```
# IMPORTANT: RUN THIS CELL IN ORDER TO IMPORT YOUR KAGGLE DATA SOURCES,
# THEN FEEL FREE TO DELETE THIS CELL.
# NOTE: THIS NOTEBOOK ENVIRONMENT DIFFERS FROM KAGGLE'S PYTHON
# ENVIRONMENT SO THERE MAY BE MISSING LIBRARIES USED BY YOUR
# NOTEBOOK.
import os
import shutil
import kagglehub
vstepanenko_disaster_tweets_path = kagglehub.dataset_download('vstepanenko/disaster-tweets')

print('Data source import complete.')
```

```
# 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 "Sav
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```

```
/kaggle/input/disaster-tweets/tweets.csv
```

```
# Load your dataset
df_tweets = pd.read_csv('/kaggle/input/disaster-tweets/tweets.csv')
```

```
import pandas as pd
# Display the first few rows of the dataset
df_tweets.head()
```

	id	keyword	location	text	target
0	0	ablaze	NaN	Communal violence in Bhainsa, Telangana. "Ston...	1
1	1	ablaze	NaN	Telangana: Section 144 has been imposed in Bha...	1
2	2	ablaze	New York City	Arsonist sets cars ablaze at dealership https:...	1
3	3	ablaze	Morgantown, WV	Arsonist sets cars ablaze at dealership https:...	1
4	4	ablaze	NaN	"Lord Jesus, your love brings freedom and pard...	0

```
# Load your dataset
df = pd.read_csv('/kaggle/input/disaster-tweets/tweets.csv')
# Show basic statistics of the dataset (mean, std, min, etc.)
df.describe()
```

	id	target
count	11370.000000	11370.000000
mean	5684.500000	0.185928
std	3282.380615	0.389066
min	0.000000	0.000000
25%	2842.250000	0.000000
50%	5684.500000	0.000000
75%	8526.750000	0.000000
max	11369.000000	1.000000

```
# Show information about the dataset (data types, non-null values)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11370 entries, 0 to 11369
Data columns (total 5 columns):
```

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        11370 non-null  int64
 1   keyword   11370 non-null  object
 2   location  7952 non-null   object
 3   text      11370 non-null  object
 4   target    11370 non-null  int64
dtypes: int64(2), object(3)
memory usage: 444.3+ KB
```

```
# Show the shape of the dataset (number of rows and columns)
df.shape
```

```
(11370, 5)
```

```
# Check for any missing values in the dataset
df.isnull().sum()
```

```
id             0
keyword        0
location    3418
text           0
target         0
dtype: int64
```

```
#Data Exploration
def explore_data(df):


    '''Input- df= pandas dataframes to be explored
       Output- print shape, info and first 5 records of the dataframe
    '''

    print("-"*50)
    print('Shape of the dataframe:',df.shape)
    print("Number of records in train data set:",df.shape[0])
    print("Information of the dataset:")
    df.info()
    print("-"*50)
    print("First 5 records of the dataset:")
    return df.head()
    print("-"*50)
```

```
#Data Exploration
explore_data(df_tweets)
```

```
--------------------------------------------------
Shape of the dataframe: (11370, 5)
Number of records in train data set: 11370
Information of the dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11370 entries, 0 to 11369
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        11370 non-null  int64
 1   keyword   11370 non-null  object
 2   location  7952 non-null   object
 3   text      11370 non-null  object
 4   target    11370 non-null  int64
dtypes: int64(2), object(3)
memory usage: 444.3+ KB
--------------------------------------------------
First 5 records of the dataset:
```

	id	keyword	location	text	target
0	0	ablaze	NaN	Communal violence in Bhainsa, Telangana. "Ston...	1
1	1	ablaze	NaN	Telangana: Section 144 has been imposed in Bha...	1
2	2	ablaze	New York City	Arsonist sets cars ablaze at dealership https:...	1
3	3	ablaze	Morgantown, WV	Arsonist sets cars ablaze at dealership https:...	1
4	4	ablaze	NaN	"Lord Jesus, your love brings freedom and pard...	0

```
import pandas as pd

# Load the tweets dataset from the specified path
df = pd.read_csv('/kaggle/input/disaster-tweets/tweets.csv')

# Perform the lowercasing step
df['lowercased_tweet'] = df['text'].str.lower()
```

```
# Display the first few rows of the dataset with the new lowercased column
print(df[['text', 'lowercased_tweet']].head())
```

```
                                                text  \
0  Communal violence in Bhainsa, Telangana. "Ston...
1  Telangana: Section 144 has been imposed in Bha...
2  Arsonist sets cars ablaze at dealership https:...
3  Arsonist sets cars ablaze at dealership https:...
4  "Lord Jesus, your love brings freedom and pard...

                                       lowercased_tweet
0  communal violence in bhainsa, telangana. "ston...
1  telangana: section 144 has been imposed in bha...
2  arsonist sets cars ablaze at dealership https:...
3  arsonist sets cars ablaze at dealership https:...
4  "lord jesus, your love brings freedom and pard...
```

```
import pandas as pd
import re

# Load the tweets dataset from the specified path
df = pd.read_csv('/kaggle/input/disaster-tweets/tweets.csv')

# Define the function to remove special characters and punctuation
def remove_special_characters(text):
    # Remove special characters and punctuation, keeping only letters and spaces
    cleaned_text = re.sub(r'[^a-zA-Z\s]', '', text)
    return cleaned_text

# Apply the function to the 'text' column (assuming the column with tweets is 'text')
df['cleaned_text'] = df['text'].apply(remove_special_characters)

# Display the first few rows of the dataset with the original and cleaned text
print(df[['text', 'cleaned_text']].head())
```

```
                                                text  \
0  Communal violence in Bhainsa, Telangana. "Ston...
1  Telangana: Section 144 has been imposed in Bha...
2  Arsonist sets cars ablaze at dealership https:...
3  Arsonist sets cars ablaze at dealership https:...
4  "Lord Jesus, your love brings freedom and pard...

                                          cleaned_text
0  Communal violence in Bhainsa Telangana Stones ...
1  Telangana Section  has been imposed in Bhainsa...
2  Arsonist sets cars ablaze at dealership httpst...
3  Arsonist sets cars ablaze at dealership httpst...
4  Lord Jesus your love brings freedom and pardon...
```

```
import pandas as pd
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

# Download required NLTK resources (run these once)
nltk.download('punkt')
nltk.download('stopwords')

# Load the dataset from the specified path
df = pd.read_csv('/kaggle/input/disaster-tweets/tweets.csv')

# Get the list of stopwords
stop_words = set(stopwords.words('english'))

# Define the function to remove stopwords from the text
def remove_stopwords(text):
    # Tokenize the text
    tokens = word_tokenize(text)

    # Remove stopwords
    filtered_tokens = [word for word in tokens if word.lower() not in stop_words]

    # Join the tokens back into a string
    return ' '.join(filtered_tokens)

# Apply stopword removal to the 'text' column (assuming 'text' is the tweet text column)
df['tweet_no_stopwords'] = df['text'].apply(remove_stopwords)

# Display the first few rows with the original and stopword-removed tweets
```

```
print(df[['text', 'tweet_no_stopwords']].head())
```

```
[nltk_data] Downloading package punkt to /usr/share/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /usr/share/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
                                                text  \
0  Communal violence in Bhainsa, Telangana. "Ston...
1  Telangana: Section 144 has been imposed in Bha...
2  Arsonist sets cars ablaze at dealership https:...
3  Arsonist sets cars ablaze at dealership https:...
4  "Lord Jesus, your love brings freedom and pard...

                                   tweet_no_stopwords
0  Communal violence Bhainsa , Telangana . `` Sto...
1  Telangana : Section 144 imposed Bhainsa Januar...
2  Arsonist sets cars ablaze dealership https : /...
3  Arsonist sets cars ablaze dealership https : /...
4  `` Lord Jesus , love brings freedom pardon . F...
```

```
import pandas as pd
import re

# Load the tweets dataset from the specified path
df = pd.read_csv('/kaggle/input/disaster-tweets/tweets.csv')

# Define the function to remove emojis
def remove_emojis(text):
    # Print the original tweet
    #print("Original Tweet:", text)

    # Remove emojis and non-ASCII characters
    cleaned_text = re.sub(r'[^\x00-\x7F]+', '', text)
    #print("After Removing Emojis:", cleaned_text)
    #print('-' * 50)  # Separator for better readability
    return cleaned_text

# Apply the function to the first 5 rows of the 'text' column
for tweet in df['text'].head(5):
    remove_emojis(tweet)

# Optionally, create a new column for cleaned tweets
df['cleaned_tweet'] = df['text'].apply(remove_emojis)

# Display the cleaned tweets for the first 5 rows
#print("\nCleaned Tweets Dataset (First 5):")
#print(df[['text', 'cleaned_tweet']].head(5))
df.head(30)
```

