University of Pisa

Department of computer science

Algorithm design

Fourth hands-on: Tweets

Domenico Erriquez

## 1. PROBLEM

Given a stream of tweets, answer the follow questions:

1. Count the percentage of happy users in the different moments of the day (morning, afternoon, evening, night)
    - Discuss what you find if compute also the percentage of unhappy users. Do the two percentages sum to 100%? Why?
2. Spell the 30 favorite words of happy users.
3. Find the number of distinct words used by happy users.
    - How could exclude words repeated only once.
4. Decide if in general happy messages are longer or shorter than unhappy messages.

## 2. Solution first question

In order to determine the percentage of happy and unhappy users in the different moments of the day, we can utilize linear counters. Specifically, we can implement the use of eight linear counters, each with a capacity equivalent to the number of users. Four of these counters will be designated for tracking happy users, one for each moment of the day (morning, afternoon, evening, night) and the remaining four will be designated for tracking unhappy users, also one for each moment of the day. When a new message arrives we set to 1 the position for that user in one of the eight linear counters based on the happy/unhappy and the moment of the day of the message. To compute the percentage of happy users in a moment of the day, it's enough to count the number of 1's in the linear counter for that moment of the day and divide by the length of the linear counter, which is the number of users. The sum between the percentages of happy and unhappy users may not always be equal to 100%. This is due to the possibility that a user may have sent both happy and unhappy messages in the same moment of the day. In such instances, the user would be included in the count for both the percentage of happy and the percentage of unhappy users, resulting in a sum greater than 100%. However, in cases where this does not occur, where a user only sends messages of a single mood in a given moment of the day, the sum of the percentages of happy and unhappy users will be equal to 100%.

## 3. Solution second question

"In order to determine the 30 most frequently used words by a happy user, we can utilize the space saving algorithm. The algorithm utilizes a table with 30 rows, representing the top 30 words, and three columns: one for the word, one for the counter and one for the error bound. For each tweet from the happy user, we will process each word w through the following algorithm:

- Check if $w$ is already in the table:
  - If it is, increment the counter for that word $count_w$.
- If it is not, and the number of rows in the table is less than 30, then add a new row to the table with the word and set $count_w$ to 1.
- If the number of rows in the table is equal to 30:
  - we replace the word $w_{min}$ that has the minimum counter $count_{w_{min}}$ with $w$
  - increment $count_w$
  - assign to the error bound column $\varepsilon_w$ the value that was in $count_{w_{min}}$

At the end, the table will contain the thirty most frequently used words by happy user.

## 4. Solution third question

To find the number of distinct words used by happy users, we can use a LogLog counter and a Bloom Filter to exclude the words that repeats once. For each word of each happy tweet, if the value in the bloom filter is zero then it is the first time we see this word, so we update the bloom filter, otherwise we update the LogLog counter.

## 5. Solution fourth question

To decide if happy messages are longer or shorter than unhappy messages, we can compute the average length of the happy messages and unhappy messages and compare them.