|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| Big Data Processing  Course Work 1 – Twitter Analysis with Map Reduce |
| |  |  |  | | --- | --- | --- | | Sarang Kharche - 150740395 | 11/9/17 | Big Data Processing | |

Table of Contents

[PART A. MESSAGE LENGTH ANALYSIS 2](#_Toc498078473)

Question 1

Answer

[PART B. TIME ANALYSIS 3](#_Toc498078474)

Question 1

Answer

Question 2

Answer

[PART C. SUPPORT ANALYSIS 5](#_Toc498078475)

Question 1

Answer

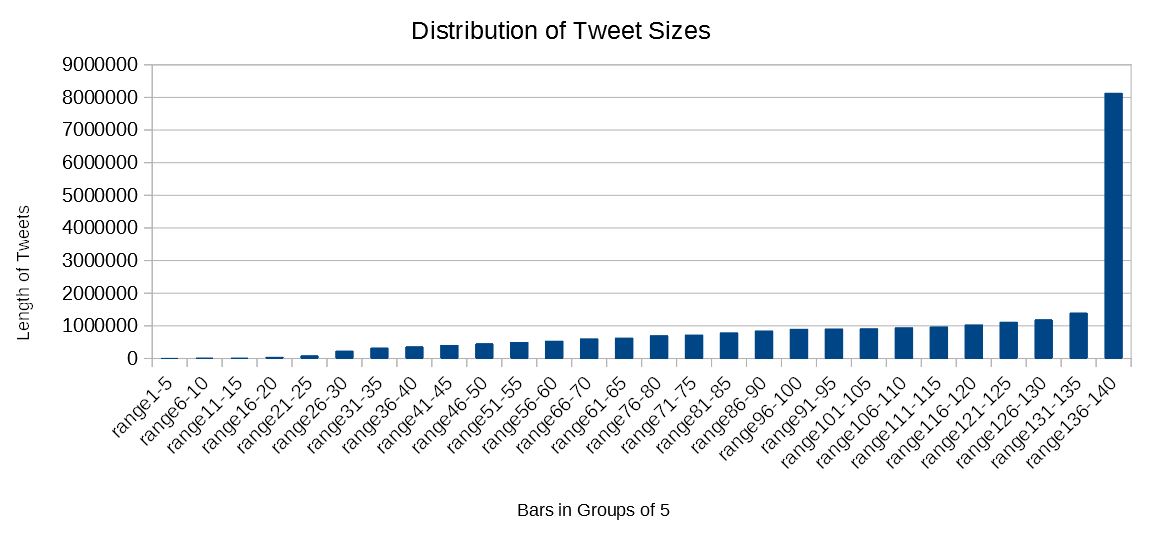
Question 2

Answer

# **PART A. MESSAGE LENGTH ANALYSIS**

1. Create a Histogram plot that depicts the distribution of tweet sizes (measured in number of characters) among the Twitter dataset. To make the data more readable, the histogram must aggregate bars in groups of 5 (that is, first bar counts tweets of length 1-5, second bar counts tweets 6-10, and so on) as part of your MapReduce job. Your MapReduce program must compute the histogram bins for a correct solution. Aggregating bins outside MapReduce will deduct marks from the complete grade.

