**Project Team-46**

**Analyzing Big Data**

**Project Step 1:** Task Selection and Data Cleaning

Aathira Vazaiel Raghavan - 1233857497

Adhithi Satish Kumar -1233343815

Ishwariya Pandi -1233885850

Vaishmitha Reddy Biyyapu- 1233298757

**Dr.Asmaa Elbadrawy**

**Due Date- 03/23/2025**

**“Data Preparation”**

**“Task 1: Preparing Data for the Recommender System”**

**“Selected Task”:**

1. “Build a recommender system that recommends books to read for every user based on their personal tastes and previous book ratings.”

**“Code:”**

import pandas as pd

from scipy.sparse import csr\_matrix

from sklearn.datasets import dump\_svmlight\_file

r\_df = pd.read\_csv('Ratings.csv', sep=';', low\_memory=False)

u\_mapng = {}

b\_mapng = {}

u\_indx = 0

for usr in r\_df["User-ID"].unique():

u\_mapng[usr] = u\_indx

u\_indx += 1

b\_indx = 0

for bk in r\_df["ISBN"].unique():

b\_mapng[bk] = b\_indx

b\_indx += 1

r\_df["User-ID"] = r\_df["User-ID"].map(u\_mapng)

r\_df["ISBN"] = r\_df["ISBN"].map(b\_mapng)

spar\_matx = csr\_matrix(

(r\_df["Rating"], (r\_df["User-ID"], r\_df["ISBN"])),

shape=(len(u\_mapng), len(b\_mapng))

)

dumy\_v = [0] \* spar\_matx.shape[0]

dump\_svmlight\_file(spar\_matx, dumy\_v, "ratings3.libsvm", zero\_based=True)

print("Successfully saved the dataset in libsvm format as 'ratings3.libsvm'")

### **“Important Note Regarding the y Vector”**

To satisfy the requirements of the dump\_svmlight\_file function, a zero-vector (dumy\_v) is used as a placeholder because the task does not involve a target variable (y). No part of the analysis or modeling makes use of this vector.

**Code Screenshot:**

A screenshot of a computer

AI-generated content may be incorrect.

**Output:**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**