CSC110 Project Phase 1: Proposal

The CSC110 Course Project is an opportunity to use what you have learned in this course and apply it in a creative, open-ended project. The final submission of your project includes a Python program and report. But before that, you must complete Phase 1: a proposal of what you plan on exploring, designing, and implementing. The CSC110 Teaching Team will give you feedback on your proposal to make sure that your idea is both sufficiently complex and can be completed by the final due date.

Logistics for Phase 1

- Due date: Friday, November 5th before 9am Eastern Time.
- This assessment can be done in groups of up to **4** students.
- You will submit your proposal on MarkUs (see submission instructions at the end of this handout).
- Please review the the Course Syllabus section on Academic Integrity.

Project topic introduction: The Impact of COVID-19

The COVID-19 pandemic is not yet over, but its impact on our lives and the globe is definitely being felt. It has impacted our environment, our healthcare systems, small and large businesses, economic markets, our memes, and virtually every facet of our society. This course is not “about” the pandemic. But the skills you’ve developed in this course, and that you will continue to develop in the rest of your computer science career, can be harnessed to (try to) answer questions about it.

And so for your final project, we ask that you investigate some aspect of the impact COVID-19 has had through a *data and computational lens*. You may investigate the effects of the pandemic on the environment, traffic, education, the potential benefits and limitations of proposed or enacted solutions, mental health, or a completely different angle. We are certain that you will be able to find something to study that is engaging and vital to you.

Project overview

Your project will focus on *answering a data-centric question related to the impact of COVID-19*. You are free to use your imagination and be creative here, and choose something that you are truly interested in answering. There are only two constraints in choosing a question to investigate: it must be meaningfully related to the impact COVID-19 has had on the world in some way, and it must be connected to some kind of real-world data.

For your project, you do the following:

- *Choose* a particular topic within the broad scope of the problem domain and *research* this topic.

- *Formulate* a specific data-centric question about this topic, informed by the research you've done.
- *Identify* one or more real-world datasets related to this topic that can help you answer this question.
- *Compute* on the dataset(s) you've found (and possibly perform other computations as well).
- *Report* the results of your computations in a visual and/or interactive way.

You have performed elements of this a number of times throughout the course already: exploring and computing on datasets (e.g., TTC subway delays, course timetables), defining data models (e.g., generative text models), and visualizations using `plotly` and `pygame`. The purpose of the project is to do something similar with a topic within the problem domain, but on a grander scale. The explorations in the assignments were, metaphorically, a small sandbox for you to play in. In the project, you will build your own, bigger sandbox and play games of your own choosing.

Example project ideas

Choosing your own topic can be hard to do, especially when given a fairly broad space like "the impact of COVID-19". Below, we've listed some example project ideas and questions to act as sources of inspiration for your own exploration and brainstorming. (Each one would need to be refined to make a good proposal.) You may not copy any of these ideas directly, but you are welcome to modify and expand on them if you find something you're particular interested in. And if you want to go in a completely different direction, that's okay too—please be creative!

- David loves models—mathematical models, that is! His question is, "*How well can simple supply-and-demand models predict real changes to prices caused by COVID-19?*" He plans to learn about some simple supply-and-demand models that estimate price and implement them in Python. Then, he will compare their estimates over the last year or so with the actual real-world data. His program will display side-by-side graphs of the model and real-world data, for different values of the model parameters.
- Jen has heard that the pandemic is significantly impacting the price of food. Her question is, *"How has the pandemic's impact on supply chains increased (or decreased) the cost of food around the world?"* (Jen might want to narrow that down to specific regions and/or specific commodities so that she isn't overwhelmed with data.) Jen plans to explore datasets on price indexes from multiple countries and see if it can be predicted by the increasing global cost of transportation (e.g., ocean freight prices, fuel prices).
- Jacqueline loves keeping up to date with the news. Her question is, *"Has the pandemic increased the discussion of mental health in the media?"* She plans to to collect articles from popular publication venues and analyse their texts for mentions of keywords relating to mental health. In addition, she wants to create her own dataset that tracks major events that occured during the pandemic. Her goal is to find out if (a) the pandemic resulted in more articles on mental health, and (b) if there are any patterns associated with major events during the pandemic and a corresponding serge surge (or dirth) of articles.

And here are some examples of **bad** ideas for a project:

- Paul absolutely *loves* cryptography. He wants to design a new cryptographic system. In order to make this relevant to the pandemic, the data he will encrypt will be related to COVID-19. His goal is to encrypt and decrypt COVID-19 data.

  (*Not meaningfully relevant to the problem domain*)

- Mario found the use of `pygame` in the tutorials fascinating. He wants to design a game using the `pygame` library. His goal is to create a game where the player repeatedly taps the screen so that a bird is able to make it through all the levels. The levels will be made relevant to the pandemic because there will be hazardous diseases to avoid.

  (*Not meaningfully relevant to the problem domain*)

- Diane is interested in the pandemic's impact on people eating in restaurants. She finds a dataset showing the average amount of money spent on restaurants each month over the last 3 years and plots it using `plotly`.

