Project Proposal

**Potential Datasets:**

1. EUR-Lex-Sum dataset - text summarization of EU legal acts (<https://huggingface.co/datasets/dennlinger/eur-lex-sum>)
2. Canadian legal dataset CAN-LII (<https://www.canlii.org/databases#cad>)
3. **BillSum - US and California congress bills - ~22,000 document/title/summary sets (Official source:** [**https://arxiv.org/abs/1910.00523**](https://arxiv.org/abs/1910.00523)**)**
4. Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation (<https://data.niaid.nih.gov/resources?id=zenodo_7151678>)

**Research Papers**

1. Clustering extractive summarization (<https://link.springer.com/article/10.1007/s10506-023-09345-y#Sec3>)
2. Extractive summarization using graph reweighting (<https://aclanthology.org/2022.nllp-1.30/>)
3. Extractive summarization techniques (<https://aclanthology.org/2022.findings-emnlp.134/>)

**Proposal Requirements**

* What do you plan to do?
* Why is it important, and why is it challenging?
* What dataset(s) will you use?
* What algorithms might you use? Are good implementations available, or will you need to write your own? (Don’t worry if you can’t answer this well at this stage of the course.)
  + BART - Repurposed BERT, BERT encoder and GPT decoder
  + Extractive - BERT (predict relevant score for a sentence, picks top sentence)
  + Hybrid - Maximum marginal relevance, select sentence most relevant to document but also scores penalty based on similarity score
  + Hybrid - Pointer generator networks (predicts whether it should use copy directly from source or generate new words - generation from softmax)
* References to at least four papers related to your proposal

**Proposal**

With our 266 NLP Final Project, we aim to develop a summarization system for U.S. and California legislative bills using the [BillSum dataset](https://arxiv.org/abs/1910.00523) (~22,000 examples of bills with human-written summaries). Legislative documents are typically very lengthy, formal, and structurally complex documents that are hard to interpret, and creating a summarization for these documents can be helpful in deciphering these documents at face value. An effective summary should be able to improve accessibility for policymakers, legal researchers, and the general public by simplifying complex bills into concise representations of what they stand for.

We are planning on exploring mostly extractive and hybrid summarization techniques. For the extractive approach, we plan on leveraging the BERT-based sentence scoring models, applying methods like cluster-based extraction, coherence, and relevance. We also plan on exploring hybrid methods like maximum marginal relevance (MMR) and pointer generator networks to balance informativeness with redundancy, picking between whether a copy from the source or generation of new content is the more effective method.

We plan on comparing our models against the human-written summaries using ROUGE and BERTScore, balancing faithfulness to the facts and readability. Ultimately, we hope our work will contribute to more interpretable summarization systems for legal frameworks.

**References:**

* [A sentence is known by the company it keeps: Improving Legal Document Summarization Using Deep Clustering (AI & Law, 2023)](https://link.springer.com/article/10.1007/s10506-023-09345-y#Sec3)
* [Computing and Exploiting Document Structure to Improve Unsupervised Extractive Summarization of Legal Case Decisions (Zhong & Litman, NLLP 2022)](https://aclanthology.org/2022.nllp-1.30/)
* [Extractive Summarization of Legal Decisions using Multi-task Learning and Maximal Marginal Relevance (Agarwal et al., Findings 2022)](https://aclanthology.org/2022.findings-emnlp.134/)
* [BillSum: A Corpus for Automatic Summarization of US Legislation (Kornilova & Eidelman 2019)](https://arxiv.org/abs/1910.00523)