PROJECT REPORT – AML 2203 -Advanced Python AI and ML Tools

Assignment 1

Team Members:

* + - Aishlee
    - Jayachandran Saravanan
    - Prashanta Timsina
    - Rohan
    - Yogita Sharma

In this assignment, first, we scraped data from the web related to E-books based on book data related to different streams. To scarp data we used the API method.

In this dataset, we have 39,990 columns and 15 features with all predictive variables.

Let us start with data description – As we scrap data with the API method and convert it into CSV format.

To start data manipulation we import all important libraries as:

import pandas as pd – used for data manipulation

import numpy as np – used to deal with numerical variables.

import matplotlip .pyplot as plt – used for data visualization

import seaborn as sns – used for data visualization

import re – used to deal with regular expression

from PIL import Image –

from word cloud import WordCloud ,Stopwords -for text manipulation

from nltk.corpus import stopwords – deal to remove stopwords from the text.

from nltk.tokenize import word\_tokenization – used for sentence tokenization

used y\_data profiling

import warnings

warnings .filterwarnings(‘ignore’)

First of all we load our dataset in csv format and then check for the number of features and columns as our data set has 15 rows and 39990 columns

Secondly in Data Wrangling, we filter the corrupted data like missing values, outliers, and duplicate variables. As in our dataset, we have different types of variables so we look at types of variables

Then we tried to find several missing variables in our dataset so as we checked there are missing values in price,s-title, edition -num, and description columns respectively and the data type is an integer so all are numerical variables.

Next, we checked the number of unique values in each column, in our dataset so we got 97 unique values in width and 150 unique values in height

Then we checked the percentage of missing values as the edition\_num and s\_title columns have more missing values as compared to other columns.

Now we are going to drop s\_title column as it has corrupted data which is not meaningful for observation and after dropping of s\_title column let us check for missing values again as checked we have 42103 missing values in total.

Let's check the number of duplicate rows in our dataset as checked we have a 9156 count of duplicate rows and we will eliminate 24 out of 9156 rows.

Next, we will find the number of numerical and categorical variables in our dataset. So, we can perform preprocessing on it. As checked our dataset has 7 numerical and 7 categorical variables.

# Text preprocessing

In this, we did text cleansing on categorical variables as we removed emojis, and meaningless words, converted data into lowercase, removed digital values, limited word repetition to 2 times, etc.

Firstly we removed emojis from our text data using the function.

Next, we clean the ‘title’ and ‘description’ columns using the function

Then , we verify the correct spellings of text using a function.

Then we verify stopwords and remove them from our text as they are meaningless.

Then we add new columns as processed title to our original dataset

Then we removed Hypen from ‘short\_pub’ column

Next we convert sale\_date to date time with proper format

Then we round off price columns to two-digit values and create new column as “updated\_price”.

Lastly we created new column as edition .

3.Step =- we performed data visualization for a better understanding of our dataset.

In this we will explore our dataset and find the distribution across variables.

First of all , we will do data visualization of ‘Books Publication Distribution’ Per ‘Year’.As we used Bar graph for our visualization and it represents in year 1750 , number of publications was 500 however after year 2000, it touched peak.

Then, We checked count of publications in each year

Then we Visualize bar graph from the year 1753 to the year 2025 and it shows the highest number of e-books published in the year ‘2022’ is 2350, followed by the year 2018 and the count was 2285 .On the other hand, least count of e-books was published in 1980 and 1987.

The most interesting fact is till now 316 books have been published and Insignificant data for year 2025 yet to be published.

Now we will plot bar graph for columns ‘number of books’ that belong to a particular price range.

Firstly , we checked list of ‘updated price’ and the we printed the minimum and maximum price .

WE plot the bar graph by defining the number of bins , and lables and creating new column with a Bin label and we found the majority of books' price fall between the range of $ 60-80 and which is inexpensive for buyers