Project Guidelines

CAPP 30255 Advanced Machine Learning for Public Policy
University of Chicago

Project Type and Scope

The project is an important component of this course. It is an opportunity for you to work on an public policy project of your choice. Course projects are also one of the favorite topics during a job interview.

You should preferably work in groups of three or four for the project, but there are no hard restrictions. Your project should apply text processing or machine learning to a task in public policy. It should also significantly advance your skills beyond what is covered in class.

Here are a few possible project ideas, some based on past projects in this course—

- Build a classifier to distinguish opinions from news, or fake news from real news.
- Compare bias (based on gender, race, etc.) in different text sources, such as user comments, transcripts of online course videos, film scripts, social media posts, blogs, court decisions, etc. Alternately, identify toxic comments in different text sources.
- Analyze tweets from political leaders and their base to determine differences in sentiment with respect to voter suppression, tax reform, living wage, etc.
- Build a structured dataset from news reports on violent conflicts. In particular, automatically identify the main actors and the type of violence.
- Predict which tweets or news stories will go viral.
- Analyze the text of legislative bills to summarize them, or to predict the number of votes it will receive, or to determine the amount of "pork" in it.
- Predict the likelihood of success of Medicare/Medicaid appeals or of complaints made to Consumer Financial Protection Bureau.

- Cluster complaints against the Chicago Police Department.
- Determine the level of correlation between politicians' statements and the campaign contributions to them.
- Determine the correlation between a politician's area of expertise—based on biographical accounts and past work experience—and the kinds of bills they author.
- Build a tool that in some manner helps citizens and organizations search or understand data or documents related to policy. For instance, the tool could regularly search a variety of government websites and determine if there are any new developments in, say, women's rights.
- Study policy applications or policy related abuse of large language models such as ChatGPT.
- Propose how governments should regulate artificial intelligence in the context of large language models.

Many of the above are topics of active research, and you will likely find research articles containing sophisticated solutions for them. Given the available time, however, you cannot expect to develop a state-of-the-art solution during this course project. But use this opportunity to delve deep into a problem of your choice, apply what you know so far in machine learning, study the literature and learn new techniques relevant to your problem, develop a serious solution, evaluate its performance, and present your experience and results. Along the way look for some insights that future interviewers would find interesting—"Horror films are the most gender neutral" or "Restaurant review scores are surprisingly uncorrelated with their level of hygiene."

Assume you'll have ample computing resources; we expect to receive AWS credits for every student by Week 5.

Time commitment

The project counts for 50% of the grade, and you should plan to work accordingly. Plan to spend a minimum of 35 hours on your project during the quarter.

Most projects will require a combination of, and several iterations of, reading up the existing literature, thinking on the problem, learning how to

use relevant software libraries, writing and executing code, and preparing a poster presentation, and report. Depending on the project you will spend more time in one activity than the other. Each team member should, however, expect to write about 1000 lines of code or more as their contribution.

Tips for a successful project

Here are some tips that may be useful as you work on your project.

- Propose a project in which you can identify a sequence of goals, including at least one that should be easy. So if you encounter unforeseen difficulties, you'll at least be able to achieve the easy goals.

  Although we'll study several advanced models, such as RNNs, CNNs, and Transformers, it is often difficult to get them to work on new datasets, particularly in a short period of time. So while it is a good idea to incorporate these models in your project, only about a third of your project should depend on these models working.

- Suppose you are studying a problem P, but don't get good results. It is often difficult to determine whether the reason is that P is an inherently hard problem, or whether your code or setup has a bug or other issues. One way to determine this is to test your setup or code on a previously-studied problem, say P', as close to P as possible. If you can get similar or comparable results as others have obtained on P', then there is a good chance your code and setup is fine, even if you don't get good results on P.

  It is best you plan for this in advance and identify P' in your proposal itself.

Deliverables

There will be three deliverables: (i) Project Proposal (10% of total course grade), (ii) Mid-quarter Report and Meetings (10%), and (iii) Final Report and Poster Presentation (30%). Your proposal and the two reports should be posted on Ed other students on Ed. You should offer constructive feedback to other teams on their proposals and reports.

Project Proposal

The proposal is due 11:59 p.m., Tuesday, April 11, on Ed. It should be 3-4 pages (per team member) long and include the following:

- Title and team members.
- Brief description of what you want to do, including why it is useful, the data and software you will use, and the software you will write, if any.
- A **detailed description of the related work**. You should search for research papers and projects that solve the same or a similar problem. For the closest two or three such papers, you should describe what methods they used for obtaining the data, preprocessing the data, learning models, choosing metrics, and evaluating their results. You should also report their results, and what implications their work has on your project. (Learn as much as you can from such related work as it will give you an idea of what you need to teach yourself—beyond the material covered in class—for your project.)
- Brief plan of action, including any insights you have, the various steps of the project, the software libraries or packages you will be using, and the software you will be developing on your own. Also include what you plan to finish by the mid-quarter meetings and by the final poster presentation.
- Brief description of how you will evaluate your work. There may be existing techniques or results you could compare to, or you could test how well your solutions performs on test data, or how well it models the available data.
- Brief description of how the work will be divided among team members.

As you work more on the project—encountering obstacles or after more thought—you might take a different path than what you had first proposed. That is fine. Yet, writing a comprehensive proposal is key to a successful project.

Mid-quarter Report and Meetings

The mid-quarter report is due 11:59 p.m., Tuesday, May 2, on Ed. It should be an update on the proposal and should include the work done and the results obtained so far, and changes to the project objectives and plan, if any. Each team should also meet with a TA by May 12 for feedback, clarifications, and if they have software to demonstrate.

Final Report and Poster Presentation

The final report is due 11:59 p.m., Monday, May 22, on Ed. You should also present a poster on your work during a session that will be organized later in the week. More details on the poster presentation will be posted around Week 7. Each team member should be prepared to talk about the project overall as well as well describe in detail the components they worked on.

The report should resemble a professional research paper and include—

- An abstract-like summary describing the work you have done.
- An introduction describing the problem and its significance.
- A detailed description of the related work.
- A detailed description of your solution and the work you have done.
- Include both what worked and what didn't, particularly if you have insights into the reason why.
- A detailed presentation of the results you have obtained and its analysis.
- Suggestions for future work along the same line.
- Description of your effort, including the relative effort on different activities. What did you have to learn for the project? What skills did you already possess? E.g., did you need to learn how to use particular libraries? Did you spend more time reading research papers, or fine tuning your parameters? If you were a team, what part of the work was done by which team member?
- Bibliography.

Please write your report in clear, concise English, and include clearly captioned, helpful figures. Unclear or sloppy reports will affect your grade.

Also submit a link to a github repository that contains all your source code. Make your repository easy to navigate and understand by adding readme files to the important directories that—

- Briefly describe the purpose of each file and the number of lines of code in it.
- List files that contain code not written by a team member, or not written as part of this project (maybe it was counted as a part of another project).
- Contain any other relevant information about your code.

Your code should be documented such that one can get a reasonable idea on how it works. E.g., each function should have a document string in the recommended style of the programming language.