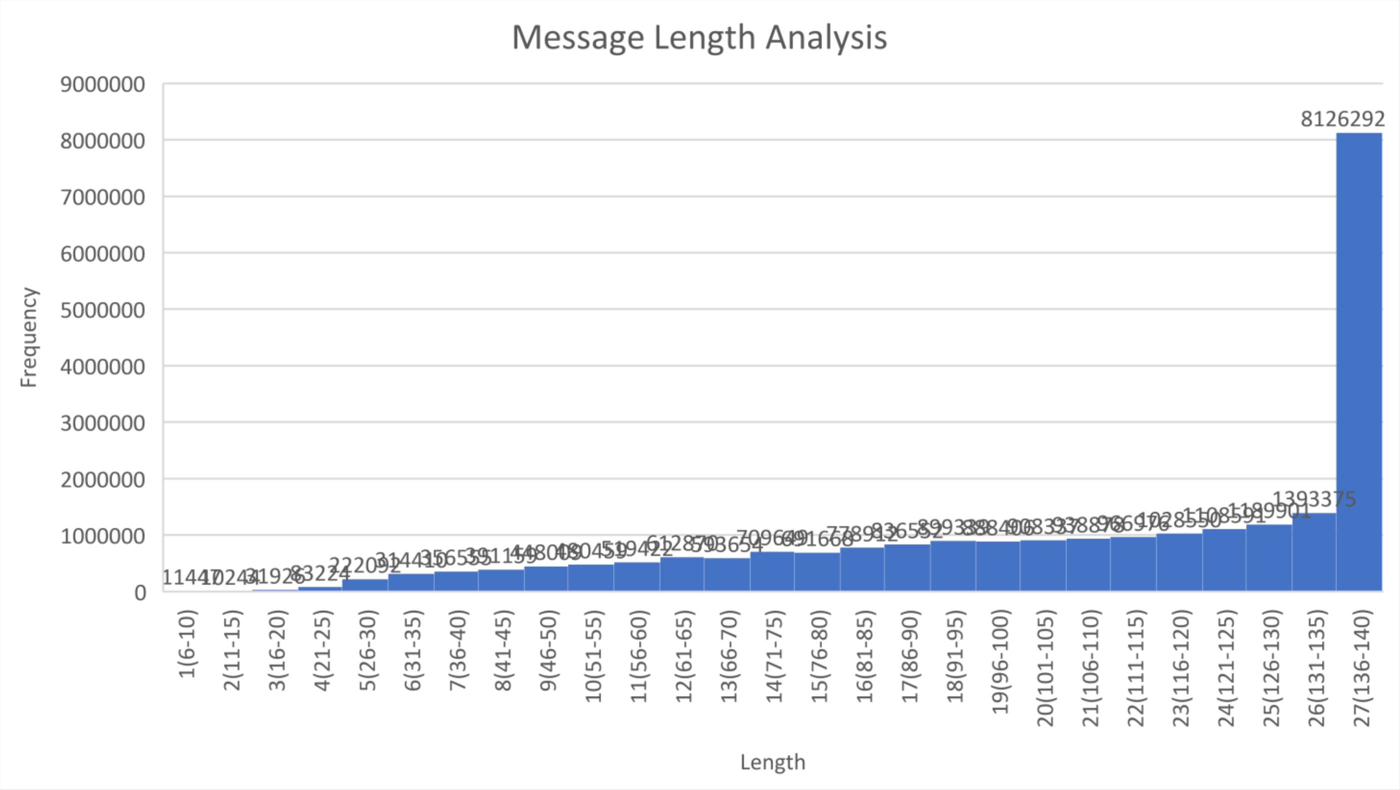
ECS640U-BIG DATA PROCESSING Assignment 1

PARTA MESSAGE LENGTH ANALYSIS (35%)

There are 4 categories. I divided it in 4 lines/fields by ";" and picking the 3rd category which is line [2] by using **`String[] line=value.toString().split(";");if(line.length== 4){String Box = line[2];`**. Then I am limiting that the length of the tweet in 140 and divided it into the range of 5 by using **`if( ( Box.length() > 0 ) & (Box.length() <= 140)){int rangenumber = (int)( ( Box.length() - 1 ) /5);`**. Therefore we have 28 group but we do not have any tweet length that between bin 0 (1 to 5). **`sum =value.get()`** to get the Length count. I also set up **`int rangestart = key.get()*5 +1`** and **`int rangeend = rangestart + 4;`** they will create the range of each bin, that would make it easier to read.

