Assignment 04: Hashing

Jackson Gregory

gregoryj17@asmsa.org

Goals

- Create a hash table so that lookups could be performed in O(1) average lookup time
- Use the table to determine the 100 most common words from a book

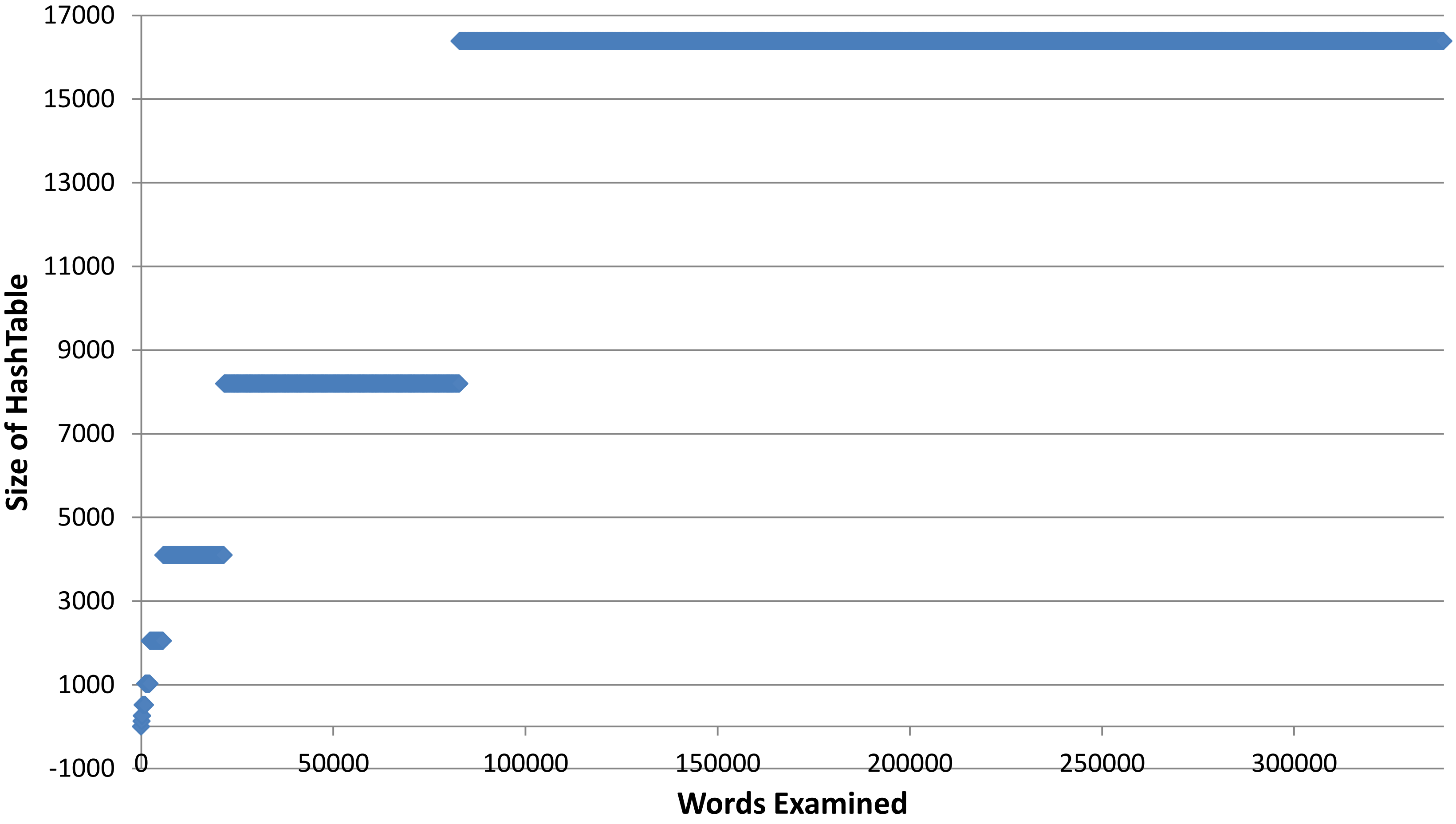
Approach

- Store keys and values
- Upsize when load is over 75%
- To store common words, store words as keys and count as values
- Increment count of existing words, add new words to table

