|  |  |
| --- | --- |
| Course- BTech | Type- Specialization Core |
| Course Code- CSET-214 | Course Name- Data Analysis using Python |
| Year- 2024-25 | Semester- Odd |
| Date- | Batch- |

**Lab # No. (3) Data manipulation using Pandas, NLTK**

CO Mapping

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lab No.** | **Name** | **CO1** | **CO2** | **CO3** |
| **3** | **Data manipulation using Pandas, NLTK** | ➹ | ➹ | ➹ |

**Introduction:**

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data". NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to [over 50 corpora and lexical resources](https://www.nltk.org/nltk_data/) such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active [discussion forum](https://groups.google.com/group/nltk-users).

1. Import NLTK and download ‘punkt’

Data= “Jack was a lively boy who loved to stay active. During the day, he focused on his schoolwork, determined to excel in his studies. He enjoyed his time at school, where he could learn new things and challenge himself. However, all work and no play made Jack feel a bit dull. So, after school, he would rush outside to play with his friends.”

1. split the data into words.
2. split the data into sentences.
3. find the 7 most common word in the above sentence.
4. Import pandas as pd and find its version.

<https://drive.google.com/drive/folders/1GVPp6rabRchPKovRA6POoBoxjzZRLlTY?usp=drive_link>

Download dataset “Housing.csv”

1. Read the csv file and display only top five rows of the data.
2. Find the shape and list all attributes in the dataset.
3. Generate the descriptive statistics and the data types of the dataset.
4. Make a new dataset with attributes [price, bedrooms, bathrooms, stories, parking, furnishing status]
5. Find total number of furnished, semi furnished, unfurnished houses.
6. Download dataset from [**https://drive.google.com/file/d/1wAp7aodlSfea1\_5aquTWyE\_3abs0CeJ5/view?usp=sharing**](https://drive.google.com/file/d/1wAp7aodlSfea1_5aquTWyE_3abs0CeJ5/view?usp=sharing)
7. Read and display the record from row 20 to 50.
8. Create a subset of the data frame using ‘discipline’ column.
9. Extract the unique values present in the ‘Rank’ column.