4 Different Experimental Scenarios

Different experimental scenarios are used to check the performance of our system.

Text Data: The data has only text values having some duplicate text data.

Numeric Data: The input data consists of numerical values with some duplicate numerical values.

Text and Numeric Data: The data is the combination of text and numerical values.

Text, Numeric, and Special characters: In this data file, there are text data, numeric data, and some special characters with duplicate data also.

For testing the algorithm for larger datasets, files with varying numbers of words from 10 to 23,598 including all data types were used.

5 Experimental Results

Our experiments were performed on jdk 1.5.0 installed on 2.1 GHz Intel Core i5 processor and Windows 7 Operating System.

5.1 Text Data

For testing the algorithm with text data, following 18 text words with four duplicates were tested and the screenshot displayed i.e. Fig. 2 shows that all the four duplicates are removed from original data.

Tarun ashish mohan tarun Amit Mehta Jangid ram atul Amit Lal
shyam mohan kumar Babu verma purbey Atul

The total number of words before cleaning was 18 in which four words were duplicate ones. After applying our method, the result shows excellent result by removing duplicate words in the data.

5.2 Numeric Data

For testing the algorithm with numerical data, the string of following 20 numeric values with three duplicates was fed as input to the algorithm and the results are shown in Fig. 3.