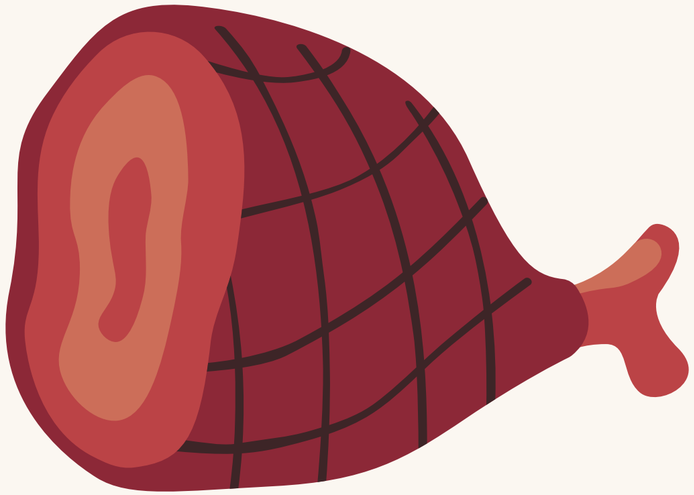
Spam OR Ham



Agenda

1. Introduction
2. Data Cleaning
3. Preprocessing&NLP Techniques
4. EDA
5. Machine learning model
6. Applications

Introduction

What is the difference between SPAM and HAM?

Spam is considered to be unsolicited messages with commercial and malicious intent compared to legitimate messages termed **ham**

ABOUT ME

Yasmin Kamal

YasminKamal-1410

Yasmin-0000


File Eat View Help
Data Cleaning

Before cleaning

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

5572 rows × 5 columns

```
v1              0
v2              0
Unnamed: 2   5522
Unnamed: 3   5560
Unnamed: 4   5566
dtype: int64
```

Number of null values

drop [Unnamed:2, Unnamed:3,Unnamed:4]

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   v1      5572 non-null   object
 1   v2      5572 non-null   object
dtypes: object(2)
memory usage: 87.2+ KB
```

dataset info

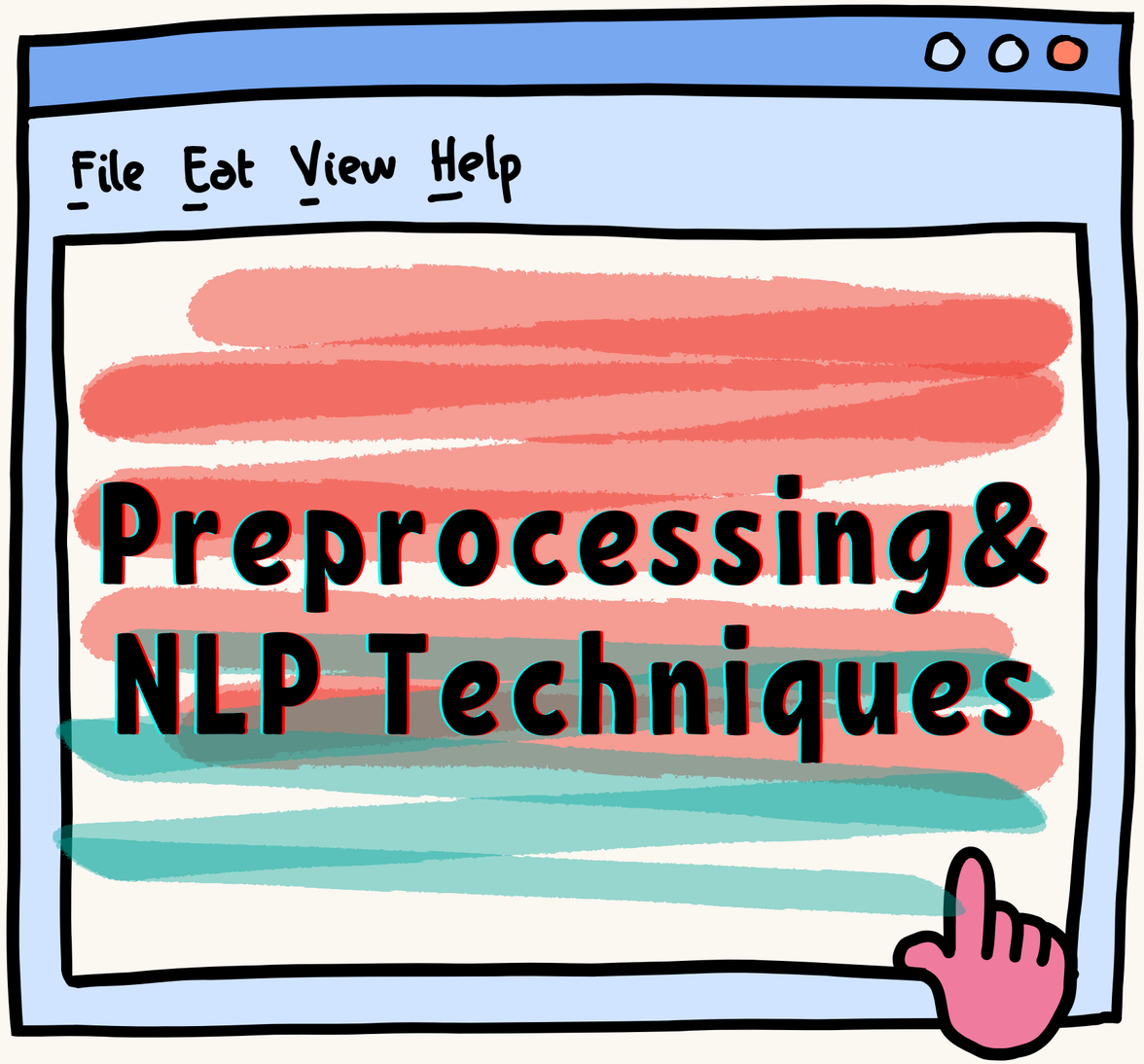
(5572, 2) Shape(Before)