  (*Too simple.*)

Project proposal instructions

Your project proposal is designed as a way to get you started working on this project early, and to give your TAs an opportunity to give you some meaningful feedback and suggestions on your ideas to make sure you are on the right track. *We expect everyone to do very well on this proposal*—our goal is to spend most of our grading time giving you feedback. This proposal does not lock you into a particular topic, and your group will be free to change your plans between this proposal and your final project submission.

Your project proposal will be a LaTeX document (using the **template `project_proposal.tex`**; see the MarkUs starter files) consisting of the following components:

1. Project title (pick something informative and professional, but you can be creative too) and name of *all* group members.
2. Brief problem description and research question. (300–400 words)
   - Give an overview of any background knowledge necessary for *the reader* to understand the problem you are studying.
   - Provide context for the problem and motivate why you have chosen your research question.
   - Your research question should be in **bold**; it should be fairly concise, but can be more than one sentence.
3. A description of at least one relevant dataset you have found. (~150 words)
   - State the source (e.g., government/organization website) and format (e.g., text, csv, json, image) of the dataset, and give some sample data contained inside that dataset.
   - Don't be afraid to cobble together your own dataset, such as creating a collection of images that are related. Or to combine two datasets from different sources.
   - You will also submit a small sample of your dataset to MarkUs along with your project proposal document. (See more below)
4. A *computational plan* for your project. (300–500 words)
   - Describe the kinds of computations you plan to perform, such as: data transformation/filtering/aggregation, computational models, and/or algorithms.

   - Explain how your program will *report* the results of your computation in a visual and/or interactive way. You don't need to go into a lot of details here, but it should be clear what you plan to do.

   Technical requirement: for your project, you **must** use at least one Python library/module that we have not covered in this course, *or* use `plotly` or `pygame` to a much larger extent than what what have given you so far in this course. (See examples and note in the next section).

   - In this part of your proposal, you should also describe one new library you intend to use, how you will use it, and why it is appropriate. Refer to specific functions, data types, and/or capabilities of the library that make it relevant for solving the problem you wish to solve.

5. A references section that lists the references you used for your proposal. This should include references from your topic research, the reference for where you obtained the dataset, and any online documentation or tutorials for the Python library you plan to use for the project.

   - You may use any academic reference style you wish, e.g., APA or MLA.

Some example sources for datasets

Finding a dataset for your idea may be difficult, depending on your topic and question. Below are some examples of websites where you can find datasets or search for datasets. But please don't limit yourselves to data from these sources.

- Google now has a Dataset Search (https://datasetsearch.research.google.com/). Learn more about this feature from their blog post (https://blog.google/products/search/discovering-millions-datasets-web/)
- Statistics Canada has a page dedicated to a data perspective on COVID-19 (https://www.statcan.gc.ca/en/covid19?HPA=1) – if you are stumped for inspiration, this may also be a good source of ideas.
- Many governments have an "open data" website where they publish data to the public that they have collected. For example, here is one for Canada (https://open.canada.ca/en/open-data), Australia (https://www.opendataaustralia.org/), and the United Kingdom (https://data.gov.uk/).

A note about originality, licensing, and references

As you do research and come across datasets, it is important to note down where a fact or piece of data came from. This is especially true with data: just because you were able to download it online does not mean you have permission to use it. You need to not only make sure that you have permission to use the data, but also must include an attribution to where you found the data. Similarly, when providing context and/or making claims in your proposal, be sure to provide a reference to where that information came from.

Example Python libraries

- `scrapy` (https://scrapy.org/): a library for extracting data from websites

- `scikit-learn` (https://scikit-learn.org/stable/index.html): a machine learning library that is (relatively) easy to use
- `scikit-image` (https://scikit-image.org/): a library that helps process image data
- Natural Language Toolkit (nltk) (https://www.nltk.org/): a library that helps with analyzing text written in a language (like English)

Also, for your reference here are links to the websites for documentation for `plotly` and `pygame`:

- `plotly` (https://plotly.com/python/)
- `pygame` (https://www.pygame.org/docs/)

Note: While you may rely on the library you choose to help you with some computations, you cannot use the library to do *all* your computations. That is, simply applying a well-known algorithm to your dataset by calling functions in the library is not a sufficiently complex project.

Similarly, if your library is responsible for visualization and not computation, your project should not simply be to load and visualize the data. The data you load must be transformed or computed upon in some way before visualization.

Submission instructions

Please **proofread** your work carefully before your final submission!

1. Login to MarkUs (https://markus.teach.cs.toronto.edu/csc110-2020-09).

2. Go to *Project Phase 1: Proposal*, and the "Submissions" tab.

3. Submit the following files: `project_proposal.tex`, `project_proposal.pdf`, and your sample dataset file(s).

    - You decide how many files are in your dataset (i.e., one or more).
    - Please don't submit the full dataset (this will likely be too large for MarkUs). Instead, extract a small part of it (e.g., just the first 100 entries) to submit.
    - Your TAs *will* be checking your data set file(s) manually, so make sure they are formatted correctly.

4. Refresh the page, and then *download each file* to make sure you submitted the right version.

Remember, you can submit your files multiple times before the due date. So you can aim to submit your work early, and if you find an error or a place to improve before the due date, you can still make your changes and resubmit your work.

After you've submitted your work, please give yourself a well-deserved pat on the back and go take a rest or do something fun or eat some chocolate!