CSC 546/746 Artificial Intelligence

Final Project guidelines

In the final project, you will apply the techniques that you have learn in CSC546/746 to a setting that you are interested in. All students must complete work in each milestone:

- **Initial discussion (Due: Sunday, September 10 at 11:00 pm):** Post your ideas about the final project in the discussion board, include a brief overview of your problem, the source of your dataset, the main approach that you plan to use to solve your problem, and the challenges you might have. At the same time, respond to at least two of your classmates' projects. You are expected to exchange ideas in the discussion board and settle your project ideas by the due date.
- **Project proposal (Due: Sunday, September 17 at 11:00 pm):** I would expect 1-2 pages (single-spaced with max 12-point font, not including the references page). The proposal should include the following items:
  - Project title
  - Problem Statement and Task Definition - What does your system do (what is its input and output)? What real-world problem does this system try to solve? Choose a problem that you could solve in 2- 3 weeks.
  - Dataset - Reference the source of the data if you do not collect the data yourself. You don't need to settle the dataset at this stage, but you need to have a plan how to collect the data you need. Describe the input/output behavior with concrete examples of both the inputs and outputs (explain what the input and output variables should look like and how they interact with the system). You can collect some preliminary data that you can use in your description of the input and output behavior.
  - An evaluation metric - How will you measure the success of your system? Why does this metric work best for this problem?
  - Related works - Search the Internet for similar projects and mention the related research and projects.
  - Description of the challenges - What are the challenges? Which might be able to address those challenges?
- **Final presentation (Due: Wednesday, Oct 04 in class):**
  - Each student will prepare for a 5–10-minute talk for the final presentation.
  - You are encouraged to use any technology to present your project, but slides may be a relatively easier option.
- **Final report (Due: Thursday, Oct 05 at 11:00 pm):** It requires a four-page manuscript in IEEE Conference format (https://www.ieee.org/conferences/publishing/templates.html) that includes abstract, introduction, methods, results, discussion & conclusion. List of information you may include in your final report:
  - Introduction - Brief overview of your problem. Why might this problem be important?
  - Literature Review - Description of other work/papers you've found that are related to your task.

- Dataset - Description of data you are using, size of the dataset, distribution of classes, any preprocessing you needed to do.
- Main Approach - Propose a model and an algorithm for tackling your task. Don't describe methods in general; describe precisely how they apply to your problem (what are the inputs/outputs, variables, factors, states, etc.)?
- Evaluation Metric - Please include what metrics, both qualitative and quantitative, you are using to evaluate the success of your problem.
- Results & Analysis - The analysis of the results is very important, and it requires you to think about what your results might mean. Describe a few experiments that you ran that show the properties (both pros and cons) of your system. Analyze the data and show either graphs or tables to illustrate your point. What's the take-away message? Were there any surprises? Are there potential errors in the method and why they may have occurred?
- Future Work - This section can include some ideas about how you could improve your model if you had more time. This can also include any challenges you're running into and how you might fix them.
- Code – Submit your code and data with the report. If the size of your dataset is too large, you may submit a small portion of your data. You need to indicate the size of your original dataset in your report.
- References - Please include a reference section with properly formatted citations (any format of your choice).

All projects require students to spend time gathering data and setting up the infrastructure to reach an end result.
You may use existing resources in your project.

- Datasets
  You are free to use existing datasets, but these might be not necessarily the best match for your problem, in which case you are probably better off making your own dataset. Here are a few places you can look to get data:
  - Popular open data repositories
    - UC Irvine Machine Learning Repository: http://archive.ics.uci.edu/ml/
      - Kaggle datasets: https://www.kaggle.com/datasets
      - Amazon's AWS datasets: https://registry.opendata.aws/
  - Meta portals (they list open data repositories)
    - Data Portals: http://dataportals.org/
    - OpenDataMonitor: http://opendatamonitor.eu/
    - Quandl: http://quandl.com/
  - Other pages listing many popular open data repositories
    - Wikipedia's list of Machine Learning datasets: https://homl.info/9
    - Quora.com: https://homl.info/10
    - The datasets subreddit: https://www.reddit.com/r/datasets
- Libraries
  You are free to use existing tools for parts of your project as long as you're clear what you used. When you use existing tools, the expectation is that you will do more on other dimensions.
  - scikit-learn: machine learning library implemented in Python

- Natural language Toolkit (NLTK): a set of tools for basic NLP in Python
- OpenCV: Python libraries for simple computer vision