# Problem Set 4

Karley Nadolski

February 2021

## 1 Web Scraping Dreams

Last semester, I took a class called Cultural Heritage Data and Social Engagement. We talked extensively about data ethics and how data scientists should incorporate certain humanities ideas into data science practices. One of the main ideas of the class was that data, especially cultural heritage data, shouldn't be "decoupled" from its origin (especially if that origin is a specific community or culture). I've always been interested in the crossroads between technology and the humanities, and that class has really changed how I think about using data (especially when finding it for myself on the internet).

Because it is situated right in the cross-hairs between humanities and data analysis, I'm most interested in scraping from classical texts from Project Gutenberg. In past classes (including the one I mentioned previously), I was introduced to both complex and relatively simple projects with textual analysis and was really interested in both.

## 2 R activity - using sparklyr

- What is the class of df? of df1?

  When working through the activity, I created a tibble containing the iris data (df1) and then copied it (df) into Spark. After doing so the classes of the two dataframes were slightly different. The class of df (the dataframe I copied into Spark) was:

  `"tbl_spark", "tbl_sql", "tbl_lazy", and "tbl".`

  The class for the df1 data frame was:

  `"tbl_df", "tbl", and "data.frame."`

- Are the column names any different across the two objects? If so, why might that be?

The column names were different in format between the two objects. For df1, the original tibble that I created directly from the iris data, the column names were:

`Sepal.Length, Sepal.Width, Petal.Width, Petal.Length, and Species`

For df, the dataframe copied into Spark, the column names were a little different:

`Sepal_Length, Sepal_Width, Petal_Width, Petal_Length, and Species`

I'm not sure exactly why the column names changed between the two objects, but it had to have been copying the dataframe into Spark that caused the change. Perhaps in Spark formatting, the underscore is more appropriate than the period between words.