Big Data Frameworks CSE3120

Lab – 2 Data Cleaning Experiment

Name: Naveen Nidadavolu
Roll No: 22MIA1049

Aim: To clean two CSV files by performing the following tasks:

1. Remove any unnecessary whitespace around the data.
2. Replace missing, `null`, or `n/a` values with a default value (`N/A`).
3. Remove duplicate rows (if a row is repeated in the file, only the first occurrence will be kept).
4. Save the cleaned data into new output files.

Algorithm

1. **Input**: Accept two CSV files as input.
2. **Open Files**: Read each file line by line.
3. **Process Header**: Retain the first line (header) as is and write it to the output file.
4. **Process Data Rows**:
5. For each row (after the header), clean it by:
    a. Removing leading and trailing whitespace from each column.
    b. Replacing any "null", "n/a", or empty values with `"N/A"`.
6. If the cleaned row has not been seen before (using a `HashSet` to track unique rows), write it to the output file.
7. **Output**: Write the cleaned data to new CSV files.
8. **End**: Save and close the files after processing.

Program:

```java
import java.io.*;
import java.util.HashSet;
public class EnhancedCSVDataCleaner {
    public static void main(String[] args) {
        String inputFilePath1 = "Table 1.csv";  // First input file
        String outputFilePath1 = "Cleaned_Table1.csv";  // First output file
        String inputFilePath2 = "Table 2.csv";  // Second input file
        String outputFilePath2 = "Cleaned_Table2.csv";  // Second output file
        try {
            cleanCSV(inputFilePath1, outputFilePath1);
            cleanCSV(inputFilePath2, outputFilePath2);
            System.out.println("CSV cleaning completed. Cleaned files saved as: " + outputFilePath1 + " and " + outputFilePath2);
        } catch (IOException e) {
            System.err.println("An error occurred while processing the files: " + e.getMessage());
        }
    }
    public static void cleanCSV(String inputFilePath, String outputFilePath) throws IOException {
        try (BufferedReader reader = new BufferedReader(new FileReader(inputFilePath));
             BufferedWriter writer = new BufferedWriter(new FileWriter(outputFilePath))) {
            HashSet<String> uniqueRows = new HashSet<>();
            String line;
            boolean isHeader = true;
            while ((line = reader.readLine()) != null) {
                if (isHeader) {
                    // Write header row as-is
                    writer.write(line);
                    writer.newLine();
                    isHeader = false;
                } else {
                    // Clean and deduplicate the rows
                    String cleanedRow = cleanRow(line);
                    if (uniqueRows.add(cleanedRow)) {
                        writer.write(cleanedRow);
                        writer.newLine();
                    }
                }
            }
        }
    }

    public static String cleanRow(String row) {
        String[] columns = row.split(regex:",");
        StringBuilder cleanedRow = new StringBuilder();

        for (int i = 0; i < columns.length; i++) {
            String column = columns[i].trim(); // Remove leading and trailing whitespace

            // Handle missing, null, or invalid data
            if (column.isEmpty() || column.equalsIgnoreCase(anotherString:"null") || column.equalsIgnoreCase(anotherString:"n/a")) {
                column = "N/A";
            }

            cleanedRow.append(column);

            // Add a comma between columns, but not after the last column
            if (i < columns.length - 1) {
                cleanedRow.append(str:",");
            }
        }

        return cleanedRow.toString();
    }
}
```

Output:

```
CSV cleaning completed. Cleaned files saved as: Cleaned_Table1.csv and Cleaned_Table2.csv

c:\Users\nvkaj\OneDrive - vit.ac.in\sem 6\BigData Frameworks\bigdata_java>
```

Cleaned Table1:

