**Final Project Proposal**

Group 3: Jehan Bugli

**Overview**

The Congressional Research Service (CRS) is a vital organization that assists legislators with [policy analysis](https://www.congress.gov/crs-products), [official bill summaries](https://www.congress.gov/help/bill-summaries), and more; however, their job is rapidly becoming more difficult due to [rapidly increasing bill introductions](https://rollcall.com/2025/03/05/publishing-pileup-congressional-bills-slow-to-reach-public/) causing significant backlogs for official text releases.

While the CRS is reportedly [exploring AI applications](https://fedscoop.com/congressional-research-service-eyes-ai-bill-summaries/), the current workflow is largely manual; when paired with growing work volume and [potential federal spending cuts](https://www.nbcnews.com/politics/congress/senate-republicans-vote-advance-massive-budget-blueprint-trumps-agenda-rcna199509) on the horizon, the utility of automated assistance is clear.

I aim to explore topic tagging automation; namely, using NLP methods to accurately tag bills with one of [32 policy area designations](https://www.congress.gov/help/field-values/policy-area). If successful, I think this could have an appreciable impact on CRS workflows by reducing monotonous work and expediting processing. If the policy area designations are changed, the training pipeline could also be reconstructed to propose tags retrospectively.

**Datasets**

To provide utility for a CRS analyst’s workflow, training data must be limited to the text one would have available. Official bill text (in XML format, including federal code citations) can be retrieved in XML format from the [GovInfo bulk data repository](https://www.govinfo.gov/bulkdata).

I plan to explore whether or not incorporating code citations could prove useful; federal code sections can be retrieved from this repository or [a House website](https://uscode.house.gov/download/download.shtml).