CSE 523: Machine Learning

Group 17 - Hardly Humans
Weekly Project Report - 2

Quora Insincere Questions Classification

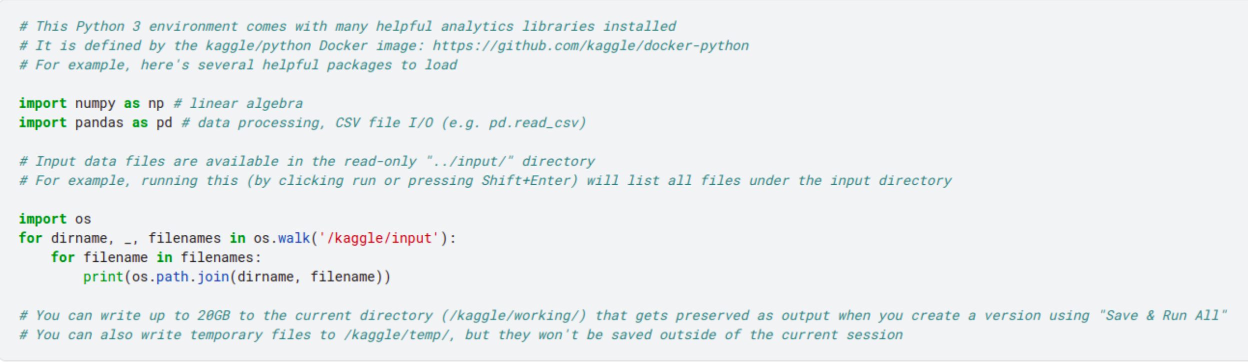
Name	Enrolment Number
Malav Doshi	AU1940017
Parth Shah	AU1940065
Sanya Zaveri	AU1920064
Mihir Pathak	AU1920138

1) Tasks Performed in the week.

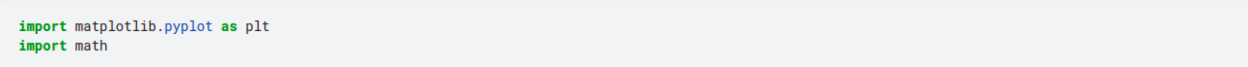
- Understanding dataset
- Importing libraries to plot data
- Reading csv files
- Analyzing data
- Data visualization by plotting bar graph

2) Outcomes of the tasks performed.

- Importing libraries

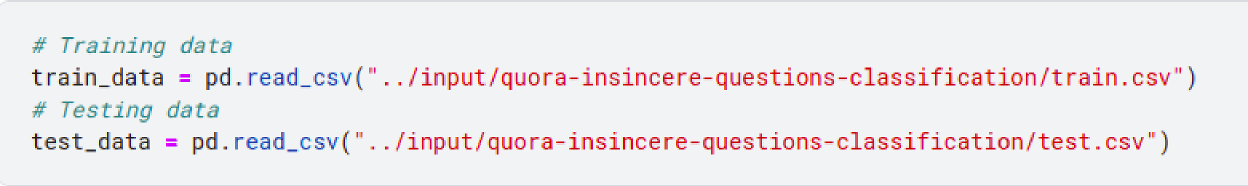


```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```



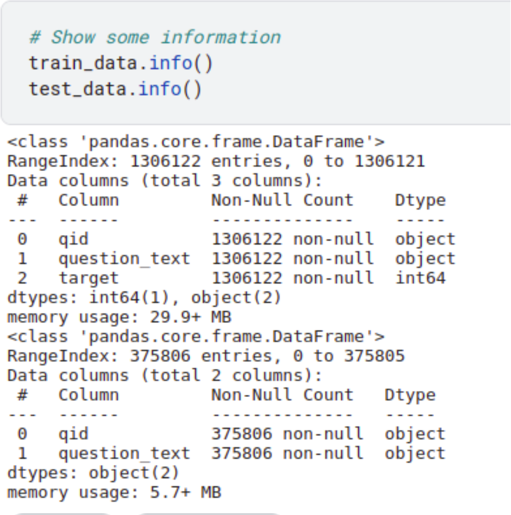
```
import matplotlib.pyplot as plt
import math
```

- Reading csv files



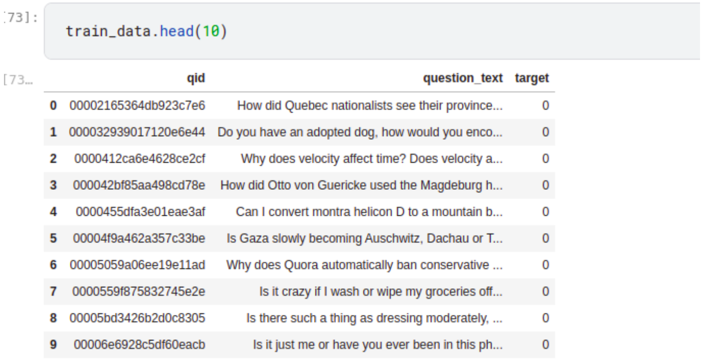
```
# Training data
train_data = pd.read_csv("../input/quora-insincere-questions-classification/train.csv")
# Testing data
test_data = pd.read_csv("../input/quora-insincere-questions-classification/test.csv")
```

- Analyzing data



