CAPSTONE Project

Industrial Safety and Health Analytics Database

Interim Report- Milestone 1& 2

Table of Contents

Overview

Goal of the project

Machine learning chatbots work using artificial intelligence. Users need not be extremely specific while talking with a bot because it can understand the natural language, not only commands. This kind of bots get continuously better or smarter as it learns from past conversations it had with people.

Problem Statement

In this project we aim to build NLP based chatbot for Industrial Safety and Health Analytics database. In this dataset, the information about accidents in 12 manufacturing plants in 3 countries are given by a brazilian company, IHM Stefanini. We will use this dataset to understand why accidents occur and discover clues to reduce tragic accidents. Below are the Data Set Columns which are defined as per the XLS sheet

Dataset columns are below

Data Set	Description
Date	timestamp or time/date information
Countries	which country the accident occurred (anonymized)
Local	The city where the manufacturing plant is located (anonymized)
Industry Sector	which sector the plant belongs to
Accident Level	from I to VI, it registers how severe was the accident (I means not severe but VI means very severe)
Potential Accident Level	Depending on the Accident Level, the database also registers how severe the accident could have been (due to other factors involved in the accident)
Genre	if the person is male of female
Employee Or Third Party	if the injured person is an employee or a third party.
Critical Risk	some description of the risk involved in the accident.
Description	Detailed description of how the accident happened

◆Usage of Pre-processing Technique

1. Step 1: Import the data
2. Step 2: Data cleansing
3. Step 3: Data pre-processing
4. Step 4: Data preparation to be used for AIML model learning.

• Data preprocessing

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put it in a formatted way. for this, we use data preprocessing tasks.

Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

It involves below steps:

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

The parts and features of the csv file

- **CSV File Header**: The header in a CSV file is used in automatically assigning names or labels to each column of the dataset. If the file does not have a header, attributes have to be manually named.
- **Comments**: one can identify comments in a CSV file when a line starts with a hash sign (#). Depending on the method chosen to load the machine learning data, one will have to determine if these comments show up, and how to identify them.
- **Delimiter**: A delimiter separates multiple values in a field and is indicated by the comma (,). The tab (\t) is another delimiter that one can use, but has to be specified clearly.
- **Quotes**: If field values in the file contain spaces, these values are often quoted and the symbol that denotes this is a double quotation mark. If chosen to use other characters, one needs to specify this in the file.

Step 1: Import the data

When running python programs, we need to use datasets for data analysis. Python has various modules which help us in importing the external data in various file formats to a python program. Below, we will see how to import data of various formats to a python program.

```
import plotly
print(plotly.__version__)
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
from sklearn.impute import SimpleImputer
```

Import csv file

The csv module enables us to read each of the row in the file using a comma as a delimiter.

```
data =
pd.read_csv("D:/DataScience/GL_Projects/CapstoneProject/IHMStefanini_industrial_safety_and_health_database_with_accidents_description.csv")
data.head(5)
```

Out[2]:

	Unnamed: 0	Data	Countries	Local	Industry Sector	Accident Level	Potential Accident Level	Genre	Employee or Third Party	Critical Risk	Description
0	0	2016-01-01 00:00:00	Country_01	Local_01	Mining	I	IV	Male	Third Party	Pressed	While removing the drill rod of the Jumbo 08 f...
1	1	2016-01-02 00:00:00	Country_02	Local_02	Mining	I	IV	Male	Employee	Pressurized Systems	During the activation of a sodium sulphide pum...
2	2	2016-01-06 00:00:00	Country_01	Local_03	Mining	I	III	Male	Third Party (Remote)	Manual Tools	In the sub-station MILPO located at level +170...
3	3	2016-01-08 00:00:00	Country_01	Local_04	Mining	I	I	Male	Third Party	Others	Being 9:45 am. approximately in the Nv. 1880 C...
4	4	2016-01-10 00:00:00	Country_01	Local_04	Mining	IV	IV	Male	Third Party	Others	Approximately at 11:45 a.m. in circumstances t...

Above is the code snippet and output which shows the importing of different packages to load the data. The major packages present here are pandas and numpy which help in importing the data where any numeric values in the file are read by numpy.