```
df.drop_duplicates(inplace=True)
```

Shape(After) (5169, 2)

After Cleaning

	v1	v2
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...


File Eat View Help
Preprocessing&
NLP Techniques

```
df['v1'] = df['v1'].map({'ham': 0, 'spam': 1})
df.head()
```

- Encoder v1 from (ham, spam) to (0 ,1)
- It is helps in machine learning model because ml deal with numeric data

v1	v2
0	Go until jurong point, crazy.. Available only ...
0	Ok lar... Joking wif u oni...
1	Free entry in 2 a wkly comp to win FA Cup fina...
0	U dun say so early hor... U c already then say...
0	Nah I don't think he goes to usf, he lives aro...

v1	v2	char_count	word_count	num_sentences
0	Go until jurong point, crazy.. Available only ...	111	24	2
0	Ok lar... Joking wif u oni...	29	8	2
1	Free entry in 2 a wkly comp to win FA Cup fina...	155	37	2
0	U dun say so early hor... U c already then say...	49	13	1
0	Nah I don't think he goes to usf, he lives aro...	61	15	1

```
df['char_count'] = df['v2'].apply(len)
#column with len of words
df['word_count'] = df['v2'].apply(lambda x: len(nltk.word_tokenize(x)))
# column with len of sentences
df['num_sentences'] = df['v2'].apply(lambda x: len(nltk.sent_tokenize(x)))
```

- create new columns('char_count','num_sentences','word_count')
- it is helps in EDA section for making analysis and charts that necessary for understanding the data

NLIK Library

1