Cleaned_Table1.csv > data

```
Order ID,Order Date,Order Quantity,Sales,Ship Mode,Profit,Unit Price,Customer Name,Customer Segment,Product Category
3,10/13/2010,6,261.54,Regular Air,-213.25,38.94,Muhammed MacIntyre,Small Business,Office Supplies
6,2/20/2012,2,6.93,Regular Air,-4.64,2.08,Ruben Dartt,Corporate,N/A
32,7/15/2011,26,2808.08,Regular Air,1054.82,107.53,Liz Pelletier,Corporate,N/A
32,7/15/2011,24,1761.4,Delivery Truck,-1748.56,70.89,Liz Pelletier,Corporate,N/A
32,7/15/2011,23,160.2335,Regular Air,-85.13,7.99,Liz Pelletier,Corporate,N/A
32,7/15/2011,15,140.56,Regular Air,-128.38,8.46,Liz Pelletier,Corporate,N/A
35,10/22/2011,30,288.56,Regular Air,60.72,9.11,Julie Creighton,Corporate,N/A
35,10/22/2011,14,1892.848,Regular Air,48.99,155.99,Julie Creighton,Corporate,N/A
36,11/2/2011,46,2484.7455,Regular Air,657.48,65.99,Sample Company A,Home Office,Technology
65,3/17/2011,32,3812.73,Regular Air,1470.3,115.79,Tamara Dahlen,Corporate,Technology
32,7/15/2008,26,N/A,Regular Air,N/A,107.53,Liz Pelletier,Corporate,Furniture
32,7/15/2008,24,N/A,Delivery Truck,N/A,70.89,Liz Pelletier,Corporate,Furniture
32,7/15/2008,23,N/A,Regular Air,N/A,7.99,Liz Pelletier,Corporate,Technology
32,7/15/2008,15,N/A,Regular Air,N/A,8.46,Liz Pelletier,Corporate,Technology
35,10/22/2008,30,N/A,Regular Air,N/A,9.11,Julie Creighton,Corporate,Office Supplies
35,10/22/2008,14,N/A,Regular Air,N/A,155.99,Julie Creighton,Corporate,Technology
36,10/22/2008,46,N/A,Regular Air,N/A,65.99,Sample Company A,Home Office,Technology
65,10/22/2008,32,N/A,Regular Air,N/A,115.79,Tamara Dahlen,Corporate,Technology
66,1/19/2009,41,108.15,Regular Air,N/A,2.88,Arthur Gainer,Consumer,Office Supplies
69,6/3/2009,42,1186.06,Regular Air,N/A,30.93,Jonathan Doherty,Corporate,Furniture
69,6/3/2009,28,51.53,Express Air,0.35,1.68,Jonathan Doherty,Corporate,Office Supplies
70,12/17/2010,48,90.05,Regular Air,-107,1.86,Helen Wasserman,Home Office,Office Supplies
70,12/17/2010,46,7804.53,Regular Air,2057.17,205.99,Helen Wasserman,Home Office,Technology
96,4/16/2009,37,4158.1235,Regular Air,1228.89,125.99,Keith Dawkins,Home Office,Technology
97,1/28/2010,26,75.57,Regular Air,28.24,2.89,Craig Yedwab,Consumer,Office Supplies
129,11/18/2012,4,32.72,Regular Air,-22.59,6.48,Pauline Chand,Corporate,Office Supplies
130,5/7/2012,3,461.89,Express Air,-309.82,150.98,Roy Collins,Corporate,Technology
130,5/7/2012,29,575.11,Regular Air,71.75,18.97,Roy Collins,Corporate,Office Supplies
130,5/7/2012,23,236.46,Regular Air,-134.31,9.71,Roy Collins,Corporate,Office Supplies
132,6/10/2010,27,192.814,Regular Air,-86.2,7.99,Emily Phan,Consumer,Technology
132,6/10/2010,30,4011.65,Delivery Truck,-603.8,130.98,Emily Phan,Consumer,Furniture
134,4/30/2012,11,1132.6,Regular Air,-310.21,95.99,Michael Dominguez,Home Office,Office Supplies
135,10/20/2011,25,125.85,Regular Air,-89.25,4.98,Anne Pryor,Consumer,Technology
166,9/11/2011,10,567.936,Express Air,-126.09,65.99,Valerie Takahito,Consumer,Technology
193,8/7/2010,14,174.89,Regular Air,-37.04,12.44,Justin Hirsh,Consumer,Office Supplies
194,4/4/2012,49,329.03,Regular Air,-197.25,7.28,Maria Zettner,Corporate,Furniture
194,4/4/2012,6,20.19,Regular Air,-13.44,3.14,Maria Zettner,Corporate,Office Supplies
195,12/27/2010,34,1315.74,Regular Air,260.87,36.55,Brad Thomas,Home Office,Office Supplies
```