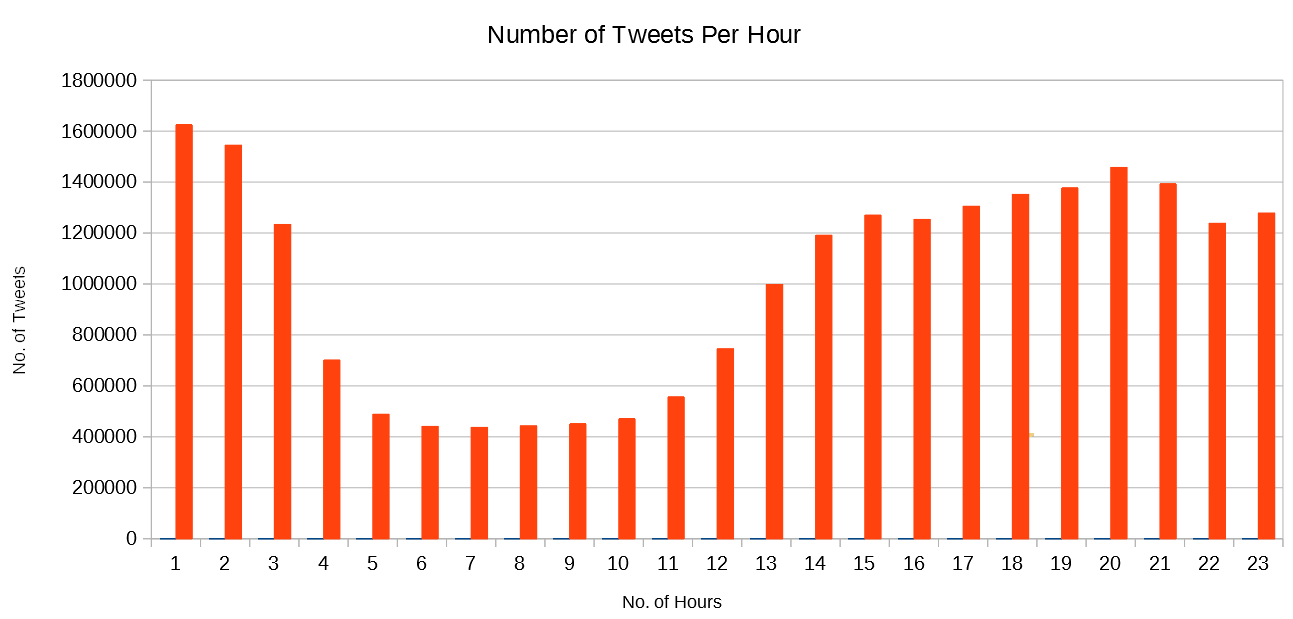
**Answer:**



# **PART B. TIME ANALYSIS**

1. Create a bar plot showing the number of Tweets that were posted each hour of the event. You should aggregate together all the messages emitted at the same hour, regardless of the day the messages were sent (hence, you will have 24 different groups). When checking the correctness of your results, keep in mind the time zone of the 2016 Olympic games, as that should give you base expectations about the prime time when the main activities occurred.

**Answer:**



1. For the most popular hour of the games, compute the top 10 hashtags that were emitted during that hour. Hashtags are words contained inside the tweet, starting with the hashcode (#) character. Does that information provide you any hint on the main events/activities that took place at peak time?

**Answer:**

|  |  |  |
| --- | --- | --- |
| 1 | #Rio2016 | 1320441 |
| 2 | #Olympics | 76598 |
| 3 | #rio2016 | 71493 |
| 4 | #Futebol | 40195 |
| 5 | #USA | 39336 |
| 6 | #Gold | 37323 |
| 7 | #BRA | 36630 |
| 8 | #CerimoniaDeAbertura | 33610 |
| 9 | #OpeningCeremony | 33493 |
| 10 | #TeamUSA | 30216 |

# **PART C. SUPPORT ANALYSIS**

1. Draw a table with the top 30 athletes in number of mentions across the dataset. For each athlete, include the number of mentions retrieved. For this question you can sort results and compute the top X outside your MapReduce code.

**Answer:**

|  |  |  |
| --- | --- | --- |
| 1 | Michael Phelps | 182398 |
| 2 | Usain Bolt | 171341 |
| 3 | Neymar | 100980 |
| 4 | Simone Biles | 80635 |
| 5 | William | 53406 |
| 6 | Ryan Lochte | 41025 |
| 7 | Katie Ledecky | 38062 |
| 8 | Yulimar Rojas | 34458 |
| 9 | Simone Manuel | 27381 |
| 10 | Joseph Schooling | 26521 |
| 11 | Michael Phelps | 182399 |
| 12 | Usain Bolt | 171342 |
| 13 | Neymar | 100981 |
| 14 | Simone Biles | 80636 |
| 15 | William | 53407 |
| 16 | Ryan Lochte | 41026 |
| 17 | Katie Ledecky | 38063 |
| 18 | Yulimar Rojas | 34459 |
| 19 | Simone Manuel | 27382 |
| 20 | Joseph Schooling | 26522 |
| 21 | Michael Phelps | 182400 |
| 22 | Usain Bolt | 171343 |
| 23 | Neymar | 100982 |
| 24 | Simone Biles | 80637 |
| 25 | William | 53408 |
| 26 | Ryan Lochte | 41027 |
| 27 | Katie Ledecky | 38064 |
| 28 | Yulimar Rojas | 34460 |
| 29 | Simone Manuel | 27383 |
| 30 | Joseph Schooling | 26523 |

1. Draw a table with the top 20 sports according to the mentions of Olympic athletes captured. For resolving athletes into sports use the medalistsrio secondary dataset. For this question you can sort results and compute the top X outside your MapReduce code**.**

**Answer:**

|  |  |  |
| --- | --- | --- |
| 1 | athletics | 460895 |
| 2 | aquatics | 453573 |
| 3 | football | 209182 |
| 4 | gymnastics | 129818 |
| 5 | judo | 97574 |
| 6 | tennis | 81522 |
| 7 | basketball | 73358 |
| 8 | cycling | 67180 |
| 9 | badminton | 61038 |
| 10 | wrestling | 34021 |
| 11 | weightlifting | 23990 |
| 12 | sailing | 23978 |
| 13 | canoe | 23975 |
| 14 | shooting | 23950 |
| 15 | equestrian | 23231 |
| 16 | boxing | 23205 |
| 17 | rowing | 17465 |
| 18 | volleyball | 17239 |
| 19 | taekwondo | 15997 |
| 20 | fencing | 12756 |