```
Tabnine | Edit | Test | Explain | Document | Ask
def transform_text(text):
    text = text.lower()
    text = nltk.word_tokenize(text)
    ps = PorterStemmer()

    y = []
    for i in text:
        if i.isalnum():
            y.append(i)

    text = y[:]
    y.clear()

    for i in text:
        if i not in stopwords.words('english') and i not in string.punctuation:
            y.append(i)

    text = y[:]
    y.clear()

    for i in text:
        y.append(ps.stem(i))

    return " ".join(y)
```

Transform words to lower case and remove stop words and special charecters

2

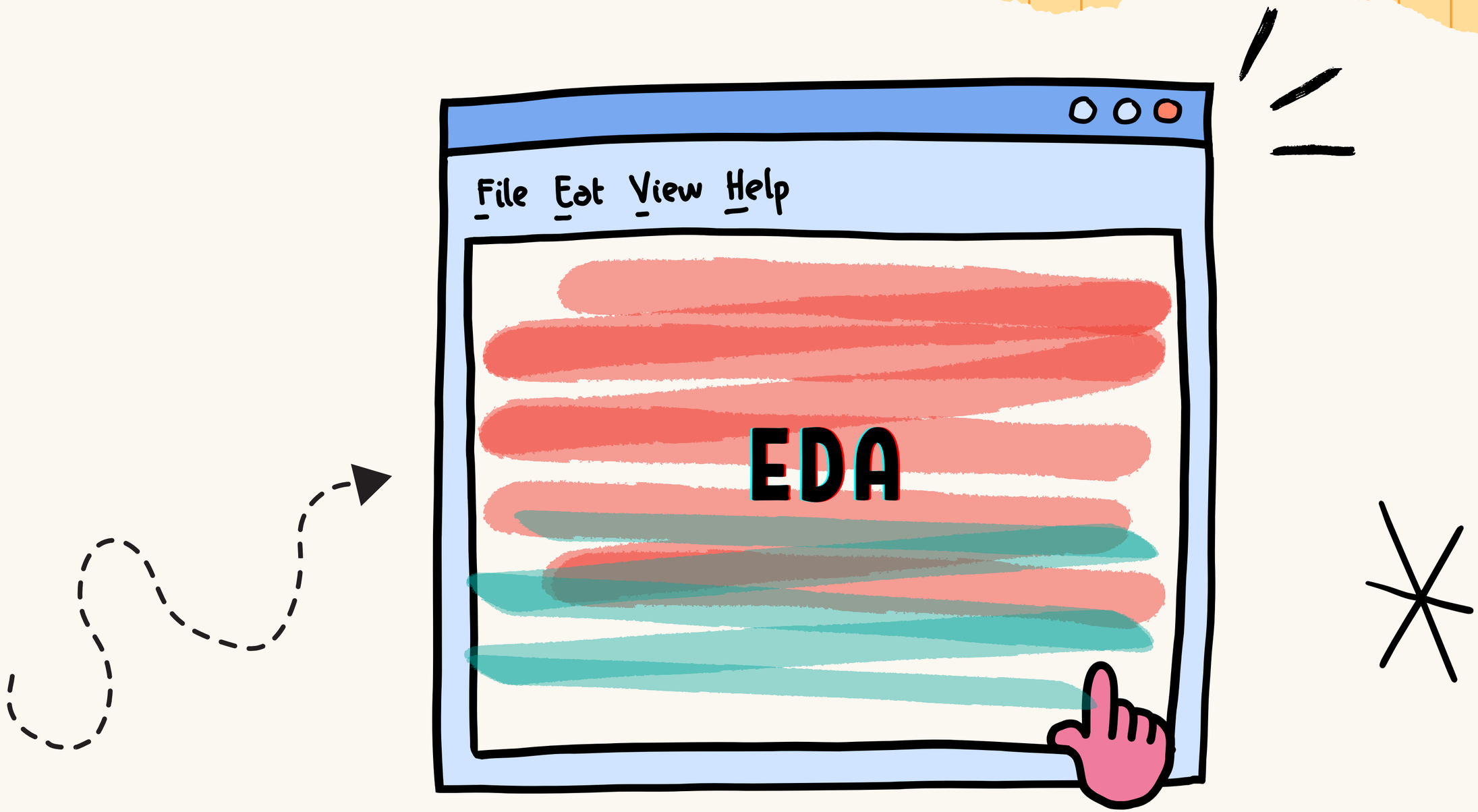
```
df['v2'][500]
```
Python
```