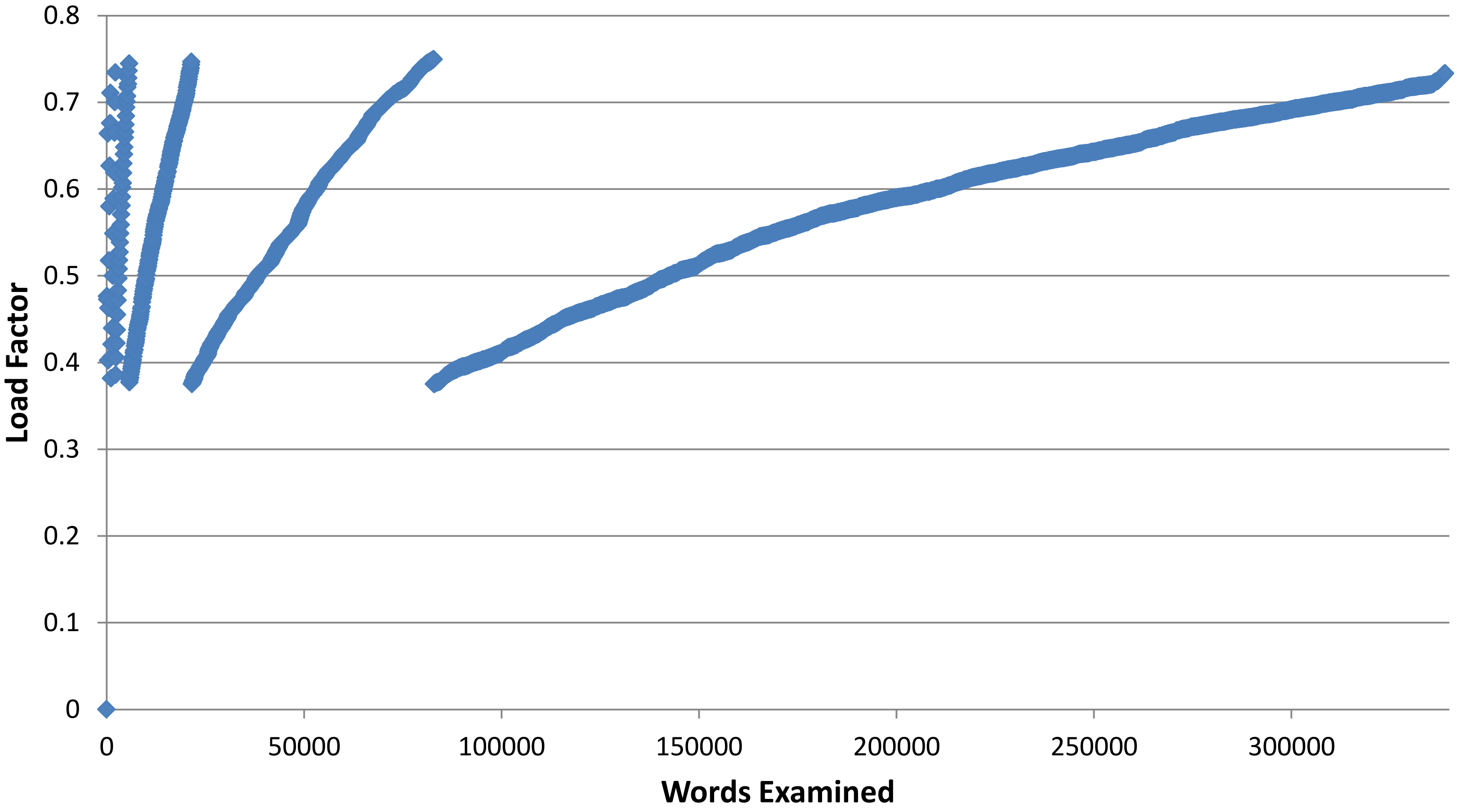
Book Used

- “Varney the Vampire; Or, the Feast of Blood”
- ~339000 words
- ~12000 unique words

