Program Structures & Algorithms FALL 2021
Final Project Report

Jingru Xiang(001586653) Yutong Wang (001530602)

- **Introduction**

  In the final project, we have implemented MSD radix sort for Chinese name sorting. We accomplish the sorting by converting Chinese, which uses Unicode characters, to pinyin with help of external library pinyin4j, then sorting by pinyin order and converting back to the corresponding Chinese name. We apply the same converting to LSD radix sort, Dual-pivot quicksort, and Huskysort, then benchmark these four sorts and compare their performance. As the sorting was done after the names were converted to pinyin, and each pinyin name performed as a string, there are two phenomena in the final name list. Firstly, different Last names may appear together as they have the same pinyin, the order depends on the order they appear in shuffledChinese.txt. Secondly, the final order is strictly based on the pinyin name as a whole instead of separated last name and first name, for instance, the rank of "Guo Jianhua" is larger than the rank of "Gu Fangfang" because "o" is bigger than "f", not because "J" is bigger than "f".

- **Sorting Comparison**

  We compare the performance of sorting Chinese in MSD radix sort, LSD radix sort, dual-pivot quicksort, and huskysort by doing benchmark. The benchmark process contains three parts: converting Chinese to pinyin, sorting pinyin with a given sort method, and converting sorted pinyin to Chinese, which records the total time of the Chinese sorting process. As converting processes mostly have a linear cost factor, the difference in time results is mainly due to the different efficiency of the sorting methods.
  It should be noted that all the reported experiments were run on an 8 GB 2133 MHz RAM running macOS.

We benchmark four sorting methods with 250K, 500K 1M, 2M, and 4M names.

	MSD radix sort	LSD radix sort	Dual-pivot quick sort	Husky sort
250K	741.8	1160.5	755.0	712.3
500K	1365.4	2361.2	1523.7	1325.8
1M	2873.4	4437.6	3214.7	2722.9
2M	6164.0	10015.1	6472.6	5779.2
4M	13173.0	23073.7	14018.8	11677.7

Comparing the performance in line chart:

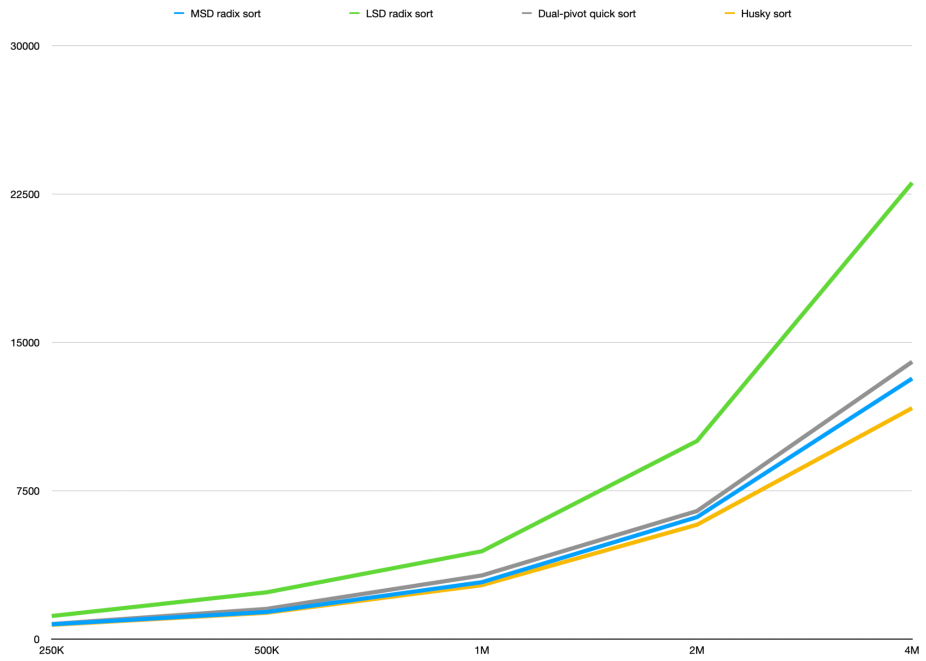
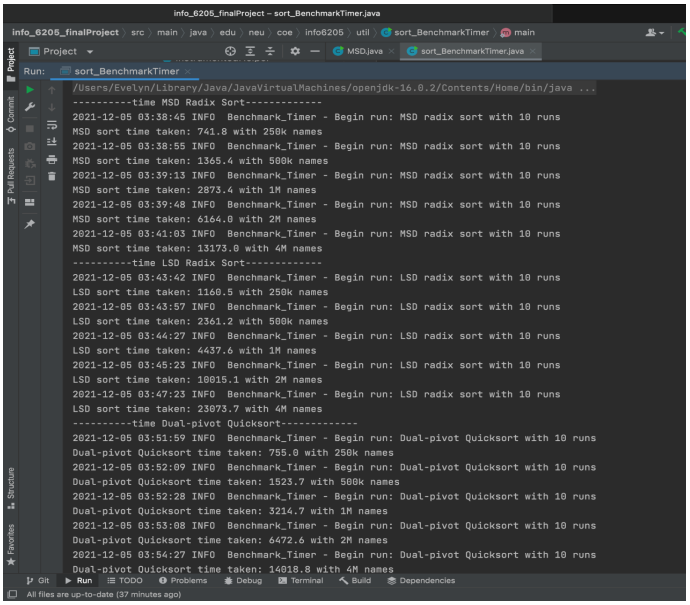


