COL761 Assignment1

Contribution:
Chapala Sriram Varma(2019CS50426) - 33.33%
Somisetty Harsha Vardhan(2020CS10390) - 33.33%
Vonteri Harshith Reddy(2019CS50450) - 33.33%

Algorithm:
1. On the given transactional database we have built FP-Tree to mine frequent item sets.(We have taken code for mining frequent items from here: https://github.com/VNSAditya02/COL761-Data-Mining/tree/main/HW1)
2. We find the frequencies of all distinct items, then considered items with highest frequency as frequent items based on a heuristic developed using frequency distribution, discarded other items. We modelled support using this heuristic.
3. After finding the support, we mine frequent item sets in the given transactional database, and assigned new symbol to frequent item set. We sorted each transaction based on frequency of items and reduced frequent items to new symbols assigned to them by maximum matching.
4. We repeated the above process for 15 iterations wherein each iteration we operate on the latest compressed file.

Results:
1. D_small.dat:
   a. Integers in original file: 118252
   b. Integers in compressed file: 36313
   c. Compression Ratio (%): 69.29%
   d. Runtime: 4 sec

2. D_medium2.dat:
   a. Integers in original file: 3960507
   b. Integers in compressed file: 2742053
   c. Compression Ratio (%): 30.76%
   d. Runtime: 320sec

3. D_medium.dat:
   a. Integers in original file: 8019015
   b. Integers in compressed file: 6152831
   c. Compression Ratio (%): 23.27%
   d. Runtime: 393sec

4. D_medium.dat:
   a. Integers in original file: 109360594
   b. Integers in compressed file: 99025304
   c. Compression Ratio (%): 9.45%
   d. Runtime: 1hour