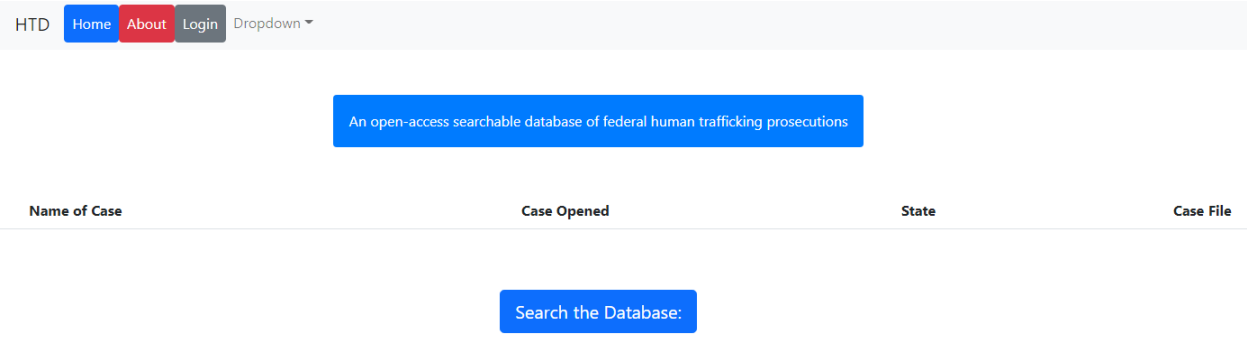

Figure 9: HTD: Krug's First Principle

Krug's second principle urges developers to utilize typical conventions and paradigms for both design and functionality. Color theory and font families are both important pieces of presentation and end user reception and thus must mesh cohesively to provide a visually pleasing user experience. Many older packages, interfaces, and applications used for research feature unintuitive, chaotic, and striking user interfaces with either too little or too many graphics, which make navigating and using them hard, if not impossible. Thus, this frontend utilizes only standard, straightforward data retrieval functionalities and mirrors the recognizable red-blue-black color scheme of Southern Methodist University's logo.

Figure 10: The Southern Methodist University Logo[1]

Further, as the entire frontend design is composed of solely Bootstrap-reliant React components, it is completely scalable and can be easily integrated with most infrastructures without customized stylesheets or functionality to facilitate it.

Finally, Krug's last three principles echo the previous definition of usability- ultimately easing end user frustration to enhance performance and experience- by both minimizing a page's busy noise and breaking content into parse-able pieces that all contribute to website's goal. All displayed information is containerized, highlighted, and separated by whitespace- allowing for straightforward query information retrieval- heralded by HTD's central purpose- "An open-access searchable database of federal human trafficking prosecution". Any further information on the project, its contents, and other metadata, is stored in its own navigable page that does not conflict with the results from a database query.

[1] https://www.smu.edu/DevelopmentExternalAffairs/MarketingCommunications/Logos

The frontend thus integrates these five principles with a streamlined and intuitive user experience to allow for any organization- inside and outside of Southern Methodist University- to seamlessly access and utilize the data curated and presented within the pipeline.

4.2 User Experience and Unit Tests

The user experience cannot be accurately captured without completely describing and delineating a typical use case. This section will walk through several sample searches using the database, from initial connection to returned case results, and highlight the uses of Krug's usability criterion and KISS principles when applicable.

All users are initially considered unauthenticated and will not be allowed to query the database. Any attempt made will prompt the server to direct them to the login portal, accessed through the site's navigation bar.