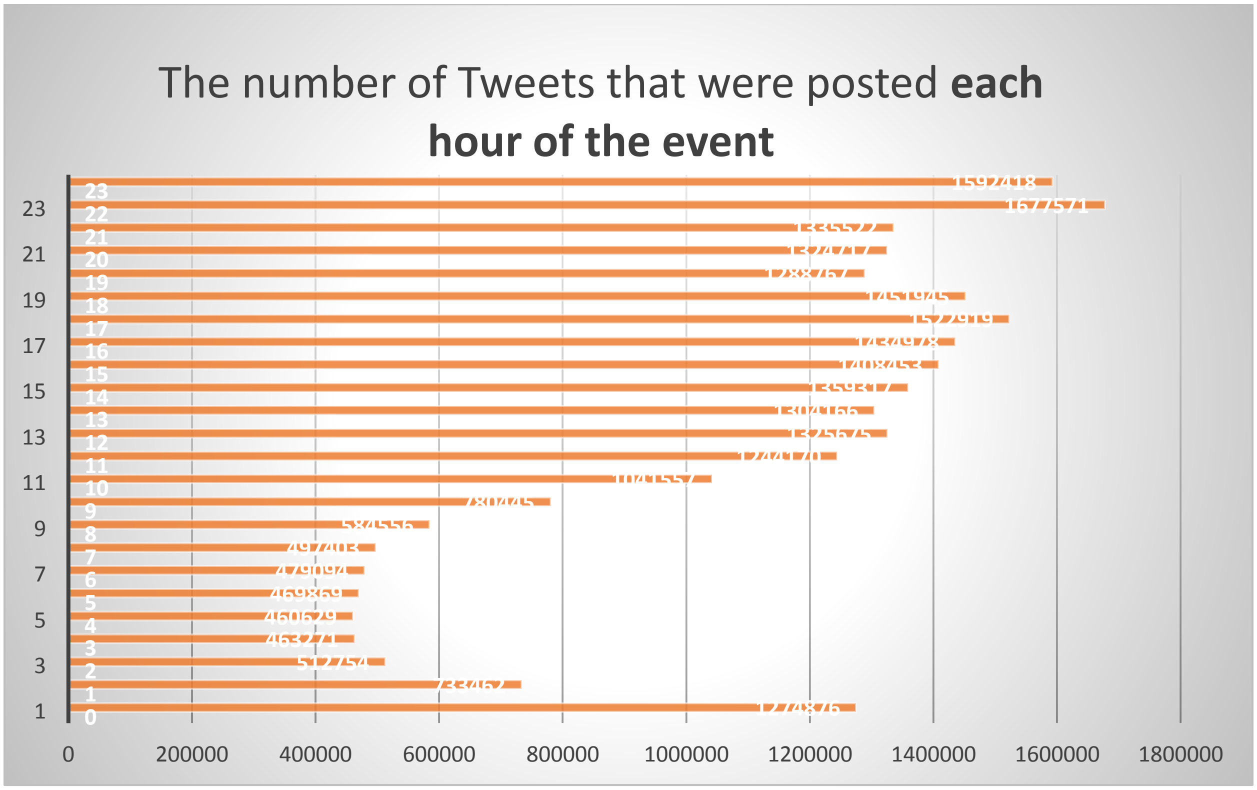
```
1(6-10)        11447
2(11-15) 10244
3(16-20) 31926
4(21-25) 83224
5(26-30) 222092
6(31-35) 314410
7(36-40) 356555
8(41-45) 391159
9(46-50) 448009
10(51-55)    480459
11(56-60)    519422
12(61-65)    612870
13(66-70)    593654
14(71-75)    709649
15(76-80)    691668
16(81-85)    778912
17(86-90)    836552
18(91-95)    899339
19(96-100)   888406
20(101-105)  908337
21(106-110)  938878
22(111-115)  966976
23(116-120)  1028550
24(121-125)  1108591
25(126-130)  1189901
26(131-135)  1393375
27(136-140)  8126292
```



The Histogram plot has clearly indicated that the length 6 to 140, and it is divided by 5. The frequency of the length that between ranges of 136-140 is the highest, it is 8,126,292.

