**Twitter Questions 1 and 2:**

**Required data to answer the questions:**

1. Number of days in data.
2. Number of days on which PrezOno twitted.
3. Number of tweets per each hour.
4. Number of different week days (example: number of Mondays in data, number of Tuesdays in data etc.)
5. Number of different week days on which PrezOno Tweeted. (Example: Number of Wednesdays on which he tweeted.)
6. Number of tweets on each week day.

From the above requirements it is clear that if we calculate Number of different week days and sum them all we get number of days. Therefore it is enough to calculate 4 to get 1 in above requirements.

Moreover, if we calculate number of different week days on which PrezOno tweeted and sum all of them we get number of days on which PrezOno tweeted. Calculating 5 gives 2.

By solving two questions together we can verify results comparing sum of 3 to sum of 6 (Total tweets by PrezOno.)

**Approach:**

We used Hadoop streaming method to solve these questions. The steps followed in mapper are as below:

1. Mapper parses each input line using json library and extracts the ScreenName, Created\_at. Created\_at is used to calculate week day (example: Wed), date (example: 24Sep2014), hour (example: 21).
2. The dates are added to a set. This set is a value in dictionary where key is week day. Since the set contains only unique values we get the unique values for each week day.

**Example**: {'mon': set(['22Sep2014', '29Sep2014']), 'wed': set(['24Sep2014', '01Oct2014'])}

1. If the user is PrezOno we do the same thing as step two to another map. This map contains the week days on which PrezOno tweeted. The key value pair (weekday+’Tweets’,1) and (hour+’Tweets’,1). Sum of these values for each key from all the mappers gives us 3 and 6 in requirements.
2. After all the lines in input to a mapper are exhausted we print all the dictitems in maps in step 2 and 3 as key ‘x’+weekday and value as the corresponding set. We prepend with ‘x’ to differentiate the key value pair containing sets to decimals.
3. Reducer add all the values corresponding to a key if the values are decimals and output the key value pairs or performs an union operation on sets when the values are sets and outputs key as key and value as size of the set.
4. We run this with single reducer and no combiner is used as we are doing much work in mapper.

**Results: Question1**

**Total Number of days in data**: 313

**Total Number of days on which PrezOno twitted**: 192

**Tweets per each hour**:

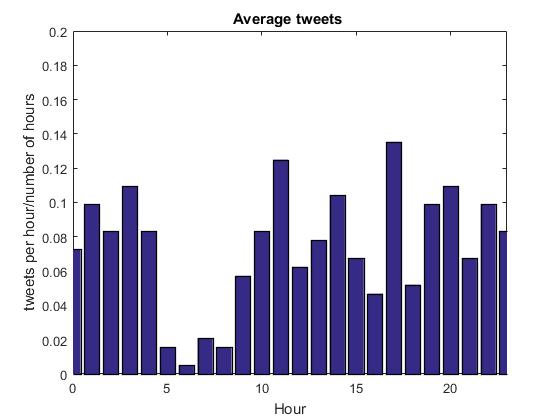
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **00** | **01** | **02** | **03** | **04** | **05** | **06** | **07** | **08** | **09** |
| 14 | 19 | 16 | 21 | 16 | 3 | 1 | 4 | 3 | 11 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** |
| 16 | 24 | 12 | 15 | 20 | 13 | 9 | 26 | 10 | 19 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **20** | **21** | **22** | **23** |  |  |  |  |  |  |
| 21 | 13 | 19 | 16 |  |  |  |  |  |  |

**So he tweets most on average at 1700-1800 or 5pm to 6pm**

Plot of the expected number of tweets for each hour of the day, for those he did tweet is as shown below: X-axis has hours and Y-axis has Average (Number of tweets in hour/Number of hours).



**Results: Question2**

**Number of week days on which PrezOno tweeted:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mon** | **Tue** | **Wed** | **Thu** | **Fri** | **Sat** | **Sun** |
| 28 | 20 | 31 | 28 | 23 | 31 | 31 |

**Total number of week days:**

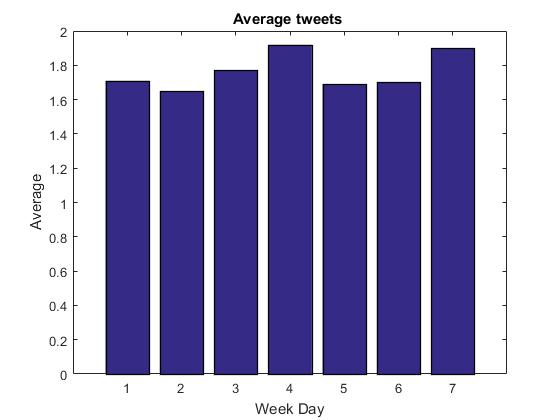
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mon** | **Tue** | **Wed** | **Thu** | **Fri** | **Sat** | **Sun** |
| 46 | 46 | 43 | 46 | 42 | 44 | 46 |

**Number of tweets per week day:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mon** | **Tue** | **Wed** | **Thu** | **Fri** | **Sat** | **Sun** |
| 48 | 33 | 55 | 54 | 39 | 53 | 59 |

**Average Tweets per week day:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mon** | **Tue** | **Wed** | **Thu** | **Fri** | **Sat** | **Sun** |
| 1.71 | 1.65 | 1.77 | 1.92 | 1.69 | 1.70 | 1.90 |

****