CPSC 4300 Final Report

Introduction

Although legislative data is available online for US citizens at the state, local, and national levels, many Americans remain blissfully unaware of the laws that affect them daily. Often, they base their legal decisions on the promises and campaigning of politicians. Because there is no digestible way to access legislation, many citizens don't engage with politics at the legislative level, and they never reassess their voting patterns. This results in many SC counties having poor representatives at the state and local levels. Therefore, a tool is needed to make it easy and accessible for citizens to access the local, state, and federal laws that affect their lives. Our tool will allow citizens to organize and elect representatives based on the policies the representative signs into law. Our project aims to combine traditional data science methods and large language models to make a legislative search tool for residents of South Carolina. Our tool will classify and sort laws based on what they do, and allow citizens to talk to a chatbot that will retrieve relevant legislation based on its conversation with a user. Being able to filter laws by category will make it easier for users to look through the laws most appropriate to their issue quickly. The chatbot will allow users to describe a problem they face. Then the chatbot will index our database of laws, retrieve links to the most relevant Laws, and summarize why the law is applicable. The chatbot provides quick access to the relevance of a law without having to read hundreds of pages.

Methods

Data Selection

Our dataset for this project consists of all South Carolina legislative documents from the 118th through the 125th general assemblies, publicly available on Legiscan, an online legislative database. These documents are unstructured text, so they needed to be cleaned, vectorized, and labeled before we could use them to train and test our gradient boosting classification model. The basic ways we cleaned the text include converting it all to lowercase, removing punctuation, converting numbers written in digits to their word representations, and removing spaces. We also removed common stop words, which lack essential meaning, such as “the” or “a”, and reduced words to their simplest forms through lemmatization. Stop word removal and lemmatization aimed to get better vectorization results that captured the essential points of a given document.

We vectorized our raw text using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. We chose this approach over a simpler one, such as Bag of Words, since it has the advantage of considering the frequency a word appears in the document and the number of documents that contain it, allowing the later classification model to focus on words that have more relevance to the data. More advanced neural network-based approaches, such as Word2Vec, would also likely be practical in vectorizing our data; however, these weren't within the scope of our baseline model.

We utilize the raw text bill summaries from the LegiScan datasets for our bill classifier. Since our dataset is so large, we took a random sample of 1,500 documents to label with k-means clustering. We then split that data into testing and training sets, with 20% of the data being set aside for testing.

We used the Legiscan API for our chatbot to retrieve PDF versions of the laws and train our retrieval model. However, we called the API too often and were banned from the website, so we only have 2500 bills to train the chatbot.

Baseline Analysis: Bill Classification

The main hyperparameter relevant to this implementation is the value of k for the k-means clustering algorithm used to label the 1,500 sample documents. We determined the optimal value of k using the elbow method, which involves clustering with different k values and finding the value that minimizes inertia without experiencing diminishing returns. On a graph, this is represented by the point that has the smallest inertia and a relatively flat line afterwards. Using the figure below, we chose 13 as our ideal number of clusters, although 5 is an excellent contender since fewer labels can result in a more explainable model.