	A	B	C	D	E	F	G	H	I	J
1	Order ID	Order Date	Order Qua	Sales	Ship Mode	Profit	Unit Price	Customer	Customer	Product Category
2	3	10/13/2010	6	261.54	Regular Air	-213.25	38.94	Muhamme	Small Busi	Office Supplies
3	6	2/20/2012	2	6.93	Regular Air	-4.64	2.08	Ruben Dar	Corporate	N/A
4	32	7/15/2011	26	2808.08	Regular Air	1054.82	107.53	Liz Pelletie	Corporate	N/A
5	32	7/15/2011	24	1761.4	Delivery Tr	-1748.6	70.89	Liz Pelletie	Corporate	N/A
6	32	7/15/2011	23	160.234	Regular Air	-85.13	7.99	Liz Pelletie	Corporate	N/A
7	32	7/15/2011	15	140.56	Regular Air	-128.38	8.46	Liz Pelletie	Corporate	N/A
8	35	10/22/2011	30	288.56	Regular Air	60.72	9.11	Julie Creig	Corporate	N/A
9	35	10/22/2011	14	1892.85	Regular Air	48.99	155.99	Julie Creig	Corporate	N/A
10	36	11/2/2011	46	2484.75	Regular Air	657.48	65.99	Sample Cc	Home Offi	Technology
11	65	3/17/2011	32	3812.73	Regular Air	1470.3	115.79	Tamara Da	Corporate	Technology
12	32	7/15/2008	26	N/A	Regular Air	N/A	107.53	Liz Pelletie	Corporate	Furniture
13	32	7/15/2008	24	N/A	Delivery Tr	N/A	70.89	Liz Pelletie	Corporate	Furniture
14	32	7/15/2008	23	N/A	Regular Air	N/A	7.99	Liz Pelletie	Corporate	Technology
15	32	7/15/2008	15	N/A	Regular Air	N/A	8.46	Liz Pelletie	Corporate	Technology
16	35	10/22/2008	30	N/A	Regular Air	N/A	9.11	Julie Creig	Corporate	Office Supplies
17	35	10/22/2008	14	N/A	Regular Air	N/A	155.99	Julie Creig	Corporate	Technology
18	36	10/22/2008	46	N/A	Regular Air	N/A	65.99	Sample Cc	Home Offi	Technology
19	65	10/22/2008	32	N/A	Regular Air	N/A	115.79	Tamara Da	Corporate	Technology
20	66	1/19/2009	41	108.15	Regular Air	N/A	2.88	Arthur Gai	Consumer	Office Supplies
21	69	6/3/2009	42	1186.06	Regular Air	N/A	30.93	Jonathan I	Corporate	Furniture
22	69	6/3/2009	28	51.53	Express Ai	0.35	1.68	Jonathan I	Corporate	Office Supplies
23	70	12/17/2010	48	90.05	Regular Air	-107	1.86	Helen Was	Home Offi	Office Supplies
24	70	12/17/2010	46	7804.53	Regular Air	2057.17	205.99	Helen Was	Home Offi	Technology
25	96	4/16/2009	37	4158.12	Regular Air	1228.89	125.99	Keith Daw	Home Offi	Technology
26	97	1/28/2010	26	75.57	Regular Air	28.24	2.89	Craig Yedv	Consumer	Office Supplies
27	129	11/18/2012	4	32.72	Regular Air	-22.59	6.48	Pauline Ch	Corporate	Office Supplies
28	130	5/7/2012	3	461.89	Express Ai	-309.82	150.98	Roy Collin	Corporate	Technology
29	130	5/7/2012	29	575.11	Regular Air	71.75	18.97	Roy Collin	Corporate	Office Supplies
30	130	5/7/2012	23	236.46	Regular Air	-134.31	9.71	Roy Collin	Corporate	Office Supplies
31	132	6/10/2010	27	192.814	Regular Air	-86.2	7.99	Emily Phar	Consumer	Technology
32	132	6/10/2010	30	4011.65	Delivery Tr	-603.8	130.98	Emily Phar	Consumer	Furniture
33	134	4/30/2012	11	1132.6	Regular Air	-310.21	95.99	Michael Dc	Home Offi	Office Supplies

