**Natural Language Processing (NLP)**

**Problem Statements-**

Q1. Why it is important to understand the business before starting the data collection?

Answer: Without the knowledge of business, one will not be able to provide an appropriate solution to the client.

Asking relevant questions and understanding the business is most important phase before starting to work on any business problem.

Q2. What is the purpose of defining project objectives?

Answer:

Project objectives define what output could be achieved. What challenges could be there and what cannot be covered in the project scope?

Q3. Why it is mandatory to define the success criteria?

Answer: Success criteria are divided into 3 sub-parts: Business success criteria, ML success criteria, and Economic success criteria.

Success criteria help to break down the project into small milestones. What do we want to achieve as a business? what accuracy we want to achieve through our model and what could be economic success criteria we want to achieve.

**Libraries-**

Q1. What are the things pandas can do?

Answer:

* Pandas is Python's powerful and widely-used open-source data manipulation and analysis library. It provides a wide range of functionality for manipulating, cleaning, and transforming data and visualizing and analyzing data.

Q2. What is the difference between matplotlib and seaborn library?

Answer:

|  |  |
| --- | --- |
| **Matplotlib** | **Seaborn** |
| Matplotlib is mainly deployed for basic plotting. | Seaborn provides a variety of visualization patterns. |
| Visualization using Matplotlib generally consists of bars, pies, lines, scatter plots and so on. | It uses less syntax and has easily interesting default themes. |

Q3. What is the other library for data visualization?

Answer:

* Plotly
* Bokeh
* pygal

**Import dataset-**

Q1. How to load an excel file in python?

Answer:

* Import necessary libraries.
* Import pandas as pd.
* df= pd.read\_excel("file path")

Q2. How to load data using an open function?

Answer:

* Import pandas as pd
* df=open(“file path”)

Q3. How to load data with multiple sheets?

Answer:

* Import pandas as pd
* Df=pd.read\_excel(“file path”, sheet\_name= “sheet title”)

**Perform EDA and derive insights**

Q1. What are the auto EDA techniques?

Answer:

* Pandas profiling
* Autoviz
* Sweetviz
* Dtale
* Data-prep

Q2. What are four business moments, and what insights we can draw from them?

Answer:

* First-moment business decision: Measures of central tendency
* Second-moment business decision: Measures of dispersion
* Third-moment business decision: Skewness
* Fourth-moment business decision: Kurtosis

Q3. Does EDA gives clean data?

Answer:

* No

**Data-preprocessing-**

Q1. What is Data Preprocessing?

Answer:

* Data Pre-processing includes the steps we need to follow to transform or encode data so that it may be easily parsed by the machine.
* The main agenda for a model to be accurate and precise in predictions is that the algorithm should be able to easily interpret the data's features.

Q2. Write the techniques in data Pre-Processing.

* Data Cleaning.
* Dimensionality Reduction.
* Feature Engineering.
* Sampling Data.
* Data Transformation.
* Imbalanced Data.

Q3. When we use label encoding and one-hot encoding?

* Label encoding- for ordinal data
* One-hot encoding- Nominal data

**Outlier treatment-**

Q1. What is the technique to remove outliers?

* Winsorizing
* Capping
* 3R method

Q2. What is the formula to find IQR, upper and lower limit for boxplot?

* IQR = Q3 – Q1
* Upper limit= Q3+1.5\*IQR
* Lower limit = Q1-1.5\*IQR

Q3. How many gauss method are there?

* The **three** operations from Gauss method are the elementary reduction operations, or row operations, or Gaussian operations.

**Zero variance-**

Q1. Why do we check zero variance?

* To check variability in the data.

Q2. What happens if the variance is zero?

* Zero variance means that there is no difference in the data values, which means that they are all the same.

**Normalization-**

Q1. What are the techniques to check whether the data is normally distributed or not?

* Q-Q Plot
* Skewness
* Histogram

Q2. How to make data scale-free?

* Standardization or normalization

**Univariate and Bivariate-**

Q1**.** What types of graphs are used to depict the bivariate analysis?

* Bivariate information is investigated utilising the scatterplot of Y against X, giving a visual image of the information's relationship.

Q2**.** What do you mean by bivariate frequency distribution?

* A dispersion demonstrating every conceivable blend of two straight-out factors as per their noticed recurrence

**Libraries-**

Q1. When a sentence is passed into Textblob it gives which two outputs?

Answer: Polarity and subjectivity.

Q2.Which Python library for Natural Language Processing is pronounced as 'pineapple'?