```
# Show some information
train_data.info()
test_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1306122 entries, 0 to 1306121
Data columns (total 3 columns):
 #   Column         Non-Null Count    Dtype 
---  ------         --------------    ----- 
 0   qid            1306122 non-null  object
 1   question_text  1306122 non-null  object
 2   target         1306122 non-null  int64 
dtypes: int64(1), object(2)
memory usage: 29.9+ MB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 375806 entries, 0 to 375805
Data columns (total 2 columns):
 #   Column         Non-Null Count   Dtype 
---  ------         --------------   ----- 
 0   qid            375806 non-null  object
 1   question_text  375806 non-null  object
dtypes: object(2)
memory usage: 5.7+ MB
```

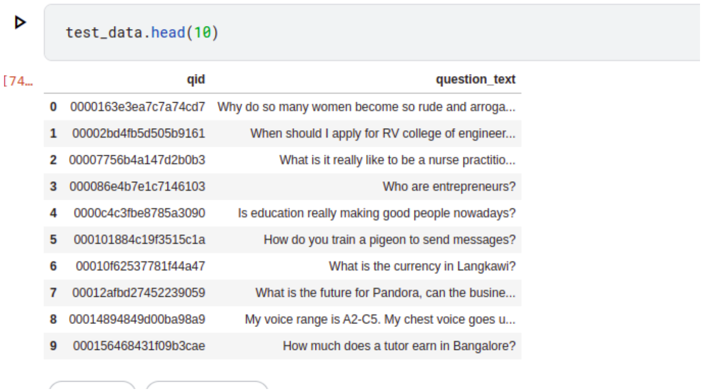


[73]:

```
train_data.head(10)
```

[73…

	qid	question_text	target
0	00002165364db923c7e6	How did Quebec nationalists see their province...	0
1	000032939017120e6e44	Do you have an adopted dog, how would you enco...	0
2	0000412ca6e4628ce2cf	Why does velocity affect time? Does velocity a...	0
3	000042bf85aa498cd78e	How did Otto von Guericke used the Magdeburg h...	0
4	0000455dfa3e01eae3af	Can I convert montra helicon D to a mountain b...	0
5	00004f9a462a357c33be	Is Gaza slowly becoming Auschwitz, Dachau or T...	0
6	00005059a06ee19e11ad	Why does Quora automatically ban conservative ...	0
7	0000559f875832745e2e	Is it crazy if I wash or wipe my groceries off...	0
8	00005bd3426b2d0c8305	Is there such a thing as dressing moderately, ...	0
9	00006e6928c5df60eacb	Is it just me or have you ever been in this ph...	0

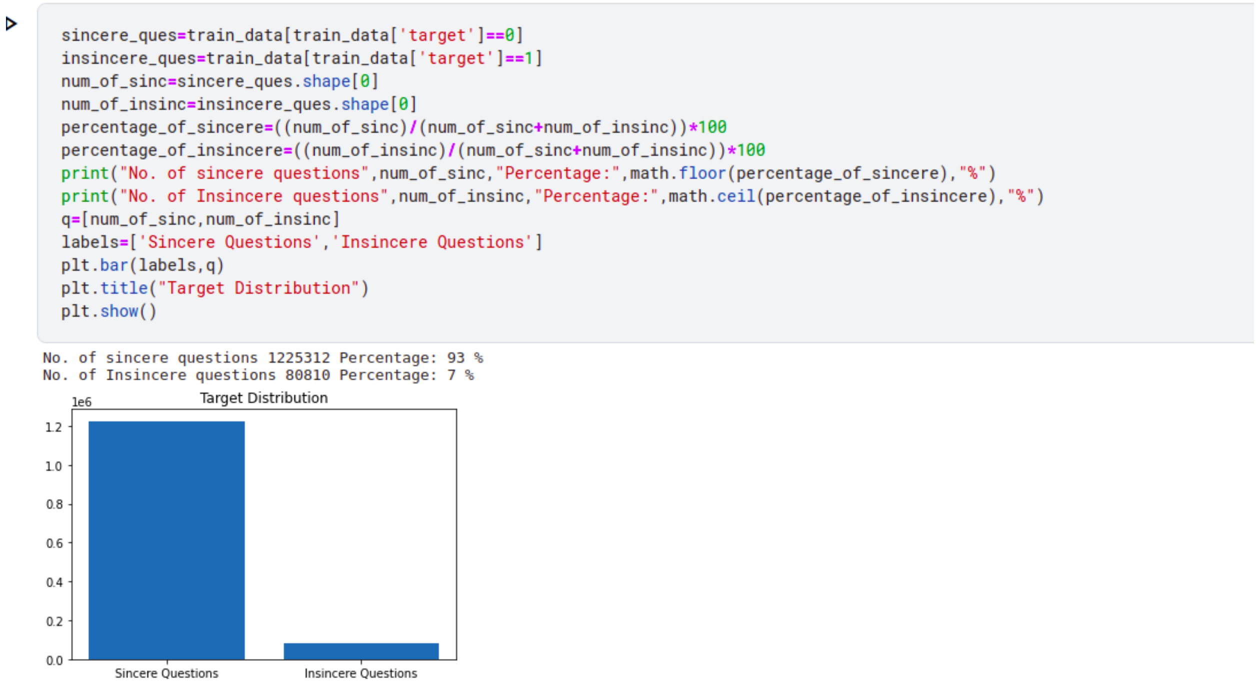


```
test_data.head(10)
```

[74…

	qid	question_text
0	0000163e3ea7c7a74cd7	Why do so many women become so rude and arroga...
1	00002bd4fb5d505b9161	When should I apply for RV college of engineer...
2	00007756b4a147d2b0b3	What is it really like to be a nurse practitio...
3	000086e4b7e1c7146103	Who are entrepreneurs?
4	0000c4c3fbe8785a3090	Is education really making good people nowadays?
5	000101884c19f3515c1a	How do you train a pigeon to send messages?
6	00010f62537781f44a47	What is the currency in Langkawi?
7	00012afbd27452239059	What is the future for Pandora, can the busine...
8	00014894849d00ba98a9	My voice range is A2-C5. My chest voice goes u...
9	000156468431f09b3cae	How much does a tutor earn in Bangalore?

- Data visualization



```
sincere_ques=train_data[train_data['target']==0]
insincere_ques=train_data[train_data['target']==1]
num_of_sinc=sincere_ques.shape[0]
num_of_insinc=insincere_ques.shape[0]
percentage_of_sincere=((num_of_sinc)/(num_of_sinc+num_of_insinc))*100
percentage_of_insincere=((num_of_insinc)/(num_of_sinc+num_of_insinc))*100
print("No. of sincere questions",num_of_sinc,"Percentage:",math.floor(percentage_of_sincere),"%")
print("No. of Insincere questions",num_of_insinc,"Percentage:",math.ceil(percentage_of_insincere),"%")
q=[num_of_sinc,num_of_insinc]
labels=['Sincere Questions','Insincere Questions']
plt.bar(labels,q)
plt.title("Target Distribution")
plt.show()
```

```
No. of sincere questions 1225312 Percentage: 93 %
No. of Insincere questions 80810 Percentage: 7 %
```

- Literature Review:
    1. http://ceur-ws.org/Vol-2517/T5-3.pdf
    2. http://ceur-ws.org/Vol-2517/T5-1.pdf
    3. https://www.researchgate.net/publication/334549103_Quora_Insincere_Questions_Classification

3) Tasks to be performed in the upcoming week.

- Data cleaning.
- Identifying the common words using Bi-gram and plotting graphs for the same.
- Data pre-processing.
- Searching for the applicable machine learning algorithms for the model.