Step 2 and 3: Data cleansing and pre-processing

Data cleansing or **data cleaning** is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

In any Machine Learning process, Data Preprocessing is that step in which the data gets transformed, or *Encoded*, to bring it to such a state that now the machine can easily parse it. In other words, the *features* of the data can now be easily interpreted by the algorithm.

Out[4]:

	Date	Country	Local	Industry Sector	Accident Level	Potential Accident Level	Gender	Employee type	Critical Risk	Description
0	2016-01-01 00:00:00	Country_01	Local_01	Mining	I	IV	Male	Third Party	Pressed	While removing the drill rod of the Jumbo 08 f...
1	2016-01-02 00:00:00	Country_02	Local_02	Mining	I	IV	Male	Employee	Pressurized Systems	During the activation of a sodium sulphide pum...
2	2016-01-06 00:00:00	Country_01	Local_03	Mining	I	III	Male	Third Party (Remote)	Manual Tools	In the sub-station MILPO located at level +170...
3	2016-01-08 00:00:00	Country_01	Local_04	Mining	I	I	Male	Third Party	Others	Being 9:45 am. approximately in the Nv. 1880 C...
4	2016-01-10 00:00:00	Country_01	Local_04	Mining	IV	IV	Male	Third Party	Others	Approximately at 11:45 a.m. in circumstances t...

In the above output we rename column names to their appropriate ones.

