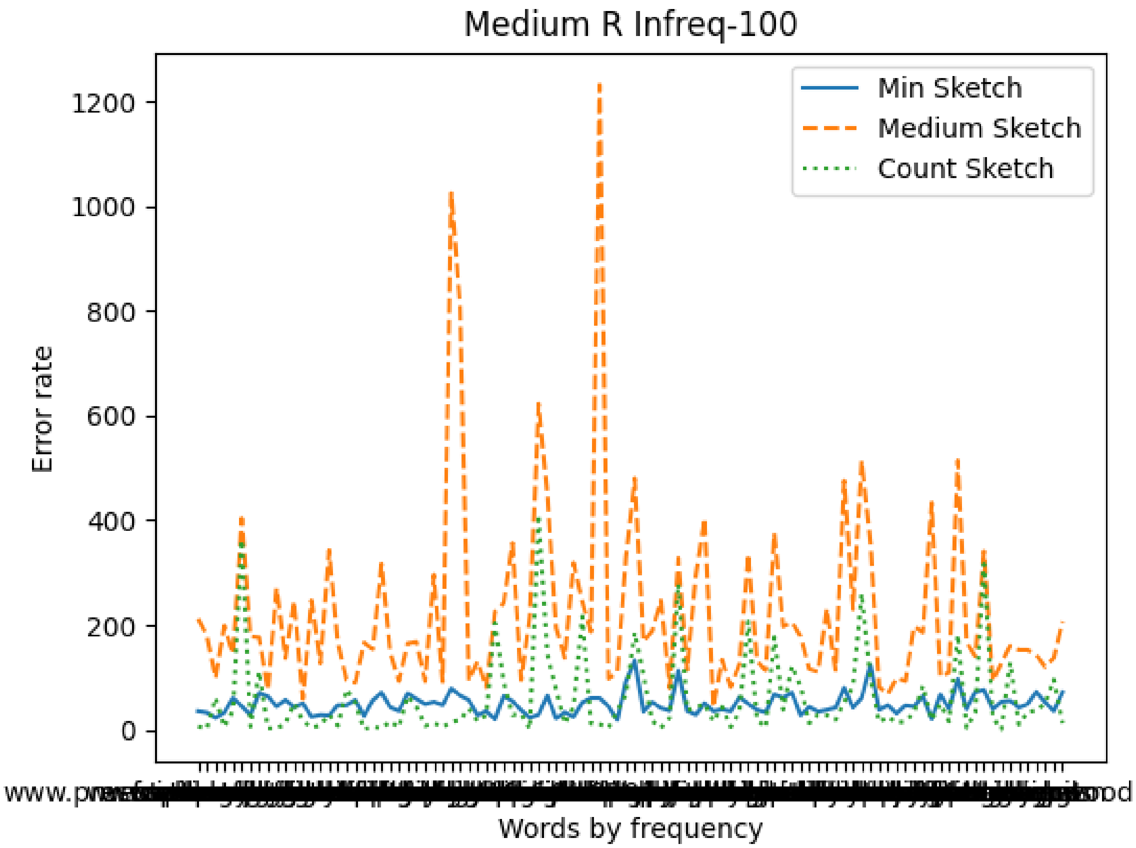
Report on minsketch and count sketch

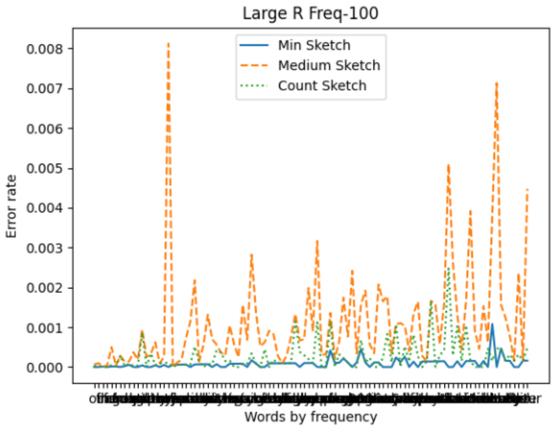
Tianling Feng

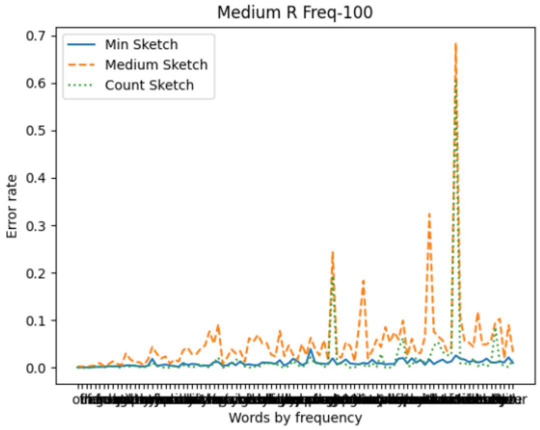
Due to the limited space, only several plots that represent some trends will be included.