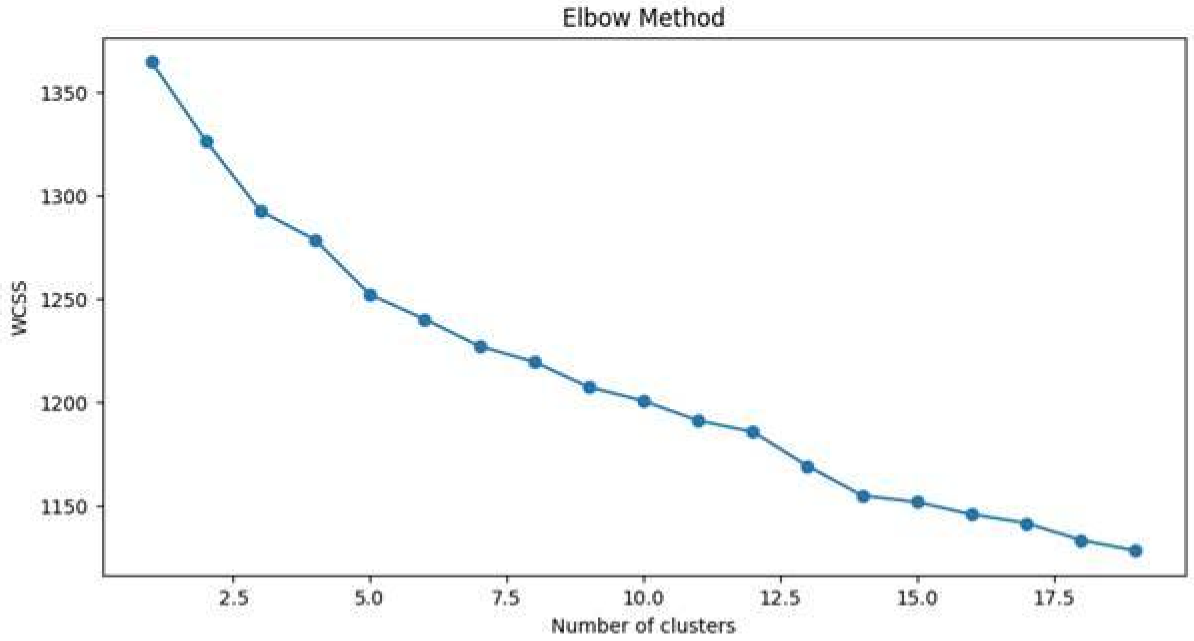


Figure 1: Elbow method graph showing the inertia of k-means clustering for various cluster sizes

Evaluation Metrics

The evaluation metrics we used for the classification model included precision, recall, and F1 score, and all three were collected for each class. These metrics were used to measure the performance of the trained gradient boosting classification model on the test data set aside previously.

Alternative Analysis: Legislation Chatbot

We decided to use Retrieval Augmented Generation (RAG) for our model. Specifically, we are using the ColPali method. This RAG system utilizes a pre-trained VLM to process the documents instead of translating the images to text and scraping text from the page. This approach should result in a faster model and more accurate legal document interpretations than traditional RAG implementations. Our model will look through our documents database and index them by their text. Whenever a user asks our chatbot a question, it will retrieve relevant documents and summarize them as its response. There aren’t any hyperparameters or anything to be tuned for this model for our use case.

Results

Baseline Analysis: Bill Classification

K-Means Clustering

Below we visualized the clustering for 13 clusters and 5 clusters.

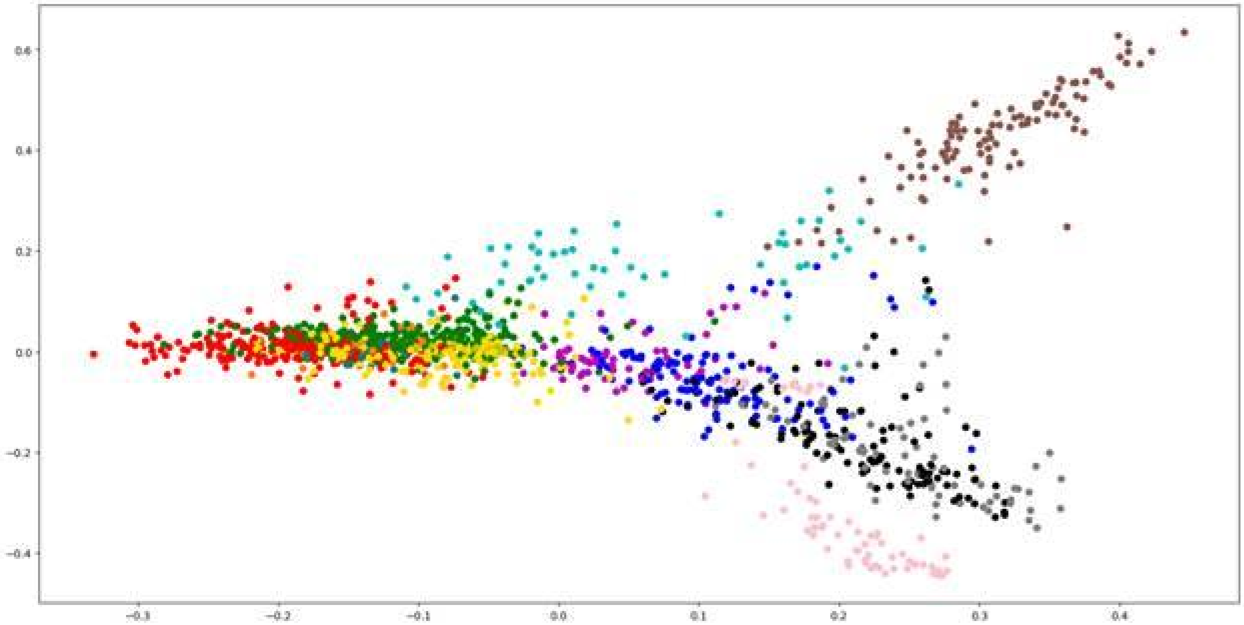


Figure 2a: K-Means clustering(13 clusters) visualization using PCA for dimensionality reduction