Once the column names are renamed, we look for data types

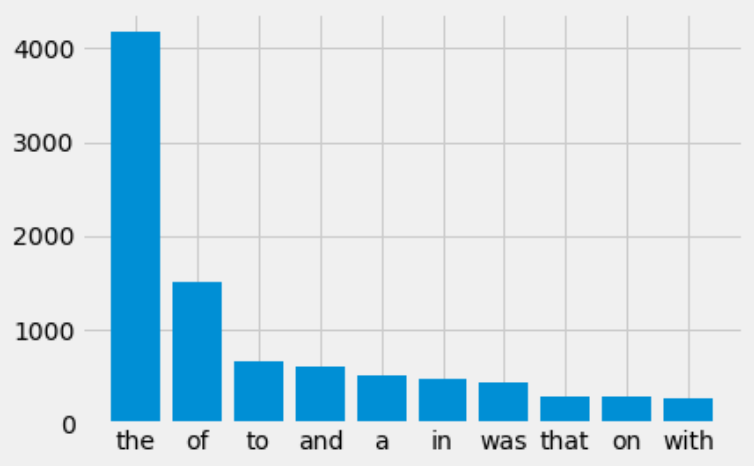
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 425 entries, 0 to 424
Data columns (total 10 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Date                      425 non-null    object
 1   Country                   425 non-null    object
 2   Local                     425 non-null    object
 3   Industry Sector           425 non-null    object
 4   Accident Level            425 non-null    object
 5   Potential Accident Level  425 non-null    object
 6   Gender                    425 non-null    object
 7   Employee type             425 non-null    object
 8   Critical Risk             425 non-null    object
 9   Description               425 non-null    object
dtypes: object(10)
memory usage: 33.3+ KB
```

Out[8]:

	Date	Country	Local	Industry Sector	Accident Level	Potential Accident Level	Gender	Employee type	Critical Risk	Description
0	2016-01-01 00:00:00	Country_01	1	Mining	0	4	Male	Third Party	Pressed	While removing the drill rod of the Jumbo 08 f...
1	2016-01-02 00:00:00	Country_02	2	Mining	0	4	Male	Employee	Pressurized Systems	During the activation of a sodium sulphide pum...
2	2016-01-06 00:00:00	Country_01	3	Mining	0	3	Male	Third Party (Remote)	Manual Tools	In the sub-station MILPO located at level +170...
3	2016-01-08 00:00:00	Country_01	4	Mining	0	0	Male	Third Party	Others	Being 9:45 am. approximately in the Nv. 1880 C...
4	2016-01-10 00:00:00	Country_01	4	Mining	4	4	Male	Third Party	Others	Approximately at 11:45 a.m. in circumstances t...

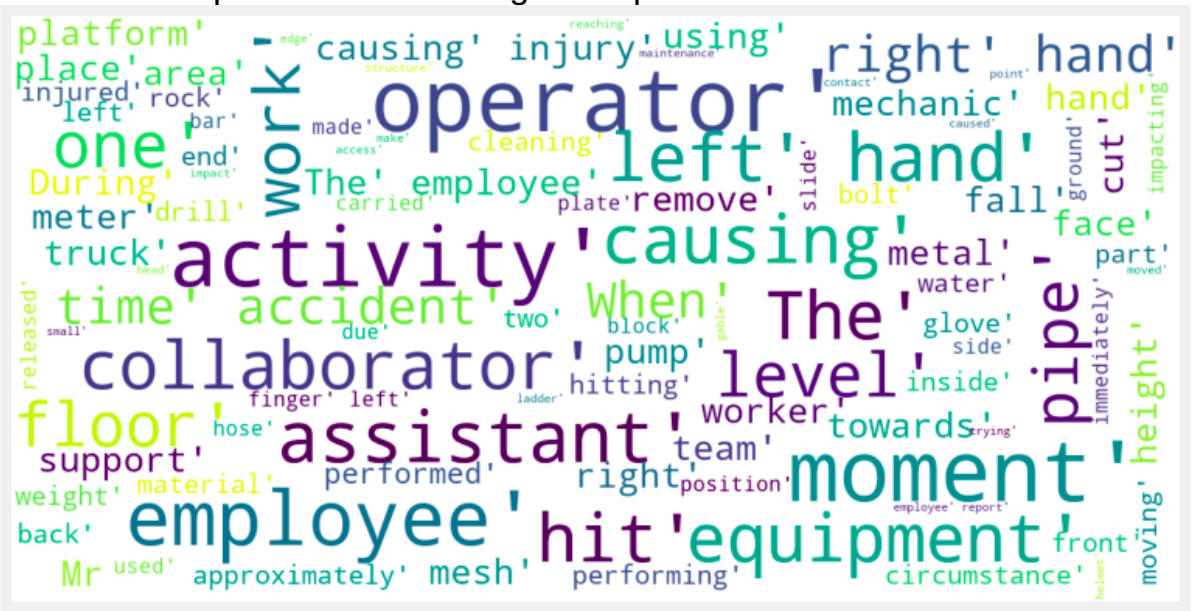
Analysis Using Stop Words in Text Mining

Stop words are basically a set of commonly used words in any language, not just English. The reason why stop words are critical to many applications is that, if we remove the words that are very commonly used in each language, we can focus on the important words instead.



Analysing the N-grams Using Word Cloud

When once the data is pre-processed, we then make usage of word cloud. The word cloud does this task according to the frequency of words in which text size tells relative importance of words of our entire dataset very quickly. This can be used where we need to quickly show how people feel about our product in presentation and grabbing attention to the important keywords that we want to represent. Below diagram represents the same.



The above output shows the scaling of the data from the word cloud. Here we randomly generate words to fit into a scale with words no greater than hundred in number. This reduces the data per frame and makes it accessible to pre-process. As shown above, randomly generated words in a frame ready for analysis.

- **Step 4: Data preparation to be used for AIML model learning.**

Data preparation may be one of the most difficult steps in any machine learning project. The reason is that each dataset is different and highly specific to the project. Nevertheless, there are enough commonalities across predictive modelling projects that we can define a loose sequence of steps and subtasks that you are likely to perform.

