BUILDING A SMARTER AI POWER SPAM CLASSIFIER

912421104036:P.Sajitha parveen
Phase 3 Development part 1

Project: To Identify the spam and ham messages using spam classifier

Content for project phase 3 :

1.Data collection:

- Gather a dataset of power system-related messages, which could include emails, text messages, or any form of communication.
- Ensure the dataset is labeled, with spam and non-spam messages correctly categorized.

Data Inspection:

- Check the format of the data, whether it's in text, CSV, or any other format.
- Examine the structure of the data to understand its attributes and labels.

Data loading:

- Use appropriate libraries (e.g., pandas for CSV data, NLTK for text data) to load the dataset into your project.

Data Cleaning:

- Handle missing values: Check for and deal with any missing data.
- Text cleaning: Preprocess the text data to remove noise, including special characters, numbers, and HTML tags. Consider using techniques like tokenization and stemming.

Data Labeling:

- Ensure that the labels for spam and non-spam are correctly assigned. You may need to review and correct the labels as necessary.

Data Splitting:

- Divide the dataset into training, validation, and test sets. A common split might be 70% for training, 15% for validation, and 15% for testing. This helps assess model performance.

2.Data preprocessing:

Data Cleaning:

- **Handling Missing Data**: Check for missing values in your dataset and decide on a strategy for dealing with them, such as imputation or removal.

Text Data Preprocessing:

- **Text Cleaning**: For text-based data, perform the following:
  - Remove special characters, numbers, and punctuation.
  - Convert text to lowercase to ensure consistency.
  - Remove HTML tags, if applicable.
  - Tokenization: Split text into individual words or tokens.
  - Stemming or Lemmatization: Reduce words to their base or root forms.

- Encoding Categorical Data: If your dataset contains categorical variables (e.g., message type, sender's location), you may need to encode them into numerical form. Use techniques like one-hot encoding or label encoding as appropriate.

Numerical Data Scaling:

- If you have numerical features, it's a good practice to scale them to the same range to avoid issues with certain machine learning algorithms. Common methods include Min-Max scaling or standardization (z-score scaling).

Data source:

Data link:(https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset)

v1	v2
ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine theregot amore wat...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, 螢 1.50 to rcv
ham	Even my brother is not like to speak with me. They treat me like aids patent.
ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been se as your callertune for all Callers. Press *9 to copy your friends Callertune
spam	WINNER!! As a valued network customer you have been selected to receivea 螢 900 pr e reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobile s with camera for Free! Call The Mobile Update Co FREE on 08002986030
ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I' cried enough today.
spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
spam	URGENT! You have won a 1 week FREE membership in our 螢 100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.

Program:

1. Import the required packages

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

2. Loading the dataset

```
raw_spam=pd.read_csv('/content/spam.csv',encoding='latin-1')
print(raw_spam)
```

Output:

```
     v1                                                 v2 Unnamed: 2  \
0       ham  Go until jurong point, crazy.. Available only ...        NaN
1       ham                      Ok lar... Joking wif u oni...        NaN
2      spam  Free entry in 2 a wkly comp to win FA Cup fina...        NaN
3       ham  U dun say so early hor... U c already then say...        NaN
4       ham  Nah I don't think he goes to usf, he lives aro...        NaN
...     ...                                              ...        ...
5567   spam  This is the 2nd time we have tried 2 contact u...        NaN
5568    ham              Will ì_ b going to esplanade fr home?        NaN
5569    ham  Pity, * was in mood for that. So...any other s...        NaN
5570    ham  The guy did some bitching but I acted like i'd...        NaN
5571    ham                        Rofl. Its true to its name        NaN

     Unnamed: 3 Unnamed: 4
0           NaN        NaN
1           NaN        NaN
2           NaN        NaN
3           NaN        NaN
4           NaN        NaN
...         ...        ...
5567        NaN        NaN
5568        NaN        NaN
5569        NaN        NaN
5570        NaN        NaN
5571        NaN        NaN

[5572 rows x 5 columns]
error
0scompleted at 12:43 PM
```

3.Removing the unwanted colomns

```
raw_spam.rename(columns = {'v1':'class_label', 'v2':'message'}, inplace = True)
raw_spam.drop(['Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'], axis = 1, inplace = True)
raw_spam[1990:2000]
```

Output:

class_label	message	
1990	ham	HI DARLIN IVE JUST GOT BACK AND I HAD A REALLY...
1991	ham	No other Valentines huh? The proof is on your ...
1992	spam	Free tones Hope you enjoyed your new content. ...
1993	ham	Eh den sat u book e kb liao huh...
1994	ham	Have you been practising your curtsey?
1995	ham	Shall i come to get pickle
1996	ham	Lol boo I was hoping for a laugh
1997	ham	\YEH I AM DEF UP4 SOMETHING SAT
1998	ham	Well, I have to leave for my class babe ... Yo...
1999	ham	LMAO where's your fish memory when I need it?