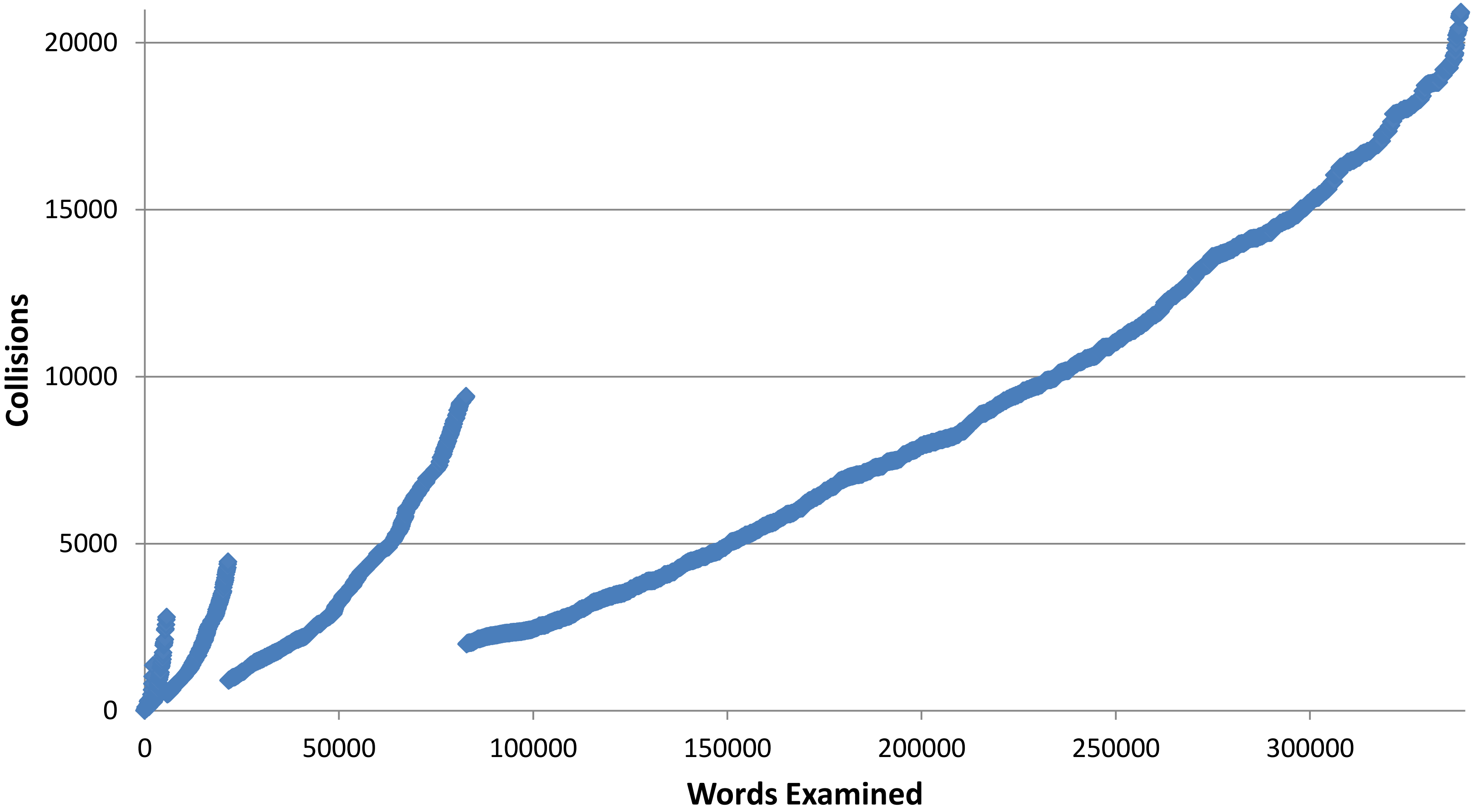
Hash Table Capacity



Load Factor

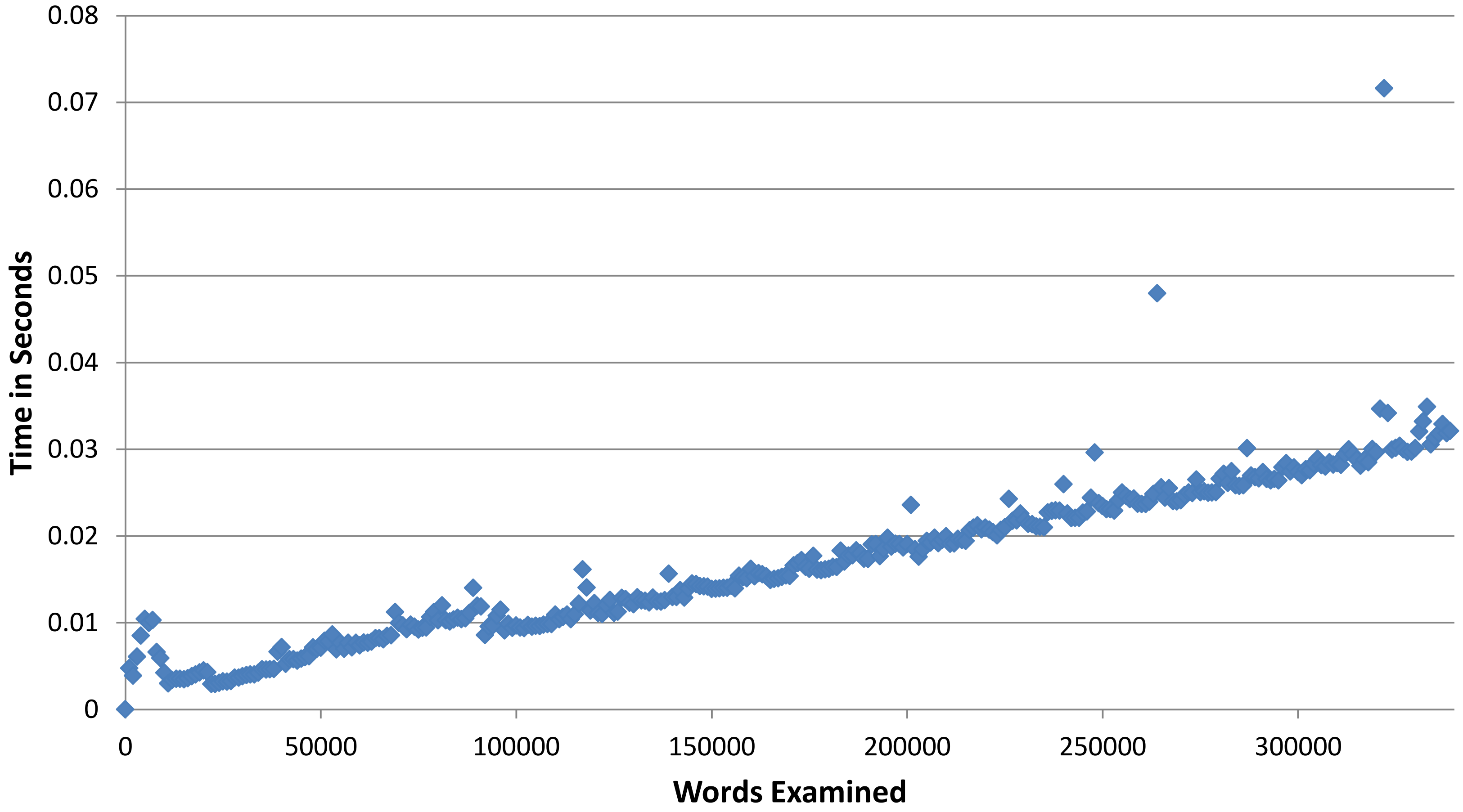

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Load Factor
0
50000
100000
150000
200000
250000
300000
Words Examined