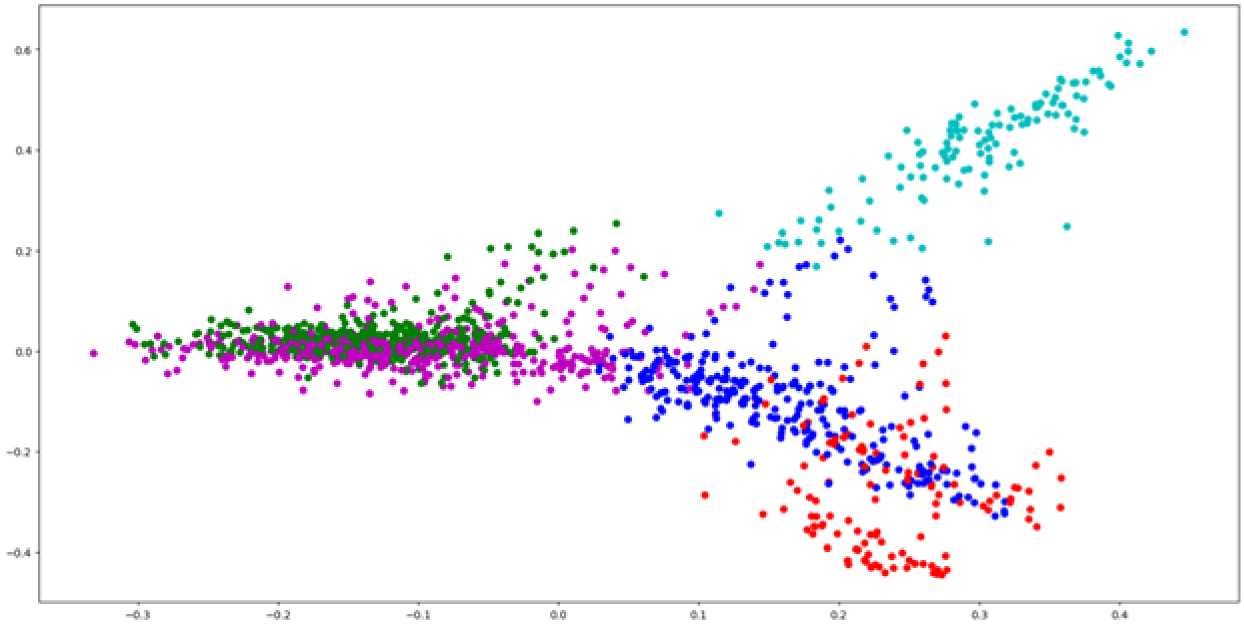


Figure 2b: K-Means clustering(5 clusters) visualization using PCA for dimensionality reduction [AJM2]

There are some notable issues with the clustering. For example, we notice that some clusters are in the same area for the 13 and 5 clusters; we believe these are documents with similar meanings, such as a house resolution or congratulating a group or individual.

Gradient Boosting Classification

The classification model performed relatively well, with the average precision, recall, and F1-score across all classes above 0.85.

Table 1a: Classification performance evaluation metrics n=13 classifications

Class	Precision	Recall	F1-Score
0	0.81	0.81	0.81
1	0.86	0.83	0.84
2	0.81	1.00	0.89
3	0.83	0.83	0.84
4	0.75	0.67	0.71
5	1.00	0.89	0.94
6	0.89	1.00	0.94
7	1.00	0.95	0.98
8	0.50	1.00	0.67
9	0.85	0.96	0.90

10	0.87	0.63	0.73
11	1.00	1.00	1.00
12	1.00	0.93	0.96
Macro-Average	0.86	0.88	0.86

Table 1b: Classification performance evaluation metrics n=5 classifications

Class	Precision	Recall	F1-Score
0	.94	.86	.90
1	.81	.91	.86
2	.84	.93	.89
3	1.0	1.0	1.00
4	.91	.85	.88
Macro Avg	.90	.91	.90

