Analytic Avengers

Limpieza de datos

- Traducir a inglés para que sea replicable
- Quitar caracteres especiales
- Quitar números
- Quitar NAs

In [11]:
```
pip install emoji
```

```
Collecting emoji
  Downloading emoji-2.11.1-py2.py3-none-any.whl (433 kB)
     ------------------------------------- 433.8/433.8 kB 3.4 MB/s eta 0:00:00
Installing collected packages: emoji
Successfully installed emoji-2.11.1
Note: you may need to restart the kernel to use updated packages.
```

In [18]:
```
pip install emoji==1.7
```

```
Collecting emoji==1.7
  Downloading emoji-1.7.0.tar.gz (175 kB)
     ------------------------------------- 175.4/175.4 kB 2.1 MB/s eta 0:00:00
  Preparing metadata (setup.py): started
  Preparing metadata (setup.py): finished with status 'done'
Building wheels for collected packages: emoji
  Building wheel for emoji (setup.py): started
  Building wheel for emoji (setup.py): finished with status 'done'
  Created wheel for emoji: filename=emoji-1.7.0-py3-none-any.whl size=171032 sha25
6=7398eb3a2276cb28a2f60365ad1a015c34c0f5b4a9f8d14e31346b0a775bc94f
  Stored in directory: c:\users\crist\appdata\local\pip\cache\wheels\37\b1\70\d87e
2dddea71a019314970e3ea065b63e27b9be29e4a579b13
Successfully built emoji
Installing collected packages: emoji
  Attempting uninstall: emoji
    Found existing installation: emoji 2.11.1
    Uninstalling emoji-2.11.1:
      Successfully uninstalled emoji-2.11.1
Successfully installed emoji-1.7.0
Note: you may need to restart the kernel to use updated packages.
```

In [20]:
```
pip install deep_translator
```

```
Collecting deep_translator
  Downloading deep_translator-1.11.4-py3-none-any.whl (42 kB)
     ---------------------------------------- 42.3/42.3 kB 1.0 MB/s eta 0:00:00
Requirement already satisfied: beautifulsoup4<5.0.0,>=4.9.1 in c:\users\crist\mini
conda3\lib\site-packages (from deep_translator) (4.11.1)
Requirement already satisfied: requests<3.0.0,>=2.23.0 in c:\users\crist\miniconda
3\lib\site-packages (from deep_translator) (2.28.1)
Requirement already satisfied: soupsieve>1.2 in c:\users\crist\miniconda3\lib\site
-packages (from beautifulsoup4<5.0.0,>=4.9.1->deep_translator) (2.3.2.post1)
Requirement already satisfied: idna<4,>=2.5 in c:\users\crist\miniconda3\lib\site-
packages (from requests<3.0.0,>=2.23.0->deep_translator) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\crist\miniconda3\lib
\site-packages (from requests<3.0.0,>=2.23.0->deep_translator) (2022.12.7)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\crist\miniconda3
\lib\site-packages (from requests<3.0.0,>=2.23.0->deep_translator) (1.26.14)
Requirement already satisfied: charset-normalizer<3,>=2 in c:\users\crist\minicond
a3\lib\site-packages (from requests<3.0.0,>=2.23.0->deep_translator) (2.0.4)
Installing collected packages: deep_translator
Successfully installed deep_translator-1.11.4
Note: you may need to restart the kernel to use updated packages.
```

In [17]:
```python
# Importar librerías

import pandas as pd
import numpy as np
from scipy import stats
from sklearn.feature_extraction.text import CountVectorizer
import matplotlib.pyplot as plt
```

Cargar datos

In [18]:
```python
df = pd.read_csv("df.csv")
df.head(3)
```

Out[18]:

	date	time	tweet
0	2023-01-01	NaN	Resuelto, muchísimas gracias , excelente servi...
1	2023-01-02	NaN	Muchas gracias, espero su dm
2	2023-01-02	NaN	Muchas gracias!

In [19]:
```python
import emoji

emoji_text = []

# Iterar sobre los emojis en la columna 'tweet' y transformarlos
for emojis in df['tweet']:
    emoji_text.append(emoji.demojize(emojis, delimiters=("","")))

# colmna emoji
df = df.assign(Emoji_Text=emoji_text)
print(df)
```

```
           date      time                                              tweet  \
0    2023-01-01       NaN  Resuelto, muchísimas gracias , excelente servi...
1    2023-01-02       NaN                      Muchas gracias, espero su dm
2    2023-01-02       NaN                                   Muchas gracias!
3    2023-01-02       NaN  Algo similar me paso. Quería renovar mi token ...
4    2023-01-02       NaN  Yeeeei! a través de mi cuenta en acabo de cont...
..          ...       ...                                               ...
807  2024-04-21  15:21:43                  Ahora confirmo, gracias estimado
808  2024-04-23  22:52:05                     😔 de acuerdo!! Muchas gracias.
809  2024-04-23  21:30:12  Tío cuando me van a graduar, tengo la TDC gara...
810  2024-04-25  00:51:24  una duda, con quién tengo que ver si tengo pro...
811  2024-04-26  13:30:56  Wey, lo mejor de es que puedo pagar con Apple ...

                                          Emoji_Text
0    Resuelto, muchísimas gracias , excelente servi...
1                        Muchas gracias, espero su dm
2                                     Muchas gracias!
3    Algo similar me paso. Quería renovar mi token ...
4    Yeeeei! a través de mi cuenta en acabo de cont...
..                                               ...
807                  Ahora confirmo, gracias estimado
808          pensive_face de acuerdo!! Muchas gracias.
809  Tío cuando me van a graduar, tengo la TDC gara...
810  una duda, con quién tengo que ver si tengo pro...
811  Wey, lo mejor de es que puedo pagar con Apple ...

[812 rows x 4 columns]
```

