DS-UA 301 Project description

Overview

Part of this course consists of building on the concepts you have learned in class to pursue a small research project of your own, using techniques from either NLP, RL or both. You will present your findings from the project as a short report (8-12 pages not including references), and to the class in “data blitz” form. You may work in groups of 2-5. Each group will submit a 2-3 sentence statement describing the contributions of each team member.

Timeline

While only the data blitzes and final report have due dates, we have also provided some milestones at which we will check in about different aspects of the project. We highly recommend you take advantage of them to discuss your ideas with us and seek out any additional resources you might need.

- **Research question** (week of March 28th)
    - Individually submit up to three research questions you are curious about
    - Get constructive feedback on feasibility from your professor and TA
    - Get connected with students with similar interests
- **Dataset/problem selection** (week of April 4th)
    - In groups, select a dataset to analyze or a problem to solve
    - Think about how to process your data for analysis
- **Analysis** (throughout April)
    - Formulate your hypothesis
    - Select analysis techniques and make predictions based on the hypothesis
    - Get feedback on which techniques might be appropriate
- **Short in-class “data blitz”**
    - ***Due date***: April 28th and May 5th in lab
    - Briefly (5-10 min) present your rationale for the project and current results
    - Get feedback from the group
- **Final report due**
    - ***Due date***: May 13th
    - Conduct any additional analysis based on feedback
    - Write and submit the report

You are also welcome to reach out to us earlier than March 28th, either at office hours or over email:

Prof. Angela Radulescu	Prof. Andrea Jones-Rooy	Johnny Ma (TA)
ar6982@nyu.edu	ajr348@nyu.edu	jlm10003@nyu.edu
OH: M 4-5pm, Zoom	OH: W 1-2pm, Zoom	OH: T 12-1pm, Zoom

Format

Project write-ups should roughly follow the standard format of a scientific research paper. You can use whatever text editor you would like. The basic outline is the following:

- **Abstract**
  - 250-word "elevator pitch" about the project communicating the question, approach and main conclusions

- **Introduction**
  - Scientific question and motivation
  - Literature review (can be short)
  - Hypothesis

- **Methods and results** (**Tip**: papers go faster when you write this part first)
  - Description of analysis techniques
  - Description, plotting, and some discussion of the results

- **Discussion**
  - Reiterate conclusions
  - Connect to previous literature
  - Limitations and future directions

Recommended workflow: conduct your analysis in a Jupyter notebook, which you can iterate on as you explore different analysis techniques. Once you have your results in figure and caption form, you can "freeze" the notebook (that is, make no more changes to the analysis), and export figures to a slideshow editor such as Keynote or Google Slides. In the editor, you can: (1) arrange figures into panels for the report; (2) turn slide decks into talks.

Note: To abide by reproducible research standards, the final submission will consist of both the report and your analysis code (submitted individually by each team member).

NLP specifics

In general, there are three broad types of projects you might carry out. If you're interested in exploring something that doesn't fit obviously into one of these three categories, just come talk with us and we'll see if it'll work (generally, we are open to ideas, but better to be safe and check first)!

(1) Use NLP to study the evolution of a component of a language in terms of the meaning or use of a word over time or by some defined population(s). In order to be hypothesis-driven, you must both demonstrate the word evolution and offer and test hypotheses about, e.g., why it changed or whether it is, indeed, used by different groups in statistically and substantively meaningful ways.

(2) Use NLP to study the relationship between language and some other phenomenon in the world. An example of this would be using Twitter data to predict or explain a political event, market change, or the popularity of an artist or work. You may also consider the other direction: Can an event explain or predict

a change in language, perhaps in meaning or frequency of use? Again, your work must be hypothesis-driven, not just descriptive.

(3) Apply NLP to a new problem. NLP has been applied far and wide in the above two types of studies as well as in the AI/language generation space. But there are many more frontiers to go -- the sky's the limit, provided it's (you guessed it) hypothesis driven! As stated above, just check with us to be sure.

RL specifics

In general, projects in reinforcement learning will come in three different flavors that we will see examples of throughout the module:

(1) Using reinforcement learning to formulate a hypothesis about the mind that you can test against human behavioral or neural data. This is an exercise in building theories about how humans behave. It would give you practice with advanced statistical hypothesis testing using time series data. It would teach you to select and work with various open datasets. And it would touch on how to design model-driven experiments.

(2) Testing one or several reinforcement learning algorithms on a benchmark problem in AI. This type of project would be mainly engineering-focused: you would start by implementing an existing algorithm for solving a problem, training it from scratch, and then extending it in interesting ways in the hope of solving the problem more efficiently.

(3) Applying reinforcement learning to a new problem. This is a neat opportunity to think about the scope of problems RL can be applied to, and the challenges that come with each. It will teach you how to think about important concepts like feature engineering, domain knowledge, and reward function design, and the constraints on each.

Because we are somewhat constrained by the arrow of time, you will see examples of each at different points throughout the RL module. It is ok, and even encouraged, to peek ahead at the readings, and talk to us in advance about which kind of project you would like to do. Once you do so, we can point you to additional resources to help you narrow your scope.

Interdisciplinary thinking

Part of the joy of learning about two different fields in the same course is that you are free to combine ideas from each in your research. Some of the most innovative work occurs where boundaries between different disciplines are crossed and re-drawn. So, while this is not explicitly required, to the extent that you want to pursue a project that mixes NLP and RL, you are welcome to do so. One thing you might consider doing is keep a running list of ideas for projects that you can populate as you advance in the course. Here are a few to start you off:

- Use RL to predict which words will "emerge" and which will "die out" as two agents communicate with each other
- Use NLP to predict affective states, like anxiety or anger, in members of a community
- Use both RL and NLP to explore different facets of a single phenomenon in which you are interested
- Use RL to better understand the process of language learning, either in terms of first or second (or third…) language acquisition, or in terms of how we “learn” new words as they emerge
- We invite you to take it from here!