**Important Note:** Python code will be considered valid only if every line of the code is explained in detail using ‘New Comment’ option in word document and it generates the same output as pasted by the participant in this document.

**Task I – Unstructured Data Analysis**

Perform this analysis on 2 excel files ‘Unstructured Data English.xlsx’ and ‘Unstructured Data Japanese.xls

1. Generate code for data cleansing in Python
2. Write a SQL Query/python to generate top 10 frequently occurring meaningful phrases in the data set
3. Using above cleansed data create a term document matrix for the given unstructured data. Use Python 3.7. The output should be in excel format explained as below,

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | S.no. (Unique identifier of a comment) | | | | |  |  |  |  |  |  | |  | |  |
| **Token** | 1 | 2 | 3 | 4 | so on…….. |  |  |  | Unique identifier of each comment | | |  | |
| Fire | 0 | 0 | 3 | 0 | ………. |  |  |  | Token |  |  | |  | |  |
| bad | 2 | 0 | 1 | 8 | ………. |  |  |  | Number of times a token appears in a comment | | | | |
| hate | 5 | 2 | 2 | 0 | ………. |  |  |  |  |  |  | |  | |  |
| shine | 1 | 1 | 2 | 3 | ………. |  |  |  |  |  |  | |  | |  |

1. Extract primary theme tags using NLP models in python.

**Task II – Structured Data Analysis**

Data Source: <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>

1. Download data from <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>
2. Do an exploratory Data Analysis (EDA) using Python
   1. Do univariate and bivariate analysis to find specific patterns and trends in data
   2. Do an outlier analysis and transform highly skewed features suitably
   3. Perform correlation analysis and remove redundant features
   4. Show the steps of feature selection and feature engineering
   5. Write interpretations & modelling data prep recommendations based on the data summary
3. Create a flag field named ‘Suspicious’ in the dataset defined as -- More than 2 failed logins and loss of data occurs when using file transfer protocol. Use R or Python or SQL
4. Train a predictive model using Tensorflow that best fits the data.
5. Do the hyperparameter tuning and show the convergence of the model using an error vs. epochs graph
6. Serialize your model and deploy your trained model using Azure DevOps (ADO)
7. Generate interpretations and recommendations for a client looking for solution/ process change requirements to deal with the network intrusion incidents.
8. You may use power point or any other visualization format to present your findings