**Program Structures & Algorithms**

**Final Project Report**

**Sorting’s for a natural language using Unicode Characters**

Santhosh Maddi (002127881)

Sai Saketh Kosuri (002199401)

Eswar Sai Nath Adusumalli (001565578)

*Information Systems & Software Engineering Systems, College of Engineering, Northeastern University, Boston*

Technical report

December 2021

# TASK

# To implement MSD radix sort for a natural language which uses Unicode characters.

# To compare your method with Timsort, Dual-pivot Quicksort, Husky sort, and LSD radix sort.

# IMPLEMENTATIONS

* Implemented MSD radix sort, LSD radix sort, Tim sort, Dual pivot sort, Husky sort for Telugu language (Unicode Range: 3072 - 3199)
* Implemented MSD and LSD radix sorts for Chinese words with comparison using pinyin conversion.
* Implemented Tim sort, Dual pivot quicksort and Husky sort for Chinese words using comparator both Collator and comparator with pinyin conversion.
* Implemented unit test cases for all sort implementations.
* Performed benchmark tests for all sorts with different types of inputs [random, reverse sorted, sorted and partial sorted] and array sizes {250k, 500k, 1M, 2M, 4M}
* Compared performance of non-comparison sorts of Chinese words with collator and pinyin Comparator implementations of comparator.

# FINDINGS:

* Irrespective of the type of the input array, husky sort performs the same
* Dual pivot sort has the worst performance for sorted and reverse sorted order, shuffling before the sorting will help
* For MSD radix sorts, using hybrid implementation i.e., switching to insertion sort with a cut off value improves the performance.
* For non-comparison sorts, collator-based comparators have better performance compared to the pinyin comparators.

**Findings for Telugu language words:**

1. Plotted Graph between time taken in milliseconds and random ordered inputs of different sizes {250000, 500000, 1000000, 2000000, 4000000} for MSD, LSD, Dual pivot quick sort, Tim sort and Husky sort.

Fig 1: Graph showing time elapsed in milli seconds for varying sizes of random ordered Telugu words in MSD, LSD, Dual pivot quick, Tim and Husky sorts.

1. Plotted graph between time taken in milliseconds and Reverse sorted inputs of different sizes{250000, 500000, 1000000, 2000000, 4000000} for MSD, LSD, Dual pivot quick sort, Tim sort and Husky sort.

Fig 2: Graph showing time elapsed in milli seconds for varying sizes of reverse sorted Telugu words in MSD, LSD, Dual pivot quick, Tim and Husky sorts.

1. Plotted Graph between time taken in milliseconds and sorted inputs of different sizes {250000, 500000, 1000000, 2000000, 4000000} for MSD, LSD, Dual pivot quick sort, Tim sort and Husky sort.

Fig 3: Graph showing time elapsed in milli seconds for varying sizes of sorted Telugu words in MSD, LSD, Dual pivot quick, Tim and Husky sorts.

1. Plotted Graph between time taken in milliseconds and partially sorted of different sizes {250000, 500000, 1000000, 2000000, 4000000} for MSD, LSD, Dual pivot quick sort, Tim sort and Husky sort.

Fig 4: Graph showing time elapsed in milli seconds for varying sizes of partially sorted Telugu words in MSD, LSD, Dual pivot quick, Tim and Husky sorts.

**Findings for Chinese language words:**

1. Plotted Graph between time taken in milliseconds and random ordered inputs of different sizes {250000, 500000, 1000000, 2000000, 4000000} for MSD, LSD, Dual pivot quick sort, Tim sort and Husky sort.

Fig 5: Graph showing time elapsed in milli seconds for varying sizes of random ordered Chinese words in MSD, LSD, Dual pivot quick, Tim and Husky sorts.

1. Plotted Graph between time taken in milliseconds and reverse sorted inputs of different sizes {250000, 500000, 1000000, 2000000, 4000000} for MSD, LSD, Dual pivot quick sort, Tim sort and Husky sort.

Fig 6: Graph showing time elapsed in milli seconds for varying sizes of reverse ordered Chinese words in MSD, LSD, Dual pivot quick, Tim and Husky sorts.

1. Plotted Graph between time taken in milliseconds and sorted inputs of different sizes {250000, 500000, 1000000, 2000000, 4000000} for MSD, LSD, Dual pivot quick sort, Tim sort and Husky sort.

Fig 7: Graph showing time elapsed in milli seconds for varying sizes of sorted Chinese words in MSD, LSD, Dual pivot quick, Tim and Husky sorts.

1. Plotted Graph between time taken in milliseconds and partially sorted inputs of different sizes {250000, 500000, 1000000, 2000000, 4000000} for MSD, LSD, Dual pivot quick sort, Tim sort and Husky sort.

Fig 8: Graph showing time elapsed in milli seconds for varying sizes of partially sorted Chinese words in MSD, LSD, Dual pivot quick, Tim and Husky sorts.

# RELATIONSHIP CONCLUSION:

# Non comparison based sort performance is better for strings than comparison based sorts.

# For radix sort the performance significantly depends on radix range.

# Hybrid sorts with cut off values have better performance.

# OUTPUT TEST CASE SCREENSHOTS:

# Dual Pivot Quick Sort:

# Text Description automatically generated

# Tim Sort test case output for Chinese and Telugu words

# Text Description automatically generated

# LSD radix sort test case output for Telugu words

# A screenshot of a computer Description automatically generated

# MSD radix sort test case output for Telugu words

# Text Description automatically generated

# LSD radix sort test case output for Chinese words

# Text Description automatically generated

# MSD radix sort test case output for Chinese words

# Text Description automatically generated