	id	keyword	location	text	target	cleaned_tweet
0	0	ablaze	NaN	Communal violence in Bhainsa, Telangana. "Ston...	1	Communal violence in Bhainsa, Telangana. "Ston...
1	1	ablaze	NaN	Telangana: Section 144 has been imposed in Bha...	1	Telangana: Section 144 has been imposed in Bha...
2	2	ablaze	New York City	Arsonist sets cars ablaze at dealership https:...	1	Arsonist sets cars ablaze at dealership https:...
3	3	ablaze	Morgantown, WV	Arsonist sets cars ablaze at dealership https:...	1	Arsonist sets cars ablaze at dealership https:...
4	4	ablaze	NaN	"Lord Jesus, your love brings freedom and pard...	0	"Lord Jesus, your love brings freedom and pard...
5	5	ablaze	OC	If this child was Chinese, this tweet would ha...	0	If this child was Chinese, this tweet would ha...
6	6	ablaze	London, England	Several houses have been set ablaze in Ngemsib...	1	Several houses have been set ablaze in Ngemsib...
7	7	ablaze	Bharat	Asansol: A BJP office in Salanpur village was ...	1	Asansol: A BJP office in Salanpur village was ...
8	8	ablaze	Accra, Ghana	National Security Minister, Kan Dapaah's side ...	0	National Security Minister, Kan Dapaah's side ...
9	9	ablaze	Searching	This creature who's soul is no longer clarent ...	0	This creature whos soul is no longer clarent b...
10	10	ablaze	NaN	Images showing the havoc caused by the #Camero...	1	Images showing the havoc caused by the #Camero...
11	11	ablaze	NaN	Social media went bananas after Chuba Hubbard ...	0	Social media went bananas after Chuba Hubbard ...
12	12	ablaze	NaN	Hausa youths set Area Office of Apapa-Iganmu L...	1	Hausa youths set Area Office of Apapa-Iganmu L...
13	13	ablaze	HYDERABAD	Under #MamataBanerjee political violence &...	1	Under #MamataBanerjee political violence &...
14	14	ablaze	Reno, NV	AMEN! Set the whole system ablaze, man. https:...	0	AMEN! Set the whole system ablaze, man. https:...
15	15	ablaze	NaN	Images showing the havoc caused by the #Camero...	1	Images showing the havoc caused by the #Camero...
16	16	ablaze	NaN	No cows today but our local factory is sadly s...	1	No cows today but our local factory is sadly s...
17	17	ablaze	NaN	Rengoku sets my heart ablaze 😔 ❤️ 🔥 P.s. I missed...	0	Rengoku sets my heart ablaze P.s. I missed thi...
18	18	ablaze	Worldwide	paulzizkaphoto: “Rundle Ablaze” Wishing you al...	0	paulzizkaphoto: Rundle Ablaze Wishing you all ...

```
import pandas as pd
import re

# Load the tweets dataset from the specified path
df = pd.read_csv('/kaggle/input/disaster-tweets/tweets.csv')

# Define the function to remove hashtags
def remove_hashtags(text):
    # Remove hashtags using regex
    cleaned_text = re.sub(r'#\w+', '', text)
    return cleaned_text.strip()  # Remove leading/trailing whitespace

# Apply the function to the entire 'text' column
df['cleaned_tweet'] = df['text'].apply(remove_hashtags)

# Print the original and cleaned tweets for the first 5 rows
print("\nCleaned Tweets Dataset (First 5):")
for original, cleaned in zip(df['text'].head(15), df['cleaned_tweet'].head(15)):
    print("Original Tweet:", original)
    print("After Removing Hashtags:", cleaned)
    #print('-' * 50)  # Separator for better readability
```

```
Cleaned Tweets Dataset (First 5):
Original Tweet: Communal violence in Bhainsa, Telangana. "Stones were pelted on Muslims' houses and some houses and vehicles were se
After Removing Hashtags: Communal violence in Bhainsa, Telangana. "Stones were pelted on Muslims' houses and some houses and vehicle
Original Tweet: Telangana: Section 144 has been imposed in Bhainsa from January 13 to 15, after clash erupted between two groups on
After Removing Hashtags: Telangana: Section 144 has been imposed in Bhainsa from January 13 to 15, after clash erupted between two g
Original Tweet: Arsonist sets cars ablaze at dealership https://t.co/gOQvyJbpVI
```

```
After Removing Hashtags: Arsonist sets cars ablaze at dealership https://t.co/gOQvyJbpVI
Original Tweet: Arsonist sets cars ablaze at dealership https://t.co/0gL7NUCPlb https://t.co/u1CcBhOWh9
After Removing Hashtags: Arsonist sets cars ablaze at dealership https://t.co/0gL7NUCPlb https://t.co/u1CcBhOWh9
Original Tweet: "Lord Jesus, your love brings freedom and pardon. Fill me with your Holy Spirit and set my heart ablaze with your l…
After Removing Hashtags: "Lord Jesus, your love brings freedom and pardon. Fill me with your Holy Spirit and set my heart ablaze wit
Original Tweet: If this child was Chinese, this tweet would have gone viral. Social media would be ablaze. SNL would have made a rac
After Removing Hashtags: If this child was Chinese, this tweet would have gone viral. Social media would be ablaze. SNL would have m
Original Tweet: Several houses have been set ablaze in Ngemsibaa village, Oku sub division in the North West Region of Cameroon by…
After Removing Hashtags: Several houses have been set ablaze in Ngemsibaa village, Oku sub division in the North West Region of Came
Original Tweet: Asansol: A BJP office in Salanpur village was set ablaze last night. BJP has alleged that TMC is behind the incident
After Removing Hashtags: Asansol: A BJP office in Salanpur village was set ablaze last night. BJP has alleged that TMC is behind the
Original Tweet: National Security Minister, Kan Dapaah's side chic has set the internet ablaze with her latest powerful video.… http
After Removing Hashtags: National Security Minister, Kan Dapaah's side chic has set the internet ablaze with her latest powerful vid
Original Tweet: This creature who’s soul is no longer clarent but blue ablaze This thing Carrying memories Memories of… https://t.co
After Removing Hashtags: This creature who’s soul is no longer clarent but blue ablaze This thing Carrying memories Memories of… http
Original Tweet: Images showing the havoc caused by the #Cameroon military as they torched houses in #Oku.The shameless military is r
After Removing Hashtags: Images showing the havoc caused by the  military as they torched houses in .The shameless military is repor
Original Tweet: Social media went bananas after Chuba Hubbard announced Monday evening his plans to return to #okstate. https://t.co
After Removing Hashtags: Social media went bananas after Chuba Hubbard announced Monday evening his plans to return to . https://t.
Original Tweet: Hausa youths set Area Office of Apapa-Iganmu Local Council Development Area ablaze. Okada Riders stormed the LG area
After Removing Hashtags: Hausa youths set Area Office of Apapa-Iganmu Local Council Development Area ablaze. Okada Riders stormed th
Original Tweet: Under #MamataBanerjee political violence &amp;amp; vandalism continues to unabated in West Bengal! office in Asanol was…
After Removing Hashtags: Under  political violence &amp;amp; vandalism continues to unabated in West Bengal! office in Asanol was…
Original Tweet: AMEN! Set the whole system ablaze, man. https://t.co/J08xHDcGbD
After Removing Hashtags: AMEN! Set the whole system ablaze, man. https://t.co/J08xHDcGbD
```

```
import pandas as pd
import re

# Load the tweets dataset from the specified path
df = pd.read_csv('/kaggle/input/disaster-tweets/tweets.csv')

# Define the function to remove URLs
def remove_urls(text):
    # Remove URLs using regex
    cleaned_text = re.sub(r'http\S+|www\S+|https\S+', '', text, flags=re.MULTILINE)
    return cleaned_text.strip()  # Remove leading/trailing whitespace

# Apply the function to the entire 'text' column
df['cleaned_tweet'] = df['text'].apply(remove_urls)

# Print the original and cleaned tweets for the first 5 rows
print("\nCleaned Tweets Dataset (First 5):")
for original, cleaned in zip(df['text'].head(5), df['cleaned_tweet'].head(5)):
    print("Original Tweet:", original)
    print("After Removing URLs:", cleaned)
    #print('-' * 50)  # Separator for better readability
```

```
Cleaned Tweets Dataset (First 5):
Original Tweet: Communal violence in Bhainsa, Telangana. "Stones were pelted on Muslims' houses and some houses and vehicles were se
After Removing URLs: Communal violence in Bhainsa, Telangana. "Stones were pelted on Muslims' houses and some houses and vehicles we
--------------------------------------------------
Original Tweet: Telangana: Section 144 has been imposed in Bhainsa from January 13 to 15, after clash erupted between two groups on
After Removing URLs: Telangana: Section 144 has been imposed in Bhainsa from January 13 to 15, after clash erupted between two group
--------------------------------------------------
Original Tweet: Arsonist sets cars ablaze at dealership https://t.co/gOQvyJbpVI
After Removing URLs: Arsonist sets cars ablaze at dealership
--------------------------------------------------
Original Tweet: Arsonist sets cars ablaze at dealership https://t.co/0gL7NUCPlb https://t.co/u1CcBhOWh9
After Removing URLs: Arsonist sets cars ablaze at dealership
--------------------------------------------------
Original Tweet: "Lord Jesus, your love brings freedom and pardon. Fill me with your Holy Spirit and set my heart ablaze with your l…
After Removing URLs: "Lord Jesus, your love brings freedom and pardon. Fill me with your Holy Spirit and set my heart ablaze with yo
--------------------------------------------------
```

```
import pandas as pd
import re

# Load the tweets dataset from the specified path
df = pd.read_csv('/kaggle/input/disaster-tweets/tweets.csv')

# Define the function to remove special characters
def remove_special_characters(text):
    # Remove special characters using regex
    cleaned_text = re.sub(r'[^a-zA-Z0-9\s]', '', text)
    return cleaned_text.strip()  # Remove leading/trailing whitespace

# Apply the function to the entire 'text' column
df['cleaned_tweet'] = df['text'].apply(remove_special_characters)
```

```
# Print the original and cleaned tweets for the first 5 rows
print("\nCleaned Tweets Dataset (First 5):")
for original, cleaned in zip(df['text'].head(15), df['cleaned_tweet'].head(15)):
    print("Original Tweet:", original)
    print("After Removing Special Characters:", cleaned)
    print('-' * 50)  # Separator for better readability
```

```
Cleaned Tweets Dataset (First 5):
Original Tweet: Communal violence in Bhainsa, Telangana. "Stones were pelted on Muslims' houses and some houses and vehicles were se
After Removing Special Characters: Communal violence in Bhainsa Telangana Stones were pelted on Muslims houses and some houses and v
--------------------------------------------------
Original Tweet: Telangana: Section 144 has been imposed in Bhainsa from January 13 to 15, after clash erupted between two groups on
After Removing Special Characters: Telangana Section 144 has been imposed in Bhainsa from January 13 to 15 after clash erupted betwe
--------------------------------------------------
Original Tweet: Arsonist sets cars ablaze at dealership https://t.co/gOQvyJbpVI
After Removing Special Characters: Arsonist sets cars ablaze at dealership httpstcogOQvyJbpVI
--------------------------------------------------
Original Tweet: Arsonist sets cars ablaze at dealership https://t.co/0gL7NUCPlb https://t.co/u1CcBhOWh9
After Removing Special Characters: Arsonist sets cars ablaze at dealership httpstco0gL7NUCPlb httpstcou1CcBhOWh9
--------------------------------------------------
Original Tweet: "Lord Jesus, your love brings freedom and pardon. Fill me with your Holy Spirit and set my heart ablaze with your l…
After Removing Special Characters: Lord Jesus your love brings freedom and pardon Fill me with your Holy Spirit and set my heart abl
--------------------------------------------------
Original Tweet: If this child was Chinese, this tweet would have gone viral. Social media would be ablaze. SNL would have made a rac
After Removing Special Characters: If this child was Chinese this tweet would have gone viral Social media would be ablaze SNL woulc
--------------------------------------------------
Original Tweet: Several houses have been set ablaze in Ngemsibaa village, Oku sub division in the North West Region of Cameroon by…
After Removing Special Characters: Several houses have been set ablaze in Ngemsibaa village Oku sub division in the North West Regic
--------------------------------------------------
Original Tweet: Asansol: A BJP office in Salanpur village was set ablaze last night. BJP has alleged that TMC is behind the incident
After Removing Special Characters: Asansol A BJP office in Salanpur village was set ablaze last night BJP has alleged that TMC is be
--------------------------------------------------
Original Tweet: National Security Minister, Kan Dapaah's side chic has set the internet ablaze with her latest powerful video.… http
After Removing Special Characters: National Security Minister Kan Dapaahs side chic has set the internet ablaze with her latest powe
--------------------------------------------------
Original Tweet: This creature who’s soul is no longer clarent but blue ablaze This thing Carrying memories Memories of… https://t.c
After Removing Special Characters: This creature whos soul is no longer clarent but blue ablaze This thing Carrying memories Memorie
--------------------------------------------------
Original Tweet: Images showing the havoc caused by the #Cameroon military as they torched houses in #Oku.The shameless military is r
After Removing Special Characters: Images showing the havoc caused by the Cameroon military as they torched houses in OkuThe shamele
--------------------------------------------------
Original Tweet: Social media went bananas after Chuba Hubbard announced Monday evening his plans to return to #okstate. https://t.c
After Removing Special Characters: Social media went bananas after Chuba Hubbard announced Monday evening his plans to return to oks
--------------------------------------------------
Original Tweet: Hausa youths set Area Office of Apapa-Iganmu Local Council Development Area ablaze. Okada Riders stormed the LG area
After Removing Special Characters: Hausa youths set Area Office of ApapaIganmu Local Council Development Area ablaze Okada Riders st
--------------------------------------------------
Original Tweet: Under #MamataBanerjee political violence &amp;amp; vandalism continues to unabated in West Bengal! office in Asanol was…
After Removing Special Characters: Under MamataBanerjee political violence amp vandalism continues to unabated in West Bengal office
--------------------------------------------------
Original Tweet: AMEN! Set the whole system ablaze, man. https://t.co/J08xHDcGbD
After Removing Special Characters: AMEN Set the whole system ablaze man httpstcoJ08xHDcGbD
--------------------------------------------------
```

```
import pandas as pd
import re

# Load the tweets dataset from the specified path
df = pd.read_csv('/kaggle/input/disaster-tweets/tweets.csv')

# Define the function to remove punctuation
def remove_punctuation(text):
    # Remove punctuation using regex
    cleaned_text = re.sub(r'[^\w\s]', '', text)
    return cleaned_text

# Apply the function to the entire 'text' column
df['cleaned_tweet'] = df['text'].apply(remove_punctuation)

# Print the original and cleaned tweets for the first 5 rows
print("\nCleaned Tweets Dataset (First 5):")
for original, cleaned in zip(df['text'].head(5), df['cleaned_tweet'].head(5)):
    print("Original Tweet:", original)
    print("After Removing Punctuation:", cleaned)
    #print('-' * 50)  # Separator for better readability
```

```
Cleaned Tweets Dataset (First 5):
Original Tweet: Communal violence in Bhainsa, Telangana. "Stones were pelted on Muslims' houses and some houses and vehicles were se
After Removing Punctuation: Communal violence in Bhainsa Telangana Stones were pelted on Muslims houses and some houses and vehicles
Original Tweet: Telangana: Section 144 has been imposed in Bhainsa from January 13 to 15, after clash erupted between two groups on
After Removing Punctuation: Telangana Section 144 has been imposed in Bhainsa from January 13 to 15 after clash erupted between two
```

```
Original Tweet: Arsonist sets cars ablaze at dealership https://t.co/gOQvyJbpVI
After Removing Punctuation: Arsonist sets cars ablaze at dealership httpstcogOQvyJbpVI
Original Tweet: Arsonist sets cars ablaze at dealership https://t.co/0gL7NUCPlb https://t.co/u1CcBhOWh9
After Removing Punctuation: Arsonist sets cars ablaze at dealership httpstco0gL7NUCPlb httpstcou1CcBhOWh9
Original Tweet: "Lord Jesus, your love brings freedom and pardon. Fill me with your Holy Spirit and set my heart ablaze with your l…
After Removing Punctuation: Lord Jesus your love brings freedom and pardon Fill me with your Holy Spirit and set my heart ablaze wit
```

```
!pip install seaborn
```

```
Requirement already satisfied: seaborn in /opt/conda/lib/python3.10/site-packages (0.12.2)
Requirement already satisfied: numpy!=1.24.0,>=1.17 in /opt/conda/lib/python3.10/site-packages (from seaborn) (1.26.4)
Requirement already satisfied: pandas>=0.25 in /opt/conda/lib/python3.10/site-packages (from seaborn) (2.2.3)
Requirement already satisfied: matplotlib!=3.6.1,>=3.1 in /opt/conda/lib/python3.10/site-packages (from seaborn) (3.7.5)
Requirement already satisfied: contourpy>=1.0.1 in /opt/conda/lib/python3.10/site-packages (from matplotlib!=3.6.1,>=3.1->seaborn)
Requirement already satisfied: cycler>=0.10 in /opt/conda/lib/python3.10/site-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (0.1
Requirement already satisfied: fonttools>=4.22.0 in /opt/conda/lib/python3.10/site-packages (from matplotlib!=3.6.1,>=3.1->seaborn)
Requirement already satisfied: kiwisolver>=1.0.1 in /opt/conda/lib/python3.10/site-packages (from matplotlib!=3.6.1,>=3.1->seaborn)
Requirement already satisfied: packaging>=20.0 in /opt/conda/lib/python3.10/site-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (
Requirement already satisfied: pillow>=6.2.0 in /opt/conda/lib/python3.10/site-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (10
Requirement already satisfied: pyparsing>=2.3.1 in /opt/conda/lib/python3.10/site-packages (from matplotlib!=3.6.1,>=3.1->seaborn)
Requirement already satisfied: python-dateutil>=2.7 in /opt/conda/lib/python3.10/site-packages (from matplotlib!=3.6.1,>=3.1->seabor
Requirement already satisfied: pytz>=2020.1 in /opt/conda/lib/python3.10/site-packages (from pandas>=0.25->seaborn) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in /opt/conda/lib/python3.10/site-packages (from pandas>=0.25->seaborn) (2024.1)
Requirement already satisfied: six>=1.5 in /opt/conda/lib/python3.10/site-packages (from python-dateutil>=2.7->matplotlib!=3.6.1,>=
```

