SUMMARY

The algorithm is divided into two python files mapper.py and reducer.py.

In the mapper.py python file, it reads the input Comments.xml file line by line. At the beginning of the loop, we create an empty list “index”. In the first for loop, we are appending the indices of double inverted commas to extract each element. Then in the next for loop, we are taking a pair of indices and checking that whether slicing the line in that range of index contains “Text” or not, if it contains “Text” then we will select the next two pairs of indices to extract the value of “Text”, and if in the next iteration of the 2nd for loop we found “CreationDate” then we extract the value of “CreationDate” and set as the value. For piping between mapper and reducer, we had printed the key and values pairs separated by single tabs, before doing this we removed the unnecessary characters from the key using the strip command. here the key is “CreationDate” and values as the length of “Text”.

In the reducer.py file, we used sys module and prettytable module to plot the table in a systematic order and defined a function named median\_calc to calculate the median in a given list. Before reading the outputs from the mapper we will initialize some variables and created one empty list median\_list to store the length of “Text” in each month of each year and created a table object using Prettytable function with two columns “Dates” and “median per months”. Then we will read the lines and split them based on the tab and store the values in the word, and length variables. And on the following code, we counted the total number of dates and appended each row of dates and their medians. In the end, plotted the table.