Task 2 Documentation

english_small.txt, Number of items = 84097

Multiplier	**1,064**	**1,091**	**1,064**	**1,091**
Table size	**300,000**	**300,000**	**300,151**	**300,151**
Load	0.28	0.28	0.28	0.28
Total Collisions	21820	11985	11976	11987
Total Probe Length	35847	17222	17252	17459
Average Probe Length	0.426	0.205	0.205	0.208

english_large.txt, Number of items = 194433

Multiplier	**1,064**	**1,091**	**1,064**	**1,091**
Table size	**300,000**	**300,000**	**300,151**	**300,151**
Load	0.648	0.648	0.648	0.648
Total Collisions	93989	64371	63623	63820
Total Probe Length	327414	209471	201338	211336
Average Probe Length	1.684	1.077	1.036	1.087

french.txt, Number of items = 202358

Multiplier	**1,064**	**1,091**	**1,064**	**1,091**
Table size	**300,000**	**300,000**	**300,151**	**300,151**
Load	0.675	0.675	0.675	0.675
Total Collisions	133171	71563	70787	71454
Total Probe Length	798228	275897	263167	268723
Average Probe Length	3.945	1.363	1.301	1.328

Test Explanations

1064 = non-prime number multiplier

1091 = prime number multiplier

300,000 = non-prime table size

300,151 = prime table size

So, combinations = [N, N], [P, N], [N, P], [P, P]

(N = Non-prime, P = Prime)

Result

In all the text files, the combination with non-prime multiplier and non-prime table size had the biggest number of collisions, probe length and average probe length as well. Other test combinations with at least one prime number showed similar trend in all text files. As a result, I found that at least one prime number must exist in either multiplier value or table size to have small number of hash value collisions.

Moreover, I found that the maximum table size must be much bigger than the number of items to avoid collisions. It is because, in English_small.txt, the biggest number of collisions is 21820, which is approximately 25% of total number of items, meaning that 25% of items had collided hash values. But in other two text files, I could see that as the number of items increase and approach to the table size, the number of collisions significantly increase. In English_large.txt, which has 2.3 times bigger number of items than English_small.txt, the biggest number of collisions is 93989, which is 48% of total number of items. Then, in French.txt, the biggest number of collisions is 133171, which is 66% of total number of items.

Plus, I could see that number of collisions and probe length have something in common. In most cases, as the number of collisions increased, the probe length increased as well.

A weird result is that two prime numbers combinations were not the most effective values to avoid collisions, which is different from the actual theory. In all text files, non-prime multiplier, 1064 and prime table size, 300151 had the least number of collisions. However, I cannot conclude that combination with non-prime multiplier and prime table size gives the least number of collisions because the number of test combination is extremely small, and the numbers of collisions have very small differences, compared to the item size of text files.

So, in my view, if a large number of test values(combinations) used, it would give more accurate results that matches to the hash table collision theory.