SUMMARY

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

In the mapper.py python file it reads the input Posts.xml file line by line, 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 “Tags” or not, if it contains “Tags” then we will select the next two pairs of indices to extract the content of “Tags”, and slice that line in that range of the index. Since Tags may have multiple Tags so we split them based on “;&lt;”. For piping between mapper and reducer, we had printed the key and value pair using another for loop to iterate upon the list formed by the split command and before doing this we removed the unnecessary characters from the key using the strip command, and then print the key spaced by a single tab and value as 1.

In the reducer.py file, we used sys and prettytable module to plot the table in a systematic order. Before reading the outputs from the mapper we will initialize some variables and define the format of the table using the Prettytable function. Then we will read the lines and split them based on the tab and store the values in the word and count variable. And on the following code, we count the number of different badges and the number of users of each badge and append them into the table then at the end print the table.