4.Exploring the dataset:

```
raw_spam['class_label'].value_counts()
```

Output:

```
ham 4825
spam 747
Name: class_label, dtype: int64
```

5. Print spam messages

```
raw_spam = raw_spam[raw_spam.class_label=='spam']
raw_spam
```

Output:

```
class_label  message
2        spam     Free entry in 2 a wkly comp to win FA Cup fina...
5        spam     FreeMsg Hey there darling it's been 3 week's n...
8        spam     WINNER!! As a valued network customer you have...
9        spam     Had your mobile 11 months or more? U R entitle...
11       spam     SIX chances to win CASH! From 100 to 20,000 po... ... ... ...
```

5537 spam Want explicit SEX in 30 secs? Ring 02073162414...

5540 spam ASKED 3MOBILE IF 0870 CHATLINES INCLU IN FREE ...

5547 spam Had your contract mobile 11 Mnths? Latest Moto...

5566 spam REMINDER FROM O2: To get 2.50 pounds free call…

5567 spam This is the 2nd time we have tried 2 contact u…

[747 rows x 2 columns]

6.prepare spam list

```
spam_list= raw_spam['message'].tolist()
print(spam_list)
```

Output:

["Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's", "FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still?”]

7.create arrray:

```
import matplotlib.pyplot as ab
import numpy as np
labels = ['ham', 'spam']
counts = [4825, 747]
ypos = np.arange(len(labels)) #converting text labels to numberic value, 0 and 1
Ypos
```

Output:

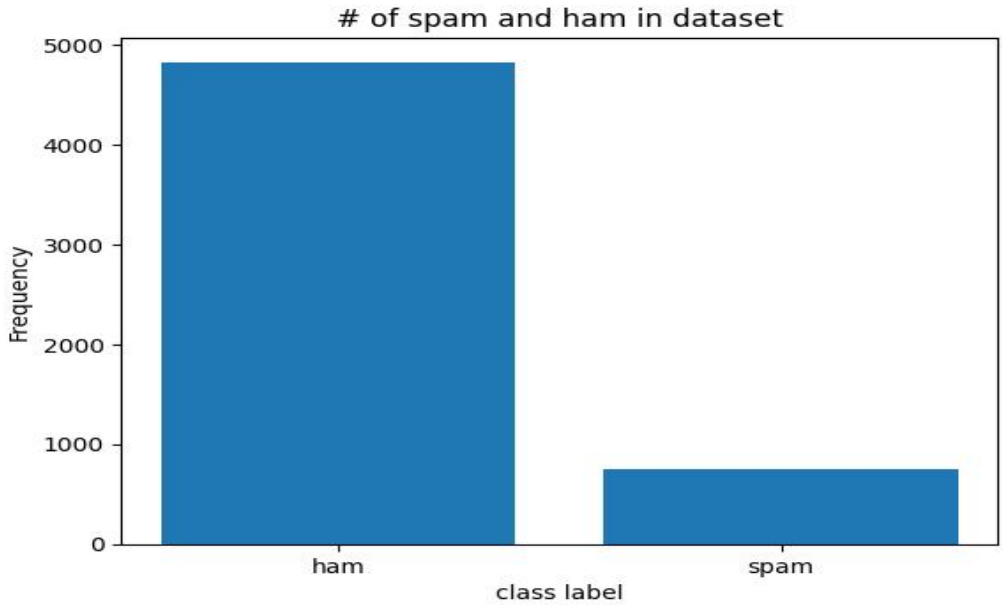
array([0, 1])

8.using graph:

```
ab.xticks(ypos, labels)
ab.xlabel("class label")
ab.ylabel("Frequency")
ab.title("# of spam and ham in dataset")
ab.bar(ypos, counts)
```

Output:

<BarContainer object of 2 artists>



9.replace the null values with a null string

mail_data=raw_spam.where((pd.notnull(raw_spam)),'')

#printing the first five rows of the dataframe

mail_data.head()

