OS 4118 **Twitter Data Exercise**  Fall AY 2020

**Introduction:** Social media data is everywhere, and lots of people think that this can become a useful source of information about people and organizations. Whether that is true or not, it is worth our while to become familiar with some of the flavors of social media data and some of the tools for handling it.

**Twitter Data:** Twitter data is widely used, maybe in large part because it is widely available. A Twitter message (a “tweet”) is very often represented in JSON format, each JSON message taking up one text line. Often these lines are grouped together into files which are then zipped using the Unix utility gzip (but in this case I’ve unzipped the file for you). The R function file() is a good tool for accessing unzipped text files, and gzfile() does the same for zipped ones. (Note: on Windows, readLines() produces warnings when reading lines from gzipped files. I suspect that’s because the read action requires a “seek” to the new-line character, and seek() is apparently problematic on Windows. The real solution is probably to unzip the file, but readLines() has worked for me so far.)

These files can be of substantial size – up to a couple of hundred megabytes zipped. I have given you a subset of 10,000 lines, figuring that if you can handle that, you’ve answered the question.

**Our plan:** Read the Twitter file (tweets.20130201\_055232\_10000lines) and extract a few fields like ID, coordinates (if present), date, user’s time zone, the language, anything else of interest, and of course the text. We will put that information into a tab-separated file for analysis. **Important**: we will need to remove from the text any special characters that will get in the way of interpreting the TSV, like tab ("\t"), newline ("\n"), or carriage return (“\r”) characters. The gsub() (“global substitute”) command can do that for you. If you’re curious I’ve put the full 230MB file up, too.

I have written some R code (Twitter1.R) to get you started on this task.

**Questions:** Now examine. What are the proportions of tweets in English? What proportion appear to come from US or Canadian time zones? Can you discover anything else interesting?