**Project#2**

**(Total Points : 100)**

**Due Date: Dec, 04 2018**

This data was collected from the City of New York’s data website, and contains all reports of vehicular incidents in New York City over a period of time. The file is roughly 303 MB in size, and contains over 1.38M records. Each record have 29 columns (more details about the data is available on the website).

Data Link: <https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95>

You can see the sample of the data from the link above.

Do the following using Map-Reduce and Spark. You can use any language of your choice (**if you are using any language other than Java/Python, please provide instructions on how to run your code**)

**Data Location on cluster:** /user/data/nypd/

**File Name (With Header):** NYPD\_Motor\_Vehicle\_WithHeader.txt

**Use this file as input (With Header):** /user/data/nypd/NYPD\_Motor\_Vehicle\_WithHeader.txt

**File Name (Without Header):** NYPD\_Motor\_Vehicle\_WithOutHeader.txt

**Use this file as input (Without Header):** /user/data/nypd/NYPD\_Motor\_Vehicle\_WithOutHeader.txt

First row of data file (NYPD\_Motor\_Vehicle\_WithHeader.txt) is the column name and **start with #** (total of 29, use link above to look at the data).

**Task#1:** Clean the input data

1. Keep following columns in the cleaned data
   1. Date
   2. Borough
   3. Zip
   4. Number Of Persons Injured
   5. Number Of Persons Killed
   6. Number Of Pedestrians Injured
   7. Number Of Pedestrians Killed
   8. Number Of Cyclist Injured
   9. Number Of Cyclist Killed
   10. Number Of Motorist Injured
   11. Number Of Motorist Killed
   12. Vehicle Type Code 1
2. Remove any row if value of any of the above mentioned column is missing (**0 is allowed**)

**Task#2:** Process the cleaned data for following information

1. Date on which maximum number of accidents took place.
2. Borough with maximum count of accident fatality
3. Zip with maximum count of accident fatality
4. Which vehicle type is involved in maximum accidents
5. Year in which maximum Number Of Persons and Pedestrians Injured
6. Year in which maximum Number Of Persons and Pedestrians Killed
7. Year in which maximum Number Of Cyclist Injured and Killed (combined)
8. Year in which maximum Number Of Motorist Injured and Killed (combined)

**Task#3:** Write code to process output file from **Task#2** and generate final output file with all the required numbers from **Task#2**

**Submission must include:**

1. Code used for cleaning the input data (either Map-Reduce or Spark)
2. Map-Reduce code along with jar file (if you are using java). Maximum number of files allowed is 3 (**code must be properly commented**)
3. Spark code (maximum number of files 3) (**code must be properly commented**)
4. Code used for **Task#3**
5. Text file with all 8 result from **Task#3**
6. Readme file with description of functionality for each of the above files

**Extra Credit : 15**

1. If you submit 1 file of Map-Reduce with all the code (+15).
2. If you submit 2 file of Map-Reduce with all the code (+5).

**Hints**:

All of the above problems are based on word count discussed in the class

**Example**: for finding Zip with maximum count of accident fatality

Use zip as key and number of accidental death (***sum of Persons Killed, Pedestrians Killed, Cyclist Killed and Motorist Killed***) as value in mapper and in reducer add all values for given zip.