Output:

class_labelmessage

2 spam Free entry in 2 a wkly comp to win FA Cup fina…

5 spam FreeMsg Hey there darling it's been 3 week's n…

8 spam WINNER!! As a valued network customer you have…

9 spam Had your mobile 11 months or more? U R entitle…

11 spam SIX chances to win CASH! From 100 to 20,000 po…

10.checking the number of rows and colomns in the dataframe

mail_data.shape

Output:

(747,2)

11.label spam mail as 0; ham mail as 1

```
mail_data.loc[mail_data['class_label'] == 'spam','class_label',] = 0


mail_data.loc[mail_data['message']=='ham','message',] = 1
```

#separating the data as texts and label

```
x=mail_data['message']
y=mail_data['class_label']
```

Output:

print(x)

```
0          Go until jurong point, crazy.. Available only ...
1          Ok lar... Joking wif u oni…
 2         Free entry in 2 a wkly comp to win FA Cup fina...
3          U dun say so early hor... U c already then say…
 4         Nah I don't think he goes to usf, he lives aro... ...
5567      This is the 2nd time we have tried 2 contact u…
5568      Will Ì_ b going to esplanade fr home?
5569      Pity, * was in mood for that. So...any other s...
5570      The guy did some bitching but I acted like i'd...
5571      Rofl. Its true to its name Name: v2, Length:
5572,     dtype: object
```

print(y)

```
2              0
5              0
8              0
9              0
11             0 ..
5537           0
5540           0
5547           0
5566           0
5567           0
Name: class_label, Length: 747, dtype: object
```

12.Spliting the data into training data and testing data

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=3)
print(x.shape)
print(x_train.shape)
print(x_test.shape)
```

Output:

(747,)

(597,)

(150,)

13.Removing punctuation and stopwords from the messages

Punctuation and stop words do not contribute anything to our model, so we have to remove them. Using NLTK library we can easily do it.

```
import nltk

nltk.download('stopwords')

from nltk.corpus import stopwords

#remove the punctuations and stopwords

import string

def message_process(message):

message = message.translate(str.maketrans('', '', string.punctuation))

message = [word for word in message.split() if word.lower()
notinstopwords.words('english')]

return " ".join(message)

raw_spam['message'] = raw_spam['message'].apply(message_process)

raw_spam.head()
```

Output:

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

	class_label	message
2	spam	Free entry 2 wkly comp win FA Cup final tkts 2...
5	spam	FreeMsg Hey darling 3 weeks word back Id like ...
8	spam	WINNER valued network customer selected receiv...
9	spam	mobile 11 months U R entitled Update latest co...

	class_label	message
11	spam	SIX chances win CASH 100 20000 pounds txt CSH1...

14.Converting words to vectors using Count Vectorizer

```
## Counting how many times a word appears in the dataset
```

we can convert words to vectors using either Count Vectorizer or by using TF-IDF Vectorizer.

TF-IDF is better than Count Vectorizers because it not only focuses on the frequency of words present in the corpus but also provides the importance of the words. We can then remove the words that are less important for analysis, hence making the model building less complex by reducing the input dimensions.

I have included both methods for your reference.

```
text = pd.DataFrame(raw_spam['message'])
label = pd.DataFrame(raw_spam['class_label'])
from collections import Counter

total_counts = Counter()
for i in range(len(text)):
    for word in text.values[i][0].split(" "):
        total_counts[word] += 1

print("Total words in data set: ", len(total_counts))
```

Output:

Total words in data set: 4313

15.sorting in decreasing order (word with highest frequency appears first)

```
vocab = sorted(total_counts, key=total_counts.get, reverse=True)
print(vocab[:60])
```

Output:

['to', 'a', 'your', 'call', 'or', 'the', '2', 'for', 'you', 'is', 'Call', 'on', 'have', 'and', 'from', 'ur', 'with', '&', '4', 'of', 'FREE', 'mobile', 'You', 'are', 'our', 'To', 'claim', 'Your', 'U', 'txt', 'text',

'in', 'now', 'Txt', 'reply', 'free', 'contact', '-', 'be', 'now!', 'u', 'just', 'send', 'this', 'won', 'get', 'only', 'Nokia', 'prize', 'per', 'been', 'service', 'STOP', 'who', 'Reply', 'new', 'cash', 'out', 'Text', 'will']

16.Mapping from words to index