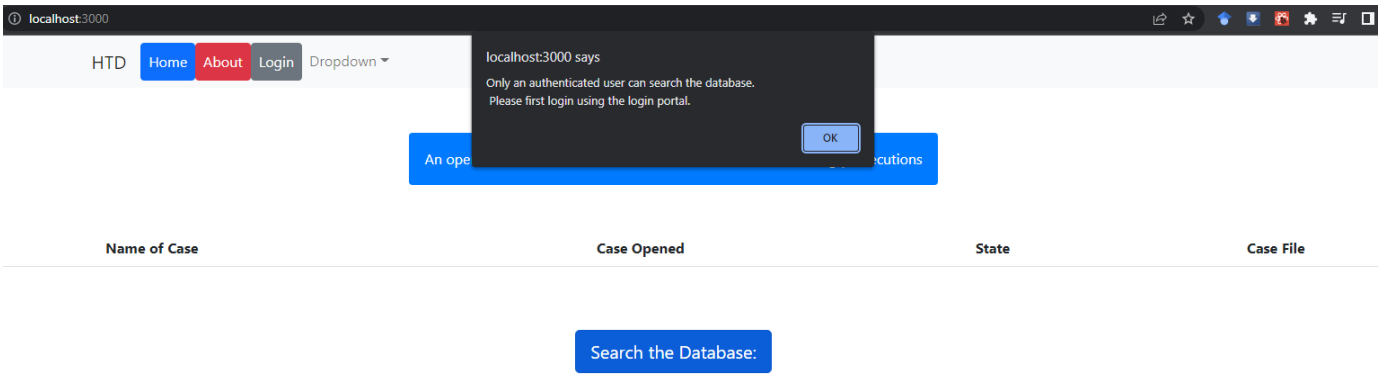

Figure 11: Step 0: Unauthenticated User Access Denial

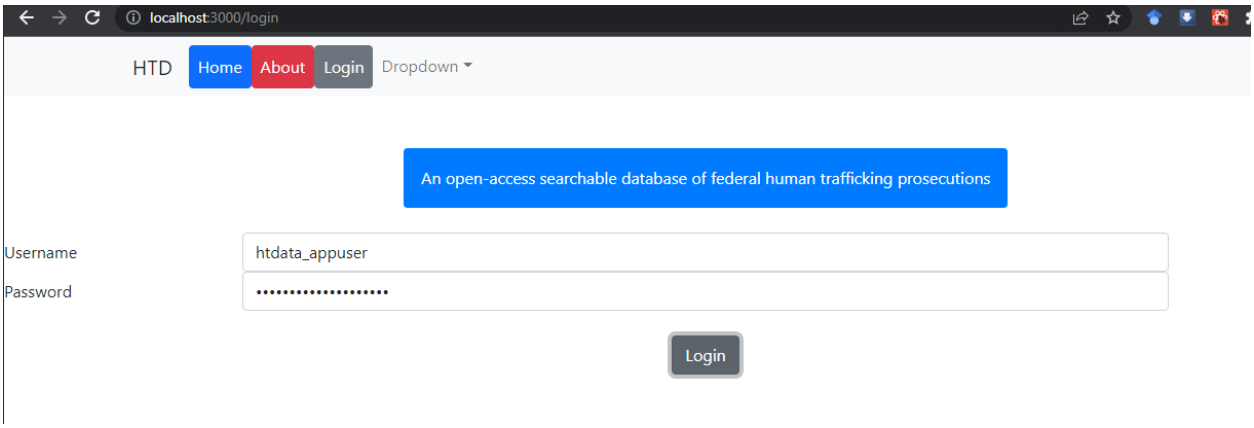

Figure 12: Step 1: User Authentication

Once successfully authenticated with the database, the user is automatically redirected to the main page and can now properly search for desired cases. The middleware server will also log any aunthentication attempts (whether succesful or unsuccessful) for record purposes. If no cases match the user's supplied search criteria, the system will appropriately prompt the user (as the page will not update when it has no cases to display).

