Lab 9: Flajolet-Martin

Ruofan Zhou

May 30, 2014

1 Questions

1.1 Q1 Compare the estimates you get using the average and the median. Both have advantages and disadvantages—describe what you observe, and explain.

The average has more statistical significance but the estimates seems absurd because small value of R can't have much weight when using average of 2^R so the values are fully discrete. While for the median, the values are more stable but what I can got is only power of 2, it should be more statistical.

1.2 Q2 Include your final version of this function.

```
public long countUniqueWords(Iterator<String> iter) {
    while (iter.hasNext()) {
        this.processWord(iter.next());
    }
    double[] arr = new double[(hashFunctions.length + 2) / 3];
    for (int i = 0; i < hashFunctions.length; ++i) {
        arr[i / 3] = 2 << maxZeros[i];
                int sum = 1;
                ++ i;
                if (i < hashFunctions.length) {
                        sum ++;
                        arr[i / 3] += 2 << maxZeros[i];
                        ++i;
                        if (i < hashFunctions.length) {
                                sum ++;
                                arr[i / 3] += 2 << maxZeros[i];
                                ++i;
                        }
                }
                arr[i / 3] = arr[i / 3] / ((double)sum * 1.0);
        }
```

```
        return (long)median(arr);
    }
```

1.3 Q3 What estimate N^ do you get? Look at the value of nbBits in the function main(), and explain what problem might arise in this case.

I got **33554432**, and it is exactly 2^{24}(which 24 is the nbBits in the function main()), so the problem is, the answer is much larger than 2^{24} while we just can get number 2^{24} because it is limited by nbBits.

1.4 Q4 What estimate N^ do you get this time? Based on this estimate, what is the smallest value that nbBits could take for this particular instance of dataset?

I got **268435456**. Based on this estimate, the smallest value of nbBits is log(268435456), is **28**. To avoid overflow, I think the smallest value we set should be **30**.

1.5 Q5 How much memory (in bytes) would you roughly need if you if you were to run ExactCount on the dataset? Justify your answer.

The space complexity is O(nlogn), while n is about **268435456**. The words's average length is about **7**, so for the data, a string covers about **7** bytes. Thus the extimate memory is about $268435456 \times log(268435456) \times 7$, is 5×10^{10} bytes.