```
vocab_size = len(vocab)
word2idx = {}
#print vocab_size
for i, word in enumerate(vocab):
    word2idx[word] = i
    # Text to Vector
def text_to_vector(text):
    word_vector = np.zeros(vocab_size)
    for word in text.split(" "):
        if word2idx.get(word) is None:
            continue
        else:
            word_vector[word2idx.get(word)] += 1
    return np.array(word_vector)
# Convert all titles to vectors
word_vectors = np.zeros((len(text), len(vocab)), dtype=np.int_)
for i, (_, text_) in enumerate(text.iterrows()):
    word_vectors[i] = text_to_vector(text_[0])

word_vectors.shape
```

Output:

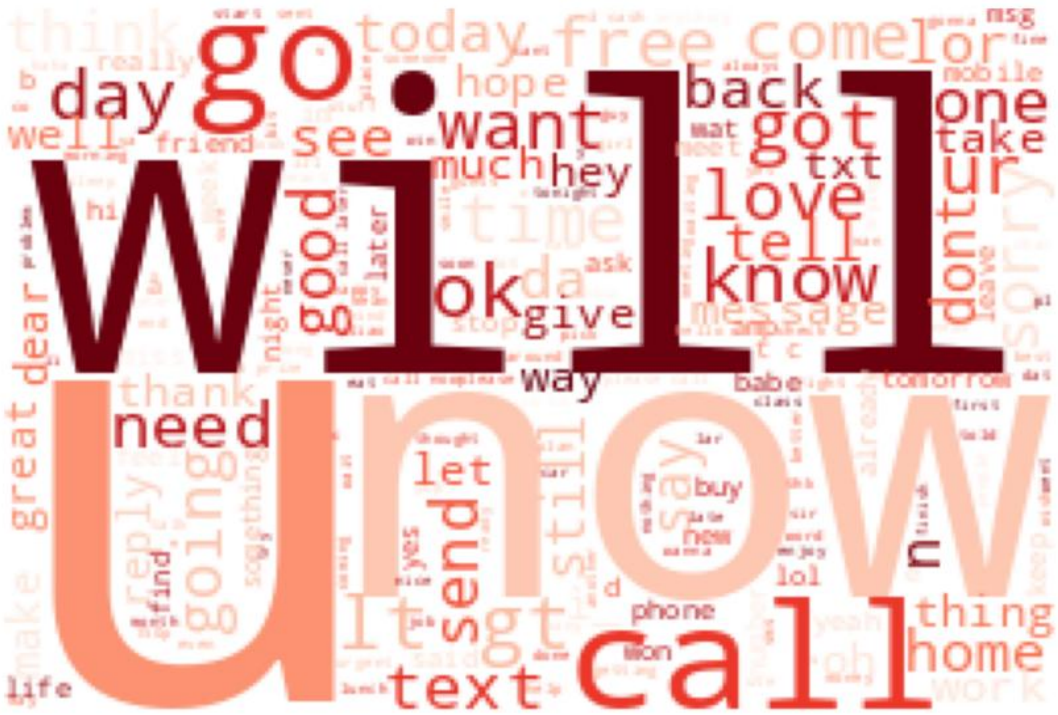
(747,3436)

17.Building word cloud to see which message is spam and which message is not

Creating spam word cloud

```
import os
import numpy as np
from wordcloud import WordCloud
from PIL import Image

# Assuming you have loaded your DataFrame 'df_spam' and extracted the 'message'
column into 'spam_list'
spam_list = raw_spam['v2'].tolist()

# Combine the text from 'spam_list' into a single string
filtered_spam = ' '.join(spam_list).lower()

# Load the comment mask image
comment_mask = np.array(Image.open("/content/comment.png"))

# Create and generate a word cloud image
wordcloud = WordCloud(
    max_font_size=160,
    margin=0,
    mask=comment_mask,
    background_color="white",
    colormap="Reds"
).generate(filtered_spam)

# Display the generated word cloud
import matplotlib.pyplot as plt

plt.figure(figsize=(8, 8), facecolor=None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad=0)
```

Output:



Creating ham word cloud:

```
import os
import numpy as np
from wordcloud import WordCloud
from PIL import Image

# Assuming you have loaded your DataFrame 'df_ham' and extracted the
'message' column into 'ham_list'
raw_ham=pd.read_csv('/content/spam.csv',encoding='latin-1')

print(raw_ham)

ham_list = raw_ham['v2'].tolist()

