Hands On 4

Algorithm Design

Ferraro Domenico
559813

1 Problem

Given a stream of tweets, answer to the following questions on twitter data:

1. Count the percentage of happy users in the different moments of the day (morning, afternoon, evening, night)
    a. 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
    a. How could exclude words repeated only once
4. Decide if in general happy messages are longer or shorter than unhappy messages

2 Solution

Percentage of happy users in the different moments of the day

Use one linear counter for each moment of the day. We have four counters for happy users and other four counters for the unhappy ones.

The 30 favorite words of happy users

We can use the space saving algorithm on the words of happy tweets.

Number of distinct words used by happy users

We can use the HyperLogLog data structure to count the distinct words used by happy users and a Bloom Filter to exclude the words that repeats once.

For each word of each 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 HyperLogLog.

Happy messages are longer or shorter than unhappy messages?

We can scan all the tweets to compute the average length of a message.