The first conclusion is that count sketch is more accurate than min sketch, while the medium sketch has the worst error rate. As we can see from the graph below, when dealing with the medium range of R and evaluate based on infrequent 100 words, count sketch will have the lowest error rate for most of the time. And the medium sketch will have the highest error rates among all three approaches.

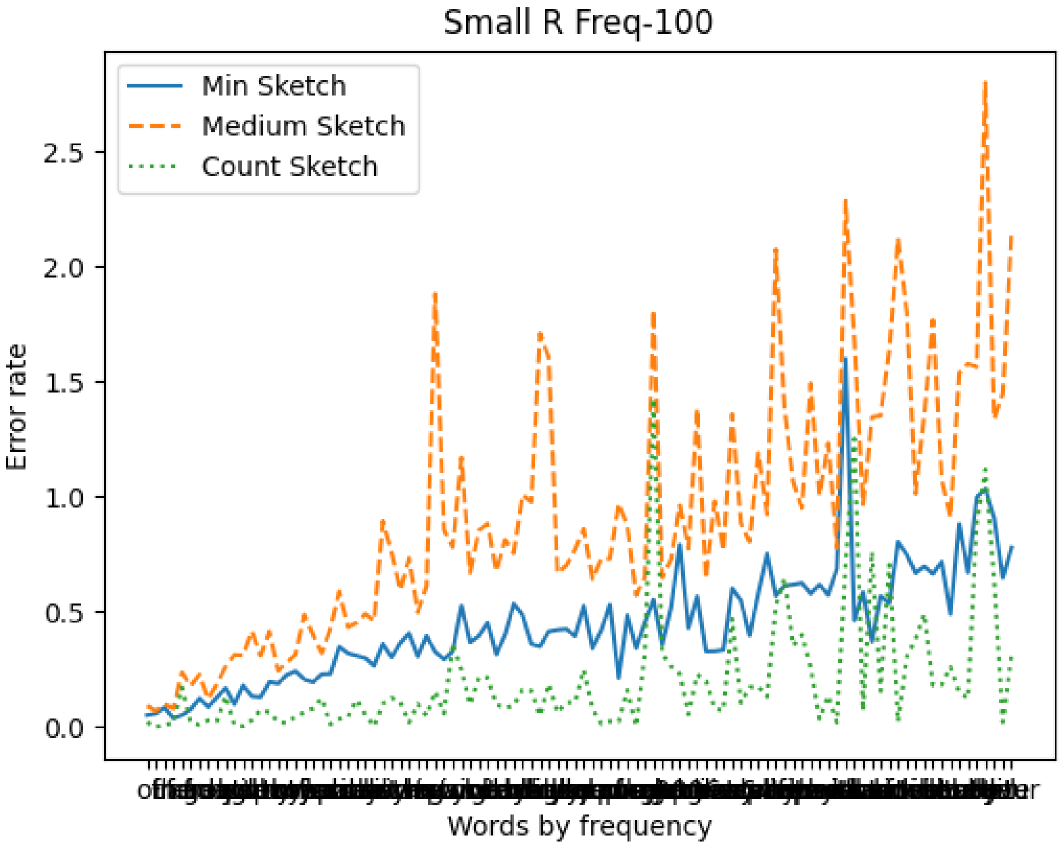


The second conclusion is that the error rate will become lower as the range of hash function grows. As we can see from the two graphs below, when R grows from 2^14 to 2^18, the error rate decreases accordingly from 0.7 to 0.007 (about 100 times). So there is the tradeoff between the space used and the accuracy of the algorithm.