# Combine the text from 'ham_list' into a single string
filtered_ham = ' '.join(ham_list).lower()

# Load the comment mask image
comment_mask = np.array(Image.open("/content/comment.png"))
```

```
# Create and generate a word cloud image for ham messages
wordcloud = WordCloud(
    max_font_size=160,
    margin=0,
    mask=comment_mask,
    background_color="white",
    colormap="Greens"  # You can choose a different colormap if desired
).generate(filtered_ham)

# Display the generated word cloud
import matplotlib.pyplot as plt

plt.figure(figsize=(8, 8), facecolor=None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad=0)

# Save the word cloud to a file (optional)
wordcloud.to_file("ham_wordcloud.png")

plt.show()
```

Output:

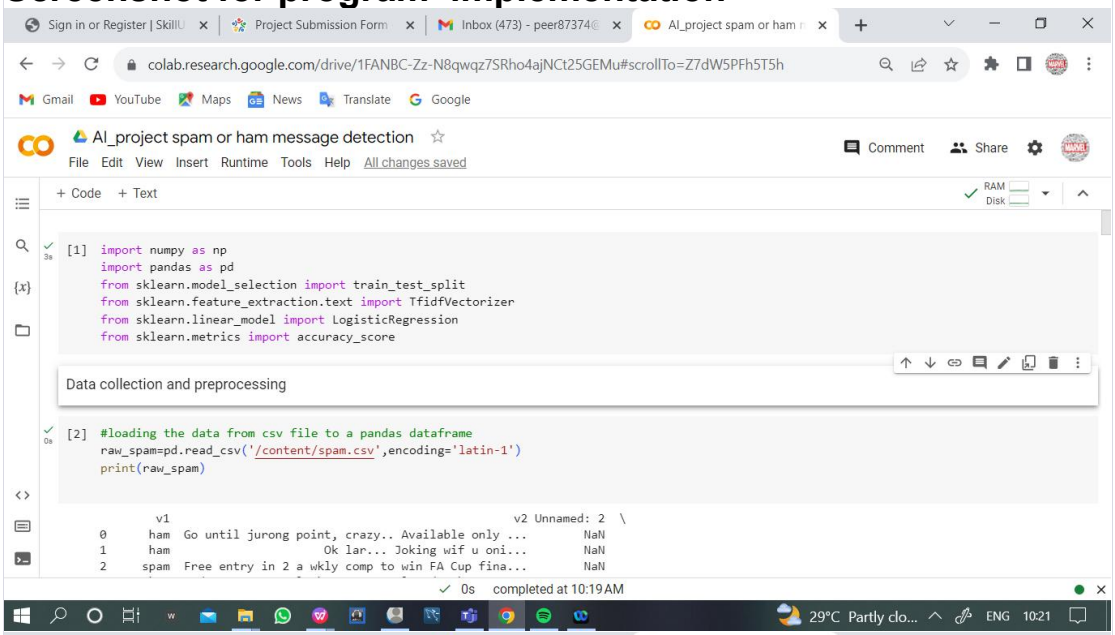
```
v1 v2 Unnamed: 2 \
 0 ham Go until jurong point, crazy.. Available only ... NaN
1 ham Ok lar... Joking wif u oni... NaN
 2 spam Free entry in 2 a wkly comp to win FA Cup fina... NaN
3 ham U dun say so early hor... U c already then say... NaN
4 ham Nah I don't think he goes to usf, he lives aro... NaN ... ... ... ...
5567 spam This is the 2nd time we have tried 2 contact u... NaN
5568 ham Will Ì_ b going to esplanade fr home? NaN
5569 ham Pity, * was in mood for that. So...any other s... NaN
5570 ham The guy did some bitching but I acted like i'd... NaN
5571 ham Rofl. Its true to its name NaN
Unnamed: 3              Unnamed: 4
```

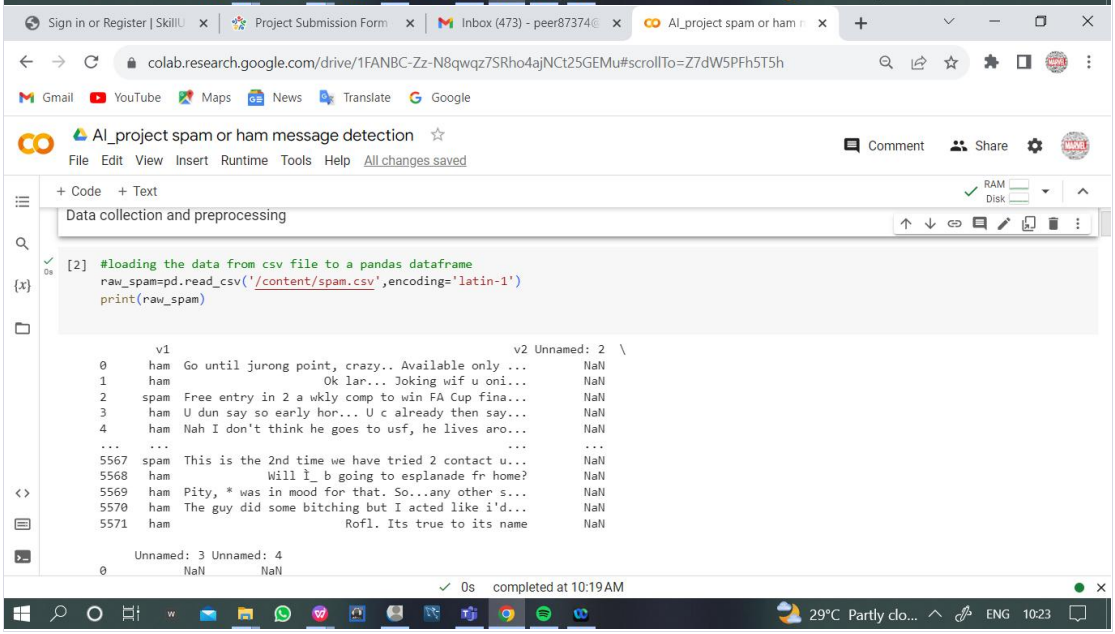
0	NaN	NaN
1	NaN	NaN
2	NaN	NaN
3	NaN	NaN
4	NaN	NaN ... ...
5567	NaN	NaN
5568	NaN	NaN
5569	NaN	NaN
5570	NaN	NaN
5571	NaN	NaN [5572 rows x 5 columns]

Project Conclusion and future work (phase3)

- In conclusion, data loading and data processing are fundamental stages in the data analysis and machine learning workflow.
- These two key steps ensure that data is collected, cleaned, and transformed into a format that is suitable for analysis or modeling.
- The quality of data loading and processing greatly impacts the accuracy and reliability of the results obtained.
- It is essential to select appropriate data sources, handle missing values, outliers, and categorical variables effectively, and conduct necessary feature engineering to extract meaningful insights from the data.
- Moreover, these processes should be well-documented for reproducibility and version control to maintain data quality throughout the analysis pipeline.
- Overall, meticulous attention to data loading and processing is critical for successful data-driven decision-making.

Screenshot for program implementation





colab.research.google.com/drive/1FANBC-Zz-N8qwqz7SRho4ajNCt25GEMu#scrollTo=g8L7ciBgcNv7

AI_project spam or ham message detection

