NAME: VIDHYA KOMADI BATMARADJY

Coursework: Big data processing

Input sanitization: It is given that each record in the input file would be of the below given format.

epoch_time; tweetId; tweet (hashtag contained); device

When inspected it was found that the input data did not have a uniform format. In order to maintain consistency and obtain valid results for all the questions, I decided to sanitize the input by instructing each mapper job to process only those records which have exactly 4 fields (i.e. if record length=4)

A.CONTENT ANALYSIS

Aim: To find the length of tweets, aggregate them in groups of 5 and find their frequency distribution(count).

tweetMapper:

1. The mapper reads the input file line by line, splits the entire line into substrings using the given delimiter (;) and uses only the tweet substring (at index position [2]) stored in an array of strings.
2. For the purpose of grouping tweets as part of MapReduce job, bin ranges (lower limit – upper limit) have been used.
3. The upper limit is obtained by dividing each tweet length by 5, rounding the resultant value to the closest integer and then multiplying it back by 5. The lower limit is obtained by subtracting a value of 4 from the upper limit.
4. The 2 numerical values thus obtained for each tweet are finally concatenated and converted to string format to be used as mapper output key (ex:1- 5).

tweetReducer:

1. The reducer processes the output data that comes from the mapper and generates the final output by summing up all the intermediate values related to each key. The bin range is used as the key and the tweet count is used as the value.

Output: The reducer output would have the below given format. The complete list can be found inside the compressed code files(tweetlength.zip).

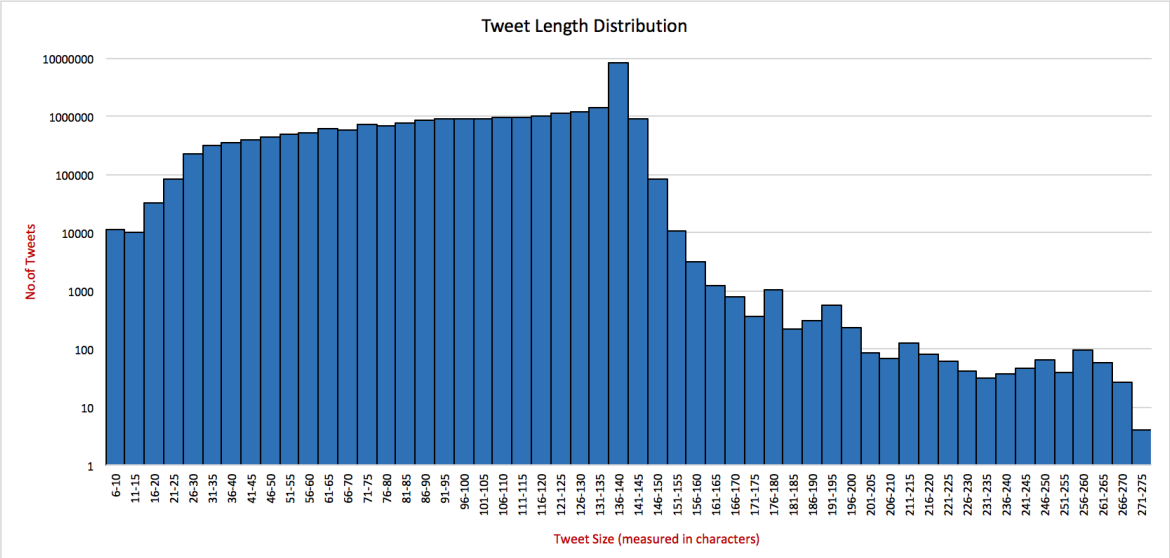
101-105 908337

106-110 938878

……

Histogram Plot: The histogram was plotted directly using the reducer output file. The respective excel file can be found under the compressed file.

I choose to plot and visualize all the tweet lengths rather than filtering out high values as it is not always the case that tweets with length greater than 140-character limit contain characters with non-standard encoding. Tweets also exceed the length of 140 due to the usernames, retweets, or URL generated by attaching a video or photos.



Aim: To find the average length of a tweet from the given dataset.

AvgMapper:

1. The mapper performs the same split operation as described for TextMapper but assigns a single common key “Tweet Average” for all of the mapper output values (tweet length).

AvgReducer:

1. The reducer computes the average length by dividing the sum of all the tweet lengths by the total no. of tweets.
2. As I did not use the custom writable (IntIntPair), I calculated the total no. of tweets using a counter value and incremented it by 1 each time a sum operation was done.

Output: The average length of a tweet was found to be 109.

B.TIME ANALYSIS

Aim: To create a time series diagram with the no. of tweets that were posted each day of the event.

DayMapper:

1. The mapper performs the same split operation but uses epoch_time substring (at index position [0]) for further processing. Although only records with 4 fields are processed, not all records contain epoch_time as the first field (i.e. there are few non-numeric values as well). So the mapper performs a numeric input validation and then converts the valid numeric data into date format using SimpleDateFormat function.
2. As the question demands only those tweets and the respective dates during which the event took place, the mapper further checks and outputs only the tweets which fall between the dates (05/08/2016 - 21/08/2016) and (07/09/2016 – 18/09/2016 for Paralympics).

DayReducer:

1. The reducer processes the output data that comes from the mapper and generates the final output by summing up all the intermediate values related to each key. In this case, the date is the key and the value is the sum of all tweets for each key.