'Fighting with the world is easy, u either win or lose bt fightng with some1 who is close to u is dificult if u lose - u lose if u win - u stil
```
```
transform_text(df['v2'][500])
```
Python
```
'fight world easi u either win lose bt fightng some1 close u dificult u lose u lose u win u still lose'
```
```
df['transformed_text'] = df['v2'].apply(transform_text)
```
Python

Check transform_text () performance

3

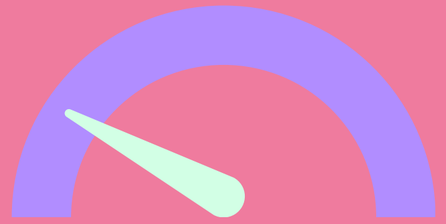
	v1	v2	char_count	word_count	num_sentences	transformed_text
0		Go until jurong point, crazy.. Available only ...	111	24	2	go jurong point crazi avail bugi n great world...
0		Ok lar... Joking wif u oni...	29	8	2	ok lar joke wif u oni
1		Free entry in 2 a wkly comp to win FA Cup fina...	155	37	2	free entri 2 wkli comp win fa cup final tkt 21...
0		U dun say so early hor... U c already then say...	49	13	1	u dun say earli hor u c alreadi say
0		Nah I don't think he goes to usf, he lives aro...	61	15	1	nah think goe usf live around though

add new column transform text

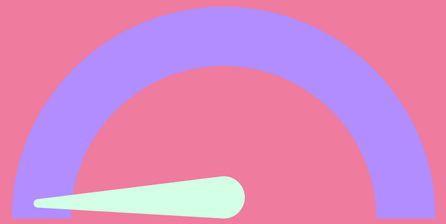

File Eat View Help
EDA

IMPORTANT KEY INSIGHTS

Avg chart number

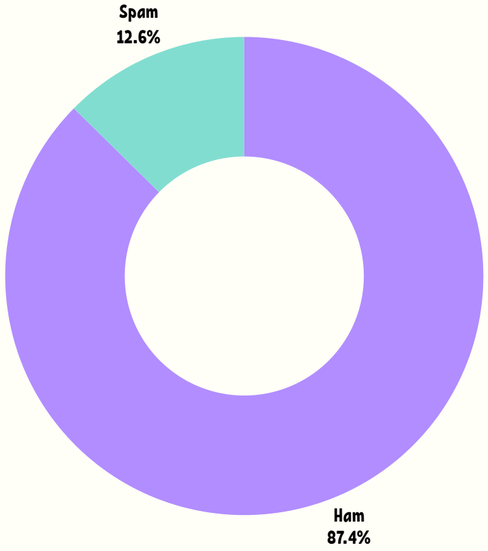


Avg word number

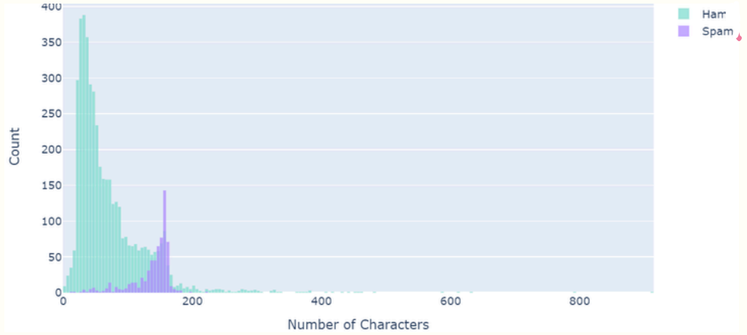


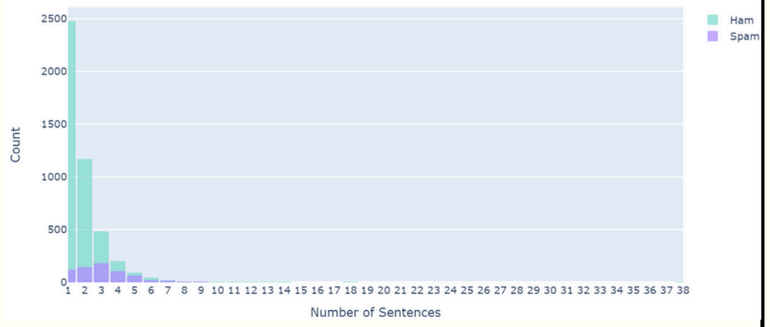
Avg sentences number

1



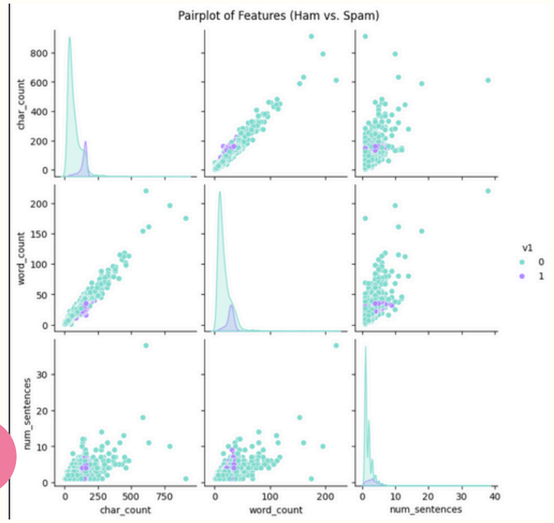
2



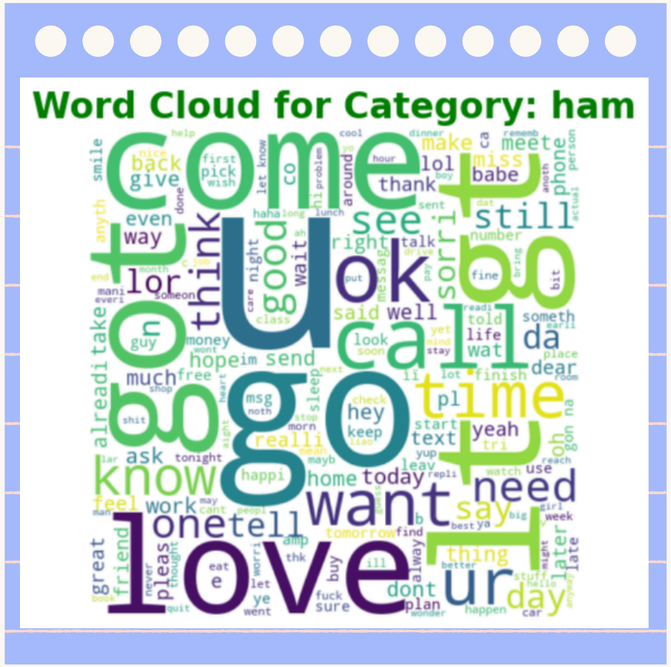


3

