University of Pisa

Department of computer science

Algorithm design

Sixth hands-on: Most frequent item in a stream

Domenico Erriquez

1. PROBLEM

Suppose to have a stream of n items, so that one of them occurs $> \frac{n}{2}$ times in the stream. Also, the main memory is limited to keeping just $O(1)$ items and their counters (where the knowledge of the value of n is not actually required). Show how to find deterministically the most frequent item in this scenario.

[Hint: the problem cannot be solved deterministically if the most frequent item occurs $\leq \frac{n}{2}$ times, so the fact that the frequency is $> \frac{n}{2}$ should be exploited.]

2. SOLUTION

The solution to find the most frequent item that occurs $> \frac{n}{2}$ keeping just the item and its counter is the following:

```
def find_most_freq_item(stream):
    counter = 0
    most_frequent_item = None
    for item in stream:
        if counter == 0:
            most_frequent_item = item
        if item == most_frequent_item:
            counter += 1
        else:
            counter -= 1
    return most_frequent_item
```

The algorithm uses just two variables: one for the element and one for its counter. Therefore, the space complexity is $O(1)$. At beginning the variable $counter$ is set to 0 and $item$ to None. After

this initialization the algorithm begins with scanning the stream and based on the values of the $counter$ and $item$ there are the following operations for execution:

- If the $counter$ is equal to 0, then the most frequent item is the current item.
- If the current item is equal to the most frequent item, then the variable $counter$ is incremented by one.
- If the current item is not equal to the most frequent item, then the variable $counter$ is decremented by one.

When the stream is consumed the algorithm returns the most frequent item.

Since the most frequent item occurs more than $\frac{n}{2}$ times, the number of increases is greater than the number of decreases, so at the end the variable $counter$ will contain a number greater than 1 and the variable $most\_frequent\_item$ will contain the most frequent item.