PartB TIME ANALYSIS (45%)

B1)This part is similar to part a, I am picking the **line[0]** since we are now more interested in the time. The difficult is to change it the time into human-readable. I used try and catch to ignore any error. By using `Date date = new Date(hours); DateFormat format = new SimpleDateFormat("HH");` to tell the hadoop we only need hours, and set the time zone to GMT-3 by `format.setTimeZone(TimeZone.getTimeZone("GMT-3"));`. We have now the time for human time zone GMT-3.



We know that the most number of Tweets that would appear at 22 at GMT-3. Although it is winter GMT-2, it is easily to compare by using summer time.

PartB2

B2) In this part, I found this there might be a little different if using the capital letter. I have decided to generalize it as one word, since there is no point to have 2 words having the same meaning. #rio2016 are the word that people were using the most.

By using that `if(token.startsWith("#")) word.set(token.toLowerCase()); all` hashtag words will be changed to lower case.

Words	Appearances
#rio2016	1404369
#olympics	88160
#gold	64916
#bra	48760
#futebol	47942
#usa	41693
#oro	38457
#cerimoniadeabertura	36141
#swimming	35841
#openingceremony	34870

Without any debut, rio2016 should have the most Appearances. In addition, swimming, futebol(football) and opening ceremony were asking place at peak time but different data. Since USA and gold were appeared quite a lot. We can conclude that ''USA earned a lot of gold medal in swimming.'' And it was early time in USA, It can be summed up that twitter is quite popular in USA. They would like to support the sport that their county good at which is swimming.

PARTC SUPPORT ANALYSIS (20%)

C1)This part does not require a reducer. We have 2 data set this part. We split the .csv by ”,”. We have 11 categories, picking line [1],[7]. And using the same code for part a and part b to convert them to lowercase, limit it in the length 140 and line[2]. **IntIntpair** to pair both input. Then I converted it back to as it was written.

Michael Phelps	187918
Neymar	173240
Usain bolt	171776
Simon Biles	79775
William	61637
Ryan Lochte	41276
Katie Ledecky	39576
Yulimar Rojas	34748
Rafaela Silva	25830
Joseph Schooling	25816
Sakshi Malik	25122
Simon Manuel	24268
Andy Murray	21612
Wayde Van Niekerk	21551
Kevin Durant	21193
Tontowi Ahmad	20911
Liliyana Natsir	20257
Andre de Grasse	17921
Penny Oleksiak	17795
Monica Puig	17484
Rafael Nadal	16138
Laura Trott	15632
Ruth Beitia	14475
Lilly king	14185
Teddy Riner	14079
Luan	13796
Shauna miller	12243
Jason Kenny	11995
Caster Semenya	11153
Allyson Felix	11114

This table has shown the top 30 athletes that was mentioned across the dataset. We should not be surprised to see Michael Phelps, Neymar and Usain Bolt to be top 3 since it is Michael Phelps and Usain Bolt last Olympic. Neymar was a young-football-superstar during this time. Since football is the most popular ball game. The football fan from whole world was watching how good could have done.

PARTC2

C2) To get top 20 sports to the mentions of Olympic athletes captured. We add this line **medalist.set(medalistInfo.get(keyMedalistName));** to the mapper.

athletics	477763
aquatics	458374
football	300297
gymnastics	134320
judo	102904
tennis	84621
basketball	74743
cycling	69062
badminton	63158
wrestling	35101
weightlifting	23577
sailing	23405
canoe	23219
shooting	23099
equestrian	22628
boxing	20963
volleyball	17614
rowing	16348
taekwondo	15533
fencing	12887

The table has shown that top 20 sports according to the mentions of Olympic athletes captured. Since Michael Phelps, Neymar and Usain Bolt, we should not be surprise that athletics, aquatics and football are the top 3 sports. This also proved that when people mentioned these 3 sportsman, they are likely to mention these 3 sports.