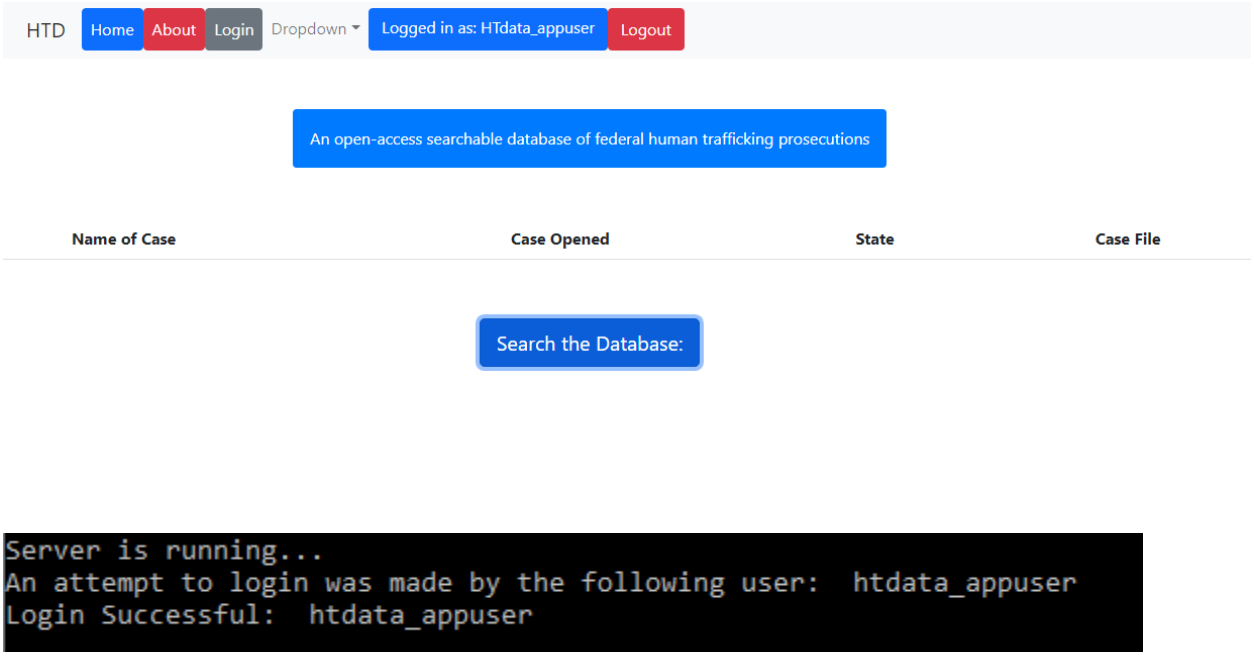

Figure 13: Step 2: Successful User Authentication

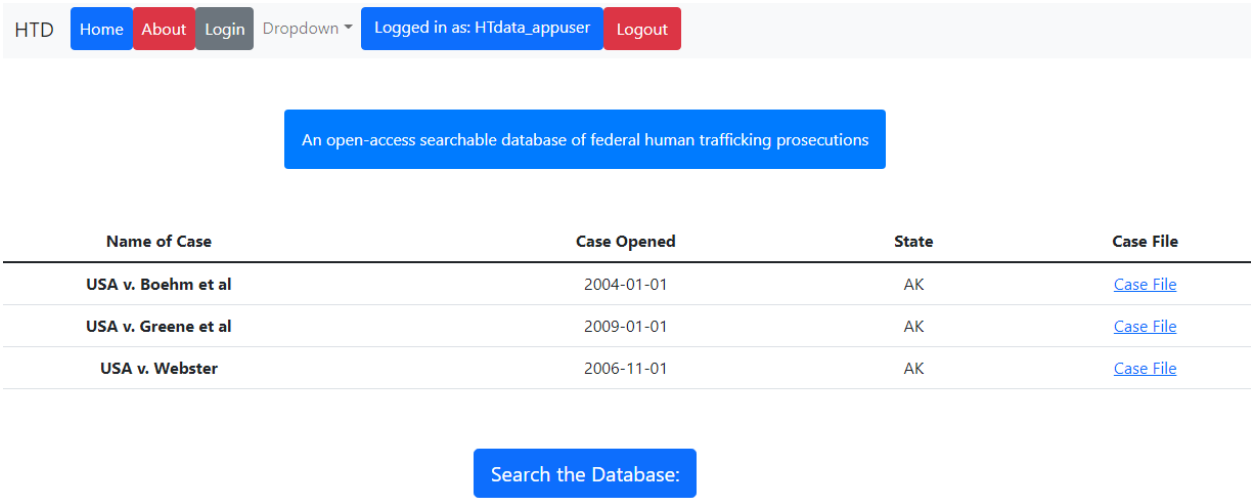

Figure 14: Step 3: Case Data Retrieval

The completed pipeline's filtering capabilities were tested on a myriad of random queries (ranging from a full scan of all current cases to very fine-grained queries that point to only one specific case in the database) to simulate all forms of expected user traffic. The results of one such query- all human trafficking cases prosecuted in Alaska from 2004 to 2010- is shown in Figure 14. Further testing queries are shown in Tables 1-5, highlighting several categories of information used by real-world human trafficking data analysts and published in reports, papers, and documents such as the annual Federal Human Trafficking Report. Each table highlights a different breakdown of cases by attribute and are listed as follows: Table Diagnostics, which tracks total summary statistics as well as top-level distinguishing factors between type of trafficking cases; Victim Genders, which divides cases by the victim or victims' gender; Victim Nationality, which divides cases by the victim's country of origin (displaying the four most common countries of origin); Defendant Nationality, which divides cases by the defendant or defendants' race (displaying the four most common nationalities among different cases); and Defendant Gender, which divides cases by the defendant or defendants' gender.

Query:	**Number of Cases Returned:**
Full Case Scan (all cases, regardless of demographics and proceedings)	1548
All Trafficking Cases of **Minors**	1181
All Trafficking Cases of **Adults**	536
All Labor Trafficking **Cases**	132

Table 1: Unit Test Cases: Table Diagnostics

Query:	Number of Cases Returned:
Trafficking Cases with **Male** Victims	32
Trafficking Cases with **Female** Victims[2]	1177

Table 2: Unit Test Cases: Victim Genders

Query:	Number of Cases Returned:
Trafficking Cases with Victims from the **United States**	1083
Trafficking Cases with Victims from **China**	10
Trafficking Cases with Victims from **Mexico**	90
Trafficking Cases with Victims from **Guatemala**	17

Table 3: Unit Test Cases: Victim Nationality

Query:	Number of Cases Returned:
Trafficking Cases with **White** Defendants	361
Trafficking Cases with **African American** Defendants	870