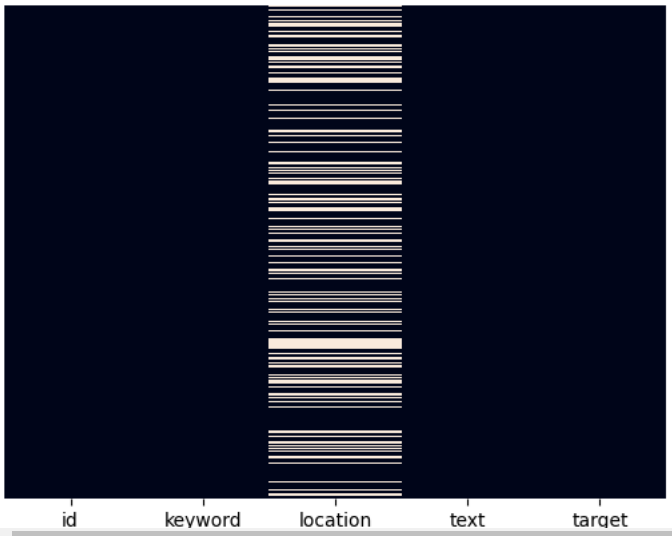
```
!pip install plotly
```

```
Requirement already satisfied: plotly in /opt/conda/lib/python3.10/site-packages (5.22.0)
Requirement already satisfied: tenacity>=6.2.0 in /opt/conda/lib/python3.10/site-packages (from plotly) (8.3.0)
Requirement already satisfied: packaging in /opt/conda/lib/python3.10/site-packages (from plotly) (21.3)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /opt/conda/lib/python3.10/site-packages (from packaging->plotly) (3.1.2)
```

```
import seaborn as sns
import pandas as py
import plotly.graph_objs as go
import plotly.offline as py
# Initialize Plotly for offline mode in Colab
py.init_notebook_mode(connected=True)
def missing_values(df):
    print('{}% of location values are missing from Total Number of Records.'.format(round((df.location.isnull().sum())/(df.shape[0])*10
    print('{}% of keywords values are missing from Total Number of Records.'.format(round((df.keyword.isnull().sum())/(df.shape[0])*100
    sns.heatmap(df.isnull(),yticklabels=False,cbar=False)
    null_feat = pd.DataFrame(len(df['id']) - df.isnull().sum(), columns = ['Count'])

    trace = go.Bar(x = null_feat.index, y = null_feat['Count'] ,opacity = 0.8, marker=dict(color = 'lightgrey', line=dict(color='#00000

    layout = dict(title =  "Missing Values")

    fig = dict(data = [trace], layout=layout)
    py.iplot(fig)
```

```
#Displays Missing values by using diagrams
missing_values(df_tweets)
```

```
30% of location values are missing from Total Number of Records.
0% of keywords values are missing from Total Number of Records.
```

