# Moody's 20191007

Please complete by October 18th, 2019. Provide code+output as part of your deliverable. Do not hesitate to reach out if you have questions.

Please read the data description in the Moody's 20191006 task.

Let's focus on two-column pages from the 1930 manual (as defined in `mastercolumns1930.csv`), and let's try to understand the page structure.

(a) Figure out how to split the words into left and right columns (each of which is meant to be read top to bottom). My algorithm is pretty fuzzy (uses quantiles instead of maxima, for example). For a given page, document when the different brightness settings give you "distinctly different" column zones/coordinates (the output from splitting the page into two columns will be something like the x-y coordinates of the two columns).

(b) Given the words in a column, classify the words into "lines." Starting with median word height is a good place where to calibrate lines. For a given page, document when the different brightness settings give you "distinctly different" lines (say just count the lines).

(c) Let's give the "swapping algorithm" a go. Settle on a line structure (a few obvious ways on how/what to choose from the previous step). Pick all word sequences within those coordinates (the ones defining the line in a given column), across all different brigthness. This is a set of 99 "guesses" as to what goes into that line. The algorithm I wrote this Spring would pick the "modal guess." Guess what that means and suggest what to pick. Document output in a sane way (it is not easy).