Data Structures Evaluation of Hash Functions

Aigerim Zhusubalieva (az2177@nyu.edu)

May 2, 2021

1. Summation Hash Function

When applied to 101 test files, Summation Hash Function produced total of 511855 collisions for 598146 unique entries to the hash table (refer to Table 1). This equates to about 5067.87 collisions on average, or collisions of 85.5% of the total entries. This function is evaluated to be the worst among all 4 functions.

2. Polynomial Hash Function

When applied to 101 test files, same as for the Summation Hash Functions, Polynomial Hash Functions produced total of 32652 collisions for 598146 unique entries to the hash table (Table 1). This equates to 323.39 collisions on average, which is equivalent to 5.46% of the total entries. Polynomial Hash Function is evaluated to be the best among the 4 tested functions.

3. Cyclic Shift Hash Function

When applied to the same 101 test files, Cyclic Shift Hash Function produced total of 34202 collisions for 598146 unique entries to the hash table (Table 1). This equates to 338.63 collisions on average or 5.72% of the total number of unique entries. Falling short from the Polynomial Hash Function, Cyclic Shift Hash Function is evaluated to be the second best hash function among the 4 functions.

4. Summation Hash Function with MAD compression

All 3 hash functions above had division as a compression method, but for the last one the MAD (multiplication-addition-division) compression method was used and produced results similar to the Summation Hash Function with division only. When tested on 101 files, the total number of collisions was 511290 for 598146 unique entries. On average, there were 5062.28 collisions per file, resulting in 85.5% collision rate. Being very close to the Summation Hash Function with division as a compression method, this function is evaluated to be the third in terms of number of collisions per unique entry to the hash table.

5. Overall Evaluation

Given the average number of collisions and the percentage of collisions per unique entries, it is concluded that the Summation Hash Functions, both with division and MAD compression method performed the worst, with slight difference (less than 1% difference). Polynomial and Cyclic Shift Hash Functions performed significantly better, even though with division compression only. Although it was close between the latter two functions with a difference of 0.26%, the Polynomial Hash Function performed the best and therefore chosen to be the default function.

file name	total words	unique words	hash 1	hash 2	hash 3	hash 4
3254	1628291	48627	46893	9803	10103	46769
2334-0	306225	22369	20944	2267	2256	20903
32046-8	240686	14369	13176	950	988	13157
31217-8	166892	14047	12837	852	948	12821
8933-0	178108	9717	8700	416	423	8697
1626-0	138892	10553	9366	520	572	9352
38531-8	130493	12144	10930	673	699	10917
34766-0	129776	11993	10827	651	719	10811
32845-8	126488	13438	12242	832	847	12221
38172-8	112371	11588	10421	626	638	10410
17669-8	109250	12766	11607	740	771	11595
10947-8	107214	12543	11344	709	776	11330
24878-8	110440	9862	8790	480	480	8781
2305-0	90448	10831	9415	500	561	9394
28726-8	104958	11661	10536	682	667	10527
5592	102189	9501	8454	425	439	8447
6696-8	90666	14363	13012	957	936	13000
5737-0	95810	9334	8277	421	436	8268
9790-8	95492	12582	11351	701	788	11329
6120-0	91704	10507	9401	490	522	9394
373-0	86140	10997	9895	572	578	9889
6073-0	86125	7973	6957	262	298	6952
13799	85021	9711	8656	391	438	8650
40745-8	83428	7515	6500	254	255	6495
58341-0	73635	10898	9793	577	571	9770
24313-8	78130	7903	6864	262	311	6857
57040-0	66876	11626	10498	646	656	10482
1944-0	76782	7188	6219	228	254	6211
55514-0	64811	13309	12122	816	882	12103
2550-0	68126	9011	7922	380	392	7915
15717-8	64075	8367	7265	312	330	7261
56870-8	66693	7948	6910	294	275	6905
18776-8	61747	8865	7808	351	391	7805
54183-0	55845	8156	7159	303	323	7157
34313-8	56074	6589	5545	182	209	5543
14744-8	58161	7338	6336	275	266	6334
pg4081	50993	9404	8385	401	419	8381
49598-8	54546	6705	5713	189	207	5707
2518	51336	6309	5394	177	203	5389
9629-8	50832	5794	4858	161	167	4856
57006-0	43803	6885	5884	186	233	5882
2429-0	43910	5681	4772	144	163	4767
2327-8	42234	5785	4783	158	169	4781
6040	41065	6810	5867	217	229	5865
8129-8	39669	4807	4004	114	100	4003
55865-0	36460	4345	3537	90	92	3536
6168	33946	3511	2817	62	65	2817
2781-0	31729	3619	2794	54	67	2794
22522-8	29367	4992	4105	120	115	4104
39706	28432	2684	2019	28	38	2018
22426-8	21923	3731	2927	63	75	2924
26772	11269	2570	1845	32	49	1845
24558	11106	2457	1727	25	33	1727

file name	**total words**	**unique words**	**hash 1**	**hash 2**	**hash 3**	**hash 4**
22897-8	10838	2408	1675	22	30	1675
51752	11193	2019	1371	18	20	1370
28698	10328	2046	1413	24	18	1412
59255	10086	2199	1497	16	31	1497
51268	10277	2355	1670	23	28	1670
3181-0	9709	2313	1587	19	21	1586
25035	10072	2111	1431	23	17	1431
51993	9697	2228	1526	16	31	1526
32067	9819	2099	1418	12	19	1418
32040	9309	1990	1301	14	16	1301
51687	9480	1988	1319	17	15	1319
23210-0	8744	2057	1395	21	21	1395
32078	9160	2155	1476	22	26	1476
51296	8905	1738	1121	8	10	1121
51699	8698	1876	1231	19	18	1231
21782	8020	2080	1390	11	15	1390
51129	8430	1932	1272	14	18	1272
28062	8659	1702	1085	11	17	1085
58743	8604	1769	1169	15	18	1169
32735	8072	2255	1554	17	29	1554
41562	8217	1914	1284	13	22	1284
877-0	8064	2205	1513	19	19	1513
58991	7705	2113	1439	14	24	1439
32104	7772	1923	1267	20	18	1267
32133	7719	1680	1078	10	11	1077
30029-8	7186	1623	994	12	14	994
22662-8	7302	2098	1402	26	24	1402
23942-8	7290	1712	1084	13	19	1083
32077	7173	1799	1152	12	17	1152
50877	7033	1761	1132	16	23	1132
58735	6507	1668	1010	17	10	1008
31840	6869	1606	999	5	12	999
29503	7005	1525	936	11	7	936
51603	6954	1446	849	7	8	848
51493	6267	1697	1056	9	17	1056
42664	6325	1602	1025	9	13	1025
58995-8	6369	1479	907	10	10	907
1982-0	3026	711	276	1	1	276
29750	5951	1523	913	6	11	913
59368	5908	1489	889	8	11	889
28650	6050	1428	875	6	7	875
51498	5572	1498	880	8	8	880
29618	5489	1417	834	7	7	834
51008	5551	1380	811	4	7	811
9205	5474	1629	1008	11	15	1008
30044	5141	1282	708	6	12	707
32347	4594	1147	605	4	9	604
23099	4354	1193	625	5	6	625
	total:	598146	511855	32652	34202	511290
	average:	5922.237624	5067.871287	323.2871287	338.6336634	5062.277228
		percentage	85.57358906	5.458867902	5.718001959	85.47913051