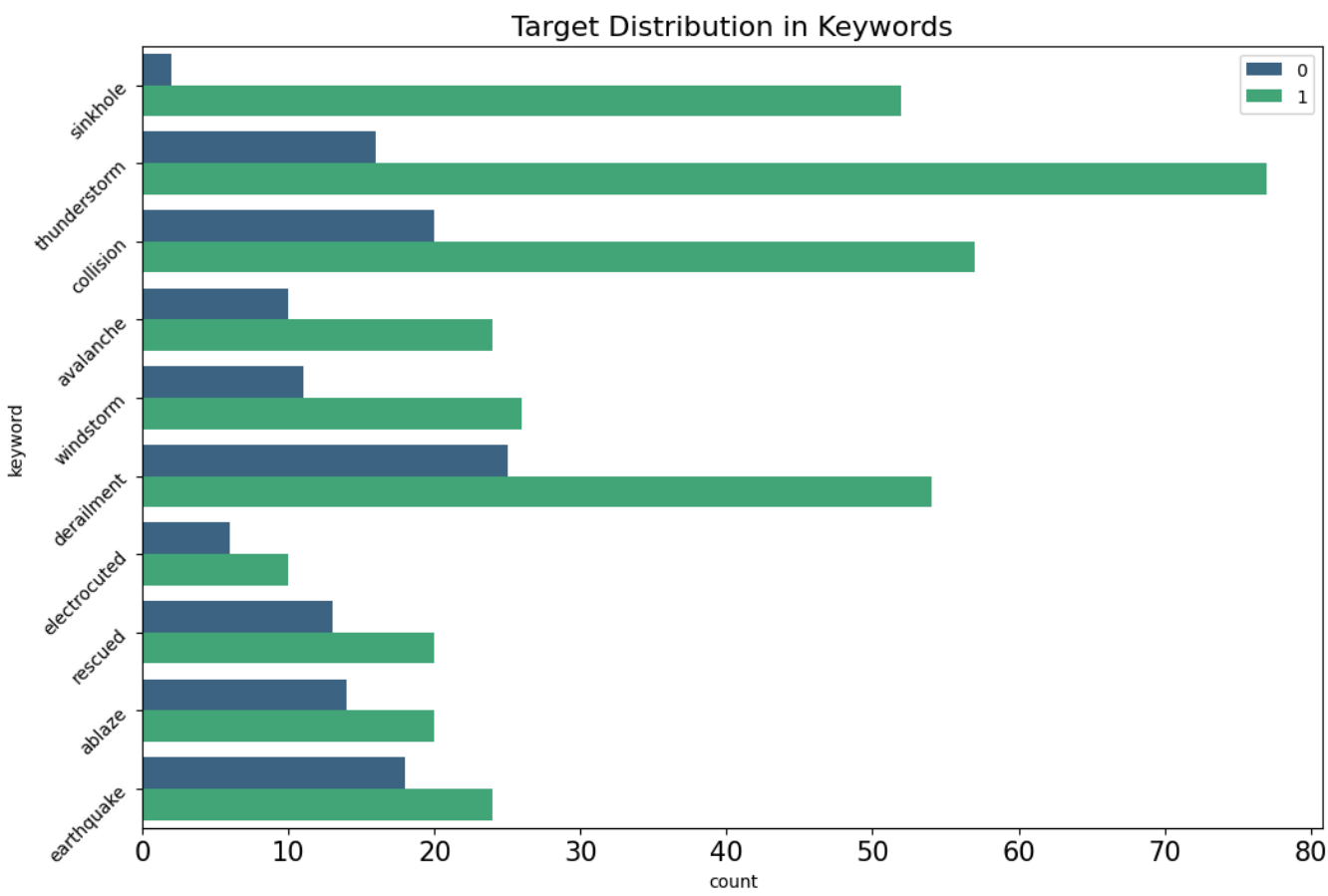



```
print(f'Number of unique values in keyword = {df_tweets["keyword"].nunique()} (Training) - {df_tweets["keyword"].nunique()} (Tweets)')
print(f'Number of unique values in location = {df_tweets["location"].nunique()} (Training) - {df_tweets["location"].nunique()} (Tweets)
```

```
Number of unique values in keyword = 219 (Training) - 219 (Tweets)
Number of unique values in location = 4504 (Training) - 4504 (Tweets)
```

