Data Collection – API And Pipelines

AP Research & Senior Research Project

Ojas Chaturvedi

https://github.com/ojas-chaturvedi

This week, my focus has centered on two main areas: understanding the ProPublica Congress API and understanding the methodology for gun control data pipelines. The ProPublica API is crucial for accessing congressional legislation in my project. I was able to create a direct system using the API to get any legislation relating to the word "gun". However, it encounters a challenge in extracting text from legislation that did not pass in Congress. Although unsuccessful legislation lacks real-world impact, its value lies in understanding legislators' framing of issues. Even failed legislation can be beneficial as it provides additional insights from legislators. To address this, I've implemented a workaround using the Playwright library to automate text copying.

Concerning gun control data pipelines, I've been reviewing papers proposing pipelines involving scraping (getting text from online sources) online news articles related to the gun control controversy. This data aims to facilitate in-context learning, enhancing accuracy by training pre-existing models on context-specific text. However, these pipelines are in the proposal stage, requiring further investigation to assess their feasibility.

Regarding my internship at the ASU Data Mining and Reinforcement Learning Lab, I have decided to work with Ph.D. student Longchao Da on a project relating to the paper "Safe Reinforcement Learning with Natural Language Constraints" [1]. Our weekly meetings involve understanding the previous paper and methodology, guiding us to formulate our original research question. This project will deepen my understanding of natural language constraints in machine learning models, potentially aiding my project if pipeline development proves unfeasible. Team lab meetings are straightforward, with each student presenting a paper related to their projects from the previous week.

Next week, I plan to continue working on the data collection. I plan to make a straightforward system using the ProPublica Congress API to obtain legislation in JSON format for easy model accessibility. Simultaneously, I aim to commence work on the pipelines, drafting an initial version and building upon it. Throughout these steps, I'll document my research process, aiding readers in replication and enhancing the methodology section of my paper.

[1] T.-Y. Yang, M. Hu, Y. Chow, P. Ramadge, and K. R. Narasimhan, "Safe Reinforcement Learning with Natural Language Constraints," *Neural Information Processing Systems*, vol. 34, May 2021.