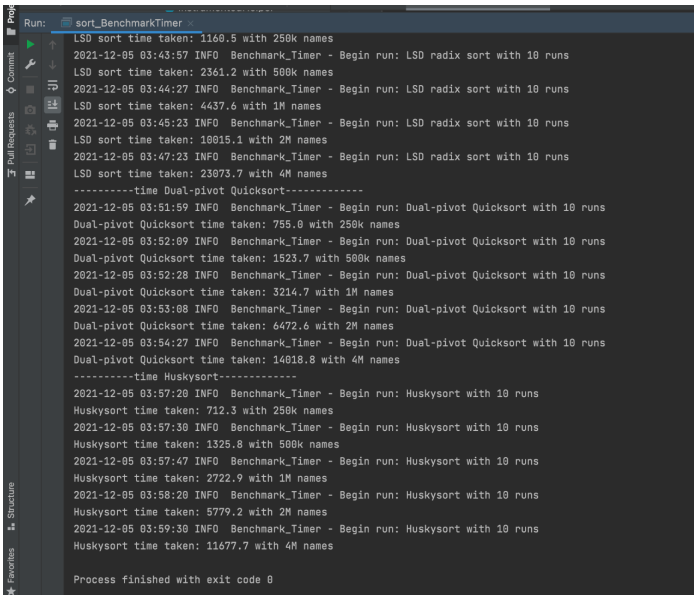
figure 1

From the graphs we can say that the huskysort is the fastest algorithm for Chinese sorting, then MSD radix sort, dual-pivot quicksort, and the LSD radix sort. Through theoretical study, we know that the time taken of LSD radix sort is O(N*M), M is the length of the longest string, in both the best and worst case. The time-taken of MSD radix sort O(N) in the best case and O(N*M) in the worst case. MSD would be more efficient as it does not need to go through every letter in each word[1]. However, a large number of elements may lower the efficiency of radix sort[2]. The time taken of quicksort is O(2NlgN) in the average case and the dual-pivot quicksort would be more efficient than this as it uses two pivots instead of one[3]. The average

time taken of Husky sort is O(NlgN) and it should always be faster than dual-pivot quicksort[4].
The test result can be explained through theoretical information. LSD radix sort is the most inefficient method used to sort Chinese because it has to go through every element in each work, dual-pivot quick sort improves a lot as it does not need to go through each element. MSD is faster because the key-indexed counting method saves the comparison time and not each element in names need to be considered. Both MSD and huskysort perform well when sorting string, however, as the dataset gets larger, the efficiency of MSD becomes lower than the huskysort.

Benchmark screenshot:

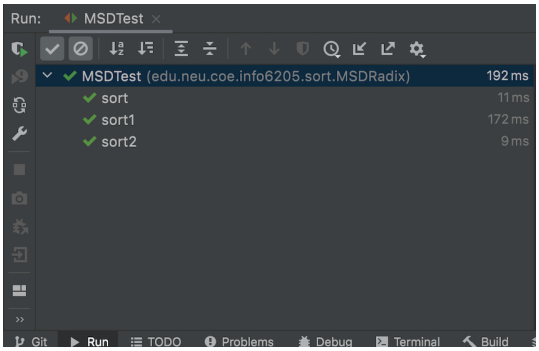


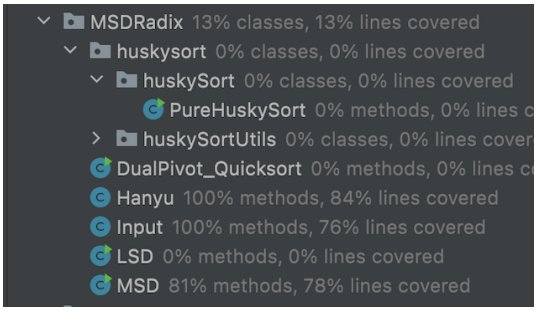


- **Unit test**

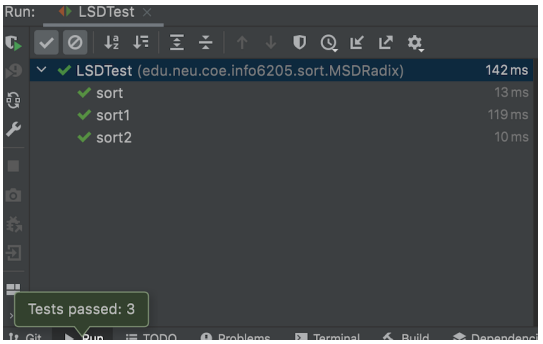
  We want to test if four algorithms can sort given Chinese names as we expected, so we provided three different lists to check if they can successfully convert Chinese to pinyin, sort pinyin in the correct order, and covert back to China.

