Daniel Griffin (5 & 3) + Sam Wenke (7)

Cloud Computing Project 3 Problem 5

Problem Statement

Given a large data set of tweets, what twitter user tweeted the most? Also what are the 5 tweeters with highest average length per tweet? Who are the bottom 5?

Approach

The most frequent tweeter was found in two stages. The first stage counted the number of tweets for each user. The map of the first stage output "<username>, 1", and the reduce summed the value for each user. The second stage calculated the top tweeter by finding the top tweeter in each map, then finding the top tweeter in each reduce. The map of the second stage found the top tweeter for its data set, and the reduce of the second stage found the final top tweeter.

The top average length tweeters were found using 3 map and reduce stages. The first map/reduce stage calculates the number of tweets, and the sum of all tweet lengths for each user. The map of the first stage emits two keys for each user. The first has the form "<username>_CNT, 1". The second has the form "<username>_SUM, len(tweet)". The map of the first stage sums up the value for each key. The second map/reduce stage calculates the average tweet length for each user. The map of the second stage determines if the key is a sum or count by checking the last 4 characters of the key, which will be either "_CNT" or "_SUM". It then outputs the key "<username>, CNT_<value>" or "<username>, SUM_<value>". The reduce of the second stage then receives both the count and sum values for each user. Using these, it calculates the average and outputs them. The third map/reduce stage performs the top 5, or bottom 5 calculation. The map of the third stage finds the top or bottom 5 tweeters for its subset. The reduce of the third stage finds the top or bottom 5 tweeters for the final data set and saves it to a file.

Results

The most frequent tweeter was:

marilyn9743 with 3419 tweets.

The top 5 most frequent tweeters was:

Huntersweat with 416.0

Daniel Griffin (5 & 3) + Sam Wenke (7)

RoyaleliteKiva with 350.0
blackxhole with 319.0
KelleeMichele with 272.0
pizzadellarry with 253.0

The bottom 5 most frequent tweeters was:
GGYuuuuuun with 0.5
nettstp with 0.5
kevin61027 with 0.5
AyoobBhz with 1.0
Abd_Alassaf with 1.0

Daniel Griffin (5 & 3) + Sam Wenke (7)

Cloud Computing Project 3 Problem 3

Problem Statement

How does @PrezOno's tweet length compare to the average of all others? What is his average length? All others?

Approach

The average tweet length for all users other than Santa Ono was found using a one stage map/reduce program. The map phase output two keys. The first was "COUNT, 1" and the second was "SUM, len(tweet)". The reduce phase summed the values for the count and sum key and output them to a file. The average tweet length was then output as the sum divided by the count. Santa Ono's average tweet length was already calculated from the output of the second stage of problem 5, and did not need to be re-calculated. The relative position of Santa Ono in relation to all other tweeters based on average tweet length was calculated by sorting the output of the second stage of problem 5 and searching for Santa Ono.

Results

All user tweets:

Total number of tweets: 6079961
Total sum of tweet lengths: 485386155
Average tweet length: 79.8333761269192351

Santa ono tweets:

Total number of tweets by Santa Ono: 341
Total sum of tweet lengths by Santa Ono: 35424
Average tweet length for Santa Ono: 130.882697947

Santa ono in relation to all others:

Total number of tweeters: 3054455
Santa Ono's Index: 2040106
Santa Ono Tweet percentages: Ono Average Tweet Length > 66.79% of tweeters in the data set.

Daniel Griffin (5 & 3) + Sam Wenke (7)

Cloud Computing Project 3 Problem 7

Problem Statement

For those tweets with For those tweets with location information, what lat/long (or city/state) is the centroid? What was the proportion of tweets with location to those without?

Approach

The centroid of all tweeter's locations was found by using a one stage map/reduce program. The first map was responsible for sorting the users who have locations, truncating the latitude and longitudes to a comma separated key that is mapped to a count; the output generated was the keys and counts associated with them. The first reduce combined the keys that were matching and summed the counts. Then a running maximum was tracked and outputted with the corresponding key and count associated with the key. The count/percentage of locations were found by a similar map/reduce program. In the first step, the users with locations were counted and a total was counted; the second step the percentage and sum was outputted.

Results

Location centroid:
Centroid: 39, -84
Total count of appearances: 320007

Count and Percentage:
Count: 1868302
Percentage of users who have location: 30.73%