LAB MANUAL
PART A
(PART A: TO BE REFFERED BY STUDENTS)

Experiment No-03

A.1 Aim:
Data Cleaning and Storage- Pre-process, filter and store social media data for business (Using Python, MongoDB, R, etc.).

Lab Objective	To understand the fundamental concepts of social media networks
Lab Outcome	Collect, monitor, store and track social media data

PART B
(PART B: TO BE COMPLETED BY STUDENTS)

(Students must submit the soft copy as per following segments within two hours of the practical. The soft copy must be uploaded on the Blackboard or emailed to the concerned lab in charge faculties at the end of the practical in case the there is no Black board access available)

Roll. No. C36	Name: Sanskruti Kadam
Class BE-C	Batch: C2
Date of Experiment:	Date of Submission:
Grade:	

B. 1Study the fundamentals of social media platform and implement data cleaning, pre-processing, filtering and storing social media data for business:
(Paste your Search material completed during the 2 hours of practical in the lab here)

- Students can use any social media data to perform cleaning, pre-processing and filtering.
- Use the chosen data to perform cleaning, pre-processing and filtering.

B.2 Input and Output:
(Command and its output)

```
def clean_text(text):
    '''Make text lowercase, remove text in square brackets,remove links,remove punctuation
    and remove words containing numbers.'''
    text = str(text).lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    return text
```

```
[ ] df_train['text'] = df_train['text'].apply(lambda x:clean_text(x))
    df_train['selected_text'] = df_train['selected_text'].apply(lambda x:clean_text(x))
```

```
[ ] # Most Common Word

    df_train['temp_list'] = df_train['selected_text'].apply(lambda x : str(x).split())
    temp = Counter(item for sublist in df_train['temp_list'] for item in sublist)
    top = pd.DataFrame(temp.most_common(20))
    top.columns = ['Common_words', 'count']
```

[ ]

	Common_words	count
0	i	7200
1	to	5305
2	the	4590
3	a	3538
4	my	2783
5	you	2624
6	and	2321
7	it	2158
8	is	2115
9	in	1986
10	for	1854
11	im	1676

[ ]
```
def remove_stopwords(x):
    return [y for y in x if y not in stopwords.words('english')]

df_train['temp_list'] = df_train['temp_list'].apply(lambda x : remove_stopwords(x))
```

[ ]
```
top = Counter(item for sublist in df_train['temp_list'] for item in sublist)
temp = pd.DataFrame(top.most_common(20))
temp.columns = ['Common_words', 'count']
temp
```

	Common_words	count
0	im	1676
1	good	1251
2	day	1058
3	love	909
4	happy	852
5	like	774

B.3 Observations and learning:

(Students are expected to comment on the output obtained with clear observations and learning for each task/ sub part assigned)

In conclusion, the exploratory data analysis (EDA) experiment has provided valuable insights into the dataset under investigation. Through visualizations and statistical analyses, we were able to identify patterns, trends, and relationships within the data. We also identified potential data quality issues, which may require further investigation and cleaning. The EDA process has helped us to better understand the dataset and formulate hypotheses for further analysis. Overall, the EDA experiment has been an important first step in the data analysis process, providing a foundation for further exploration and modeling.

B.4 Conclusion:

(Students must write the conclusion as per the attainment of individual outcome listed above and learning/observation noted in section B.3)

Hence exploratory data analysis was conducted on twitter dataset using python.

B.5 Question of Curiosity

(To be answered by student based on the practical performed and learning/observations)

Q1. What is data cleaning? Explain its importance?

Ans:- Data cleaning is the process of detecting and correcting or removing errors, inconsistencies, and inaccuracies in data. It involves identifying and addressing problems such as missing values, duplicate records, outliers, incorrect data formats, and other data anomalies that can affect the quality of data.

Data cleaning is important for several reasons:

Ensures accuracy: Clean data ensures that the insights derived from the data are accurate and reliable. Data that is not clean can result in inaccurate or unreliable conclusions.

Improves decision-making: Clean data helps decision-makers make informed decisions based on accurate and reliable insights.

Increases efficiency: Cleaning data before analysis saves time and resources, as it eliminates the need to redo analysis or correct errors after analysis.

Facilitates data integration: Data cleaning helps to integrate multiple data sources that may have different formats and standards, making it easier to combine and analyze data from different sources.

Enhances data quality: Data cleaning improves the overall quality of data, making it more valuable for various purposes such as research, business intelligence, and data analytics.

Overall, data cleaning is an essential step in data management that ensures data accuracy, reliability, and consistency, which are crucial for making informed decisions and deriving insights.

Q2. What are the steps involved in data cleaning?
Ans:-There are several steps involved in data cleaning. The specific steps may vary depending on the nature and complexity of the data, but the following are some common steps that are typically involved in data cleaning:

Identify data quality issues: The first step is to identify the quality issues in the data. This involves examining the data for missing values, duplicate records, incorrect data formats, and other anomalies.

Define data cleaning rules: Once the data quality issues are identified, data cleaning rules need to be defined. These rules specify how to address the issues identified in step 1. For example, a rule may be to delete duplicate records or to fill in missing values using statistical methods.

Data profiling: Data profiling involves analyzing the data to understand its structure, relationships, and distributions. This helps to identify any patterns or anomalies that may require attention.

Data validation: Data validation involves checking the data against external sources or standards to ensure that it is accurate and complete. For example, data may be validated against a set of predefined rules or against data from a different source.

