Graphs

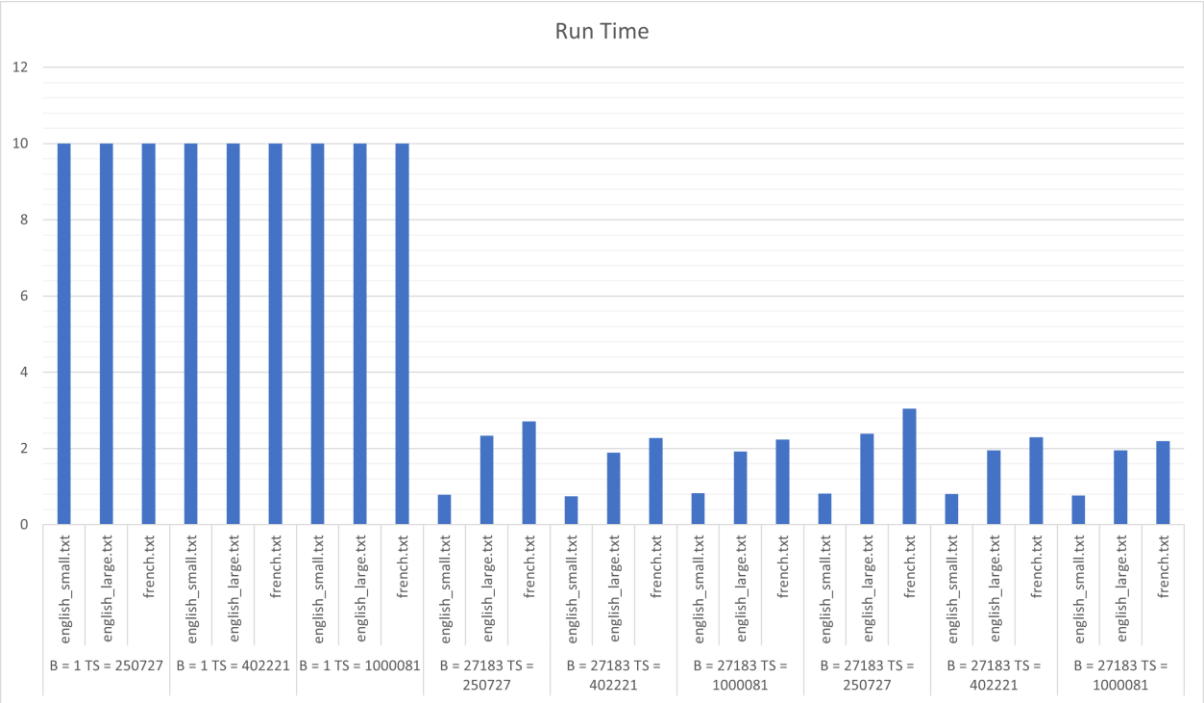

Figure 1: Run Time

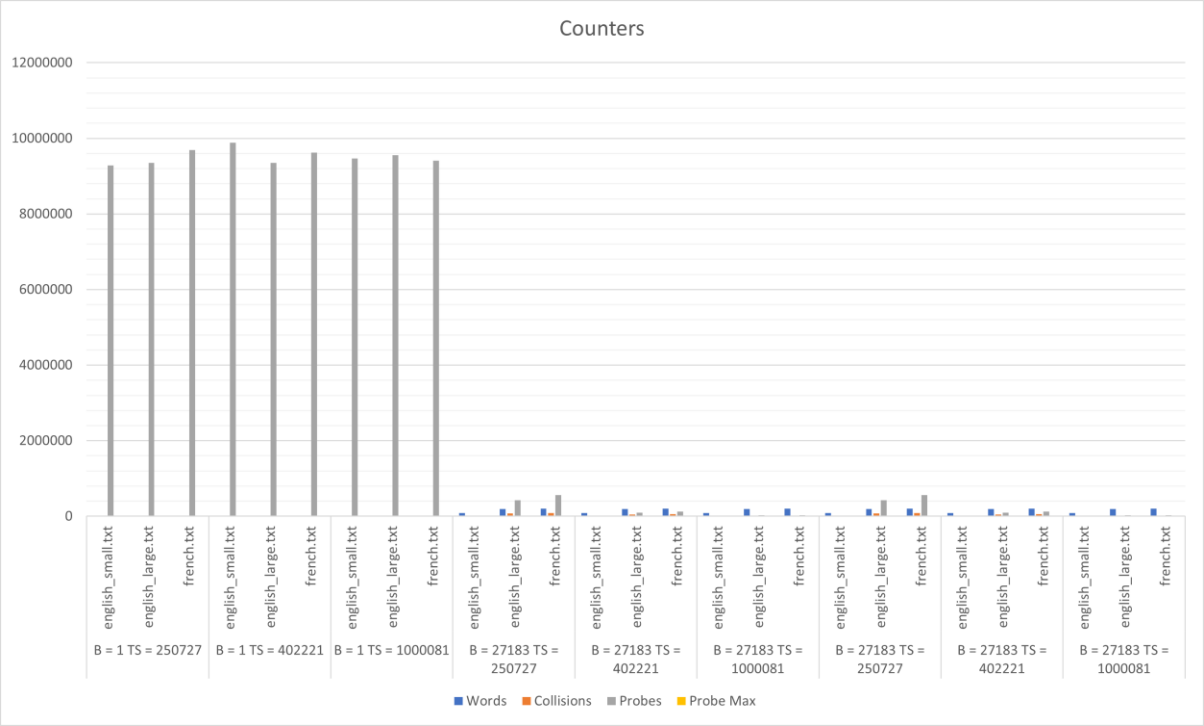