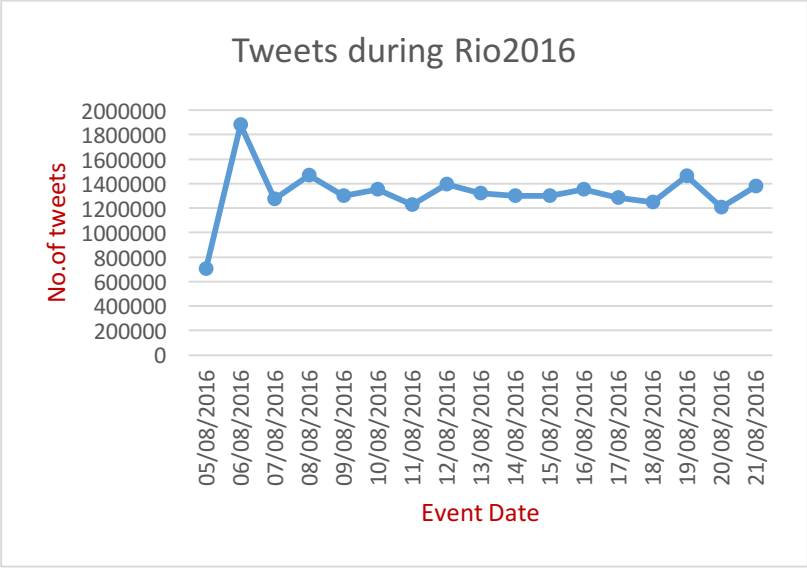
Output: The reducer output would have the below given format. The complete list can be found inside the compressed code files(DayTweet.zip).

2016-08-05 702749

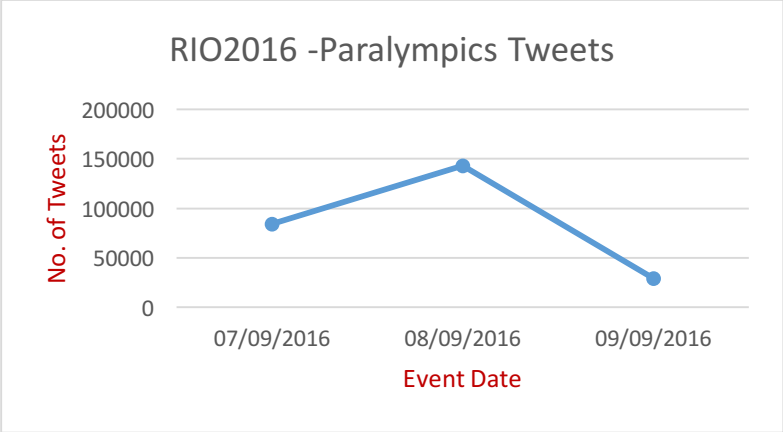
2016-08-06 1880213

……..

Time Series Diagram: From the diagram it is clear that there was a significant increase in tweets between 5th and 6th of August after which tweets dropped and started to fluctuate for the following days until the end of event (21st of August).



Time Series Diagram (Paralympics): From the diagram it is clear that there was a gradual increase in tweets from 7th until 8th of September after which it dipped heavily exhibiting a decreasing trend towards the end of the event.



C.Hashtag Analysis

Aim: To classify the support tweet results into different countries as part of MapReduce job and estimate the countries that received the highest support from Twitter messages.

HashMapper:

1. The mapper performs the same split operation and takes the tweet substring (at index position [2]). In order to maintain uniformity and ease out the country classification, the entire tweet is converted to lowercase.
2. The hashtags are then extracted using pattern matching and sanitized to contain only words (any numeric, special or non-standard encoded characters in the word is excluded).
3. The words are then checked against the below given list of team affiliations using appropriate conditions.

   Example: (teamusa, teamusago), (gousa, gousago, usago), (goteamusa, goteamusago), (supportusa, supportteamusa).
4. For each match found, the affiliation keywords are filtered out from the input and the resultant text is compared against the list of country codes (ex: US or USA) and names (United States of America) with the help of `Locale.getISOCountries utility.`The ISOCountries are converted to lowercase to facilitate uniform string comparision.
5. For words that match the ISOCountries list, the respective ISOcountry names are emitted as the mapper output key.
6. **Note:** There were other team affiliation keywords like “Herewegousa”, “Letsgousa”, ”Letsgoteamusa”, etc. But I choose to exclude the same as they did not contribute to significant flips between the country positions except that the count for each country increased by few numbers but its support position remained the same.

HashReducer:

1. The reducer processes the output data that comes from the mapper and generates the final output by summing up all the intermediate values related to each key. In this case, the country name is the key and the value is the sum of all support tweets for each key.

Reducer Output: The reducer output would have the below given format. The complete list can be found inside the compressed code files (HashTweet.zip).

afghanistan 49
albania 433
algeria 514
american samoa 13
angola 476
……..

Consolidated Output: It is found that **USA** received the highest support. A list of top 15 countries is given below. The complete list is formatted (based on support count) in countrysupport.xls which could be found in the compressed files.

No.	Countries	No. of Support Tweets
1	**united states**	**370394**
2	united kingdom	173389
3	canada	81962
4	spain	77926
5	netherlands	66492
6	jamaica	36600
7	malaysia	20646
8	serbia	20515
9	kenya	14082
10	ireland	13546
11	france	12960
12	austria	9786
13	nigeria	9655
14	india	8765
15	saudi arabia	8097