Extracting data from HTML

Code documentation

# Abstract

This project work is about collecting and extracting information about consumer pouches from online shopping websites. The codding work is very similar for all three datasets. The code can collect information detailed about purchases, product names, the number of products, price of the product, month of assets, and location in the form of zip code. Based on this information, data analysis and machine learning-based prediction analysis for the sales prediction analysis. Also, based on the product name collected Name entity recognition NLP algorithm is also utilized to understand what type of products are more selling online.

Table of Contents

[Abstract 1](#_Toc110282389)

[Codding information 1](#_Toc110282390)

[Dataset information 1](#_Toc110282391)

[Function information 1](#_Toc110282392)

[Store information in CSV 2](#_Toc110282393)

[Data visualization 2](#_Toc110282394)

[Discussion on the use case of a machine learning model for this dataset 2](#_Toc110282395)

[References 3](#_Toc110282396)

# Codding information

## Dataset information

The dataset utilized in the project work is available on the JSON files. Each JSON file contained multiple numbers of HTML information about various consumers. Each consumer information is extracted and stored in the CSV file.

## Function information

* In the code, first load the pre-required libraries
* After those multiple functions were created to process the data and convert it into a data frame
* The first function created is a ‘json\_reader’: this function can load and read the JSON file from the directory, and these JSON files collect the information about the HTML codes of the webpage. The function is returning the HTML list information from the JSON files.
* The following function is ‘clean\_capture’: This function contains the HTML-based information about the final single invoice from the HTML list. This function is to find the table structure from the HTML page and pass it as a list. Once the table list is collected, process and filter out the space from the table, and final product information is given.
* The following function is ‘create\_dict’: this function is utilized for cleaning and information and storing the data in the dictionary. The first of an empty dictionary is created with information about the product name; quantity ordered, location of the consumer (zip code), payment mode, and the product's price. The product name is collected from the list. The quantity of the products is stored as an integer value, the cost of the product is cleaned and removed extra symbols from the text, the price is stored as a float value in the data frame, and then the location of the order needed to be delivered are stored into a dictionary. At last, the payment method was held in the directory. It helps to analyze how consumers use various payment methods and whether they are any other, allowing us to target many additional payment companies’ data analyses.
* After storing all this information in the dictionary, the next step is to check whether there is any empty set in the dataset. If the group is open, it fills with their previous order, and if the last information of the consumer is also blank, then it is set to the default location of Mumbai.
* In the following function, all the about function and objects created are called, and the dataset was passed. In the class named ‘dataframe\_creation,’ the HTML code is processed from the part named ‘json\_reader,’ and the returned list is further passed into this function. In this function, first, the length of the HTML doc is checked if it is less than 1000, which is just a receipt or purchase and doesn’t have any other information about the consumer or products. This information is just passed.
* Proper HTML information is passed, and the function ‘clean\_capture’ is called back and stored all data into a new variable. Then the part ‘create\_dict’ is called back and stored in the processed data frame into new variables. And at last, the one entire JSON files HTML information is parsed and stored into data frames. This function returns the list of data frames for each JSON file.

## Store information in CSV

Once all the functions are developed, all the JSON files in the folder are called and passed through the part ‘json\_reader’ and ‘dataframe\_creation,’ which contain multiple process functions for cleaning the dataset and store into the data frame.

# Data visualization

* For the data analysis of pandas, libraries were utilized to group the dataset into the various categories analyzed. For data visualization, seaborn and matplotlib libraries are used.
* From the dataset information, the first of analyzing the various payment methods used by consumers. For almost all the websites, credit card users are more significant than UPI payment users and, after all, others.
* And for further data analyzed in which month most frequent payment method and monthly orders from various locations.
* Based on the monthly dataset trying to visualize the ordered place location in ZIP code and the order price in INR.

# Discussion on the use case of a machine learning model for this dataset

* From this dataset, we get inflammation about the consumer shopping habits and, from which particular location more people are purchasing from the website, their payment methods to analyze the type of consumer. In this dataset, we can explore that more consumers are utilizing credit cards and UPI payment methods. From the monthly sales data analysis, more peoples are interested in buying stuff during the Indian festivals in October. During the Diwali festival, more people buy their property from various locations. During the off-season, most of the tear-one and tear-two city people are buying online.
* Further, this data can be utilized for the time series analysis to analyze in which month and location people purchase stuff. From this predictive information, the online store can buy their goods from the industry at a lower price and store; During the festive and pick demand, they can sell these goods, earn good profits, and provide better customer service by delivering on time.
* For a better product prediction analysis, the NER-based NLP method is needed to collect more information about the product name and most repetitive brands, which consumers like to purchase more often. Based on this information machine learning model can be developed. For example, more people buy AC and Refrigerators during summer. So based on the time-series analysis, we can understand which brand product was on hot sell the previous year. Based on this information, shopping website organization can order and purchases item in advance at a lower price, and during the season, they can sell at a reasonable price. Their profit margin will be on the higher side.
* One of the drawbacks of the code is it requires a new set of configurations for the mining and scrapping of new websites. Hence, every time we vising to a different shopping website, it requires a new keyword to locate the information. For example, the code developed for the Amazon website data scrapping is not utilized for Flipkart or any other website.

# References

1. The Hitchhiker’s Guide to python, article on HTML scrapping using python, link: [HTML Scraping — The Hitchhiker's Guide to Python (python-guide.org)](https://docs.python-guide.org/scenarios/scrape/)
2. Jarmul, K. and Lawson, R., 2017. *Python Web Scraping*. Packt Publishing Ltd.
3. Thomas, D.M., and Mathur, S., 2019, June. Data analysis by web scraping using python. In *2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 450-454). IEEE.
4. Gull, M. and Pervaiz, A., 2018, April. Customer behavior analysis towards online shopping using data mining. In *2018 5th International Multi-Topic ICT Conference (IMTIC)* (pp. 1-5). IEEE.