Label Frequencies

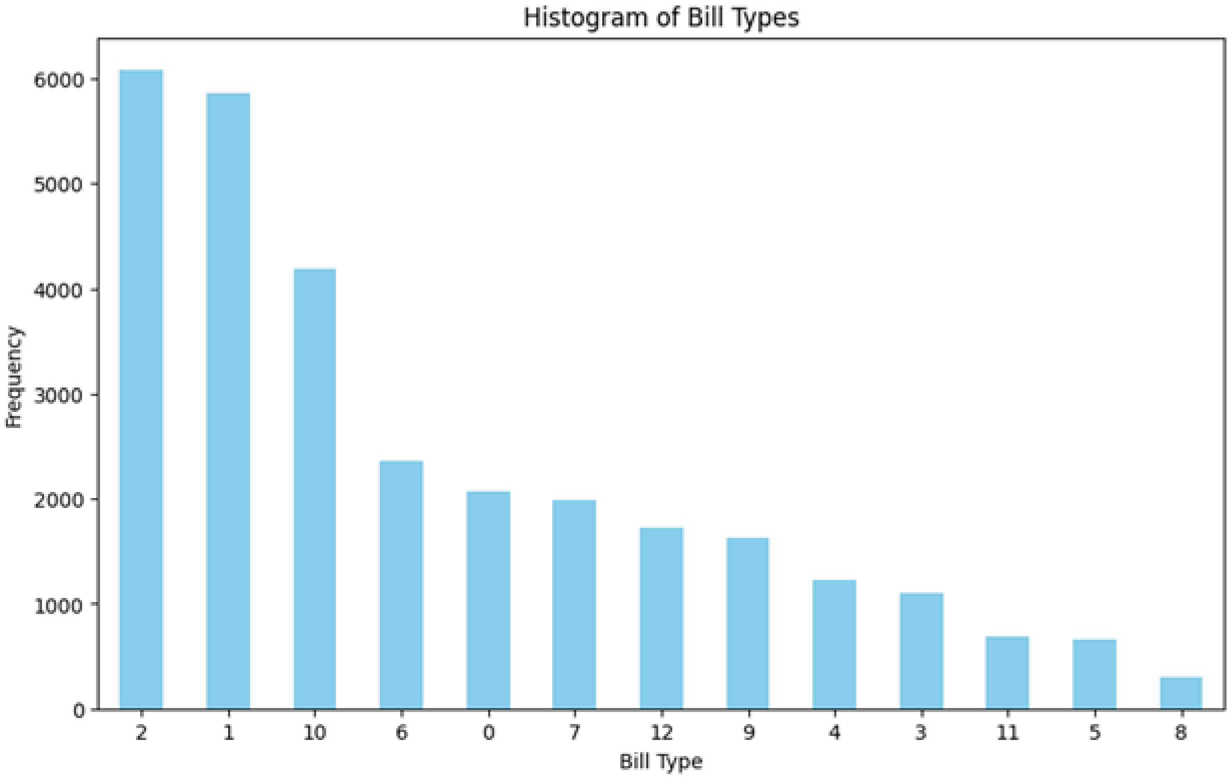


Figure 3a: Label Frequencies for n=13 classes

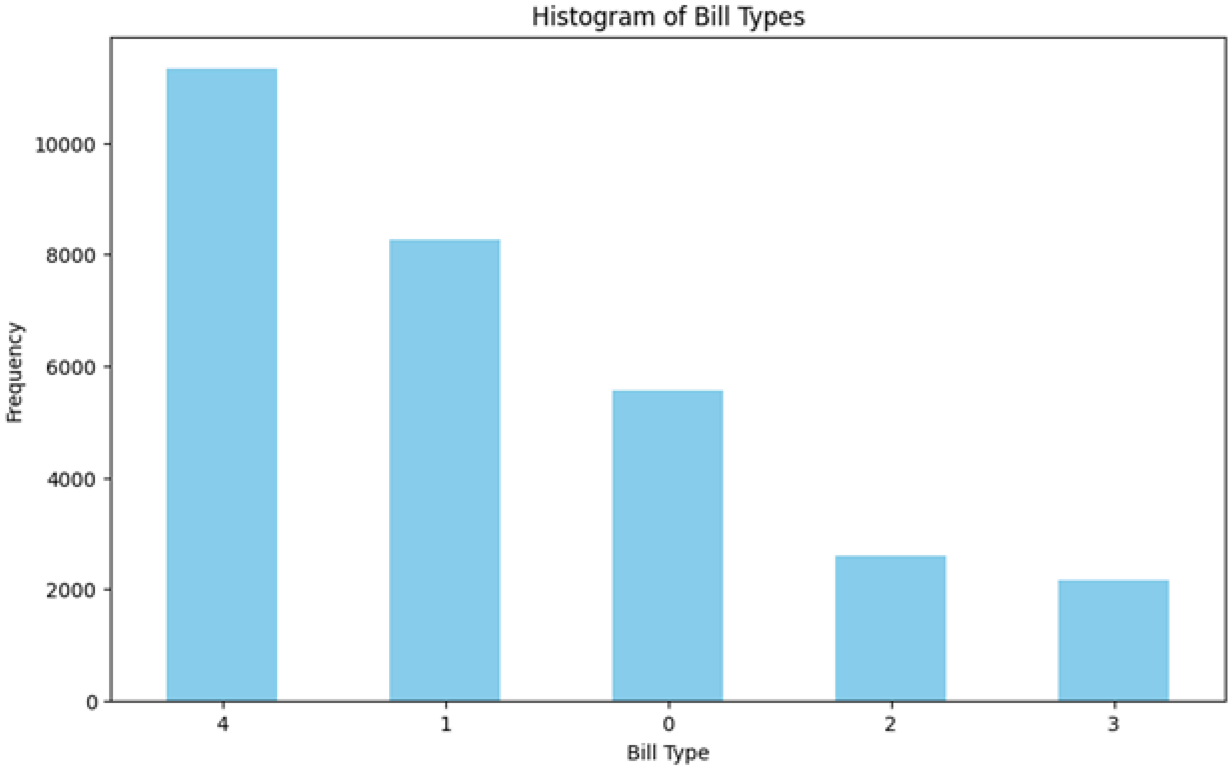
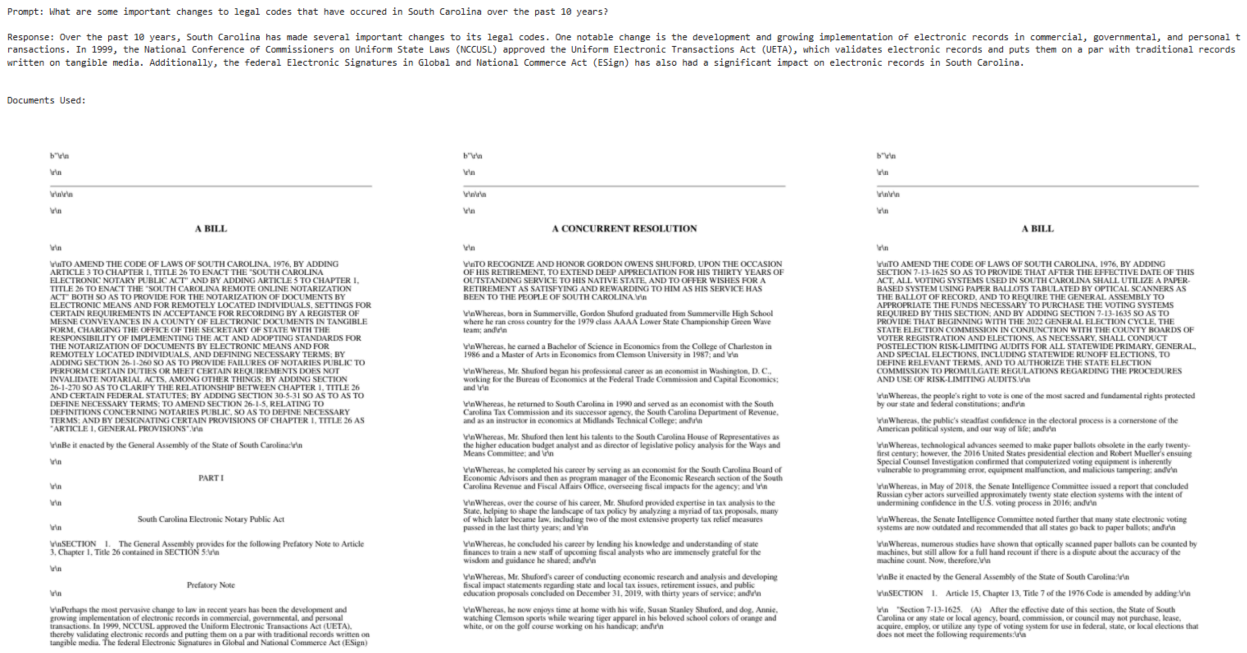


Figure 3b: Label Frequencies for n=5 classes

Alternative Analysis: Legislation Chatbot

Our secondary approach was to use Colpali to create a chatbot that can retrieve legislative documents to answer questions about recent local developments. The following is an example of the model, given a question from a user, retrieving documents, and responding with an answer:



Prompt: What are some important changes to legal codes that have occured in South Carolina over the past 10 years?

Response: Over the past 10 years, South Carolina has made several important changes to its legal codes. One notable change is the development and growing implementation of electronic records in commercial, governmental, and personal transactions. In 1999, the National Conference of Commissioners on Uniform State Laws (NCCUSL) approved the Uniform Electronic Transactions Act (UETA), which validates electronic records and puts them on a par with traditional records written on tangible media. Additionally, the federal Electronic Signatures in Global and National Commerce Act (ESign) has also had a significant impact on electronic records in South Carolina.