Trafficking Cases with **Hispanic** Defendants	209
Trafficking Cases with **Asian** Defendants	74

Table 4: Unit Test Cases: Defendant Nationality

[2] In some cases, the gender of victims was either not known or neither male nor female. Therefore, these two numbers do not add up to the total 1548 cases currently in the database.

Query:	Number of Cases Returned:
Trafficking Cases with **Male** Defendants	1395
Trafficking Cases with **Female** Defendants[3]	515

Table 5: Unit Test Cases: Defendant Gender

[3] Many cases have multiple defendants (ie: a case with a male defendant and a female defendant will show up under both these categories, leading to the number to be higher than the total 1548 cases currently in the database.

CHAPTER 5: CONCLUSIONS

This pipeline lays the groundwork for expansion of the human trafficking research sphere with its simplistic, lightweight nature and its containment within Southern Methodist University, a leading figure in today's computing era and thus a potential research partner for many different institutions. A singular, expandable source of human trafficking data as a central point for analysis and research directly contributes to expediting any vein of research and analytics- from law agencies, student research groups, to any other type of analytical research project.

The pipeline addresses both chief problems in today's human trafficking prevention efforts- a lack of time and a lack of resources. A centralized, easily-queryable human trafficking data warehouse abates both problems. Because the pipeline must be used by many different organizations, the end user is at the forefront of every design choice to remove as much wait time and frustration as possible. Further, the specific query terms used break down human trafficking into smaller, less overwhelming pieces- such as a comparison between defendant nationalities, a breakdown of cases by the victim(s)' gender(s), or by country of initial first contact with the victim – which can then be more easily analyzed to quickly gather intel regarding human trafficking trends or patterns around the world. This allows investigations to proactively target areas that may warrant more resources and dissuade against areas that may require fewer resources. It is this dynamic allocation of time and resources that revolutionizes the human trafficking prosecution process.