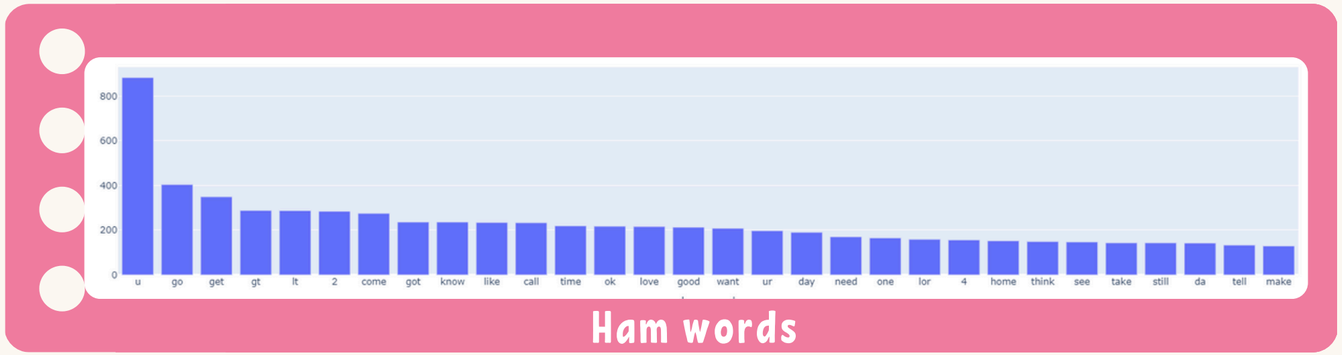
4

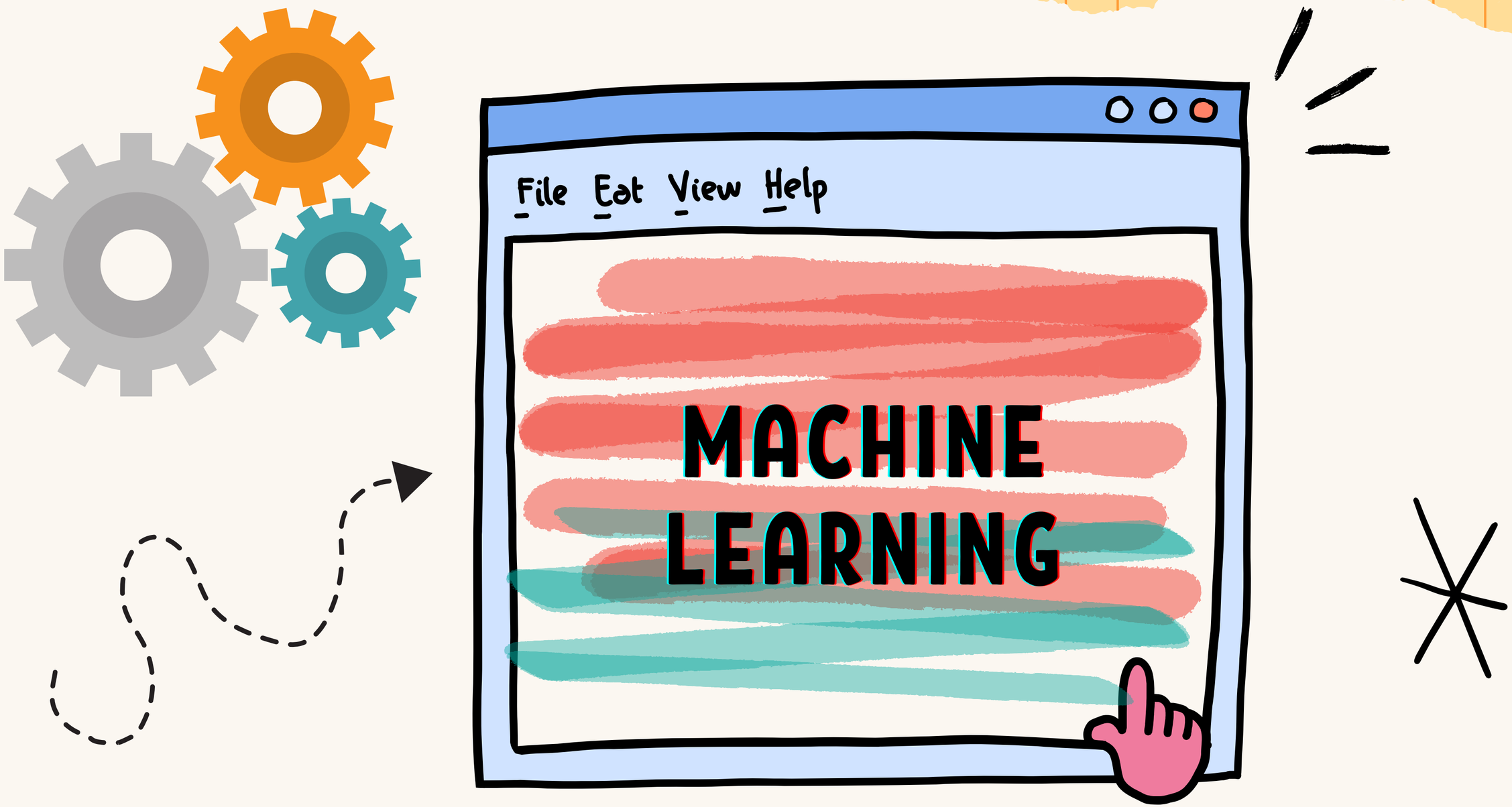


Most common words for Ham



Most common words for spam



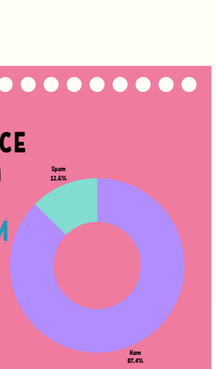

File Eat View Help
MACHINE LEARNING

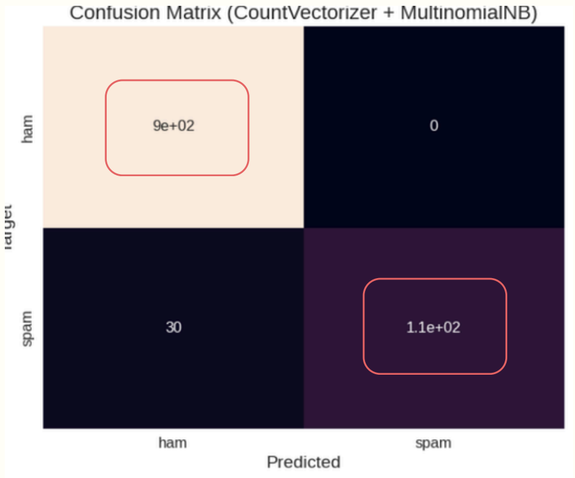
MULTINOMIALNB MODEL (BEFORE CROSS - VALIDATION)

	precision	recall	f1-score	support
0	0.97	1.00	0.98	896
1	1.00	0.78	0.88	138
accuracy			0.97	1034
macro avg	0.98	0.89	0.93	1034
weighted avg	0.97	0.97	0.97	1034

RECALL NOT THAT BAD BUT AFTER CROSS VALIDATION IT WOULD BE BETTER THAN BEFORE

AS WE CAN SEE, THERE IS NO BALANCE BETWEEN SPAM AND HAM. THE DATA SET SHOWS THAT HAM PRESENTS 87.4% FROM THE DATA, WHILE SPAM PRESENTS 12.6%.