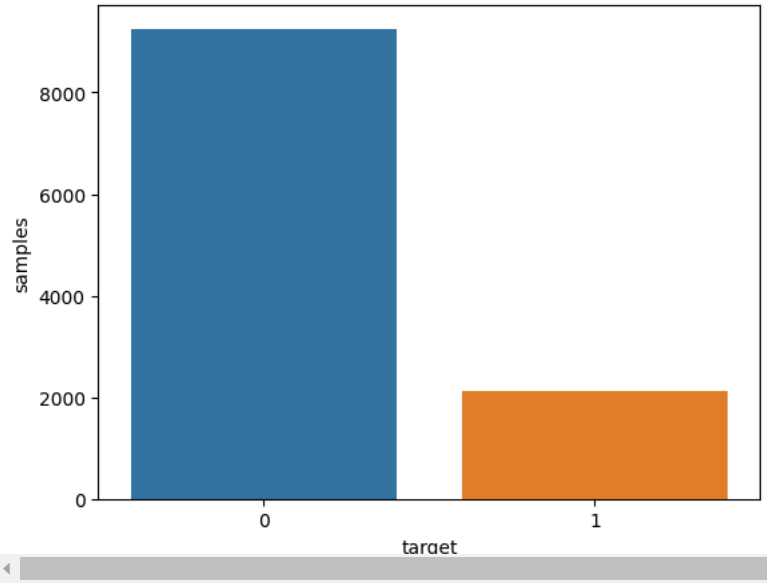
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the tweets dataset from the specified path
df_tweets = pd.read_csv('/kaggle/input/disaster-tweets/tweets.csv')

# Calculate the mean target value grouped by keyword
df_tweets['target_mean'] = df_tweets.groupby('keyword')['target'].transform('mean')

# Set the number of top keywords to display
top_n = 10  # You can adjust this number
top_keywords = df_tweets.groupby('keyword')['target'].mean().nlargest(top_n).index
filtered_df = df_tweets[df_tweets['keyword'].isin(top_keywords)]

# Set the figure size for the plot
fig = plt.figure(figsize=(12, 8), dpi=100)  # Increased size

# Create a count plot with sorted keywords based on target_mean
sns.countplot(y=filtered_df.sort_values(by='target_mean', ascending=False)['keyword'],
              hue=filtered_df.sort_values(by='target_mean', ascending=False)['target'],
              palette='viridis')  # Optional color palette
```

```
# Customize the plot
plt.tick_params(axis='x', labelsize=15)
plt.tick_params(axis='y', labelsize=12)
plt.legend(loc='upper right')
plt.title('Target Distribution in Keywords', fontsize=16)
plt.yticks(rotation=45, fontsize=10)  # Rotate y-axis labels for better readability

# Show the plot
plt.show()

# Drop the 'target_mean' column after plotting
df_tweets.drop(columns=['target_mean'], inplace=True)
```

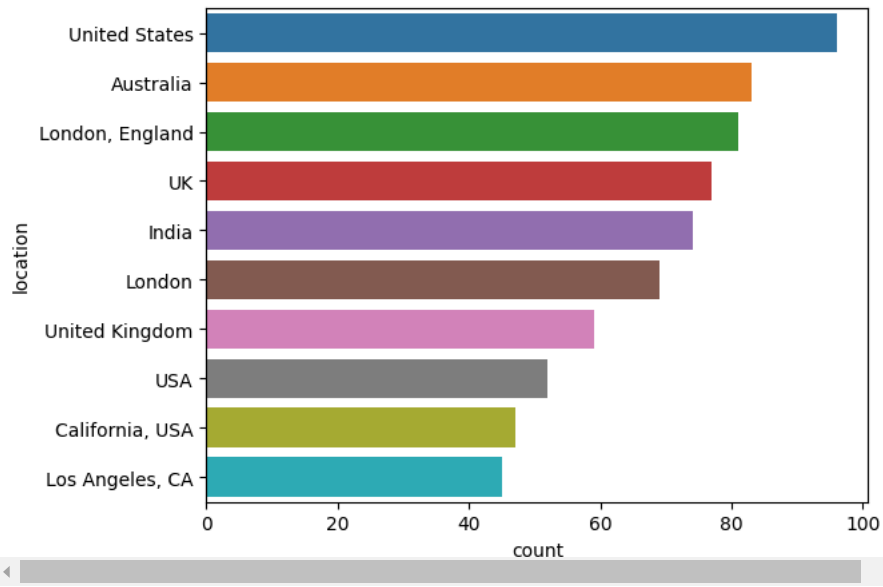



```
import seaborn as sns
import matplotlib.pyplot as plt

# Print target distribution
print('Target of 0 is {} % of total'.format(round(df_tweets['target'].value_counts()[0] / len(df_tweets['target']) * 100)))
print('Target of 1 is {} % of total'.format(round(df_tweets['target'].value_counts()[1] / len(df_tweets['target']) * 100)))