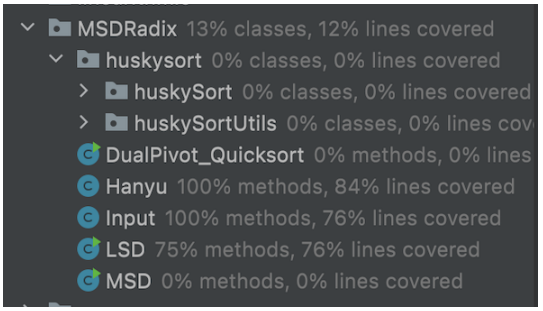
  MSD radix sort



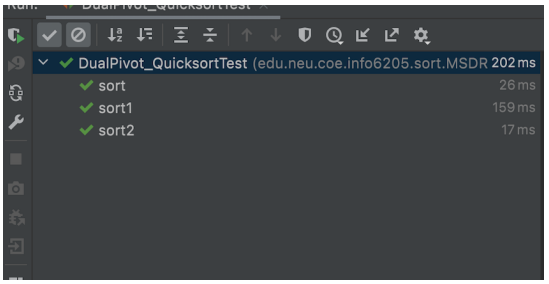


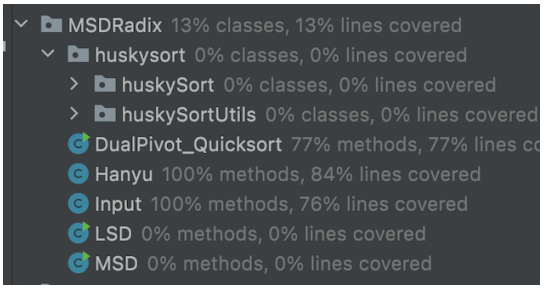
  LSD radix sort



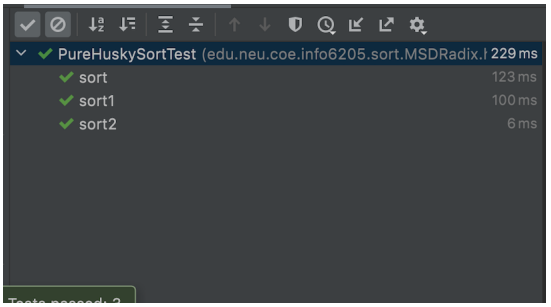


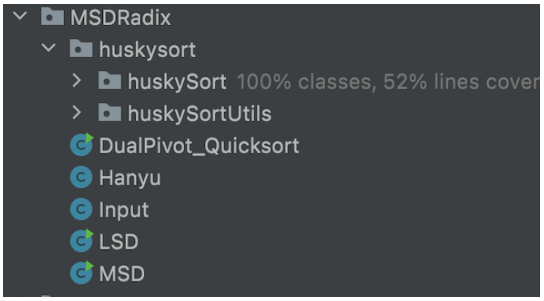
  Dual-pivot quicksort





  Huskysort





- **Conclusion**

  The efficiency of four sorting methods sorting Chinese in pinyin order from the lowest to the highest is LSD radix sort, Dual-pivot quicksort, MSD radix sort, and huskysort.

- **Recommendation**

  Based on the experiment results, we would recommend using either MSD radix sort or huskysort when sorting a small number of strings and using huskysort if the list gets very large.

- **References:**

[1] MSD(Most Significant Digit) Radix Sort. (2021, August 11). Retrieved from https://www.geeksforgeeks.org/msd-most-significant-digit-radix-sort/

[2] Shukla, A., Saxena, A.K., Baruah, P.K., Bandyopadhyay, S., Sahni, S., Jiménez-González, D., Navarro, J.J., Larrba-Pey, J., Cederman, D., & Tsigas, P. (2012). ' Review of Radix Sort & Proposed Modified Radix Sort for Heterogeneous Data Set in Distributed Computing Environment '.

[3] QuickSort. (2021, August 10). Retrieved from https://www.geeksforgeeks.org/quick-sort/

[4] Hillyard, R.C., Liaozheng, Y., & SaiVineethK., R. (2020). Huskysort. *ArXiv, abs/2012.00866*.