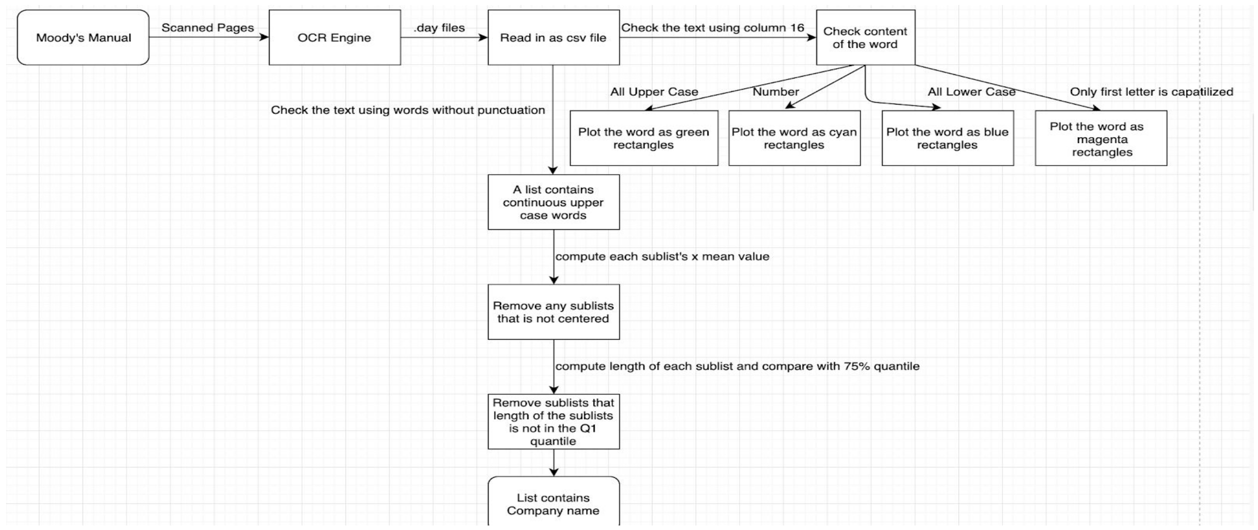
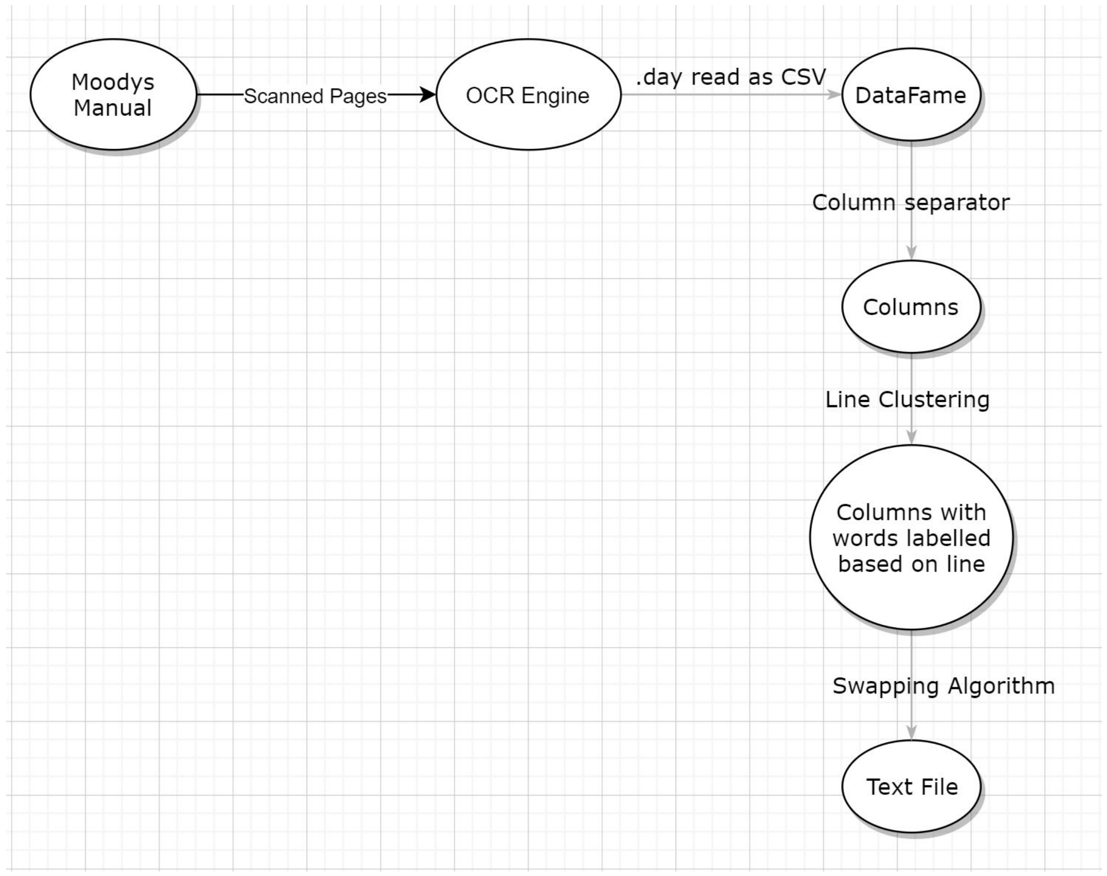
Task 1 Architecture


Moody's Manual
Scanned Pages
OCR Engine
.day files
Read in as csv file
Check the text using column 16
Check content of the word
All Upper Case
Number
All Lower Case
Only first letter is capatilized
Plot the word as green rectangles
Plot the word as cyan rectangles
Plot the word as blue rectangles
Plot the word as magenta rectangles
Check the text using words without punctuation
A list contains continuous upper case words
compute each sublist's x mean value
Remove any sublists that is not centered
compute length of each sublist and compare with 75% quantile
Remove sublists that length of the sublists is not in the Q1 quantile
List contains Company name

First we have the Moody's Manual where the OCR Engine is used to convert the Moody's Manual into .day files, we used python3 to read in .day files as csv file, then we check the content using column 16 which is words without punctuation, we plotted all words as rectangles and color the rectangles based on the content. If the content is all upper case, we plotted the word as green rectangles, if the content is all number, we plotted the content as cyan rectangles, if the content is all lower case words, we plotted the words as blue rectangles, if only the first letter of the content is capitalized, we plotted the words as magenta rectangles. Secondly, we grabs the word sequences that look like a company name where the company name will be all upper case words, in this case,we were using the column 16 which is the words without punctuation, we first checked for all continues upper case words, then compare each sublist x mean value which is the x location of each word in page with the overall page x location mean, we removed any sublists that is not centered next we check the remaining list's y location for its font, and compare the length of each sublists with overall length of the page's word with 75% quantile, remove sublist that length of the sublist which is not in the Q1 quartile afterward to get a list of word sequence that looks like a company name.

Task 2 Architecture



We are given pages from the Moodys Manuals, which are scanned in and input into our OCR engine. The .day file is a journal file, which we read as a CSV and input into a DataFrame for processing. From this DataFrame, we need to find the location of the columns and lines of the page. In order to find the columns, we look at the location of blank spaces where there is no text, and draw a line to indicate the location of the column.To find the lines, we perform clustering based on the y-axis value (the height) of the text, and again draw a line indicating the location of the lines. Finally, we use a swapping algorithm which selects the most accurate page and swaps that in, which we then output as a text file as the final product.