Part 1

1) How long does it take using the provided, brute force code to generate order-5 text using Romeo and Juliet (romeo.txt) and generating 100, 200, 400, 800, and 1600 random characters. Do these timings change substantially when using order-1 or order-10 Markov models? Why?

romeo.txt		Random characters				
153,000 characters		100	200	400	800	1600
Time	order-5	0.093	0.172	0.266	0.484	0.985
	order-1	0.156	0.25	0.438	0.813	1.562
	order-10	0.078	0.172	0.282	0.516	0.937

When using order-10, the timings do not change substantially, but using order-1, the timings increase substantially. Because when a random string containing 5 or 10 char is generated, the occurrence of this gram is at similar frequency. However, order-1 will have higher frequency, and the storage of next letter takes longer time.

2) Romeo has roughly 153,000 characters. Hawthorne's Scarlet Letter contains roughly 500,000 characters. How long do you expect the brute force code to take to generate order-5 text when trained on hathorne.txt given the timings you observe for romeo when generating 400, 800, 1600 random characters? Do empirical results match what you think? How long do you think it will take to generate 1600 random characters using an order-5 Markov model when the King James Bible is used as the training text --- our online copy of this text contains roughly 4.4 million characters. Justify your answer -- don't test empirically, use reasoning.

hawthorne.txt	Random characters		
500,000 characters Order-5	400	800	1600
expect Time	0.869	1.58	3.21
empirical Time	0.672	1.25	2.594

Since in hawthorne, a file containing 500,000 words, it takes 2.594 to generate 1600 characters using order-5 Markov model, thus, in King James Bible, which contains 4.4 million characters, it would takes (4,400,000/500,000)=8.8 folds of the time, which is 8.8*2.594=22.827.

3) Provide timings using your Map/Smart model for both creating the map and generating 200, 400, 800, and 1600 character random texts with an order-5 Model and romeo.txt. Provide some explanation for the timings you observe.

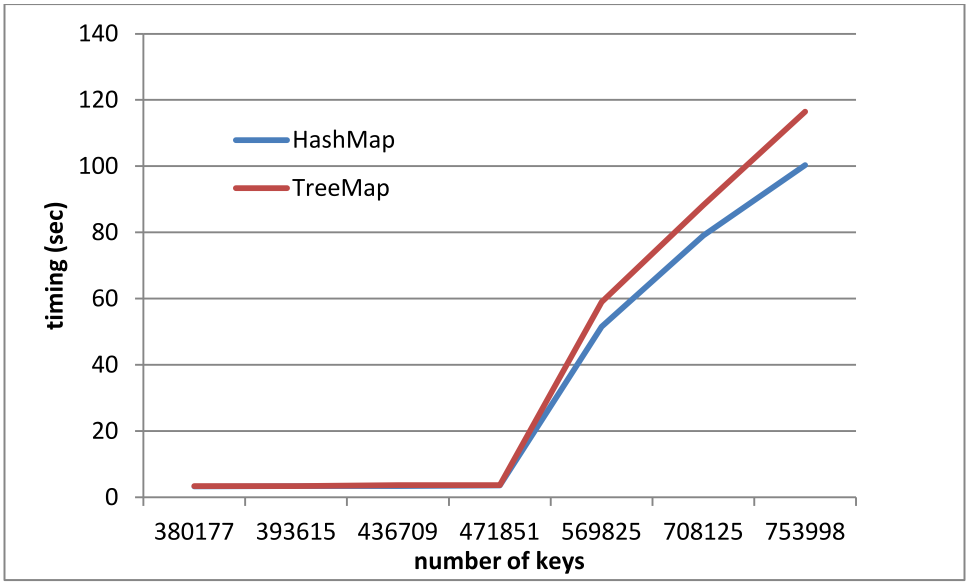
romeo.txt		Random characters				
153,000 characters		200 (first run)	200	400	800	1600
Time	order-5 MapMarkov	0.36	0	0	0	0.015

In Map/smart model, the first run takes longer because it is generating the Map. However, after the Map is constructed, as long as the order number (k value) does not change, the text can be generated directly from the Map without looping through the whole string, so it saves time.

Part 2

In analysis of the performance of WordMarkovModel using a HashMap and a TreeMap, big files such as kjv10 (containing 753998 keys) were chosen. Order-5 and 200 random words generation were used for all the test runs. In order to make the analysis more representative, each file was run three times.

When the number of keys is below 50,000, the performance by both HashMap and TreeMap are very stable, and are similar in time. However, as the number of keys goes above 50,000, the running time for both model increase dram atically. Also, TreeMap becomes slower compared to HashMap.



number of keys	HashMap			Ave Time	Tree Map			Ave Time
380177	3.281	3.344	3.328	3.317667	3.438	3.469	3.453	3.453333
393615	3.328	3.453	3.406	3.395667	3.469	3.438	3.437	3.448
436709	3.406	3.375	3.328	3.369667	3.437	4.25	3.453	3.713333
471851	3.359	3.297	4.109	3.588333	3.437	4.329	3.453	3.739667
569825	51.766	50.719	52.094	51.52633	58.343	59.016	59.609	58.98933
708125	76.922	82.109	78.25	79.09367	88.047	87.719	88.953	88.23967
753998	101.063	100.187	99.813	100.3543	115.969	116.312	117.062	116.4477