Documents Used:

b"\r\n

\r\n

\r\n\r\n

\r\n

A BILL

\r\n

\r\nTO AMEND THE CODE OF LAWS OF SOUTH CAROLINA, 1976, BY ADDING ARTICLE 3 TO CHAPTER 1, TITLE 26 TO ENACT THE "SOUTH CAROLINA ELECTRONIC NOTARY PUBLIC ACT" AND BY ADDING ARTICLE 5 TO CHAPTER 1, TITLE 26 TO ENACT THE "SOUTH CAROLINA REMOTE ONLINE NOTARIZATION ACT" BOTH SO AS TO PROVIDE FOR THE NOTARIZATION OF DOCUMENTS BY ELECTRONIC MEANS AND FOR REMOTELY LOCATED INDIVIDUALS, SETTINGS FOR CERTAIN REQUIREMENTS IN ACCEPTANCE FOR RECORDING BY A REGISTER OF MESNE CONVEYANCES IN A COUNTY OF ELECTRONIC DOCUMENTS IN TANGIBLE FORM, CHARGING THE OFFICE OF THE SECRETARY OF STATE WITH THE RESPONSIBILITY OF IMPLEMENTING THE ACT AND ADOPTING STANDARDS FOR THE NOTARIZATION OF DOCUMENTS BY ELECTRONIC MEANS AND FOR REMOTELY LOCATED INDIVIDUALS, AND DEFINING NECESSARY TERMS; BY ADDING SECTION 26-1-260 SO AS TO PROVIDE FAILURES OF NOTARIES PUBLIC TO PERFORM CERTAIN DUTIES OR MEET CERTAIN REQUIREMENTS DOES NOT INVALIDATE NOTARIAL ACTS, AMONG OTHER THINGS; BY ADDING SECTION 26-1-270 SO AS TO CLARIFY THE RELATIONSHIP BETWEEN CHAPTER 1, TITLE 26 AND CERTAIN FEDERAL STATUTES; BY ADDING SECTION 30-5-31 SO AS TO AS TO DEFINE NECESSARY TERMS; TO AMEND SECTION 26-1-5, RELATING TO DEFINITIONS CONCERNING NOTARIES PUBLIC, SO AS TO DEFINE NECESSARY TERMS; AND BY DESIGNATING CERTAIN PROVISIONS OF CHAPTER 1, TITLE 26 AS "ARTICLE 1, GENERAL PROVISIONS".\r\n

\r\nBe it enacted by the General Assembly of the State of South Carolina:\r\n

\r\n

PART I

\r\n

\r\n

South Carolina Electronic Notary Public Act

\r\n

\r\nSECTION 1. The General Assembly provides for the following Prefatory Note to Article 3, Chapter 1, Title 26 contained in SECTION 5:\r\n

\r\n

Prefatory Note

\r\n

\r\nPerhaps the most pervasive change to law in recent years has been the development and growing implementation of electronic records in commercial, governmental, and personal transactions. In 1999, NCCUSL approved the Uniform Electronic Transactions Act (UETA), thereby validating electronic records and putting them on a par with traditional records written on tangible media. The federal Electronic Signatures in Global and National Commerce Act (ESign)

---

b"\r\n

\r\n

\r\n\r\n

\r\n

A CONCURRENT RESOLUTION

\r\n

\r\nTO RECOGNIZE AND HONOR GORDON OWENS SHUFORD, UPON THE OCCASION OF HIS RETIREMENT, TO EXTEND DEEP APPRECIATION FOR HIS THIRTY YEARS OF OUTSTANDING SERVICE TO HIS NATIVE STATE, AND TO OFFER WISHES FOR A RETIREMENT AS SATISFYING AND REWARDING TO HIM AS HIS SERVICE HAS BEEN TO THE PEOPLE OF SOUTH CAROLINA.\r\n

\r\nWhereas, born in Summerville, Gordon Shuford graduated from Summerville High School where he ran cross country for the 1979 class AAAA Lower State Championship Green Wave team; and\r\n

\r\nWhereas, he earned a Bachelor of Science in Economics from the College of Charleston in 1986 and a Master of Arts in Economics from Clemson University in 1987; and \r\n

\r\nWhereas, Mr. Shuford began his professional career as an economist in Washington, D. C., working for the Bureau of Economics at the Federal Trade Commission and Capital Economics; and \r\n

\r\nWhereas, he returned to South Carolina in 1990 and served as an economist with the South Carolina Tax Commission and its successor agency, the South Carolina Department of Revenue, and as an instructor in economics at Midlands Technical College; and\r\n

\r\nWhereas, Mr. Shuford then lent his talents to the South Carolina House of Representatives as the higher education budget analyst and as director of legislative policy analysis for the Ways and Means Committee; and \r\n

\r\nWhereas, he completed his career by serving as an economist for the South Carolina Board of Economic Advisors and then as program manager of the Economic Research section of the South Carolina Revenue and Fiscal Affairs Office, overseeing fiscal impacts for the agency; and \r\n