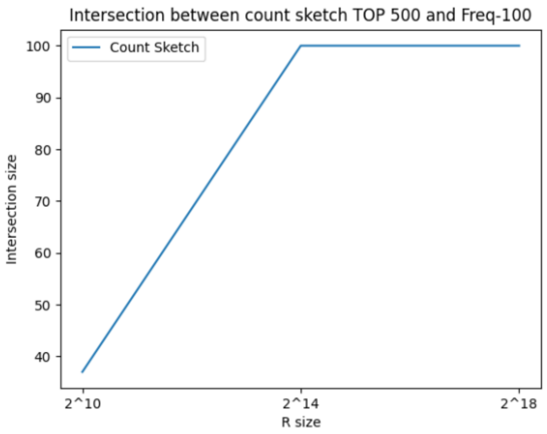


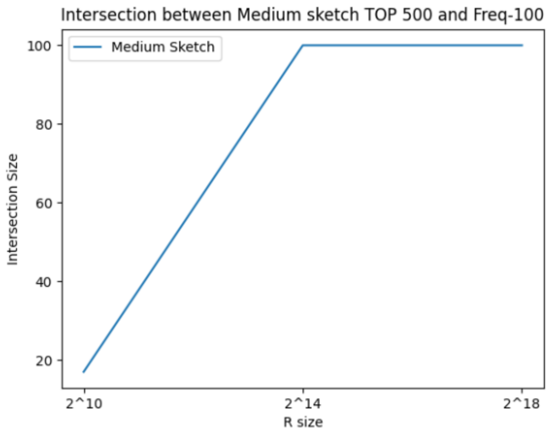
The third conclusion is that the error rate is lower for words with higher counts than for words with lower counts. The following graph (x-axis is organized in a descending order of word frequencies) clearly shows this trend. It means that the accuracy for heavy hitters is higher.

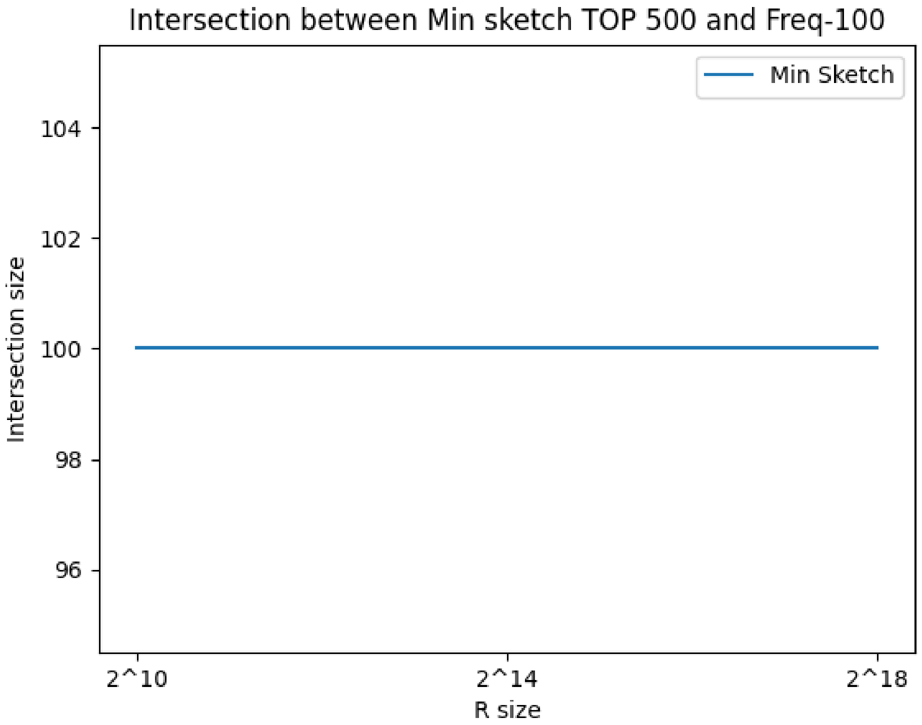


The final conclusion is that the intersection between actual frequent 100 queries and the top is relatively high (especially for size of R above 2^14). For all three methods, the frequent

100 queries can be found in their heap of 500 queries (for size of R above 2^14). Overall, the medium still has the worst accuracy among all three methods.


Intersection between count sketch TOP 500 and Freq-100
Count Sketch
Intersection size
2^10
2^14
2^18
R size


Intersection between Medium sketch TOP 500 and Freq-100
Medium Sketch
Intersection Size
2^10
2^14
2^18
R size


Intersection between Min sketch TOP 500 and Freq-100
Min Sketch
Intersection size
2^10
2^14
2^18
R size