DSCI 510: Principles of Programming for Data Science: Final Project Guidelines

In the final project for this class, you will have the opportunity to apply the knowledge and programming skills you have learned to a real-world problem. Your project should focus on web scraping (or collection data through APIs), data cleaning, analysis, and visualization using Python.
You can work on the project on your own or can form a team of maximum 2 students.

Project Deadline: Dec 15, 2025 11:59 PM
You must upload all files to a folder in a GitHub repository and submit the link to it on Brightspace.

In the **README** file (described below), you must mention who the team members are (including name, email, Github username and USC ID) for this project.

Project Proposal

Upload to Brighspace a **one page** proposal describing the project. This proposal should include the following:

1. Name of the project and team members
2. What problem are you trying to solve?
3. How will you collect data and from where?
4. What analysis will you do and what visualizations will you create?

The deadline for this proposal is Nov 21, 2025. We will provide feedback and suggest changes if required.

Project Goals and Steps

1. **Data Collection (20%):** You should identify websites or web resources from which you will get the raw data for your project. You can either web-scrape data or collect data using available APIs. This could include news articles, e-commerce websites, social media posts, weather data, or any other publicly available web content.
   You should not be simply downloading a file from a website for example. This step should be fairly sophisticated and demonstrate the techniques you learnt in the class.
   Using Python libraries like `BeautifulSoup` and `requests`, you should write scripts to scrape data from the chosen websites. This step includes handling HTML parsing, making HTTP requests, and extracting relevant information.
   We recommend that you scrape data from static websites, or use APIs provided by the source. If you scrape data from dynamically generated pages, you might see unforeseen complications. **Note:** Some APIs are paid, and you might have limited access only, without paying for it.

2. **Data Cleaning (20%):** After data collection, you will need to clean the data and preprocess it. This will involve handling missing values, cleaning HTML tags, removing duplicates, and converting data into a **structured format** for analysis. **If your raw data is not in English, you must translate the data to English in this step.**

3. **Data Analysis (20%):** You will analyze the scraped data to gain insights or answer specific questions. You should perform statistical analyses, generate descriptive statistics, using libraries such as NumPy and Pandas (or any other library you wish).

4. **Data Visualization (20%)** : You should create plots, graphs, or charts using Matplotlib or any other library, to effectively communicate your findings.

5. **Final Report (20%):** You will submit a final report, describing the problem you are trying to solve, what data you collected, how you collected it and what type of data cleaning you performed. Explanation about analysis and Visualization.

The grading shown here is **a general guideline**, but it can be changed based on your project. If your data collection is simple but the analysis is fairly complicated, we will adjust the rubric accordingly.
Final grading rubric: TBD

Project Deliverables

Github Repository

We will create an assignment for the final project. You will accept the assignment and commit your code to the repository. Here is the generic structure of the repository:

```
github_repo_structure/
├── README.md
├── requirements.txt
├── data/
│   ├── processed/
│   └── raw/
├── project_proposal.pdf
├── results/
│   └── final_report.pdf
└── src/
        ├── clean_data.py
        ├── get_data.py
        ├── run_analysis.py
        ├── utils/
        └── visualize_results.py
```

Here is what each of the folders/files should contain / mean:

1. **project_proposal.pdf:** The project proposal file.

2. **requirements.txt:** This file lists all the external libraries you have used in your project. To install all the required libraries, you can run
   `pip install -r requirements.txt`

3. **README.md**: Documentation on how to install the requirements for your project. How to run your code, explain how to get data, how to clean data, how to run analysis code and finally how to produce the visualizations.
   We have created sections in the README.md file for you to fill in. Make sure you fill all the sections.

4. **data/ folder**: This folder contains the data used in this project.
   a. The **data/raw** folder will have the raw files you downloaded from the web. It could contain (not exhaustive) html, csv, xml or json files.
   b. The **data/processed** folder will contain your structured files after data cleaning. You may clean the data and convert them to JSON files for example. Your analysis and visualization code will perform operations on the files from this folder.

Note: Make sure your individual files are less than 100MB.

5. **results/ folder**: This folder will contain the final report and any other files you might have as part of your project. You may choose to create a jupyter notebook for visualizations, this notebook will be present in this folder.

6. **src/ folder:** This folder contains the source code for your project.
    a. **get_data.py** will download the data and store the data in the data/raw folder.
    b. **clean_data.py** will clean the data, transform the data and store the files in the data/processed folder.
    c. **run_analysis.py** will have code to analyze the data to answer the project specific questions.
    d. **visualize_results.py** will create visualization using matplotlib or any other library to conclude the result of the analysis you performed.
    e. **utils/** folder may contain any utility code you write.

The directory structure provided is a basic template. Replace placeholder files, such as project_proposal.pdf, with your actual files, like your proposal PDF. You can create more files in this repository as you require.

Final Report

The final report will have these sections.

1. What is the name of your project and who is in the team? Please describe it as a research question and provide a short description.
2. What data did you collect? How did you collect it? How many data samples did you collect?
    a. Specify exact data sources and your approach.
    b. Describe what has been changed from your original plan, what challenges you encountered or resolved.
3. What kind of analysis and visualizations did you do?
    a. What analysis techniques did you use, and what are your findings?
    b. Describe the figures you made. Explain its setup, meaning of each element.
    c. Describe your observations and conclusion.
    d. Describe the impact of your findings.
4. Future Work
    a. Given more time, what direction would you take to improve your project?

The final project report should contain (2-5) pages, both inclusive. It may not be too short or too long. Please spend a decent amount of time on the report. Your report is the first file we read. We won't know how great your project is if you can't explain it clearly.