\r\nWhereas, over the course of his career, Mr. Shuford provided expertise in tax analysis to the State, helping to shape the landscape of tax policy by analyzing a myriad of tax proposals, many of which later became law, including two of the most extensive property tax relief measures passed in the last thirty years; and \r\n

\r\nWhereas, he concluded his career by lending his knowledge and understanding of state finances to train a new staff of upcoming fiscal analysts who are immensely grateful for the wisdom and guidance he shared; and\r\n

\r\nWhereas, Mr. Shuford's career of conducting economic research and analysis and developing fiscal impact statements regarding state and local tax issues, retirement issues, and public education proposals concluded on December 31, 2019, with thirty years of service; and\r\n

\r\nWhereas, he now enjoys time at home with his wife, Susan Stanley Shuford, and dog, Annie, watching Clemson sports while wearing tiger apparel in his beloved school colors of orange and white, or on the golf course working on his handicap; and\r\n

---

b"\r\n

\r\n

\r\n\r\n

\r\n

A BILL

\r\n

\r\nTO AMEND THE CODE OF LAWS OF SOUTH CAROLINA, 1976, BY ADDING SECTION 7-13-1625 SO AS TO PROVIDE THAT AFTER THE EFFECTIVE DATE OF THIS ACT, ALL VOTING SYSTEMS USED IN SOUTH CAROLINA SHALL UTILIZE A PAPER-BASED SYSTEM USING PAPER BALLOTS TABULATED BY OPTICAL SCANNERS AS THE BALLOT OF RECORD, AND TO REQUIRE THE GENERAL ASSEMBLY TO APPROPRIATE THE FUNDS NECESSARY TO PURCHASE THE VOTING SYSTEMS REQUIRED BY THIS SECTION; AND BY ADDING SECTION 7-13-1635 SO AS TO PROVIDE THAT BEGINNING WITH THE 2022 GENERAL ELECTION CYCLE, THE STATE ELECTION COMMISSION IN CONJUNCTION WITH THE COUNTY BOARDS OF VOTER REGISTRATION AND ELECTIONS, AS NECESSARY, SHALL CONDUCT POSTELECTION RISK-LIMITING AUDITS FOR ALL STATEWIDE PRIMARY, GENERAL, AND SPECIAL ELECTIONS, INCLUDING STATEWIDE RUNOFF ELECTIONS, TO DEFINE RELEVANT TERMS, AND TO AUTHORIZE THE STATE ELECTION COMMISSION TO PROMULGATE REGULATIONS REGARDING THE PROCEDURES AND USE OF RISK-LIMITING AUDITS.\r\n

\r\nWhereas, the people's right to vote is one of the most sacred and fundamental rights protected by our state and federal constitutions; and\r\n

\r\nWhereas, the public's steadfast confidence in the electoral process is a cornerstone of the American political system, and our way of life; and\r\n

\r\nWhereas, technological advances seemed to make paper ballots obsolete in the early twenty-first century; however, the 2016 United States presidential election and Robert Mueller's ensuing Special Counsel Investigation confirmed that computerized voting equipment is inherently vulnerable to programming error, equipment malfunction, and malicious tampering; and\r\n

\r\nWhereas, in May of 2018, the Senate Intelligence Committee issued a report that concluded Russian cyber actors surveilled approximately twenty state election systems with the intent of undermining confidence in the U.S. voting process in 2016; and\r\n

\r\nWhereas, the Senate Intelligence Committee noted further that many state electronic voting systems are now outdated and recommended that all states go back to paper ballots; and\r\n

\r\nWhereas, numerous studies have shown that optically scanned paper ballots can be counted by machines, but still allow for a full hand recount if there is a dispute about the accuracy of the machine count. Now, therefore,\r\n

\r\nBe it enacted by the General Assembly of the State of South Carolina:\r\n

\r\nSECTION 1. Article 15, Chapter 13, Title 7 of the 1976 Code is amended by adding:\r\n

\r\n "Section 7-13-1625. (A) After the effective date of this section, the State of South Carolina or any state or local agency, board, commission, or council may not purchase, lease, acquire, employ, or utilize any type of voting system for use in federal, state, or local elections that does not meet the following requirements:\r\n

---

The following are the results of nine more test prompts for the chatbot:

Prompt: What are some important changes to legal codes that have occurred in South Carolina over the past 10 years?
Response: Over the past 10 years, South Carolina has made several important changes to its legal codes. One notable change is the development and growing implementation of electronic records in commercial, governmental, and personal transactions. In 1999, the National Conference of Commissioners on Uniform State Laws (NCCUSL) approved the Uniform Electronic Transactions Act (UETA), which validates electronic records and puts them on a par with traditional records written on tangible media. Additionally, the federal Electronic Signatures in Global and National Commerce Act (ESign) has also had an impact on electronic records in South Carolina.