# Get the value counts of target
x = df_tweets.target.value_counts()

# Correct usage of sns.barplot with named arguments
sns.barplot(x=x.index, y=x.values)

# Set label for y-axis
plt.gca().set_ylabel('samples')

# Display the plot
plt.show()
```

```
Target of 0 is 81 % of total
Target of 1 is 19 % of total
```




```
sns.barplot(y=df_tweets['location'].value_counts()[:10].index,x=df_tweets['location'].value_counts()[:10],orient='h');
```




```
# Drop the column 'location' from the training dataset
df_train=df_tweets.drop(['location'],axis=1)
```

```
# A disaster tweet exmaple
df_train[df_tweets['target']==1]['text'][10:20]
```

```
16    No cows today but our local factory is sadly s...
19    French cameroun set houses ablaze in Ndu and r...
20    Cameroon's #BIR soldiers on the 05/01/2020 inv...
21    As fires ablaze throughout the land/as the pro...
24    Originally they were intended to be fired at b...
26    Another arson in Njikom,Boyo,NWR. The ambazomb...
27    Another public market in #Haiti mysteriously s...
30    Marivan, Kurdistan Province Monday, Jan 13th, ...
31    Marivan, Kurdistan Province Monday, Jan 13th, ...
32    How can you turn a blind eye to the icident of...
Name: text, dtype: object
```

```
#A non-disaster tweet example
df_tweets[df_train['target']==0]['text'][10:20]
```

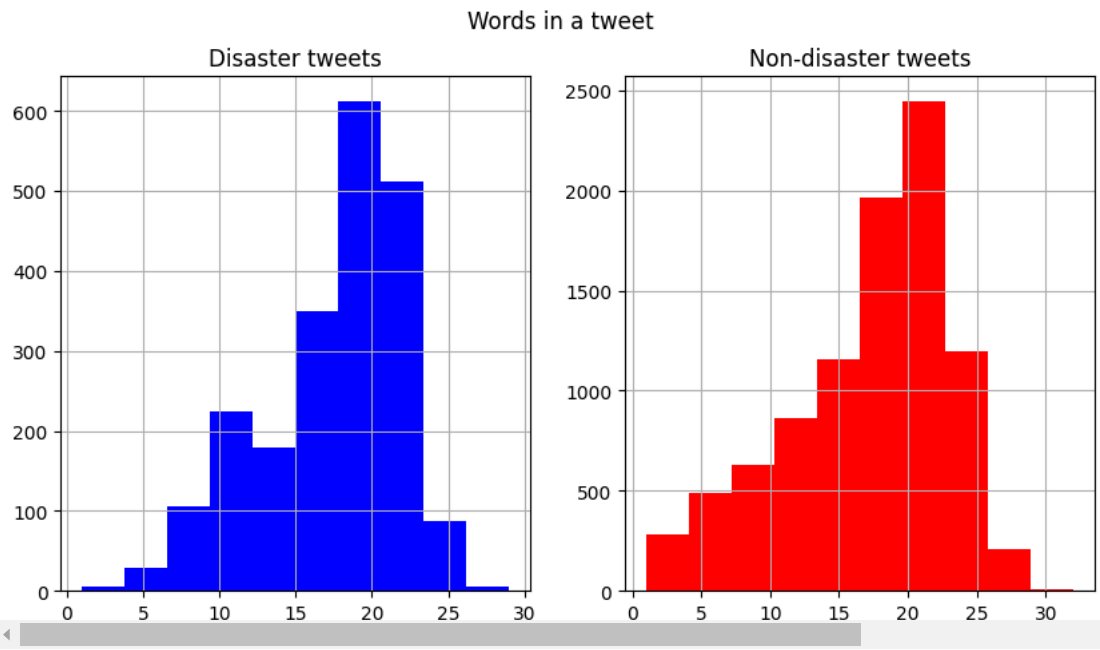
```
25    Warm greetings to all on the occasion of #Lohr...
28                                that is kind true sadly
29           I swear that jam will set the world ablaze
33    This love is so completely crazy. You've been ...
34    Terms in A Demon Burning Dark: The Ruined: Peo...
35    📷 Heartfelt appreciation to Prime Minister YAB...
37    ❤️❤️❤️ he gave us everything... He had a horri...
38    😁yeah! His new swag is on point 100%, since th...
39    This is cool and all these days I have been do...
```

```
41    my back and neck are still fucked up from the ...
Name: text, dtype: object
```

```
#Words count
df_tweets['words_count'] = df_tweets['text'].str.split().map(lambda x: len(x))
df_tweets.head()
```

	id	keyword	location	text	target	words_count
0	0	ablaze	NaN	Communal violence in Bhainsa, Telangana. "Ston...	1	19
1	1	ablaze	NaN	Telangana: Section 144 has been imposed in Bha...	1	23
2	2	ablaze	New York City	Arsonist sets cars ablaze at dealership https:...	1	7
3	3	ablaze	Morgantown, WV	Arsonist sets cars ablaze at dealership https:...	1	8
4	4	ablaze	NaN	"Lord Jesus, your love brings freedom and pard...	0	23

```
#Create visualization of the distribution of the word counts in comparision to target feature
import matplotlib.pyplot as plt
fig,(ax1,ax2)=plt.subplots(1,2,figsize=(10,5))
dis_tweet=df_tweets[df_tweets['target']==1]['words_count']
ax1.hist(dis_tweet,color='blue')
ax1.set_title('Disaster tweets')
ax1.grid()
nondis_tweet=df_tweets[df_tweets['target']==0]['words_count']
ax2.hist(nondis_tweet,color='red')
ax2.set_title('Non-disaster tweets')
ax2.grid()
fig.suptitle('Words in a tweet')
plt.show()
```



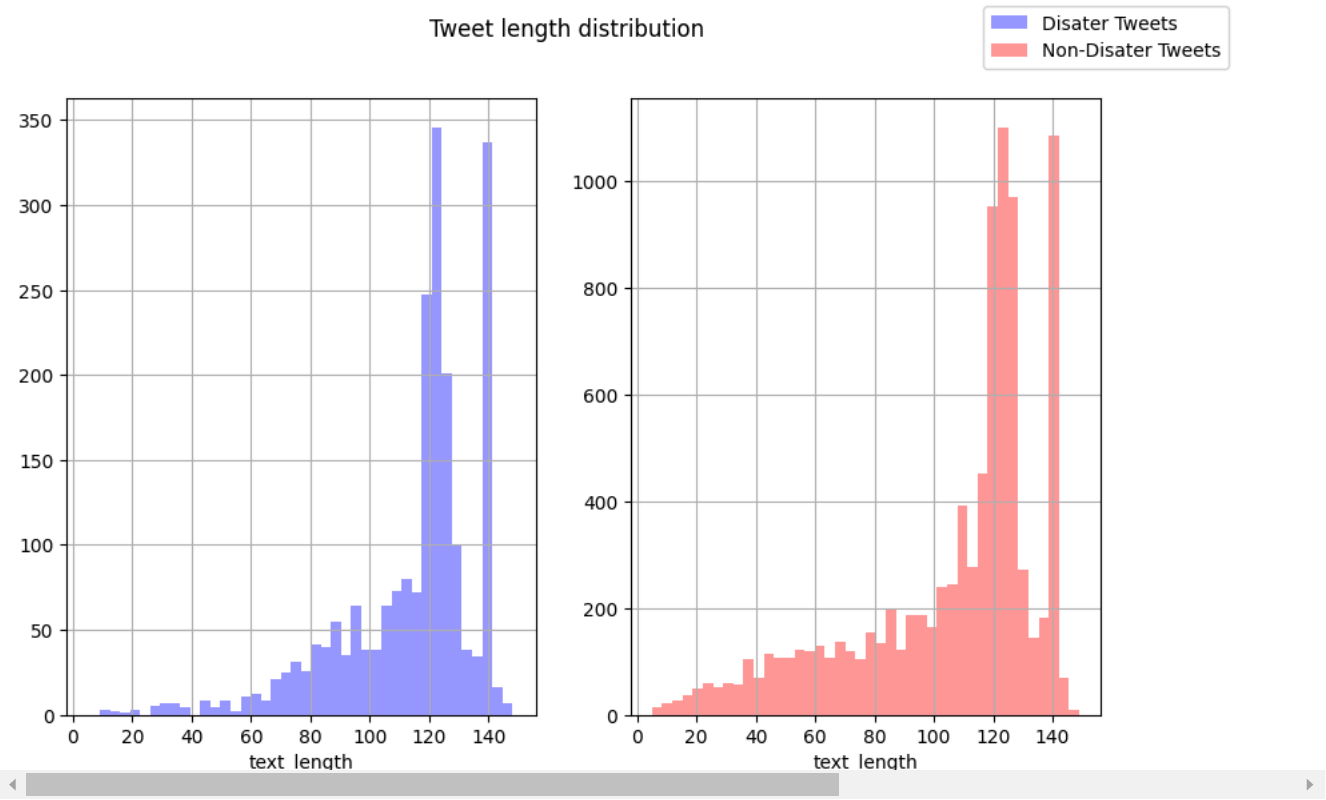

