Thomson Reuters Data Science Challenge

Welcome to the Thomson Reuters Data Science Challenge!

Scientists at TR Labs work with a diverse group of stakeholders on applied research projects. We value rigor and reproducibility, as well as the ability to present a compelling analysis. For this data challenge, you will develop an exploratory model on a real dataset. You should envision handing off the project to a colleague without the opportunity to communicate. This scenario emphasizes the importance of clearly documenting your ideas and observations.

The following criteria will be considered.

1. Reproducibility
2. Clear, concise explanations of key decisions and findings
3. Accurate characterizations of expected performance
4. Discussion of alternative approaches you might try given more time

If you utilize any found code or analysis in submission, please cite your sources appropriately.

Finally, to ensure a fair and equal recruiting process, please don't share the challenge or your solution with anyone else.

About the Task

This data challenge is designed to emulate the kinds of analyses we do fairly often in Labs. Since Thomson Reuters' Westlaw is a premier legal research platform, we are often asked to model legal concepts in judicial opinions or briefs. Your assignment will focus on a particular concept known as “procedural posture.”

The procedural posture of a case is a summary of how the case arrived before the court. It describes the procedural history including any prior decisions under appeal.

The dataset you will download contains judicial decisions which have been manually annotated with labels to indicate the procedural posture. The table below shows the top ten most frequent postures in this dataset, as well as a measure of annotator agreement for each category.

Posture	**Kappa**
Appellate Review (Criminal)	0.93
Motion for Attorney's Fees	0.86
Sentencing or Penalty Phase Motion or Objection	0.84
On Appeal (Civil)	0.84
Review of Administrative Decision	0.83
Motion to Dismiss for Lack of Subject Matter Jurisdiction	0.81
Motion for Preliminary Injunction	0.80
Motion to Dismiss	0.79
Post-Trial Hearing Motion	0.77
Trial or Guilt Phase Motion or Objection	0.63

Please answer the following questions and include your report, all code, documentation, and instructions required to run your solution in the uploaded zip file. You do not need to upload data. Please also delete your local copy of the data when you are done.

Question 1

- Download the data here: TRDataChallenge2023.zip. Note that this zip contains a single text file in JSON Lines format.

- Programmatically read the data into your preferred analytical environment.

- Report the number of documents, number of postures, and the number of paragraphs in the dataset.

- Describe the data and any aspects relevant to the modeling task in Question 2.

Question 2

Our business partners would like to automate the labeling of judicial opinions with procedural postures. You are being asked to perform an initial exploration to inform feasibility.

- Given the dataset provided, build a model that achieves the desired automation.

- Analyze the performance of your model , including strengths/weaknesses.

- Make a recommendation to the business about the feasibility of this task based on your observations in the analysis.

Please bear in mind: This intended to be an exploratory model. State of the art results are not expected. Characterizing how the model performs is more important than achieving the highest possible performance.

Question 3

Regardless of the feasibility determination in Question 2, assume your stakeholders are satisfied with the initial results and want to proceed with this project. Now they are asking for next steps. You can either decide to (a) conduct additional experiments to improve / validate the model OR (b) start the process to put this model in production. Either choice is acceptable for this question. Whichever you choose, think through the next steps in some detail, as though you were planning to complete the work.

- Describe the next steps for this project in a manner accessible to both technical and non-technical audiences.

- Be sure to motivate each step you recommend.

- Identify potential challenges that may need to be addressed for the next steps to be successful.

Thank you for taking the time to complete this assignment!

You should have received a task in your candidate home account to upload your zip file with your report, all code, documentation, and instructions. Click here to Sign In and Submit Data Science Challenge.