**Project 1: Data Exploration and Social Network Analysis with Smogon Competitive Pokemon data (R)**

**Learning Goals**

* Refine skills in data wrangling of .json, text, and potentially web data
* Practice data wrangling and inference skills related to social network analysis with a large dataset
* Learn to draw applicable, meaningful insights from processing of large datasets
* Learn to work with data where information and documentation about the data is scarce

**About the Data**

Smogon is a competitive Pokemon community that hosts a popular online Pokemon battle simulator. In a Pokemon game, 2 players form a team of 6 Pokemon that battle in a turn-based roster format. Every “official” game (as opposed to private games or games with custom rules) played on the servers is logged. Unfortunately, raw logs are not published, but a few community members pre-process the data and publish it monthly to provide information on trends in tactics. This includes:

* Usage – proportion of teams containing a certain pokemon
* Proportion of teams relying on identifiable tactics
* Common teammates and counters to pokemon

Stats are released monthly and come mostly in a pre-formatted .txt file for viewing and discussion. However, more data savvy users are pointed to a “chaos” folder containing .json files which are said to contain more information that could be incorporated into the text files. An example of data from Jan 2020 can be found at <https://www.smogon.com/stats/2020-01/>.

Despite not having all the raw data, what is available is already a rich resource.

First, I think the data can provide (or better present) data of general interest, such as:

* What are the most common Pokemon used?
* What are the most common types of Pokemon used?
* How these have changed over time given rule changes, bans, or evolution in tactics?
* Most used moves, and their types
* Most used items given pokemon
* Most used moves given pokemon

Second, given the highly relational nature of Pokemon (there are complementary and threatening combinations), I think Social Network Analysis and Network Visualisations can be useful for players to understand the prevailing relational patterns across games, and hence better prepare themselves for likely scenarios when facing other players. These include:

* Which Pokemon are popular and hence important to consider when teambuilding?
* Which Pokemon tend to be used together, or tend to face-off against each other?
* Which Pokemon is opponent likely to switch into to counter a play?
* What/Are there team compositions that pose a significant threat to your team?

The drawback is that the data is not at all documented, save for a couple of forum threads where the people who maintain the data reply to queries. Worse, the person who wrote the code that processes the raw logs is apparently no longer involved with maintaining the project.

**Project Objectives**

To answer some of the questions above, and others I find meaningful, and communicate them with visualisations and exposition in an Rmd document (or better, a webpage, if I get round to it). Ideally this will involve 2 iterations of data sharing with the community.

* Initial - early/mid-April
* 2nd iteration including feedback and more information the community is interested in

**Why this project**

The objective here is to ask good questions and present the answers I find to them well. Context matters for this, and as much as I’d like to work with a more practical dataset for the CV, as a humble student, I’m finding it difficult thinking about what I can do with datasets from places like OECD, the Federal government, etc. that people would actually care about. But as an occasional Smogon player, I think I can make good use of the available data, and I already have a good impression of what’s available. I’m thinking that I should take the experience of doing this and work on a more substantive dataset the next time.

**Project 2: Kaggle Competition using the Titanic dataset (Python)**

**Learning Goals**

* Refine skills in data wrangling
* Practice skills related to machine learning, particularly feature engineering and model refinement
* Inculcate good data processing practices for machine learning like
  + Exploring the data
  + Checking for missing values
  + Centring and scaling continuous features where advisable
* Gain fluency with the machine learning workflow
* Get comfortable with the constant refinement and trial-and-error involved in building a good machine learning model

**About the Data**

The Titanic dataset is a publicly available dataset containing information about the passengers on the Titanic. As a machine learning exercise, the objective is to predict who survived the disaster using other features like gender, ticket class, age, etc..

Kaggle runs a permanent “competition” (rankings reset every 2 months and there’s no prize) using the Titanic dataset as an entryway to get beginners started with machine learning and/or the idea of prediction competitions.

**Project Objectives**

* Make a submission to Kaggle, then improve model performance as far as possible. Get my name on the board (anywhere, no hard expectations here).
* To gain enough confidence with machine learning and the Kaggle framework to take on another competition with a less friendly dataset.

**Why this project**

My impression of prediction tasks is that getting started is rather simple – you split the data, fit the model, predict, get accuracy score. You can try a little harder and do cross validation or a tuning parameter search. The hard part is the ‘art’ of prediction – what tends to work and what doesn’t (and when), how do you get round annoying problems in different situations etc. I think this is probably what comes with practice and experience. Kaggle has been recommended to me by a few people to try my hand at ML “for real”, which apparently is good for the CV. Kaggle recommends the titanic as a starting point., and hopefully I feel good enough to move on to a more challenging dataset and problem after that.

**Timeline**

The general plan is to do most of the heavy lifting of the projects, particularly the prediction project, during/after Spring break.

Before Spring break

* Learn and practice the relevant techniques in classes
* Try to get as much of the 4051 assignments done before/over Spring break to create time for these projects after the break
* Work through these
  + <https://kateto.net/network-visualization>
    - To see how I can create better, more visible network graphs than the base igraph plots. Likely important for the Pokemon network, which can be quite big depending on what is included.
  + <https://www.coursera.org/learn/competitive-data-science>
    - To see what people think about when trying to improve model accuracy
* Get and process the raw data (particularly for the Pokemon data)
* See how far I can get with the Pokemon data (which I'm more confident with and have a rough idea of what I want to do and how to do it)
* Run an initial model with the Titanic data