**Project 3: Web APIs & Classification**

**Description**

In week four we've learned about a few different classifiers. In week five we'll learn about webscraping, APIs, and Natural Language Processing (NLP). Now we're going to put those skills to the test.

For project 3, your goal is two-fold:

1. Using Reddit's API, you'll collect posts from two subreddits of your choosing.
2. You'll then use NLP to train a classifier on which subreddit a given post came from. This is a binary classification problem.

**About the API**

Reddit's API is fairly straightforward. For example, if I want the posts from [/r/boardgames](https://www.reddit.com/r/boardgames), all I have to do is add .json to the end of the url: <https://www.reddit.com/r/boardgames.json>

To help you get started, we have a primer video on how to use Reddit's API: <https://www.youtube.com/watch?v=5Y3ZE26Ciuk>

**Requirements**

* Gather and prepare your data using the requests library.
* **Create and compare two models**. One of these must be a Bayes classifier, however the other can be a classifier of your choosing: logistic regression, KNN, SVM, etc.
* A Jupyter Notebook with your analysis for a peer audience of data scientists.
* An executive summary of the results you found.
* A short presentation outlining your process and findings for a semi-technical audience.

**Pro Tip 1:** You can find a good example executive summary [here](https://www.proposify.biz/blog/executive-summary).

**Pro Tip 2:** Reddit will give you 25 posts **per request**. To get enough data, you'll need to hit Reddit's API **repeatedly** (most likely in a for loop). *Be sure to use the time.sleep() function at the end of your loop to allow for a break in between requests.****THIS IS CRUCIAL***

**Pro tip 3:** The API will cap you at 1,000 posts for each subreddit (assuming the subreddit has that many posts).

**Pro tip 4:** At the end of each loop, be sure to save the results from your scrape as a csv: JSON from Reddit > Pandas DataFrame > CSV. That way, if something goes wrong in your loop, you won't lose all your data.

**Necessary Deliverables / Submission**

* Code and executive summary must be in a clearly commented Jupyter Notebook.
* You must submit your slide deck.
* Materials must be submitted by **9:00 AM on Monday, May 18st**.

**Rubric**

Your local instructor will evaluate your project (for the most part) using the following criteria. You should make sure that you consider and/or follow most if not all of the considerations/recommendations outlined below **while** working through your project.

For Project 3 the evaluation categories are as follows:  
**The Data Science Process**

* Problem Statement
* Data Collection
* Data Cleaning & EDA
* Preprocessing & Modeling
* Evaluation and Conceptual Understanding
* Conclusion and Recommendations

**Organization and Professionalism**

* Organization
* Visualizations
* Python Syntax and Control Flow
* Presentation

**Scores will be out of 30 points based on the 10 categories in the rubric.**  
*3 points per section*

| **Score** | **Interpretation** |
| --- | --- |
| **0** | *Project fails to meet the minimum requirements for this item.* |
| **1** | *Project meets the minimum requirements for this item, but falls significantly short of portfolio-ready expectations.* |
| **2** | *Project exceeds the minimum requirements for this item, but falls short of portfolio-ready expectations.* |
| **3** | *Project meets or exceeds portfolio-ready expectations; demonstrates a thorough understanding of every outlined consideration.* |

**The Data Science Process**

**Problem Statement**

* Is it clear what the goal of the project is?
* What type of model will be developed?
* How will success be evaluated?
* Is the scope of the project appropriate?
* Is it clear who cares about this or why this is important to investigate?
* Does the student consider the audience and the primary and secondary stakeholders?

**Data Collection**

* Was enough data gathered to generate a significant result?
* Was data collected that was useful and relevant to the project?
* Was data collection and storage optimized through custom functions, pipelines, and/or automation?
* Was thought given to the server receiving the requests such as considering number of requests per second?

**Data Cleaning and EDA**

* Are missing values imputed/handled appropriately?
* Are distributions examined and described?
* Are outliers identified and addressed?
* Are appropriate summary statistics provided?
* Are steps taken during data cleaning and EDA framed appropriately?
* Does the student address whether or not they are likely to be able to answer their problem statement with the provided data given what they've discovered during EDA?

**Preprocessing and Modeling**

* Is text data successfully converted to a matrix representation?
* Are methods such as stop words, stemming, and lemmatization explored?
* Does the student properly split and/or sample the data for validation/training purposes?
* Does the student test and evaluate a variety of models to identify a production algorithm (**AT MINIMUM:** Bayes and one other model)?
* Does the student defend their choice of production model relevant to the data at hand and the problem?
* Does the student explain how the model works and evaluate its performance successes/downfalls?

**Evaluation and Conceptual Understanding**

* Does the student accurately identify and explain the baseline score?
* Does the student select and use metrics relevant to the problem objective?
* Does the student interpret the results of their model for purposes of inference?
* Is domain knowledge demonstrated when interpreting results?
* Does the student provide appropriate interpretation with regards to descriptive and inferential statistics?

**Conclusion and Recommendations**

* Does the student provide appropriate context to connect individual steps back to the overall project?
* Is it clear how the final recommendations were reached?
* Are the conclusions/recommendations clearly stated?
* Does the conclusion answer the original problem statement?
* Does the student address how findings of this research can be applied for the benefit of stakeholders?
* Are future steps to move the project forward identified?

**Organization and Professionalism**

**Project Organization**

* Are modules imported correctly (using appropriate aliases)?
* Are data imported/saved using relative paths?
* Does the README provide a good executive summary of the project?
* Is markdown formatting used appropriately to structure notebooks?
* Are there an appropriate amount of comments to support the code?
* Are files & directories organized correctly?
* Are there unnecessary files included?
* Do files and directories have well-structured, appropriate, consistent names?

**Visualizations**

* Are sufficient visualizations provided?
* Do plots accurately demonstrate valid relationships?
* Are plots labeled properly?
* Are plots interpreted appropriately?
* Are plots formatted and scaled appropriately for inclusion in a notebook-based technical report?

**Python Syntax and Control Flow**

* Is care taken to write human readable code?
* Is the code syntactically correct (no runtime errors)?
* Does the code generate desired results (logically correct)?
* Does the code follows general best practices and style guidelines?
* Are Pandas functions used appropriately?
* Are sklearn and NLTK methods used appropriately?

**Presentation**

* Is the problem statement clearly presented?
* Does a strong narrative run through the presentation building toward a final conclusion?
* Are the conclusions/recommendations clearly stated?
* Is the level of technicality appropriate for the intended audience?
* Is the student substantially over or under time?
* Does the student appropriately pace their presentation?
* Does the student deliver their message with clarity and volume?
* Are appropriate visualizations generated for the intended audience?
* Are visualizations necessary and useful for supporting conclusions/explaining findings?

**Why we choose this project for you?**

This project covers three of the biggest concepts we cover in the class: Classification Modeling, Natural Language Processing and Data Wrangling/Acquisition.

Part 1 of the project focuses on **Data wrangling/gathering/acquisition**. This is a very important skill as not all the data you will need will be in clean CSVs or a single table in SQL. There is a good chance that wherever you land you will have to gather some data from some unstructured/semi-structured sources; when possible, requesting information from an API, but often scraping it because they don't have an API (or it's terribly documented).

Part 2 of the project focuses on **Natural Language Processing** and converting standard text data (like Titles and Comments) into a format that allows us to analyze it and use it in modeling.

Part 3 of the project focuses on **Classification Modeling**. Given that project 2 was a regression focused problem, we needed to give you a classification focused problem to practice the various models, means of assessment and preprocessing associated with classification.