This process provides a context in which we can consider the data preparation required for the project, informed both by the definition of the project performed before data preparation and the evaluation of machine learning algorithms performed after.

Out[11]:

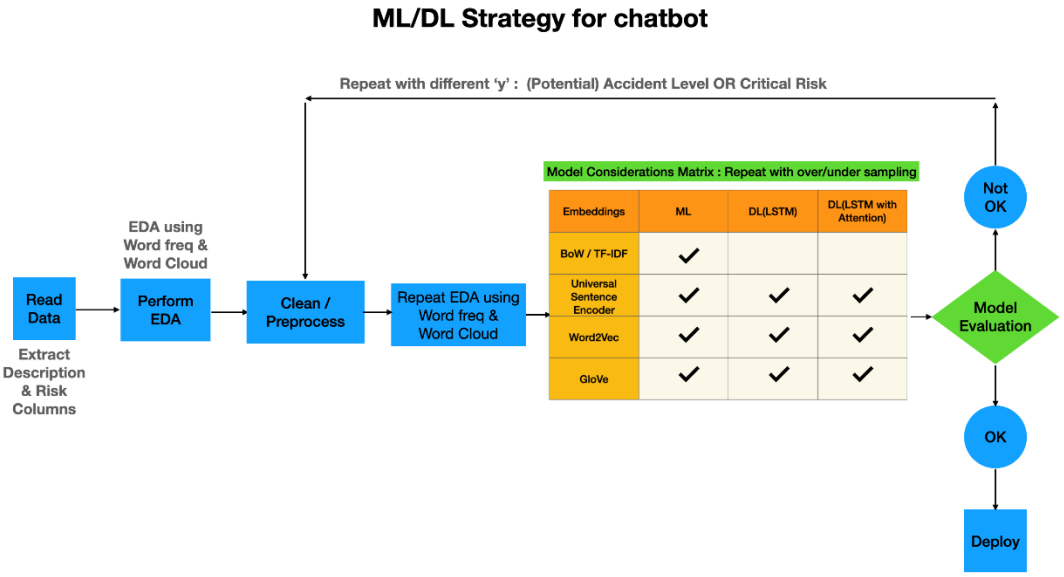
	Date	Country	Local	Industry Sector	Accident Level	Potential Accident Level	Gender	Employee type	Critical Risk	Description	Year	Month	Day	Weekday	WeekofYear
0	2016-01-01	Country_01	1	Mining	0	4	Male	Third Party	Pressed	While removing the drill rod of the Jumbo 08 f...	2016	1	1	Friday	53
1	2016-01-02	Country_02	2	Mining	0	4	Male	Employee	Pressurized Systems	During the activation of a sodium sulphide pum...	2016	1	2	Saturday	53
2	2016-01-06	Country_01	3	Mining	0	3	Male	Third Party (Remote)	Manual Tools	In the sub-station MILPO located at level +170...	2016	1	6	Wednesday	1
3	2016-01-08	Country_01	4	Mining	0	0	Male	Third Party	Others	Being 9:45 am. approximately in the Nv. 1880 C...	2016	1	8	Friday	1
4	2016-01-10	Country_01	4	Mining	4	4	Male	Third Party	Others	Approximately at 11:45 a.m. in circumstances t...	2016	1	10	Sunday	1

Out[12]:

	Date	Country	Local	Industry Sector	Accident Level	Potential Accident Level	Gender	Employee type	Critical Risk	Description	Year	Month	Day	Weekday	WeekofYear	Q
0	2016-01-01	Country_01	1	Mining	0	4	Male	Third Party	Pressed	While removing the drill rod of the Jumbo 08 f...	2016	1	1	Friday	53	Fi
1	2016-01-02	Country_02	2	Mining	0	4	Male	Employee	Pressurized Systems	During the activation of a sodium sulphide pum...	2016	1	2	Saturday	53	Fi
2	2016-01-06	Country_01	3	Mining	0	3	Male	Third Party (Remote)	Manual Tools	In the sub-station MILPO located at level +170...	2016	1	6	Wednesday	1	Fi
3	2016-01-08	Country_01	4	Mining	0	0	Male	Third Party	Others	Being 9:45 am. approximately in the Nv. 1880 C...	2016	1	8	Friday	1	Fi
4	2016-01-10	Country_01	4	Mining	4	4	Male	Third Party	Others	Approximately at 11:45 a.m. in circumstances t...	2016	1	10	Sunday	1	Fi

In the above output tables, we can see that the data is being prepared for various analyses as per requirement for AIML model learning. The data is rearranged accordingly where the null values are eliminated during the cleaning phase. We use an elif statement loop to achieve this task.

■ . Machine Learning Strategy for Chatbot



The above flowchart shows the various steps involved in building a strategy for chatbot. In the first few steps we perform data preprocessing and loading. We then repeat EDA with over and under sampling. In the final phase we evaluate the model. If the model runs well, we deploy it else the process is repeated.

● Summary of Milestone 1

Data preprocessing is an important step when it comes to machine learning. Any dataset given, generally has many null entries and the data is scrambled. Without proper preprocessing analysis on the dataset is impossible. The techniques mentioned above help in sorting and scrumming data according to the needs of the user making it possible for performing analysis and different operations as per the problem statement of the project. We hereby conclude that all the data pre-processing techniques for the given problem statement have been applied and successfully executed.