```
[3] raw_spam.rename(columns = {'v1':'class_label', 'v2':'message'}, inplace = True)
    raw_spam.drop(['Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'], axis = 1, inplace = True)
    raw_spam[1990:2000]
```

	class_label	message
1990	ham	HI DARLIN IVE JUST GOT BACK AND I HAD A REALLY...
1991	ham	No other Valentines huh? The proof is on your ...
1992	spam	Free tones Hope you enjoyed your new content. ...
1993	ham	Eh den sat u book e kb liao huh...
1994	ham	Have you been practising your curtsey?
1995	ham	Shall i come to get pickle
1996	ham	Lol boo I was hoping for a laugh
1997	ham	\YEH I AM DEF UP4 SOMETHING SAT
1998	ham	Well, I have to leave for my class babe ... Yo...

0s completed at 10:24 AM

colab.research.google.com/drive/1FANBC-Zz-N8qwqz7SRho4ajNCt25GEMu#scrollTo=5HX1ecSWlCre

AI_project spam or ham message detection

```
#exploring the dataset
raw_spam['class_label'].value_counts()
```

```
ham     4825
spam     747
Name: class_label, dtype: int64
```

```
raw_spam = raw_spam[raw_spam.class_label=='spam']
raw_spam
```

	class_label	message
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
5	spam	FreeMsg Hey there darling it's been 3 week's n...
8	spam	WINNER!! As a valued network customer you have...
9	spam	Had your mobile 11 months or more? U R entitle...

0s completed at 10:26 AM

Sign in or Register | SkillU | Project Submission Form | Inbox (473) - peer87374@ | AI_project spam or ham

colab.research.google.com/drive/1FANBC-Zz-N8qwqz7SRho4ajNCt25GEMu#scrollTo=T5ChWvRCnv76

Gmail YouTube Maps News Translate Google

AI_project spam or ham message detection

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

Connecting