Answer: PyNLPl

Q3. nlp = sp.load("en\_core\_web\_sm") in the code sp word refer to which library?

Answer: In the above code the sp word relates to spacy and the previous code of the above code is “import spacy as sp”.

**Import dataset**

# Q1. [How to read xml file using python?](https://stackoverflow.com/questions/57351894/how-to-read-xml-file-using-python)

Ans:

* import xml.etree.ElementTree as ET
* tree = ET.parse(''my\_file.xml')
* root = tree.getroot()

**Perform EDA and derive insights**

Q1. How to import stopwords?

Answer:

* from nltk.corpus import stopwords

Q2. How to import FreqDist?

Answer:

* from sklearn. from nltk.probability import FreqDist

Q3. Which process is used in NLP to split paragraphs and sentences into smaller units that can be more easily assigned meaning?

Answer: Tokenization

**Data-preprocessing-**

Q1.Write a code for punctuation removal?

Answer:

* import string
* a\_string = '!hi. wh?at is the weat[h]er lik?e.'
* for character in string.punctuation:

a\_string = a\_string.replace(character, '')

* print(a\_string)

.

Q2.Why we are using “.lower()” in NLP Data Preprocessing.

# Answer: Convert Text data to lowercase.

# Q3.How to import TfidfVectorizer ?

Answer:

* from sklearn.feature\_extraction.text import TfidfVectorizer

**Plotting-**

Q1. Tag cloud also referred to as………

Answer: Word Cloud

Q2. Which two libraries are required for creating word cloud?

Answer: wordcloud and matplotlib

Q3.  Word cloud is also known as…………

Answer:Text Cloud.

## NLP -

Q1. Handling Sentence Ambiguity is main challenge for ……………

Answer:NLP

## Q2. Conversion one human language to another is called as…………

## Answer: Machine Translation

Q3.What does "morphological segmentation" mean??

## Answer: Separate words into individual morphemes and identify the class of the morphemes.

## Model-

Q1. By using which term in NLP we can extract the useful information from crucial documents by statistical data?

Answer: TF-IDF

Q2. What is Naive Bayes algorithm, when we can use this algorithm in NLP?

Answer: [Naive Bayes algorithm](https://www.mygreatlearning.com/blog/introduction-to-naive-bayes/) is a collection of classifiers which works on the principles of the Bayes’ theorem. This series of NLP model forms a family of algorithms that can be used for a wide range of classification tasks including sentiment prediction, filtering of spam, classifying documents.

Q3.Which is the process of shortening a long piece of text with its meaning and effect intact?

### Answer: Text Summarization.

Q4.Which technique in NLP refers to extracting structured information automatically from unstructured sources to ascribe meaning to it?

### Answer:  Information extraction

Q5.Tagger Module & Relation Extraction Module is Belong to which model?

Answer:Information extraction

Q6.What is the model commonly used which depends on word frequencies or occurrences to train a classifier?

### Answer: Bag of Words

## Q7.Which model help learners to understand deep representations in downstream tasks by taking an output from the corrupt input.

### Answer: Masked language

### Q8. Cosine Similarityis  used to compute the distance between Two-………………….. in NLP.

Answer: word vectors

### Q9. Dissimilarity between words expressed using cosine similarity will have values significantly higher than ……

### Answer:0.5

Q10. BERT is a ………..  NLP library.

Answer:Word embedding

**Flask deployment:**

Q1. What are the files required for flask deployment?

Ans.

-          templets folder (.html file)

-          app.py

-          Saved model file

Q2. What is necessary libraries for flask?

Ans:

-          Pandas

-          Pickle

-          Flask

-          Render\_template

-          Request

Q3. Apart from pickle which library we can use for saving and loading the model?

Ans:JSON, JOB LIB, Stats model

**@app.route:**

Q1. Why we are using @app.route?

Ans: App Routing means mapping the URLs to a specific function that will handle the logic for that URL.

Q2. Why we are using ‘/’ this inside the root?

Ans: to justify the home page.

Q3. What is GET & POST request I flask?

Ans: Used to send HTML form data to the server. The data received by the **POST method** is not cached by the server. HEAD, Same as GET method, but no response body.

**INPUTS AND OUTPUTS:**

Q1. How you collect the input data in flask?

Ans: By initializing the form inside the HTML file and collecting those inputs by using post method.

Q2. How would you call HTML file in flask?

Ans: By using return and render\_templets.

Q3. Why we use debug= true?

Ans: to save the changes automatically into the server.