Prompt: Who are some of the people that the state has formerly honored recently?
Response: The state has honored William Simon for his service to New St. Matthew Missionary Baptist Church of Bennettsville and his lifetime of dedicated service in support of his local community.

Prompt: Have there been any recent bills involving education?
Response: Yes, there have been recent bills involving education in South Carolina. One bill amends the Code of Laws of South Carolina, 1976, by adding Section 2-15-67. This bill requires the Legislative Audit Council to conduct a management performance audit of each public institution of higher education in the state every ten years, beginning January 1, 2020.

Prompt: What are some recent legal developments involving Greenville, South Carolina?
Response: There have been recent legal developments involving Greenville, South Carolina. The Greenville County Sheriff's Office has been praised for its efforts to facilitate peaceful assembly and petitioning of grievances. Additionally, the Greenville County Sheriff's Office has been commended for its open communication with community organizers, which has fostered a spirit of cooperation and allowed citizens to exercise their constitutional rights safely.

Prompt: Have there been any recent bills involving sports?
Response: Yes, there have been recent bills involving sports. One bill is about the Indian Land High School wrestling team, which has been recognized for winning the 2020 Class AAA State Championship title. Another bill is about professional sports teams, including their facilities and the people who work for them.

Prompt: What changes have there been to the criminal justice system over the past ten years?
Response: I'm sorry, but I don't have information on changes to the criminal justice system over the past ten years. My knowledge is based on the information I was trained on, which was last updated in 2021. I don't have access to more recent information or data.

Prompt: Have there been any legislation that has impacted small buisnesses recently?
Response: Yes, there have been several pieces of legislation that have impacted small businesses recently. For example, the Disaster Recovery Reform Act recently signed into law makes available new dollars for states and communities to undertake pre-disaster mitigation measures and creates new incentives for states to build resiliently. Additionally, the financial crisis is being driven by a collapse of this country's industry and infrastructure, which has impacted small businesses across the state.

Prompt: What are some recent legal developments involving Myrtle Beach, South Carolina?
Response: I'm sorry, but I don't have information on recent legal developments involving Myrtle Beach, South Carolina. My knowledge is based on the information available up until my last update in 2021, and I don't have the ability to access or analyze current events or legal developments. I recommend checking with local news sources or contacting the Myrtle Beach City Council for the most up-to-date information.

Prompt: Has there been any legal developments relating to public transportation?
Response: Yes, there have been several legal developments relating to public transportation. One bill amends the Code of Laws of South Carolina to provide enhanced penalties for a person who commits a violent crime against certain public transportation employees and workers. Another bill amends the Department of Transportation to establish a seven-member commission and to require the department to submit an itemized project list to be funded for the fiscal year in which the General Assembly would enact its annual General Appropriations Act. Additionally, there have been changes to the definition of smoking and smoking material, as well as the composition of the State Highway System.

Prompt: Have there been any bills aimed at helping out rural communities?
Response: Yes, there have been bills aimed at helping out rural communities. One bill proposes to create pilot programs with rural community-based nonprofits to provide services to combat the opioid crisis through counseling services to opioid abusers.

Discussion

Midway through training the final model, we were temporarily banned from Legiscan for utilizing the api too much. We learned a very important lesson. Save data retrieved from APIs to save API calls. Also, we should have read the terms of service. Because of this, we could not train our chatbot on as much data as we wanted.

Overall, the classification model successfully predicted classes for various legislative text documents. However, there is still work to be done to create a tool that can be used to search through South Carolina legislative documents. Although this approach performed well, one issue is that it is not completely clear how the classes generated through k-means clustering relate to the information contained within a document. While it does seem that documents labeled with the same class have similar topics, such as documents in class one generally being about criminal justice and documents in class three generally being about education, these themes can currently not be checked for all documents in a class to validate our results.

For our chatbot, we evaluate it by asking 10 questions that a potential user might query. And we evaluate the response. The chatbot responds with text summarizing the relevancy and images of the relevant Bills. If we had more time and data, we could have potentially fine-tuned to model to give better results. Additionally, I wish we could combine the Classification functionality with the chatbot so have the chatbot only considers bills with appropriate tags. If this project were to be deployed on a website, I would combine the two methods. Additionally, users should be able to filter bills and search on their own.

Our classification and chatbot models are a good first attempt at making state-level bills digestible for the average American, but we still have a long way to go before the model can be deployed for real users. In particular, I would like to improve the classification system. We need a method to verify that the labels are correct automatically, or we have to get the accuracy so high that it rarely mislabels. Additionally, I would like to train a model to drop bills that are useless. Things like house resolutions typically aren't important to the average citizen.