```
[ ]
print(raw_spam)
```

```
     class_label                                            message
2           spam  Free entry in 2 a wkly comp to win FA Cup fina...
5           spam  FreeMsg Hey there darling it's been 3 week's n...
8           spam  WINNER!! As a valued network customer you have...
9           spam  Had your mobile 11 months or more? U R entitle...
11          spam  SIX chances to win CASH! From 100 to 20,000 po...
...          ...                                                ...
5537        spam  Want explicit SEX in 30 secs? Ring 02073162414...
5540        spam  ASKED 3MOBILE IF 0870 CHATLINES INCLU IN FREE ...
5547        spam  Had your contract mobile 11 Mnths? Latest Moto...
5566        spam  REMINDER FROM O2: To get 2.50 pounds free call...
5567        spam  This is the 2nd time we have tried 2 contact u...

[747 rows x 2 columns]
```

```
spam_list= raw_spam['message'].tolist()
```

Connecting to Python 3 Google Compute Engine backend

27°C Mostly cl... ENG 10:45

Sign in or Register | SkillU | Project Submission Form | Inbox (473) - peer87374@ | AI_project spam or ham

colab.research.google.com/drive/1FANBC-Zz-N8qwqz7SRho4ajNCt25GEMu#scrollTo=oUptc1wOIKCh

Gmail YouTube Maps News Translate Google

AI_project spam or ham message detection

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk

```
[7] spam_list= raw_spam['message'].tolist()
```

```
[8] print(spam_list)
```

```
["Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 0845
```

```
import matplotlib.pyplot as ab
import numpy as np
labels = ['ham', 'spam']
counts = [4825, 747]
ypos = np.arange(len(labels)) #converting text labels to numberic value, 0 and 1
ypos
```

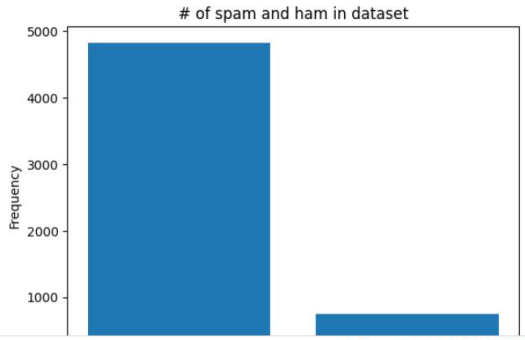
```
array([0, 1])
```

0s completed at 10:46 AM

Result ENG 10:47

```
ab.ylabel("Frequency")
ab.title("# of spam and ham in dataset")
ab.bar(ypos, counts)
```

```
<BarContainer object of 2 artists>
```



0s completed at 10:49 AM

```
[11] #replace the null values with a null string
     mail_data=raw_spam.where((pd.notnull(raw_spam)),'')
```

```
#printing the first five rows of the dataframe
mail_data.head()
```

	class_label	message
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
5	spam	FreeMsg Hey there darling it's been 3 week's n...
8	spam	WINNER!! As a valued network customer you have...
9	spam	Had your mobile 11 months or more? U R entitle...
11	spam	SIX chances to win CASH! From 100 to 20,000 po...

```
#chacking the number of rows and colomns in the dataframe
mail_data.shape
```

```
(747, 2)
```

0s completed at 10:50 AM

```
[15] mail_data.loc[mail_data['message']=='ham','message',] = 1
```

```
[16] #separating the data as texts and label
     x=mail_data['message']
     y=mail_data['class_label']
```

```
print(x)
```

```
2       Free entry in 2 a wkly comp to win FA Cup fina...
5       FreeMsg Hey there darling it's been 3 week's n...
8       WINNER!! As a valued network customer you have...
9       Had your mobile 11 months or more? U R entitle...
11      SIX chances to win CASH! From 100 to 20,000 po...
                                   ...
5537    Want explicit SEX in 30 secs? Ring 02073162414...
5540    ASKED 3MOBILE IF 0870 CHATLINES INCLU IN FREE ...
5547    Had your contract mobile 11 Mnths? Latest Moto...
5566    REMINDER FROM O2: To get 2.50 pounds free call...
5567    This is the 2nd time we have tried 2 contact u...
Name: message, Length: 747, dtype: object
```

splitting the data into training data and testing data

Double-click (or enter) to edit

