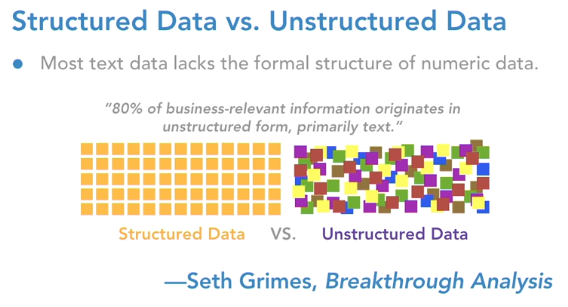
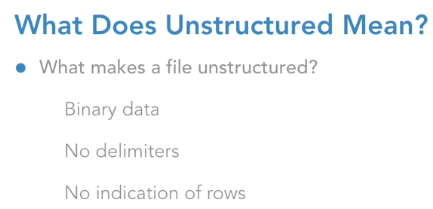
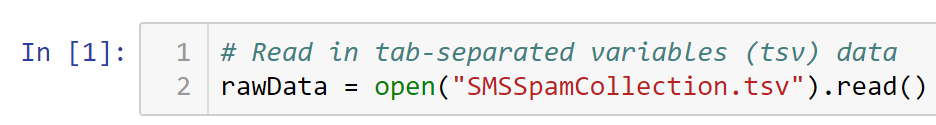
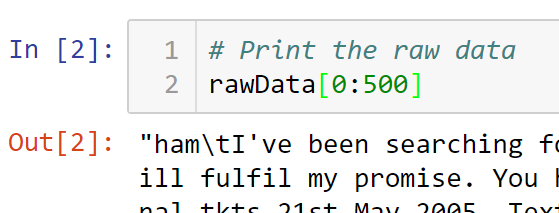
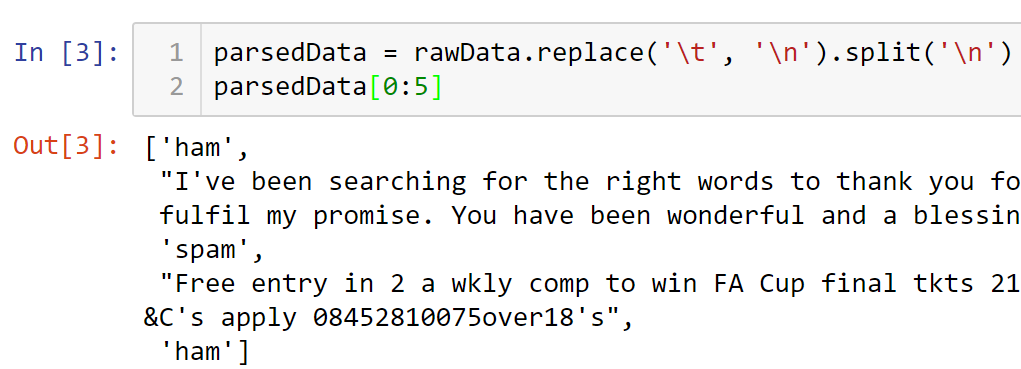
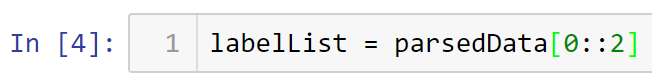
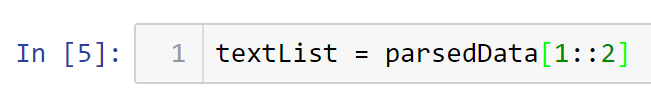
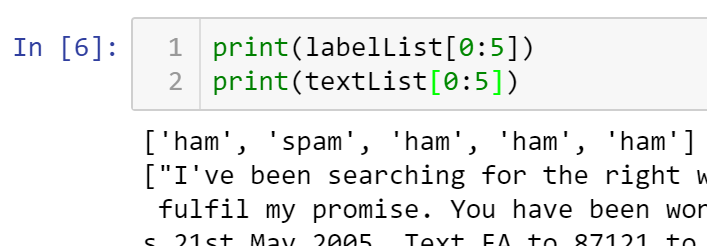
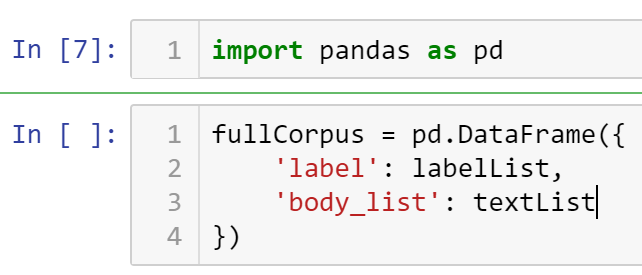
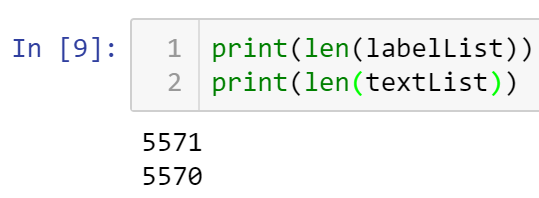
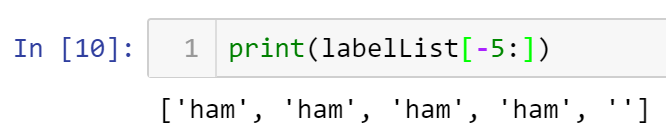
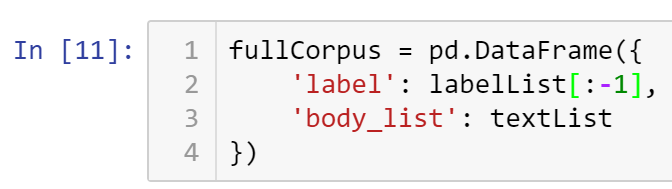
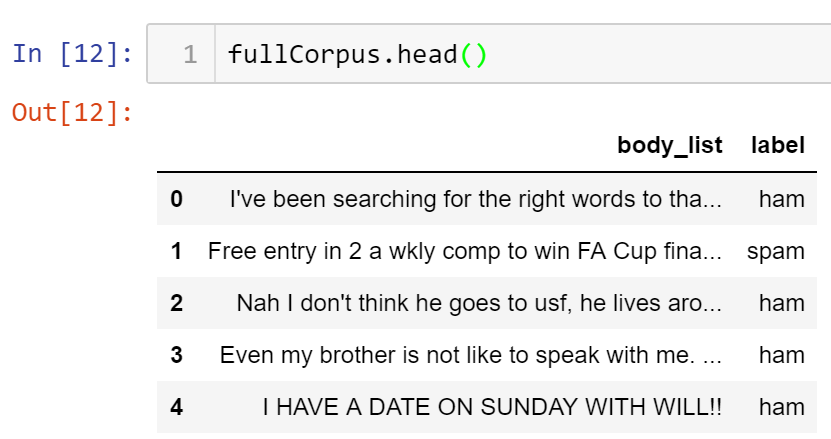
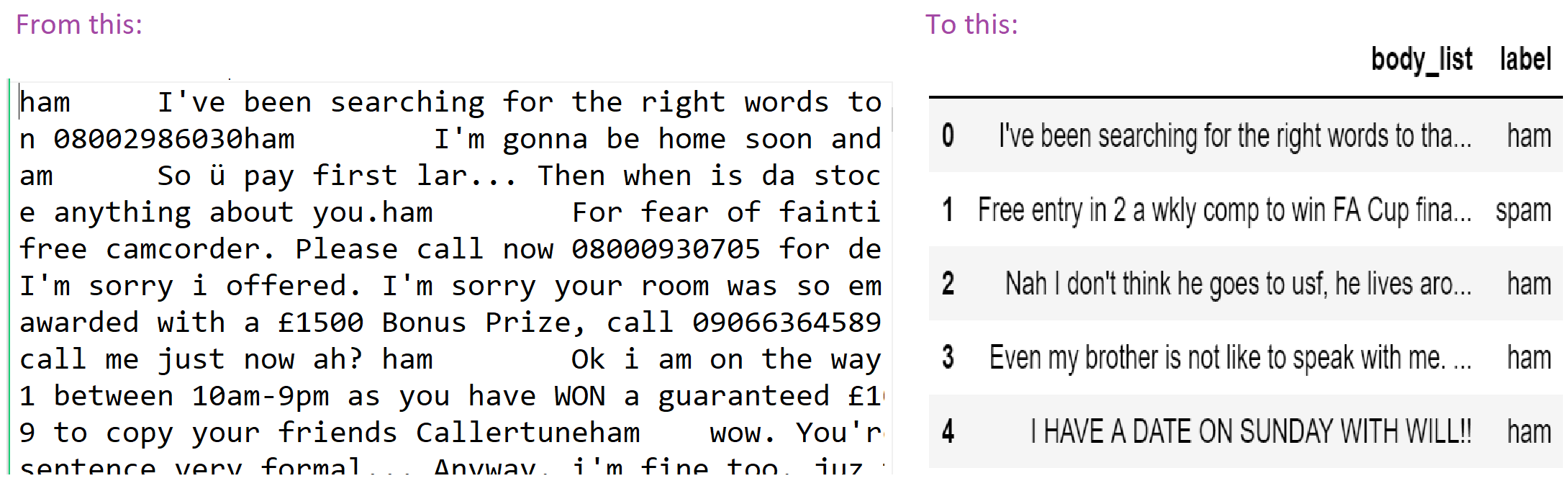
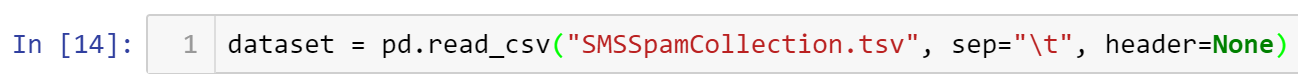
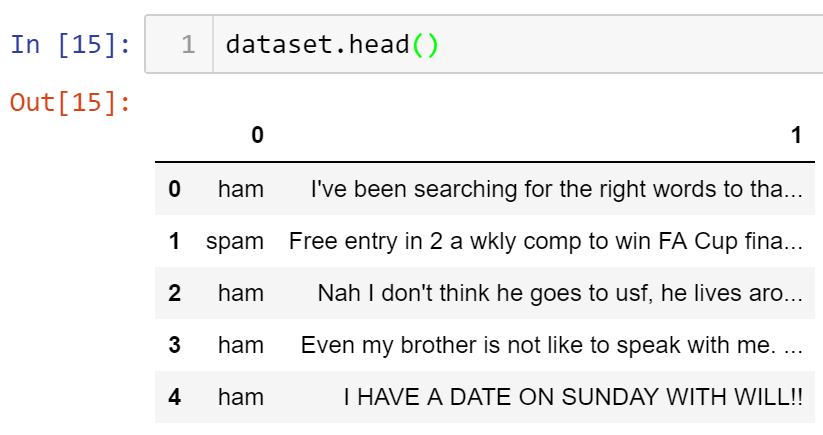
*Start a new Python project folder … do not reuse the previous workspace … As with any keyboard-driven console-like environment, developing muscle -memory for the common commands is also part of the learning curve.*





* Processing unstructured data
* 
* Print out the first 500 characters
* 
* View the “.tsv” file in a notepad, and note how unstructured it is.
* Note the \t and \n separators. \t = tab, \n=new line, or carriage return line feed
* There are many ways to tackle unstructured data. One way is to replace \t with \n, then split it into a list.
* 
* Printing the first five in the list parsedData, showed alternating label and associated value.