Figure 2: Counters (Including Hash Base 1)

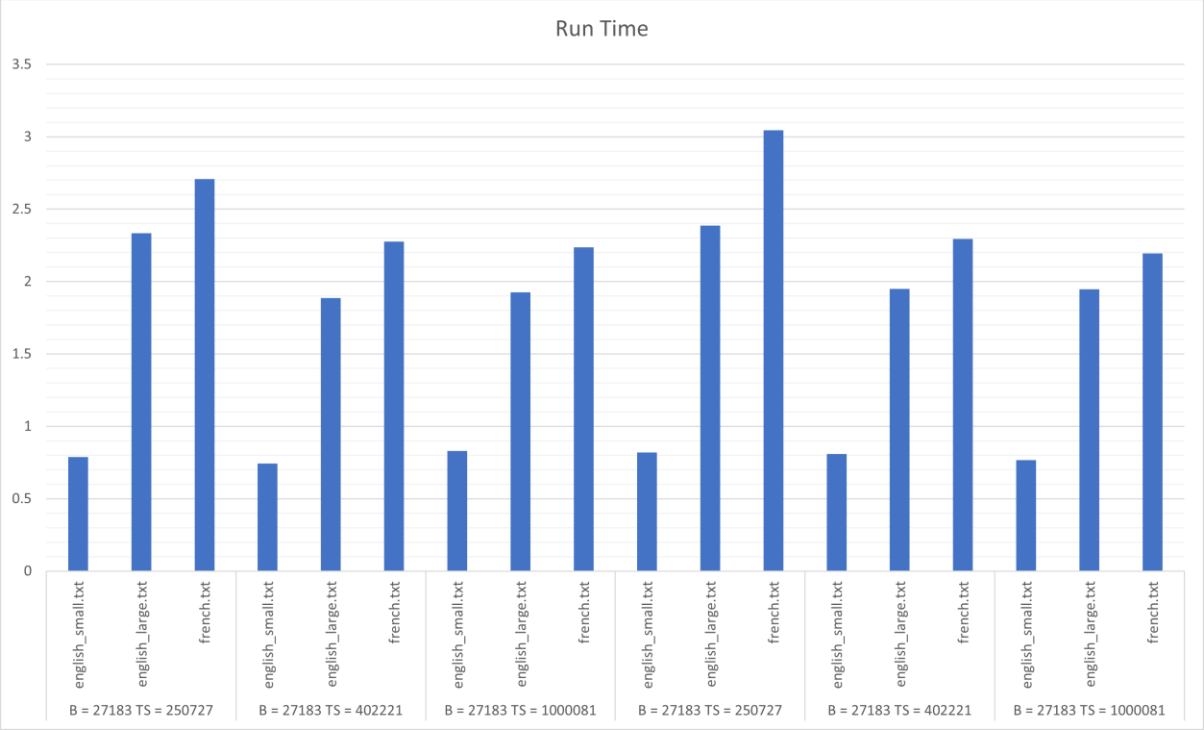

Figure 3: Run Time (excluding Hash Base 1)

