Group 1

Task 3

For task 3, first we created a master and an app, local and My App respecify by using spark configuration that allow us to do so. Then we saved these configurations in a variable named sc. then using stop words we listed all the stop words such as punctuation marks and digits.

Then by using if function we saved the data line by line in a temporary file named as tem. What we need to do is that: first, we need to determine which category this news is in. So here, I check if this line is URL and if so, we just split it by "\" and save its category. And if not, we assume it is the review line and we saved it as message which only includes low case letters. Then we paired the data in class and message. And by using remover function from spark library, we removed the stop words and got output which is saved as message clean. Then we paired class and our new message clean that we got after removing the stop words and counted the words in that new data set using count and grouped it as class and word. Then we separated the data by class and count and finally ranked the words and filtered the top 10 most words, which save in a csv file as our final result.