BDA ASSIGNMENT -4

NAME---- PRIYANKA SINGH
ROLL NO. ---- 2018174

NAME -------- KIRAN SETHI
ROLL NO.---- MT20211

METHODOLOGY

APPROACH AND REASON:

The assignment aimed at making the use of two algorithms Jagdish et als. And Guha's algorithm to create a V-optimal histogram.

For the offline setting we crawled tweets regarding covid and covid vaccines we extracted the most popular hashtags and stored them in a csv file.

We then implemented the Jagdish et als algorithm for offline setting . different hashtags were mapped to integer values and then used as x-axis values and the frequency these hashtags appear in the csv file were used as the y-axis.

We had 280 unique hashtags which were divided into 150 buckets of different sizes.

And the total error obtained using jagdish's algorithm is reported below:

B. For online settings we used kafka. The extracted hashtags were sent through producer and the resulting message was received at consumer end then we performed guhas algorithm to obtain the histogram ranges.

When the data was completely unseen then we update only the last column for each of the bucket but if the data streaming in was seen previously then whole histogram ranges were recomputed agin.

Sample hashtags values

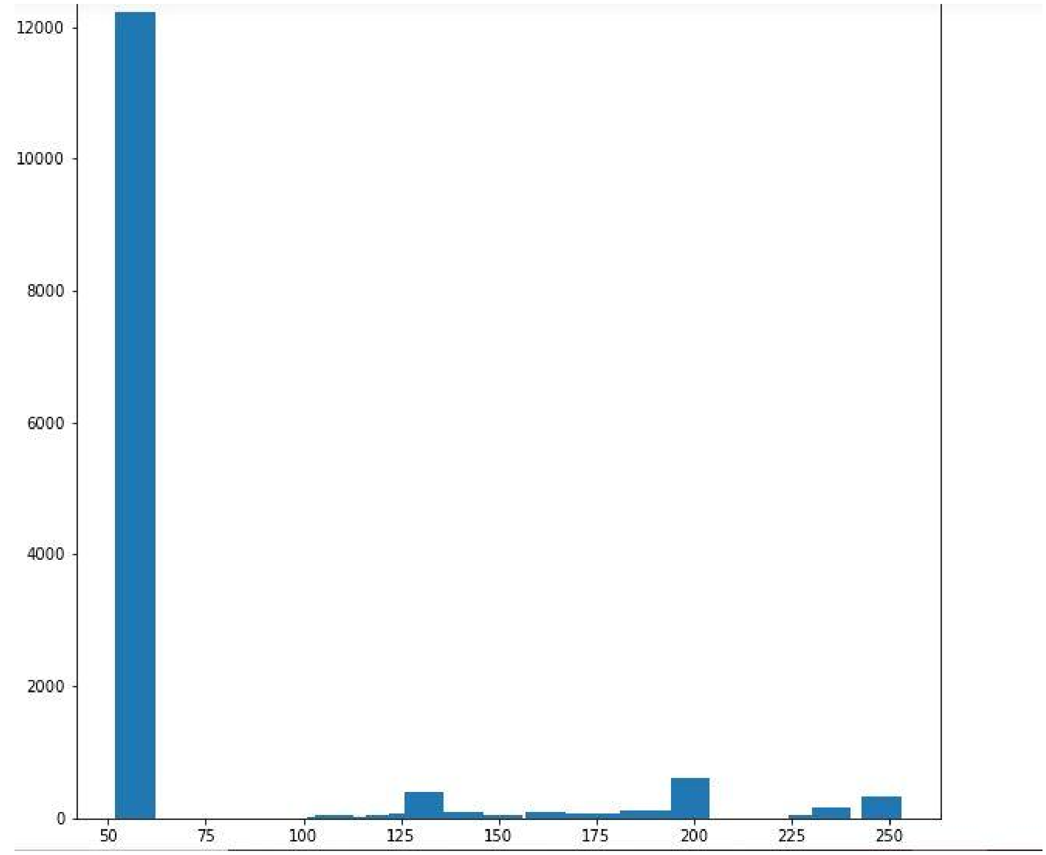
```
oxygen
```

```
icu
oxygencylinder
delhincr
covid
verified
oxygencylinders
gurgaon
sos
remdisivir
```

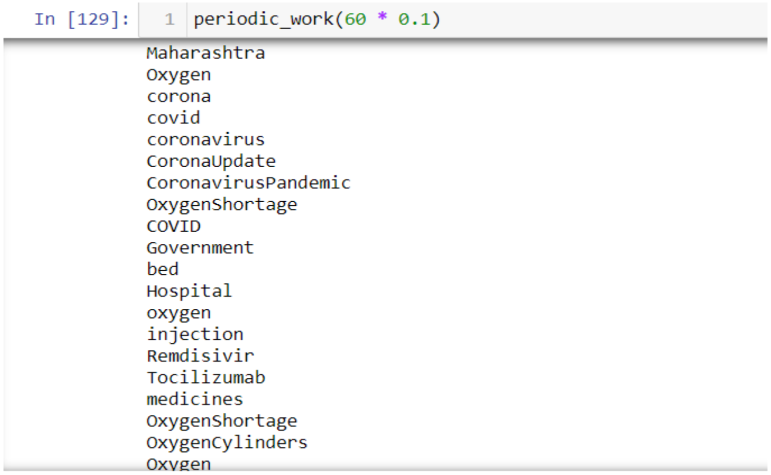
ERROR obtained

```
72.089
```

HISTOGRAM OBTAINED:
TASK-1
In the offline setting using jagdish-et-al



Streaming Data at the Kafka Producer end from tweepy api. This goes to



```
In [129]:  periodic_work(60 * 0.1)
```

```
Maharashtra
Oxygen
corona
covid
coronavirus
CoronaUpdate
CoronavirusPandemic
OxygenShortage
COVID
Government
bed
Hospital
oxygen
injection
Remdisivir
Tocilizumab
medicines
OxygenShortage
OxygenCylinders
Oxygen
```

Bucet ranges obtained:

```
[57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74
  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92
  93  94  95  96  97  98  99 100 101 102 103 104 105 106 108 111 112 113
 115 116 117 118 119 120 121 123 124 125 126 127 130 131 140 141 148 149
 150 151 155 156 157 158 159 160 161 162 171 172 177 178 179 182 183 184
 186 190 198 199 228 229 233 235 247 248]
```

ERROR obtained

```
232.05317460317343
```

Learning :

We learnt several techniques like extracting data from twitter using tweepy . setting up kafka and using different algorithms to obtain v-optimal histogram and how the messages are sent and received between the consumer and producer.