Data transformation: Data transformation involves converting the data into a format that is consistent and usable for analysis. This may involve converting data types, standardizing values, or normalizing data.

Data enrichment: Data enrichment involves adding additional information to the data to enhance its value. For example, adding geographical information to customer data can help to analyze the customer behavior across different locations.

Data integration: Data integration involves combining data from different sources into a unified dataset. This may involve standardizing data formats, resolving data conflicts, and matching records across different sources.

Data documentation: Data documentation involves creating documentation that describes the data, its sources, and any changes made during the data cleaning process. This helps to ensure that the data is well-documented and easy to understand for future use.

Overall, data cleaning is a complex process that requires careful planning and execution to ensure the data is clean, accurate, and usable for analysis.

Q3. Explain various data cleaning techniques.
Ans:-There are several data cleaning techniques that can be used to address different types of data quality issues. Some of the most common techniques include:

Removing duplicates: Duplicate records can occur due to data entry errors, system bugs, or merging of datasets. Removing duplicates can improve data accuracy and reduce the risk of errors in analysis.

Handling missing values: Missing values can be handled in several ways, including imputation (filling in missing values with a reasonable estimate), deletion (removing rows with missing values), or interpolation (estimating missing values based on the values of neighboring data points).

Standardizing data: Standardizing data involves converting data values to a common scale or format. For example, converting dates to a standardized date format, or converting units of measurement to a common unit.

Correcting data format errors: Data format errors can include incorrect data types, inconsistent date formats, or inconsistent naming conventions. Correcting data format errors ensures data consistency and accuracy.

Handling outliers: Outliers are extreme values that are significantly different from the majority of data points. Outliers can be handled by either removing them or transforming them to a more appropriate value.

Handling inconsistent data: Inconsistent data refers to data that does not conform to established standards or rules. This can include misspelled names, inconsistent addresses, or inconsistent data formats. Handling inconsistent data involves identifying and correcting errors to ensure data consistency.

Resolving data conflicts: Data conflicts can occur when data from different sources or systems contain conflicting information. Resolving data conflicts involves identifying the source of the conflict and determining the correct value based on established rules or standards.

Data profiling: Data profiling involves analyzing the data to identify patterns, distributions, and relationships. This can help identify potential quality issues and inform data cleaning strategies.

Overall, data cleaning techniques are used to ensure that data is accurate, complete, and consistent. The specific techniques used depend on the nature and complexity of the data and the quality issues that need to be addressed.

Q4. What are the steps involved in data pre-processing? Enlist the data pre-processing techniques.
Ans:- Data pre-processing refers to the process of preparing raw data for analysis. It involves several steps to transform and clean the data to make it usable for further analysis. The following are some common steps involved in data pre-processing:

Data collection: This step involves collecting raw data from various sources such as databases, spreadsheets, or text files.

Data cleaning: Data cleaning involves identifying and correcting errors, inconsistencies, and inaccuracies in the data.

Data integration: Data integration involves combining data from different sources to create a unified dataset.

Data transformation: Data transformation involves converting the data into a format that is consistent and usable for analysis. This may involve converting data types, standardizing values, or normalizing data.

Data reduction: Data reduction involves reducing the size of the dataset while retaining the most important information. This may involve sampling, feature selection, or feature extraction techniques.

Data discretization: Data discretization involves converting continuous data into discrete categories or ranges. This can be useful for some analysis techniques.

Data normalization: Data normalization involves scaling the data to a common range or format. This is useful for analysis techniques that are sensitive to the scale of the data.

Data aggregation: Data aggregation involves combining data at a higher level of granularity, such as summing sales data by month or year.

Data balancing: Data balancing involves addressing imbalances in the data distribution, such as oversampling minority classes or undersampling majority classes.

Data pre-processing techniques can be categorized into two types: numerical data pre-processing techniques and categorical data pre-processing techniques. Some common data pre-processing techniques include:

Numerical data pre-processing techniques:

- Scaling
- Standardization
- Binning
- Log transformation

- Outlier removal

Categorical data pre-processing techniques:

- One-hot encoding
- Label encoding
- Binary encoding
- Frequency encoding

Overall, data pre-processing is a critical step in the data analysis process, as it ensures that the data is accurate, consistent, and usable for analysis.

Q5. Explain the difference between data cleaning and data pre-processing techniques. Ans:- Data cleaning and data pre-processing are both important steps in preparing data for analysis, but they are different in nature and scope.

Data cleaning is the process of identifying and correcting errors, inconsistencies, and inaccuracies in the data. This may involve removing duplicate records, handling missing values, correcting data format errors, and resolving data conflicts. Data cleaning is focused on improving the quality of the data by ensuring that it is accurate, complete, and consistent.

Data pre-processing, on the other hand, involves a broader set of activities aimed at preparing raw data for analysis. This may involve several steps, including data cleaning, data integration, data transformation, data reduction, data discretization, data normalization, data aggregation, and data balancing. The goal of data pre-processing is to transform raw data into a format that is usable for analysis, by addressing issues such as data quality, data format, data size, and data type.

In summary, data cleaning is a subset of data pre-processing, which focuses specifically on addressing issues related to data quality. Data pre-processing is a more comprehensive process that involves a range of techniques and activities aimed at transforming raw data into a format that is usable for analysis. Both data cleaning and data pre-processing are important steps in the data analysis process, as they ensure that the data is accurate, consistent, and usable for analysis.