ECON 8320 – Tools for Data Analysis
Assignment 12 [25 points]

When attempting to model text, it is important to be able to extract features of the text for analysis. A powerful tool for extracting text features (as well as for searching, replacing, and performing other text-based tasks) is Regular Expression (regex).

For this assignment, your job is to process features of emails sent and received by Enron employees during the investigation into fraudulent activity by Enron executives in the early 2000's using regular expressions. Given emails from a single employee, generate the following information for each email:

1. Extract the text of the email into one feature
2. Create a column labelled "worry" containing the following values: 1 if some form of the word "worry" is contained in the text of the email, 0 otherwise
3. Create a column labelled "trouble" with the following values: 1 if some form of the word "trouble" is contained in the text of the email, 0 otherwise
4. Create a column containing the number of recipients of the email
5. Create a column labelled "sent" with the following values: 1 if the email was sent by the user, 0 otherwise

Submit your code, as well a csv containing the new columns, to Canvas when you complete this assignment.