Collisions



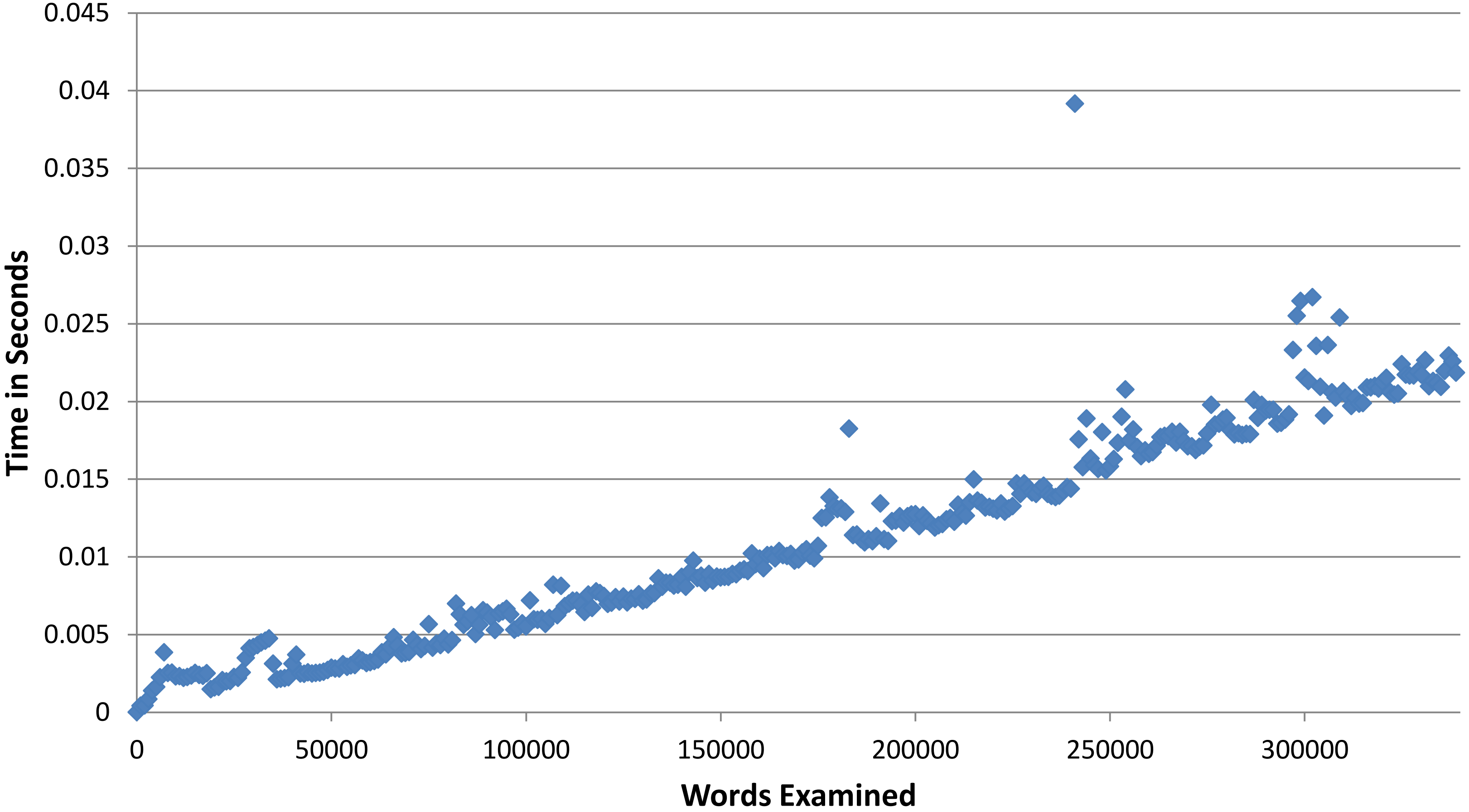
Collisions
20000
15000
10000
5000
0
0
50000
100000
150000
200000
250000
300000
Words Examined

Time



0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Time in Seconds
0
50000
100000
150000
200000
250000
300000
Words Examined

Reference Implementation Time



Most Common Words

100-91	90-81	80-71	70-61	60-51	50-41	40-31	30-21	20-11	10-1
marchdale	should	out	may	could	were	do	me	for	it
bannerworth	did	much	charles	varney	on	my	what	said	he
jack	francis	up	more	or	who	there	all	is	that
house	vampyre	has	admiral	an	would	we	which	his	you
yes	here	mr	when	such	then	this	so	have	i
their	time	flora	she	henry	some	no	him	not	a
into	am	man	can	are	now	one	at	be	of
shall	say	know	your	been	upon	will	with	as	to
see	than	them	her	well	if	from	had	was	and
about	come	very	any	sir	by	they	but	in	the

List as .csv

- The list of the top 100 words is available as a .csv file, which includes the frequency of each word, at the following link: https://github.com/gregoryj17/CP3-Stuff/blob/master/A04/top100.csv

Reflections

- Although my hash table scales as the same rate as Java’s HashMap, it is a constant factor slower. Although the difference is small, it adds up. This shows the table could be further optimized for lowered time complexity
- The count for words differs slightly depending on the way plaintext is obtained. Different spacing on imports may lead to slightly different results, if words are affected

Assignment 04: Hashing

Jackson Gregory

gregoryj17@asmsa.org