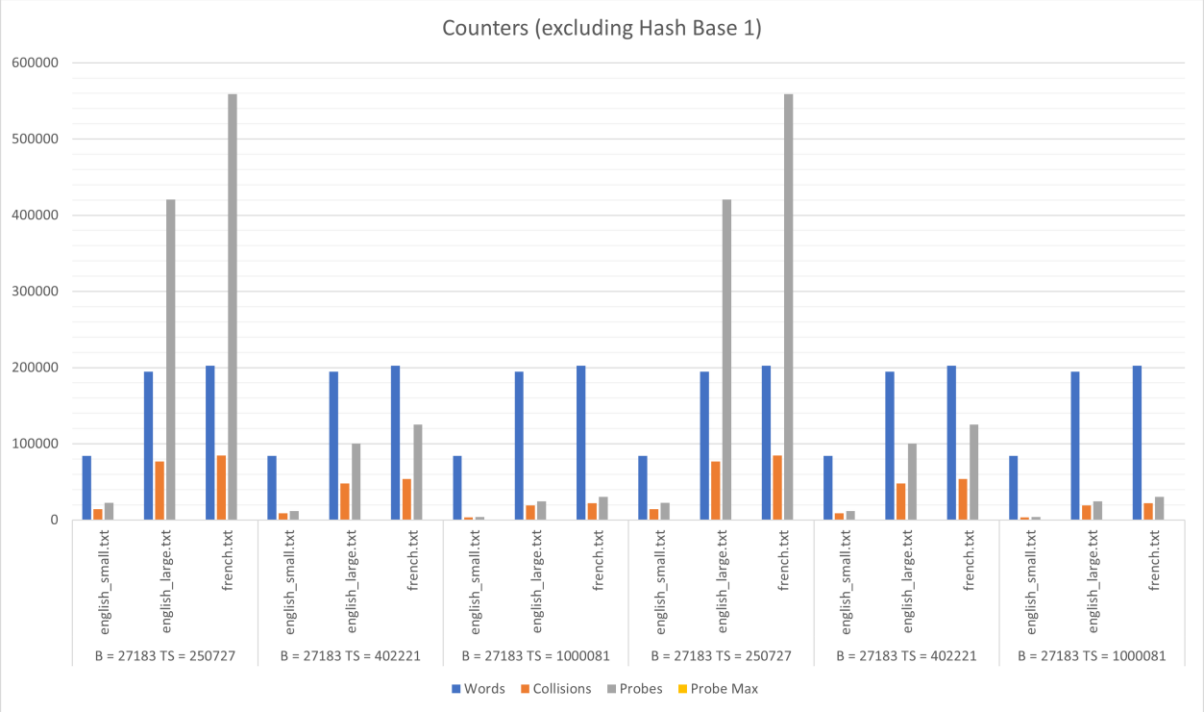

Figure 4: Counters (excluding Hash Base 1)

Analysis

What values work well, which work poorly, and why.

From looking at the output generated, we cans see that the worst output would be using a hash base of 1. Hash bases should be large and a prime number to minimize clustering that would lead to low performance. This means that with a hash base of only 1 would result in a lot of big clusters that would slow down run time as evident in Figure 1.

The better values used are B = 27183 TS = 402221, B = 27183 TS = 1000081, B = 27183 TS = 402221, and B = 27183TS = 1000081 which all have large hash bases and table size.

The relationship between the counters and the runtime.

The relationship between the counters and the runtime can be seen most clearly with the outputs using hash base 1. The large number of probes have greatly increased the run time to its maximum of 10 seconds. This can also be seen in B = 27183 TS = 250727 and B = 27183TS = 250727 in Figure 3 and 4 where the higher number of probes resulted in a slightly higher run time compared to the rest. We can conclude that the higher the number of probes would result in a higher run time.

The length of the longest probe chain and the promise of O(1) time complexity.

The longest probe chain in the generated output is at 4976 with B = 1 TS = 402221 and the file english_small.txt. The promise of O(1) time complexity comes from the fact that inserting and finding data to and from a hash table has an O(1) time complexity. The problem with this method is that the hash function might give a position that has already been occupied (collision) and to circumvent this we either have to resort to separate chaining or open addressing to store the new data. Both separate chaining and open addressing also has a linear O(1) time complexity and the hope is that the extra time taken to locate a new position is relatively short. But as shown in the output with a longest probe chain of 4976, this is not always the case.

Explain why rehash_count is 0 in all runs.

Because the load factor did not exceed 50%.