Twitter data questions: ([schema](https://github.com/episod/twitter-api-fields-as-crowdsourced/wiki)) (hdfs://data/twitter)

Problem Statement:

1. What hour of the day does @PrezOno’s tweet the most on average, using every day we have twitter data?  Include a plot of the expected number of tweets for each hour of the day, for those he did tweet.  For example if Ono tweeted once every day at 12:30PM, his expected number of tweets between 12 and 1 would be 1.  If he alternates between 2 and 3 tweets per day, his average would be 2.5.
2. What day of the week does @PrezOno tweet the most on average?  Use the same example as in #1 but for days of the week.

**Solution to 1st One.**

**Approach:**

I chose streaming-mode in hadoop. I wrote Map and Reduce jobs in python. mapOno.py and reduceOne.py are the map and reduce python files.

Mapper explanation:

Mapper reads each tweet from HDFS file system and checks whether it has tweet from PrezOno. If it has PrezOno it emits the time, date he tweeted and one.

Tweets are JSON strings. When a tweet with screen name Prezone is detected, complete tweet created time JSON string value is sent as key and ‘1’ as value.

*Mapper output <time, 1>*

Reducer explanation:

Reducer gets tweet ‘created time’ and ‘1’ as input. Created time has information about date, day and time.

Date field is used to calculate number of days in reducer. Time field is used to calculate number of tweets for each hour. Reducer adds the number of one’s for each hour and determines the total number of tweets for each hour, then it divides with number of days. Reducer output will be average number of tweets for each hour.

*Sample Reducer output <time, Average nu>*

Reducer output for first 3 hours:

0 0.0732984293194

1 0.0994764397906

2 0.0837696335079

3 0.109947643979

Plot for output data: At 5PM (17:00) on average there are more number of tweets.

2).

**Approach:**

I chose streaming-mode. I wrote Map and Reduce jobs in python. mapDay.py and reduceDay.py are the map and reduce python files.

Mapper explanation:

Mapper reads each tweet from HDFS file system and checks whether it has tweet from PrezOno. If it has PrezOno it emits the time (includes date, day and time) he tweeted and one.

Tweets are JSON strings. When a tweet with screen name Prezone is detected, complete tweet created time JSON string value is sent as key and ‘1’ as value.

*Mapper output < {time}, 1>*

Reducer explanation:

Reducer gets ‘Day (along with time)’ and ‘1’ as input. It adds the number of one’s for each day and determines the total number of tweets for each day, then it divides with number of days.

Number of days are calculated by using date and month fields.Reducer output will be average number of tweets for each day

*Reducer output <Day, average number of tweets>*

Reducer output: (Output is present in Output folder of github)

*Wed 0.287958115183*

*Sun 0.30890052356*

*Thu 0.282722513089*

*Tue 0.17277486911*

*Mon 0.251308900524*

*Fri 0.204188481675*

*Sat 0.277486910995*

**Plot based on reducer output.**

On Sunday President Ono tweeted most on average.