CS333 Homework 6 Report
Nathan Kim (njk24) and Jeffrey Luo (jl834)

1.
    a. [44, 242, 77]
    b. See code
    c.
        i. ‘the’ appears 8191 times in stream.txt
        ii. ‘are’ appears 333 times in stream.txt
        iii. ‘sydney’ appears 72 times in stream.txt
        iv. ‘london’ appears 27 times in stream.txt
    d. 128124 bytes are used from 1c
2.
    a. See code
    b.
        i. ‘the’ is estimated to appear 8359 times according to CMS. The ratio to the actual value is 1.02
        ii. ‘are’ is estimated to appear 597 times according to CMS. The ratio to the actual value is 1.79
        iii. ‘sydney’ is estimated to appear 240 times according to CMS. The ratio to the actual value is 3.33
        iv. ‘london’ is estimated to appear 183 times according to CMS. The ratio to the actual value is 6.78
    c. We would never underestimate the count because the minimum is being raised across the board for the bytes of an individual word. This concept can most easily be seen at the beginning, when everything is 0. If we were to say that the first word in the stream were 5 letters, all the byte counts for the word would be raised by 1. In this way, there may be overestimates but there will never be underestimates because all ties are raised evenly as well. This explains the conservative CMS but would thus also apply to the regular CMS as it raises all byte counts for the word evenly no matter what (unless the word length is shorter than the amount of rows, which would then contribute to 255, which is never going to be the minimum)
    d.
        i. ‘the’ is estimated to appear 8191 times according to CMS. The ratio to the actual value is 1.000
        ii. ‘are’ is estimated to appear 334 times according to CMS. The ratio to the actual value is 1.003
        iii. ‘sydney’ is estimated to appear 98 times according to CMS. The ratio to the actual value is 1.361
        iv. ‘london’ is estimated to appear 81 times according to CMS. The raio to the actual value is 3.000

e. Since CMS typically artificially boosts the estimated frequency of each item, if a word is common it will always appear to have a high frequency when you get the estimate from the CMS, but if a word is not as common, it will have a larger but still relatively low frequency estimate. This makes the ratio for the more common words much more closer to the actual count whereas the ratio for a less common word might have more collisions than actual appearances, thus making the ratio seem much larger than a more common word.

f. 5120 bytes would be used by CMS when using 5 rows. The ratio of this to our original hash table from 1d is 0.03996

3.

a.

$$ed \bmod (p-1)(q-1) = 1 \quad \text{by definition}$$

$$ed = 1 + k(p-1)(q-1) \; where\; k \rightarrow x^{ed} \bmod N = x(x^{(p-1)(q-1)})^{k}$$

Due to Fermat's Little Theorem, we know that if x is prime, then for every $1 \leq a \leq x$, $a^{x-1} \bmod x = 1$.

Since p and q are prime, we can say:

$$x^{p-1} \bmod p = x^{q-1} \bmod q = 1$$

Knowing this, it is now trivial to say:

$$x^{ed} \bmod N = x^{1+k(p-1)(q-1)} \bmod N = (x)x^{k(p-1)(q-1)} \bmod N = x$$

b. The algorithm itself is O(log(m)) where *m* is the size of the exponent. However, since we are dealing with very large numbers, the multiplication steps are *O*(n^2) time. This results in a time complexity of *O*(log(m)n^2). Since *m* is extremely large, we can change the equation to *O*(n^3).

c. First message: hello, is anyone listening?
Verification: false
Second message: use the utf-8 encoding of this message to seed rc4
Verification: true
Third message: happy halloween!
Verification: false

d. the project gutenberg ebook of a tale of two cities by charles dickens this ebook is for the use of