Project Milestone 3

Group 6: Search Wizards

The Status since the Last Milestone

We assessed the eCFR Title 12 dataset by employing BERTopic to analyze the initial 1000 rows. Our evaluation revealed the dataset's robustness; no missing data was observed. Subsequently, we deliberated on potential applications for the dataset. Unanimously, we concluded that named entity recognition (NER) stands out as an excellent use case for further exploration.

Failed Ideas/Experiments

Initially, our efforts were directed towards customizing the SpaCy for the identification and categorization of distinct legal entities, encompassing regulatory bodies, banking terminology, and financial instruments. Regrettably, constraints such as time limitations and the impracticality of collective annotation hindered our progress in generating annotated data sets and further training the whole dataset. Consequently, we opted to defer this task to our forthcoming stages and chose to stay with the labels that SpaCy can recognize.

Blockers

- Lack of project management led to burst capacity and rushed deliverables:
    - Limited stand-ups
    - No delegation
    - No timelines
- Lack of domain expertise
    - Difficulty in interpreting and drawing insights from the dataset
- Varying programming knowledge and proficiencies
    - Harder to communicate ideas and delegate work
- No labeled data to train or test on

Preliminary Results

We stuck with the labels that SpaCy can recognize. Based on our dataset, we decided to use the following three included labels:

- ORG (an organization, such as a company or institution);
- LAW (references to law and regulations);
- NORP (nationalities or religious or political groups).

We examined the output in our notebook, and the labels need to be more specific to be sufficient. Therefore, we decided to leave it to milestone 4. For example, in the outputs, we see most of the entities are labeled as ORG, but for the LAW and NORP, we didn't see lots of entities labeled. Also, some entities formed with the combination of the name of the organization and the act (e.g., The Federal Credit Union Act), were still labeled as ORG. Due to the restrictions of the labels from SpaCy that are general, we need to fine-tune the original model to adapt to the specific dataset that we have.

Next Steps / Timeline

The main next step is to revisit our methodology and use case. We need to understand *what* we want to do (goal) and *how* we want to do it (execution). We want to create a simple but strong NER model using resources and materials from the course without super-advanced techniques.

We will want to go back to **Homework 1** and review materials and resources relevant to that assignment. We will replicate that methodology, adapted for our dataset and use case, and then improve upon that model using techniques and tips learned up until this point in the semester.

Timeline

Deliverable/Goals	Deadline
• Annotate data, create training, validation, and test sets • Create annotation guidelines • Review Homework 1 and any relevant course materials and resources	4/9/24
• Implement baseline models and review metrics • Review and address any feedback or concerns	4/16/24
• Fine-tune model • Add any additional functionality or techniques to improve the model	4/23/24
Project Milestone 4	4/30/24
Final Project Deliverable	5/7/24