IMPORTANT COLOUMNS

Y <- v1

v2

char_count

word_count

num_sentences

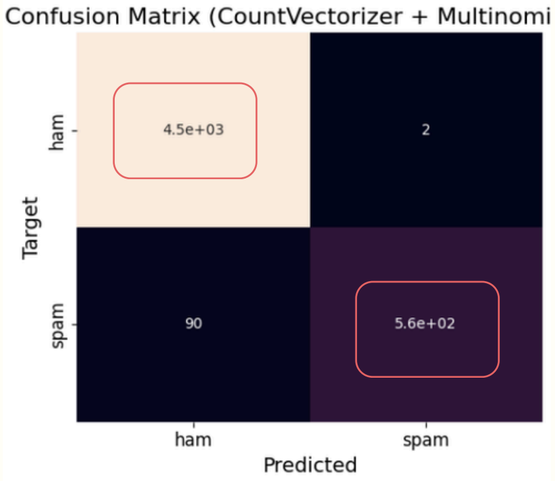
X <- transformed_text

MULTINOMIALNB MODEL (AFTER CROSS - VALIDATION)

	precision	recall	f1-score	support
ham	0.98	1.00	0.99	4507
spam	1.00	0.86	0.92	653
accuracy			0.98	5160
macro avg	0.99	0.93	0.96	5160
weighted avg	0.98	0.98	0.98	5160

RECALL AFTER USIND CROSS-VALIDATION (Stratified K-Fold)

AS WE CAN SEE, AFTER CROSS-VALIDATION (STRATIFIED K-FOLD), THE BALANCE IS ACHIEVED BETWEEN HAM AND SPAM.



IMPORTANT COLOUMNS

Y <- v1

X <-
- v2
- char_count
- word_count
- num_sentences
- transformed_text

APPLICATIONS

GITHUB

STREAMLIT

HUGGING FACE

THANK YOU!