CS 6350 - ASSIGNMENT 1b

Please read the instructions below before starting the assignment.

- This assignment consists of two parts. Please create separate folders named **parti** and **partii** and zip them together for submission.

- You should use a cover sheet, which can be downloaded at: http://www.utdallas.edu/~axn112530/cs6350/CS6350_CoverPage.docx

- You are allowed to work in pairs i.e. a group of two students is allowed. Please write the names of the group members on the cover page. Only one submission per team is required.

- The deadline for this assignment is Wednesday September 21 at 11:59 PM. No extensions are allowed.

- You have a total of 4 free late days for the entire semester. You can use at most 2 days for any one assignment. After that, there will be a penalty of 10% for each late day. The submission for this assignment will be closed 2 days after the due date.

- Please ask all questions on Piazza, and not through email to the instructor or TA.

ASSIGNMENT 1b

Part 1 (25 points)

In the first part you uploaded 6 large files on HDFS. In this part, you will run WordCount algorithm on them using MapReduce. You will need to remove stop words before running the algorithm. You are free to use any reasonable list of stop words that you can find online. Be sure to mention the source in the readme file.

Deliverable:

Only one zipped file containing your source code for the project. It should be turned in through eLearning only.
Readme file containing the source of the stop words that you used.

Part 2 (25 points)

In the first part, you downloaded 6 files containing tweets. In this part, you will run a WordCount program on them using MapReduce.. The MapReduce program will look for Hashtags i.e. those words starting with "#" and output the count of such Hashtags.

What to submit:
Only one zipped file containing your source code for the project. It should be turned in through eLearning only.