In this assignment the objective is to write a spark code that will take an SMS spam dataset (link: https://www.kaggle.com/uciml/sms-spam-collection-dataset) and will perform the following two things. You must submit two separate code files for the two problems below.

i) It will produce word co-occurrence statistics from spam and non-spam SMS in two separate files. The output would be a three column file of the form <w1 w2 count>, where w1 and w2 are lexicographically ordered and the count is the number of SMSes that contains both w1 and w2. You must not include the stopwords. The stopwords list can be found here: https://gist.github.com/sebleier/554280

ii) Given a word, produce the most frequently occurring five words from the spam and non-spam SMSes.

You must submit the code files as well as the output files. For the problem 2, use the following words: bitter, candid and super.

IMPORTANT:

Put all your files in a folder, zip it and upload it in moodle. Submit exactly one file for each group. The filename must be in this format: bdp-assignment-1-rollno.zip, where the rollno denotes the roll no of the group leader.

Deadline: 05.04.22, 11.55 pm (Indian time).