```
#Text Length
df_tweets['text_length'] = df_tweets['text'].apply(lambda x : len(x))
df_tweets.head()
```

	id	keyword	location	text	target	words_count	text_length
0	0	ablaze	NaN	Communal violence in Bhainsa, Telangana. "Ston...	1	19	125
1	1	ablaze	NaN	Telangana: Section 144 has been imposed in Bha...	1	23	131
2	2	ablaze	New York City	Arsonist sets cars ablaze at dealership https:...	1	7	63
3	3	ablaze	Morgantown, WV	Arsonist sets cars ablaze at dealership https:...	1	8	87
4	4	ablaze	NaN	"Lord Jesus, your love brings freedom and pard...	0	23	140

```
#Create visualization of the distribution of text length in comparision to target feature
f, (ax1, ax2) = plt.subplots(1, 2, sharex=True,figsize=(10,6))
sns.distplot(df_tweets[(df_tweets['target'] == 1)]['text_length'], ax=ax1, kde=False, color='blue',label='Disater Tweets')
sns.distplot(df_tweets[(df_tweets['target'] == 0)]['text_length'],ax=ax2, kde=False, color='red',label='Non-Disater Tweets');
f.suptitle('Tweet length distribution')
f.legend(loc='upper right')
ax1.grid()
```

```
ax2.grid()
plt.show()
```

```
/tmp/ipykernel_442/2003380046.py:3: UserWarning:


`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751


/tmp/ipykernel_442/2003380046.py:4: UserWarning:


`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751
```




```
import string
import pandas as pd
def remove_punctuation(text):
    no_punct=[words for words in text if words not in string.punctuation ]
    words_wo_punct=''.join(no_punct)
    return words_wo_punct

# Remove punctuation from both train and test dataset
df_tweets['text_wo_punct']=df_tweets['text'].apply(lambda x: remove_punctuation(x))
#df_tweets['text_wo_punct']=df_tweets['text'].apply(lambda x: remove_punctuation(x))

df_train.head()
```

	id	keyword	text	target
0	0	ablaze	Communal violence in Bhainsa, Telangana. "Ston...	1
1	1	ablaze	Telangana: Section 144 has been imposed in Bha...	1
2	2	ablaze	Arsonist sets cars ablaze at dealership https:...	1
3	3	ablaze	Arsonist sets cars ablaze at dealership https:...	1
4	4	ablaze	"Lord Jesus, your love brings freedom and pard...	0

```
import re
from nltk.stem import PorterStemmer, WordNetLemmatizer

wn = WordNetLemmatizer()
from nltk.tokenize import word_tokenize
def tokenize(text):
    split=re.split("\W+",text)
    return split
df_tweets['text_wo_punct_split']=df_tweets['text_wo_punct'].apply(lambda x: tokenize(x.lower()))
#df_tweets['text_wo_punct_split']=df_tweets['text_wo_punct'].apply(lambda x: tokenize(x.lower()))

df_train.head()
```

	id	keyword	text	target
0	0	ablaze	Communal violence in Bhainsa, Telangana. "Ston...	1
1	1	ablaze	Telangana: Section 144 has been imposed in Bha...	1
2	2	ablaze	Arsonist sets cars ablaze at dealership https:...	1
3	3	ablaze	Arsonist sets cars ablaze at dealership https:...	1
4	4	ablaze	"Lord Jesus, your love brings freedom and pard...	0

```
import nltk
from string import punctuation
import re
from nltk.stem import PorterStemmer, WordNetLemmatizer
wn = WordNetLemmatizer()
from nltk.tokenize import word_tokenize

stopword = nltk.corpus.stopwords.words('english')
print(stopword[:11])
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've"]
```

```
def remove_stopwords(text):
    text=[word for word in text if word not in stopword]
    return text

df_tweets['text_wo_punct_split_wo_stopwords']=df_tweets['text_wo_punct_split'].apply(lambda x: remove_stopwords(x))
#df_tweets['text_wo_punct_split_wo_stopwords']=df_tweets['text_wo_punct_split'].apply(lambda x: remove_stopwords(x))
df_tweets.head()
```

	id	keyword	location	text	target	words_count	text_length	text_wo_punct	text_wo_punct_split	text_wo_punct_split_wo_
0	0	ablaze	NaN	Communal violence in Bhainsa, Telangana. "Ston...	1	19	125	Communal violence in Bhainsa Telangana Stones ...	[communal, violence, in, bhainsa, telangana, s...	[communal, violenc telanga
1	1	ablaze	NaN	Telangana: Section 144 has been imposed in Bha...	1	23	131	Telangana Section 144 has been imposed in Bhai...	[telangana, section, 144, has, been, imposed, ...	[telangana, section, 14 b

```
ps = PorterStemmer()

from catboost import CatBoostClassifier
print(ps.stem('believe'))
print(ps.stem('believing'))
print(ps.stem('believed'))
print(ps.stem('believes'))
```

```
believ
believ
believ
believ
```

```
!pip install nltk
```

```
Requirement already satisfied: nltk in /opt/conda/lib/python3.10/site-packages (3.2.4)
Requirement already satisfied: six in /opt/conda/lib/python3.10/site-packages (from nltk) (1.16.0)
```

```
import pandas as pd

tweets = pd.read_csv('/kaggle/input/disaster-tweets/tweets.csv')

# Drop unnecessary columns
df = tweets.drop(['id', 'keyword', 'location'], axis=1)

# Display the first 5 rows
df.head()
```

	text	target
0	Communal violence in Bhainsa, Telangana. "Ston...	1
1	Telangana: Section 144 has been imposed in Bha...	1
2	Arsonist sets cars ablaze at dealership https:...	1
3	Arsonist sets cars ablaze at dealership https:...	1
4	"Lord Jesus, your love brings freedom and pard...	0

```
df['target'].value_counts()
```

```
target
0    9256
1    2114
Name: count, dtype: int64
```

```
#Balancing the Dataset
df_0_class = df[df['target']==0]
df_1_class = df[df['target']==1]
df_0_class_undersampled = df_0_class.sample(df_1_class.shape[0])
df = pd.concat([df_0_class_undersampled, df_1_class], axis=0)
df['target'].value_counts()
```

```
target
0    2114
1    2114
Name: count, dtype: int64
```

```
from sklearn.model_selection import train_test_split

# Splitting the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(
    df['text'],
    df['target'],
    stratify=df['target'],
    test_size=0.2,  # 20% data for testing
    random_state=42  # Ensures reproducibility
)

# Verifying the split
print(f"Training set size: {len(X_train)}")
print(f"Test set size: {len(X_test)}")
```

```
Training set size: 3382
Test set size: 846
```

```
import string

# Defining the function to remove punctuation
def remove_punctuation(text):
    punctuationfree = "".join([i for i in text if i not in string.punctuation])
    return punctuationfree

# Applying the function to the text column and storing the cleaned text
tweets['clean_msg'] = tweets['text'].apply(lambda x: remove_punctuation(x))
```

```
# Display the DataFrame
tweets.head()
```

	id	keyword	location	text	target	clean_msg
0	0	ablaze	NaN	Communal violence in Bhainsa, Telangana. "Ston...	1	Communal violence in Bhainsa Telangana Stones ...
1	1	ablaze	NaN	Telangana: Section 144 has been imposed in Bha...	1	Telangana Section 144 has been imposed in Bhai...
2	2	ablaze	New York City	Arsonist sets cars ablaze at dealership https:...	1	Arsonist sets cars ablaze at dealership httpst...
3	3	ablaze	Morgantown, WV	Arsonist sets cars ablaze at dealership https:...	1	Arsonist sets cars ablaze at dealership httpst...
4	4	ablaze	NaN	"Lord Jesus, your love brings freedom and pard...	0	Lord Jesus your love brings freedom and pardon...