5.1 Future Work

As the study of human trafficking and its various trends is always expanding, this pipeline must be able to facilitate deeper and more complex queries to better understand the serviced data when such data is scraped and curated. While this pipeline currently only allows

for queries that explore case-level, defendant-level, and victim-level demographics- its React component-based frontend and robust backend can be expanded to allow for more complete mapping of trends, as analysis of human trafficking is significantly more complex than cases, defendants, and victims.

Trends can also be more easily understood through visualizations- such as heat maps to differentiate types of trafficking based on geographic area, choropleth maps of trafficking frequency by country or state of origin, or even simple histograms of various statistical data points- and the pipeline can be updated to summarize data trends of captured and retrieved data as such. Overall, "[v]isualizations are a key piece of understanding human trafficking trends", says Ms. Vanessa Bouche, and this pipeline can be expanded to encompass this visual aspect of parsing human trafficking data.

Further, if sufficient computing power is used to continually create and update such visualizations, they can be used to generate real-time feedback, which allows for present-day trend mapping to a somewhat accurate degree, even when the number of records is small, instead of simply extrapolating trends from only previous data.

APPENDIX

A.1 Data Query Keywords

Case information

- State prosecuted
- Year prosecuted

Trafficking information

- Type of trafficking (minor sex trafficking/adult sex trafficking/labor trafficking)

Victim information

- Victim nationality
- Victim gender
- Victim age

Defendant information

- Defendant gender
- Defendant race
- Defendant nationality
- Defendant age

A.2 Express.js Endpoints

- **GET: "/"** - A sanity test endpoint to ensure that connection to the database is secure and that the database is currently online. It returns a JSON object of all minor sex trafficking cases, a grouping which is very often investigated by authorities due to the many statutes each case must be prosecuted under.

- **POST: "/" –** An endpoint that allows for query submissions. Any content currently within the submission form will be sent to the Express.js server for processing and dynamically constructed into an object that is automatically queried against the database.

- **POST: "/login"**- The endpoint that governs login procedures. The contents of the login form are sent to the Express.js server and validated against the user credentials stored in the database.

BIBLIOGRAPHY

Ayden, Balki. “How Cryptocurrency and Human Trafficking Collide.” Guidehouse- Mekong Club Blog, https://guidehouse.com/insights/financial-crimes/2021/cryptocurrency-human-trafficking-mekong-club.

Bouche, Vanessa. “HTData Prep Meeting- March 21st, 2022.” 22 Mar. 2022.

Cornell Law School. “18 U.S. Code § 1584 - Sale into Involuntary Servitude.” *Legal Information Institute*, Legal Information Institute, https://www.law.cornell.edu/uscode/text/18/1584.

Feehs, K. and Wheeler, A., 2022. *Federal Human Trafficking Report | Human Trafficking Institute*. [online] Traffickinginstitute.org. Available at: <https://traffickinginstitute.org/federal-human-trafficking-report/> [Accessed 15 February 2022].

“Human Trafficking Statistics & Facts.” Safe Horizon. 29 Apr. 2020, https://www.safehorizon.org/get-informed/human-trafficking-statistics-facts/#definition/.

Lawrimore, John. “CSE3345 – Don’t Make Me Think” 20 Apr. 2020, https://smu.instructure.com/courses/68608/files/3059000?module_item_id=559825

Li, Lin, et al. "Detection and characterization of human trafficking networks using unsupervised scalable text template matching." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.

Mukherjee, Sumit, et al. “A machine learning pipeline for aiding school identification from child trafficking images.” Proceedings of the Conference on Information Technology for Social Good. 2021

Portnoff, Rebecca Sorla. *The Dark Net: De-Anonymization, Classification and Analysis.* University of California, Berkeley, 2017.

VanMSFT. “Choose an Authentication Mode – SQL Server.” *SQL Server | Microsoft Docs,* https://docs.microsoft.com/en-us/sql/relational-databases/security/choose-an-authentication-mode?view=sql-server-ver15

“What is Human Trafficking?” *What is Human Trafficking | Homeland Security,* https://www.dhs.gov/blue-campaign/what-human-trafficking

Zhu, Jessica, Lin Li, and Cara Jones. "Identification and detection of human trafficking using language models." *2019 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 2019.