```
[19] x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=3)
```

```
print(x.shape)
print(x_train.shape)
print(x_test.shape)
```

```
(747,)
(597,)
(150,)
```

```
[ ] import nltk
    nltk.download('stopwords')
    from nltk.corpus import stopwords

    #remove the punctuations and stopwords
    import string
```

```python
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords

#remove the punctuations and stopwords
import string
def message_process(message):

    message = message.translate(str.maketrans('', '', string.punctuation))
    message = [word for word in message.split() if word.lower() not in stopwords.words('english')]

    return " ".join(message)

raw_spam['message'] = raw_spam['message'].apply(message_process)
raw_spam.head()
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

	class_label	message
2	spam	Free entry 2 wkly comp win FA Cup final tkts 2...
5	spam	FreeMsg Hey darling 3 weeks word back Id like ...

```
11        spam   SIX chances win CASH 100 20000 pounds txt CSH1...
```

```
[ ]
```

```python
text = pd.DataFrame(raw_spam['message'])
label = pd.DataFrame(raw_spam['class_label'])
from collections import Counter

total_counts = Counter()
for i in range(len(text)):
    for word in text.values[i][0].split(" "):
        total_counts[word] += 1

print("Total words in data set: ", len(total_counts))
```

```
Total words in data set:  3436
```

```python
vocab = sorted(total_counts, key=total_counts.get, reverse=True)
print(vocab[:60])
```

```
['call', '2', 'Call', '4', 'FREE', 'ur', 'mobile', 'U', 'claim', 'prize', 'text', 'txt', 'Txt', 'free', 'u', 'STOP', 'reply', 'contact', '16',
```

```python
vocab_size = len(vocab)
word2idx = {}
#print vocab_size
for i, word in enumerate(vocab):
    word2idx[word] = i
    # Text to Vector
def text_to_vector(text):
    word_vector = np.zeros(vocab_size)
    for word in text.split(" "):
        if word2idx.get(word) is None:
            continue
        else:
            word_vector[word2idx.get(word)] += 1
    return np.array(word_vector)
# Convert all titles to vectors
word_vectors = np.zeros((len(text), len(vocab)), dtype=np.int_)
for i, (_, text_) in enumerate(text.iterrows()):
    word_vectors[i] = text_to_vector(text_[0])

word_vectors.shape
```

```
(747, 3436)
```

```python
import os
import numpy as np
from wordcloud import WordCloud
from PIL import Image

# Assuming you have loaded your DataFrame 'df_spam' and extracted the 'message' column into 'spam_list
spam_list = raw_spam['message'].tolist()

# Combine the text from 'spam_list' into a single string
filtered_spam = ' '.join(spam_list).lower()

# Load the comment mask image
comment_mask = np.array(Image.open("/content/comment.png"))

# Create and generate a word cloud image
wordcloud = WordCloud(
    max_font_size=160,
    margin=0,
    mask=comment_mask,
    background_color="white",
    colormap="Reds"
).generate(filtered_spam)
```


```
        v1                                                v2 Unnamed: 2  \
0      ham  Go until jurong point, crazy.. Available only ...        NaN
1      ham                      Ok lar... Joking wif u oni...        NaN
2     spam  Free entry in 2 a wkly comp to win FA Cup fina...        NaN
3      ham  U dun say so early hor... U c already then say...        NaN
4      ham  Nah I don't think he goes to usf, he lives aro...        NaN
...    ...                                                ...        ...
5567  spam  This is the 2nd time we have tried 2 contact u...        NaN
5568   ham              Will Ì_ b going to esplanade fr home?        NaN
5569   ham  Pity, * was in mood for that. So...any other s...        NaN
5570   ham  The guy did some bitching but I acted like i'd...        NaN
5571   ham                        Rofl. Its true to its name        NaN

     Unnamed: 3 Unnamed: 4
0           NaN        NaN
1           NaN        NaN
2           NaN        NaN
3           NaN        NaN
4           NaN        NaN
...         ...        ...
5567        NaN        NaN
5568        NaN        NaN
5569        NaN        NaN
5570        NaN        NaN
5571        NaN        NaN

[5572 rows x 5 columns]
```

```
5569        NaN        NaN
5570        NaN        NaN
5571        NaN        NaN

[5572 rows x 5 columns]
```