```
#Lowering the Text
tweets['text']= tweets['clean_msg'].apply(lambda x: x.lower())
tweets.head()
```

	id	keyword	location	text	target	clean_msg
0	0	ablaze	NaN	communal violence in bhainsa telangana stones ...	1	Communal violence in Bhainsa Telangana Stones ...
1	1	ablaze	NaN	telangana section 144 has been imposed in bhai...	1	Telangana Section 144 has been imposed in Bhai...
2	2	ablaze	New York City	arsonist sets cars ablaze at dealership httpst...	1	Arsonist sets cars ablaze at dealership httpst...
3	3	ablaze	Morgantown, WV	arsonist sets cars ablaze at dealership httpst...	1	Arsonist sets cars ablaze at dealership httpst...
4	4	ablaze	NaN	lord jesus your love brings freedom and pardon...	0	Lord Jesus your love brings freedom and pardon...

```
#Tokenization
import re
def tokenization(text):
    tokens = re.split('W+',text)
    return tokens
#applying function to the column
tweets['msg_tokenied']= tweets['text'].apply(lambda x: tokenization(x))
tweets.head()
```

	id	keyword	location	text	target	clean_msg	msg_tokenied
0	0	ablaze	NaN	communal violence in bhainsa telangana stones ...	1	Communal violence in Bhainsa Telangana Stones ...	[communal violence in bhainsa telangana stones...
1	1	ablaze	NaN	telangana section 144 has been imposed in bhai...	1	Telangana Section 144 has been imposed in Bhai...	[telangana section 144 has been imposed in bha...
2	2	ablaze	New York City	arsonist sets cars ablaze at dealership httpst...	1	Arsonist sets cars ablaze at dealership httpst...	[arsonist sets cars ablaze at dealership https...

```
#Stop Word Removal
import nltk
#Stop words present in the library
stopwords = nltk.corpus.stopwords.words('english')
stopwords[0:10]
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're"]
#defining the function to remove stopwords from tokenized text
def remove_stopwords(text):
    output= [i for i in text if i not in stopwords]
    return output
#applying the function
tweets['no_stopwords']= tweets['msg_tokenied'].apply(lambda x:remove_stopwords(x))
tweets.head()
```

	id	keyword	location	text	target	clean_msg	msg_tokenied	no_stopwords
0	0	ablaze	NaN	communal violence in bhainsa telangana stones ...	1	Communal violence in Bhainsa Telangana Stones ...	[communal violence in bhainsa telangana stones...	[communal violence in bhainsa telangana stones...
1	1	ablaze	NaN	telangana section 144 has been imposed in bhai...	1	Telangana Section 144 has been imposed in Bhai...	[telangana section 144 has been imposed in bha...	[telangana section 144 has been imposed in bha...
2	2	ablaze	New York City	arsonist sets cars ablaze at dealership httpst...	1	Arsonist sets cars ablaze at dealership httpst...	[arsonist sets cars ablaze at dealership https...	[arsonist sets cars ablaze at dealership https...
3	3	ablaze	Morgantown, WV	arsonist sets cars ablaze at dealership httpst...	1	Arsonist sets cars ablaze at dealership httpst...	[arsonist sets cars ablaze at dealership https...	[arsonist sets cars ablaze at dealership https...

```
import pandas as pd
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize
import nltk

# Download the NLTK tokenizer models (if not already downloaded)
nltk.download('punkt')

# Load the dataset from the provided path
tweets = pd.read_csv('/kaggle/input/disaster-tweets/tweets.csv')

# Initialize the Porter Stemmer
stemmer = PorterStemmer()

# Defining the function for stemming
def stemmer_func(text):
    tokens = word_tokenize(text)  # Tokenize the input text
    stemmed_text = [stemmer.stem(token) for token in tokens]  # Stem each token
    return ' '.join(stemmed_text)  # Join tokens back into a sentence

# Applying the stemmer function on the 'text' column
tweets['msg_stemmed'] = tweets['text'].apply(lambda x: stemmer_func(str(x)))

# Display the first few rows
tweets.head()
```

```
[nltk_data] Downloading package punkt to /usr/share/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

	id	keyword	location	text	target	msg_stemmed
0	0	ablaze	NaN	Communal violence in Bhainsa, Telangana. "Ston...	1	commun violenc in bhainsa , telangana . `` sto...
1	1	ablaze	NaN	Telangana: Section 144 has been imposed in Bha...	1	telangana : section 144 ha been impos in bhain...
2	2	ablaze	New York City	Arsonist sets cars ablaze at dealership https:...	1	arsonist set car ablaz at dealership http : //...
3	3	ablaze	Morgantown, WV	Arsonist sets cars ablaze at dealership https:...	1	arsonist set car ablaz at dealership http : //...
4	4	ablaze	NaN	"Lord Jesus, your love brings freedom and pard...	0	`` lord iesu , your love bring freedom and par...

```
import spacy
import pandas as pd

# Load the spaCy English language model
nlp = spacy.load('en_core_web_sm')

# Load the dataset from the provided path
tweets = pd.read_csv('/kaggle/input/disaster-tweets/tweets.csv')

# Defining the function for lemmatization
def lemmatizer(text):
    doc = nlp(text)
    lemm_text = [token.lemma_ for token in doc]
    return ' '.join(lemm_text)  # Join tokens back into a sentence

# Applying the lemmatizer function on the 'text' column
tweets['msg_lemmatized'] = tweets['text'].apply(lambda x: lemmatizer(str(x)))

# Display the first few rows
tweets.head()
```

	id	keyword	location	text	target	msg_lemmatized
0	0	ablaze	NaN	Communal violence in Bhainsa, Telangana. "Ston...	1	communal violence in Bhainsa , Telangana . " s...
1	1	ablaze	NaN	Telangana: Section 144 has been imposed in Bha...	1	Telangana : section 144 have be impose in Bhai...
2	2	ablaze	New York City	Arsonist sets cars ablaze at dealership https:...	1	arsonist set car ablaze at dealership https://...
3	3	ablaze	Morgantown, WV	Arsonist sets cars ablaze at dealership https:...	1	arsonist set car ablaze at dealership https://...
4	4	ablaze	NaN	"Lord Jesus, your love brings freedom and pard...	0	" Lord Jesus , your love bring freedom and par...

```
!jupyter nbconvert --to html /content/KNN.ipynb
```

```
[NbConvertApp] WARNING | pattern '/content/KNN.ipynb' matched no files
This application is used to convert notebook files (*.ipynb)
        to various other formats.

        WARNING: THE COMMANDLINE INTERFACE MAY CHANGE IN FUTURE RELEASES.

Options
```

```
=======
The options below are convenience aliases to configurable class-options,
as listed in the "Equivalent to" description-line of the aliases.
To see all configurable class-options for some <cmd>, use:
    <cmd> --help-all

--debug
    set log level to logging.DEBUG (maximize logging output)
    Equivalent to: [--Application.log_level=10]
--show-config
    Show the application's configuration (human-readable format)
    Equivalent to: [--Application.show_config=True]
--show-config-json
```