**Data Set:**

**Large Language Models: the tweets** <https://www.kaggle.com/datasets/konradb/chatgpt-the-tweets/versions/169?resource=download>

I have used the dataset from Kaggle, The dataset is about the Large Language Model: Tweets. Published on December 20, 2022.

This data set has a collection of tweets with the hashtag #chatgpt : discussions about the chatgpt language model, sharing experiences with using chatgpt, or asking for help with chatgpt-related issues. The tweets could also include links to articles or websites related to chatgpt, as well as images, videos, or other media. Overall, a collection of tweets with the hashtag #chatgpt would provide a glimpse into the online conversation surrounding chatgpt."

**Data cleaning:**

**I have used python to clean the data using jupyter note book.**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

import pandas as pd

import re

# Define column names and their corresponding data types

columns = {

"user\_name": str,

"text": str,

"user\_location": str,

"user\_description": str,

"user\_created": str,

"user\_followers": str,

"user\_friends": str,

"user\_favourites": str,

"user\_verified": str,

"date": str,

"hashtags": str,

"source": str

}

# Load the dataset with specified column names and data types

file\_path = '/Users/saipreethamvudutha/Downloads/tweets.csv'

df = pd.read\_csv(file\_path, names=columns.keys(), dtype=columns, skiprows=1)

# Display the first few rows to understand the structure of the data

print(df.head())

# Handle missing values

df.dropna(inplace=True)

# Remove duplicates

df.drop\_duplicates(inplace=True)

# Define a function to remove Chinese characters from a string

def remove\_chinese(text):

return re.sub(r'[\u4E00-\u9FFF]+', '', str(text))

# Define a function to remove URLs from a string

def remove\_urls(text):

return re.sub(r'http\S+', '', str(text))

# Define a function to eliminate '////' from user\_name

def remove\_slashes(text):

return text.replace('////', '')

# Columns where Chinese characters might be present

columns\_to\_clean = ["user\_name", "text", "user\_location", "user\_description", "hashtags", "source"]

# Clean Chinese characters and URLs from specific columns

for col in columns\_to\_clean:

df[col] = df[col].apply(remove\_chinese)

df[col] = df[col].apply(remove\_urls)

# Remove '////' from user\_name column

df['user\_name'] = df['user\_name'].apply(remove\_slashes)

# Save the cleaned data to a new CSV file without index

cleaned\_file\_path = '/Users/saipreethamvudutha/Downloads/cleaned\_tweets.csv'

df.to\_csv(cleaned\_file\_path, index=False)

**OUTPUT:**

username \

0 Walee MENA

1 Dataiku

3 Lithium Systems

4 Paramendra Kumar Bhagat

text \

0 #OpenAI has revealed its plan to launch #ChatG...

1 What are #LargeLanguageModels, how are they de...

2 apodecisionacious\natappear\nhe \ #Cha...

3 Business owner? Stay informed with the latest ...

4 ChatGPT: Motorbike For The Mind (13) #ChatGPT ...

user\_location \

0 UAE

1 New York, NY

2 Fayetteville

3 Based in Central Scotland

4 NY

user\_description \

0 OFFICIALLY IN MENA! We are the region's larges...

1 Dataiku is the only AI platform that connects ...

2 TG：<https://t.co/C2kHIu7BKg> 官网：<https://t.co/56a>...

...

1 ['LargeLanguageModels'] HubSpot

2 ['黑客', '合约', 'ChatGPT'] Twitter Web App

3 NaN Hootsuite Inc.

4 ['ChatGPT', 'GPT4', 'AI', 'ArtificialIntellige... Hootsuite Inc.

*Output is truncated. View as a* [*scrollable element*](command:cellOutput.enableScrolling?598a371d-54e0-44f7-a2d4-501e260b38cf) *or open in a* [*text editor*](command:workbench.action.openLargeOutput?598a371d-54e0-44f7-a2d4-501e260b38cf)*. Adjust cell output* [*settings*](command:workbench.action.openSettings?%5B%22%40tag%3AnotebookOutputLayout%22%5D)*...*

**POWER BI DASH BOARDS:**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Key Findings:**

**Hong Kong has the most numbers of verified twitter (now ‘x’) users.**

**Times Now is the most followed twitter handle.**