Cleaned Table2:

	A	B	C	D	E	F
1	Ship Mode	Profit	Unit Price	Shipping Cost	Customer Name	
2	Regular Air	-213.25	38.94	35	Muhammed MacIntyre	
3	Delivery Truck	457.81	208.16	68.02	Barry French	
4	Regular Air	46.71	8.69	2.99	Barry French	
5	Regular Air	1198.97	195.99	3.99	N/A	
6	Regular Air	-4.72	5.28	2.99	N/A	
7	Regular Air	782.91	39.89	3.04	N/A	
8	Regular Air	93.8	15.74	1.39	N/A	
9	Delivery Truck	440.72	100.98	26.22	N/A	
10	Regular Air	N/A	100.98	69	Sylvia Foulston	
11	Regular Air	N/A	65.99	5.26	Jim Radford	
12	Regular Air	N/A	155.99	8.99	Jim Radford	
13	Express Air	N/A	3.69	0.5	Carlos Soltero	
14	Regular Air	-5.77	4.71	0.7	Carlos Soltero	
15	Regular Air	-172.88	15.99	13.18	Carl Ludwig	
16	Regular Air	-144.55	4.89	4.93	Carl Ludwig	
17	Regular Air	5.76	2.88	0.7	Don Miller	
18	Regular Air	252.66	40.96	1.99	Jack Garza	
19	Delivery Truck	-1766.01	95.95	74.35	Julia West	
20	Regular Air	-236.27	3.89	7.01	Eugene Barchas	
21	Delivery Truck	-236.27	120.98	30	Eugene Barchas	
22	Regular Air	118.94	500.98	5.76	Eugene Barchas	
23	Delivery Truck	3424.22	500.98	26	Edward Hooks	

Cleaned_Table2

Cleaned_Table2.csv > data

```
1   Ship Mode,Profit,Unit Price,Shipping Cost,Customer Name
2   Regular Air,-213.25,38.94,35,Muhammed MacIntyre
3   Delivery Truck,457.81,208.16,68.02,Barry French
4   Regular Air,46.71,8.69,2.99,Barry French
5   Regular Air,1198.97,195.99,3.99,N/A
6   Regular Air,-4.72,5.28,2.99,N/A
7   Regular Air,782.91,39.89,3.04,N/A
8   Regular Air,93.8,15.74,1.39,N/A
9   Delivery Truck,440.72,100.98,26.22,N/A
10  Regular Air,N/A,100.98,69,Sylvia Foulston
11  Regular Air,N/A,65.99,5.26,Jim Radford
12  Regular Air,N/A,155.99,8.99,Jim Radford
13  Express Air,N/A,3.69,0.5,Carlos Soltero
14  Regular Air,-5.77,4.71,0.7,Carlos Soltero
15  Regular Air,-172.88,15.99,13.18,Carl Ludwig
16  Regular Air,-144.55,4.89,4.93,Carl Ludwig
17  Regular Air,5.76,2.88,0.7,Don Miller
18  Regular Air,252.66,40.96,1.99,Jack Garza
19  Delivery Truck,-1766.01,95.95,74.35,Julia West
20  Regular Air,-236.27,3.89,7.01,Eugene Barchas
21  Delivery Truck,-236.27,120.98,30,Eugene Barchas
22  Regular Air,118.94,500.98,5.76,Eugene Barchas
23  Delivery Truck,3424.22,500.98,26,Edward Hooks
24
```

Result:
Given datasets successfully cleaned using java.