Convertir columna 'date' a formato fecha y quitar guión bajo de palabras unidas

In [20]:
```
# # df columnas con solo valores NA
# df = df.dropna(axis=1, how='all')

# Convertir la columna "date" a formato de fecha
df['date'] = pd.to_datetime(df['date'])

# Reemplazar guión bajo con espacio en blanco en la columna 'Emoji_Text'
df['Emoji_Text'] = df['Emoji_Text'].str.replace('_', ' ')
```

In [21]:
```
df = df[["date","time","Emoji_Text"]]
df.head()
```

Out[21]:

	date	time	Emoji_Text
0	2023-01-01	NaN	Resuelto, muchísimas gracias , excelente servi...
1	2023-01-02	NaN	Muchas gracias, espero su dm
2	2023-01-02	NaN	Muchas gracias!
3	2023-01-02	NaN	Algo similar me paso. Quería renovar mi token ...
4	2023-01-02	NaN	Yeeeei! a través de mi cuenta en acabo de cont...

Traducir a inglés la base de datos

In [24]:
```
# # Ejemplo de traducción usando esta librería

# from deep_translator import GoogleTranslator

# text = 'Hola mundo'
```

```
# translated = GoogleTranslator(source='spanish', target='english').translate(text)

# print(f'Texto original: {text}')
# print(f'Texto traducido: {translated}')
```

```
Texto original: Hola mundo
Texto traducido: Hello World
```

In [ ]:
```
from deep_translator import GoogleTranslator

# Función para traducir texto utilizando Google Translator
def translate_text(text):
    translated_text = GoogleTranslator(source='auto', target='english').translate(t
    return translated_text

# Aplicar la función de traducción a la columna "tweet" y asignar los resultados a
df['translated_tweet'] = df['Emoji_Text'].apply(translate_text)
```

Pruebas con la columna 808 que tiene emoji para revisar que la traducción y emoji-palabra funcione

In [58]:
```
# Acceder al valor de la columna 'Emoji_Text' en el registro número 808
emoji_text_808 = df.loc[808, 'translated_tweet']  # Recordando que los índices en P

print("Valor en la fila 808 de la columna 'Emoji_Text':", emoji_text_808)
```

```
Valor en la fila 808 de la columna 'Emoji_Text': pensive face agree!! thank you so
much.
```

In [59]:
```
df.head()
```

Out[59]:

	date	time	translated_tweet
0	2023-01-01	NaN	solved, thank you very much, excellent service...
1	2023-01-02	NaN	thank you very much, i await your dm
2	2023-01-02	NaN	thank you so much!
3	2023-01-02	NaN	something similar happened to me. i wanted to ...
4	2023-01-02	NaN	yeeeei! through my account i just signed up fo...

In [27]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 812 entries, 0 to 811
Data columns (total 4 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   date              812 non-null    datetime64[ns]
 1   time              424 non-null    object
 2   tweet             812 non-null    object
 3   translated_tweet  812 non-null    object
dtypes: datetime64[ns](1), object(3)
memory usage: 25.5+ KB
```

Crear dataframe solo con columnas de interés

In [60]:
```
df1 = df[["date","time","translated_tweet"]]
df.columns
```

Out[60]:
```
Index(['date', 'time', 'translated_tweet'], dtype='object')
```

In [62]:
```
# Convertir el texto en la columna "tweet" a minúsculas
df1['translated_tweet'] = df1['translated_tweet'].str.lower()
```

In [63]:
```
df_traducido = pd.read_csv("df_traducido.csv")
df.head()
```

Out[63]:

	date	time	translated_tweet
0	2023-01-01	NaN	solved, thank you very much, excellent service...
1	2023-01-02	NaN	thank you very much, i await your dm
2	2023-01-02	NaN	thank you so much!
3	2023-01-02	NaN	something similar happened to me. i wanted to ...
4	2023-01-02	NaN	yeeeei! through my account i just signed up fo...

Eliminar caracteres especiales

In [65]:
```
import re

# Función para eliminar caracteres especiales
def remove_special_characters(text):
    # Utilizamos una expresión regular para eliminar caracteres especiales y dejar
    text = re.sub(r'[^a-zA-Z0-9\s]', '', text)
    return text

# Aplicar la función de eliminación de caracteres especiales a la columna "translat
df1['translated_tweet'] = df1['translated_tweet'].apply(remove_special_characters)

df1.head()
```

Out[65]:

	date	time	translated_tweet
0	2023-01-01	NaN	solved thank you very much excellent service a...
1	2023-01-02	NaN	thank you very much i await your dm
2	2023-01-02	NaN	thank you so much
3	2023-01-02	NaN	something similar happened to me i wanted to r...
4	2023-01-02	NaN	yeeeei through my account i just signed up for...

Guardar dataframe final

In [66]:
```
# Guardar el DataFrame en un archivo CSV
df.to_csv("df_traducido_emojis.csv", index=False)
```