Literature Review & Data Collection – Legislation Duplicates

AP Research & Senior Research Project

Ojas Chaturvedi

https://github.com/ojas-chaturvedi

This week, I successfully finalized my data collection script to gather congressional gun legislation texts. However, integrating this script with ASU's supercomputer Sol presented unexpected challenges. Connecting to the ASU VPN to access Sol encountered technical issues, prompting me to submit a support ticket to the ASU IT Department while awaiting a resolution to resume data collection. Below are the links for my data collection scripts:

Overall script code:
https://github.com/ojas-chaturvedi/NLP-Gun-Legislation/blob/master/data_collection/main.py

Web Scraper code:
https://github.com/ojas-chaturvedi/NLP-Gun-Legislation/blob/master/web-scraper.py

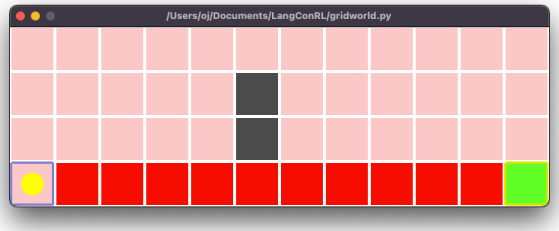
However, during further examination of the gathered data, I stumbled upon a significant issue: multiple versions of the same bill. Although this is a common occurrence due to legislation originating from various bodies like the Senate, House of Representatives, and joint/concurrent committees, it posed a problem as Congress.gov did not address this duplication during data retrieval. I cannot keep the duplicate legislation, as this would essentially be the same as the sentiment analysis models giving the same score twice even when there was only one bill with that score. Initially, I planned to remove duplicate legislation to avoid redundancy, therefore preventing repeated scores in sentiment analysis. However, I soon realized that these duplicates often differ due to committee amendments. Therefore, I need to devise a method to identify and retain only the latest version of each bill for accurate sentiment analysis.

While awaiting the completion of my data collection step, I've started revising my literature review, a section already drafted during my time in AP Research. Previously, my sentiment analysis project dived into gun control court cases within the D.C. Court of Appeals, mainly due to its data accessibility. Unfortunately, a similar paper emerged in late December, forcing me to modify my project to maintain originality, prompting a shift towards analyzing congressional legislation. Remarkably, this transition allowed me to retain much of my existing literature review, primarily focusing on the rationale behind analyzing the gun control

debate and identifying gaps in the current field. Yet, with the transition to congressional legislation, my research direction has pivoted towards exploring papers more directly related to legislative analysis for inclusion in my sentiment analysis. Notably, I've encountered a lack of Natural Language Processing (NLP) work within the realm of congressional legislation—a revealing observation highlighting a significant gap in current research. Nonetheless, I've uncovered several invaluable papers discussing trends in English language usage at the federal level, a consideration that could potentially reshape the scope of my project. Previously, my analysis focused on legislation spanning from 2001 to 2023—a period marked by escalating debates following tragic events such as school shootings. Upon completing the review of acquired literature, if findings indicate minimal evolution in English language usage within a specific timeframe, I will have the flexibility to adjust the project's temporal scope accordingly. As training the sentiment analysis models on different English wording styles would lead to decreased accuracy, it underscores the importance of this comprehensive review process.

In my internship role at the ASU Data Mining and Reinforcement Learning (RL) Lab, I started in the initial stages of developing the testing environment and RL agent. Collaborating with PH.D. student Longchao Da, we've leveraged OpenAI's gym toolkit, tailored for developing and comparing reinforcement learning algorithms. Utilizing pre-existing environments available online, we've established a foundational setup, as depicted in the accompanying image to construct and evaluate a new RL agent. The yellow circle represents the RL agent, while the green box signifies the desired endpoint, aimed to be reached via the most optimized route. However, given the project's emphasis on constraints within RL agents, additional elements such as gray boxes, impassable obstacles, and red boxes, serving as deterrents with negative rewards, have been introduced. Upon finishing the environment, my focus shifted toward implementing a Deep Q-network (DQN) to train a policy guiding the RL agent through the environment efficiently. DQN, a neural network architecture combining deep learning principles with Q-learning, a classic reinforcement learning algorithm, demonstrates particular efficacy in handling high-dimensional state spaces, such as those found in numerous real-world scenarios.

Figure 1: Simple environment to test basic RL agent



In the upcoming week, my agenda is packed with addressing the challenges posed by duplicate congressional legislation, as discussed earlier. While devising a solution for this issue, I'll ensure

access to Sol and proceed to submit my script for data retrieval. I also intend to continue refining my literature review for my final research paper, specifically diving into papers concerning English word style to further concentrate the project's scope. Additionally, I'll tackle the pending task from last week—exploring the feasibility of broadening my project scope beyond 'gun' legislation by incorporating relevant keywords like 'firearm' to capture a holistic view of the gun control discourse. Amidst advancing my individual project, I'll continue contributing to my internship responsibilities, focusing on completing the development of the DQN policy for the RL agent to establish a foundational simple environment before delving into more intricate algorithms and environments.