* So, what we’ve just done is to create some structure to the original unstructured data.
* Next is to create another list where we store every other item from the parsedData list. Starting from position zero to the end (no ending number), and every other item as indicated by 2.
* 
* We’ll do the same with text list so we can get the text body in a different list. We will start at position 1 which is the second element of the list.
* 
* Verifying the contents of our two new lists
* 
* Now, it is time to combine both so we can do more analysis
* One way of doing this is to combine both list into a dictionary
* 
* If you run this code now, you will get an error because the list have different number of rows
* Checking the length of each list
* 
* Note that labelList has one more extra entry
* The inference is that it may have pick up something at the last entry
* List display the last few entries … top five backwards ...
* 
* Note that the very last entry is empty; that is causing an extra count.
* Drop that entry and you will have same size lists.
* So, one way of dropping the last entry is to capture all from the beginning except the last one (indicated as -1). Run the code below and the error is gone.
* 
* Printing the first five entries of fullCorpus dictionary …
* 
* So now you have a nice clean version (structured) compared to what we have in the beginning (unstructured).
* 
* The process that you just went through is the manual and winded way.
* There is a shortcut you can use after you understand the manual process.
* The moment you see in a file \t, that is your indicator that you can read the file as a tab-separated variable, or tsv way. You will use the read\_csv() function, but you will declare that the separator is a ‘\t’.
* Here is the shortcut code:
* 
* Also, declare that there is no header because the raw data set don’t have column names. If you don’t add this parameter, it will take the first entry and assume that that is the header.
* 
* Printing the dataset
* 
* As you can see, the output above is very close to the manual operation you did earlier. The only difference is the column titles; here it is “0” and “1” while the previous is “body\_list” and “label”. Also, the order of the column is different.
* All submissions should be separate from other exercises and quests. Please do not lump all your answers into one document and re-using that same workspace to gain multiple points. Thanks.
* Place your name at the bottom of your code, download your Python program in html format, and submit your work in Canvas.