Tecnológico de Monterrey

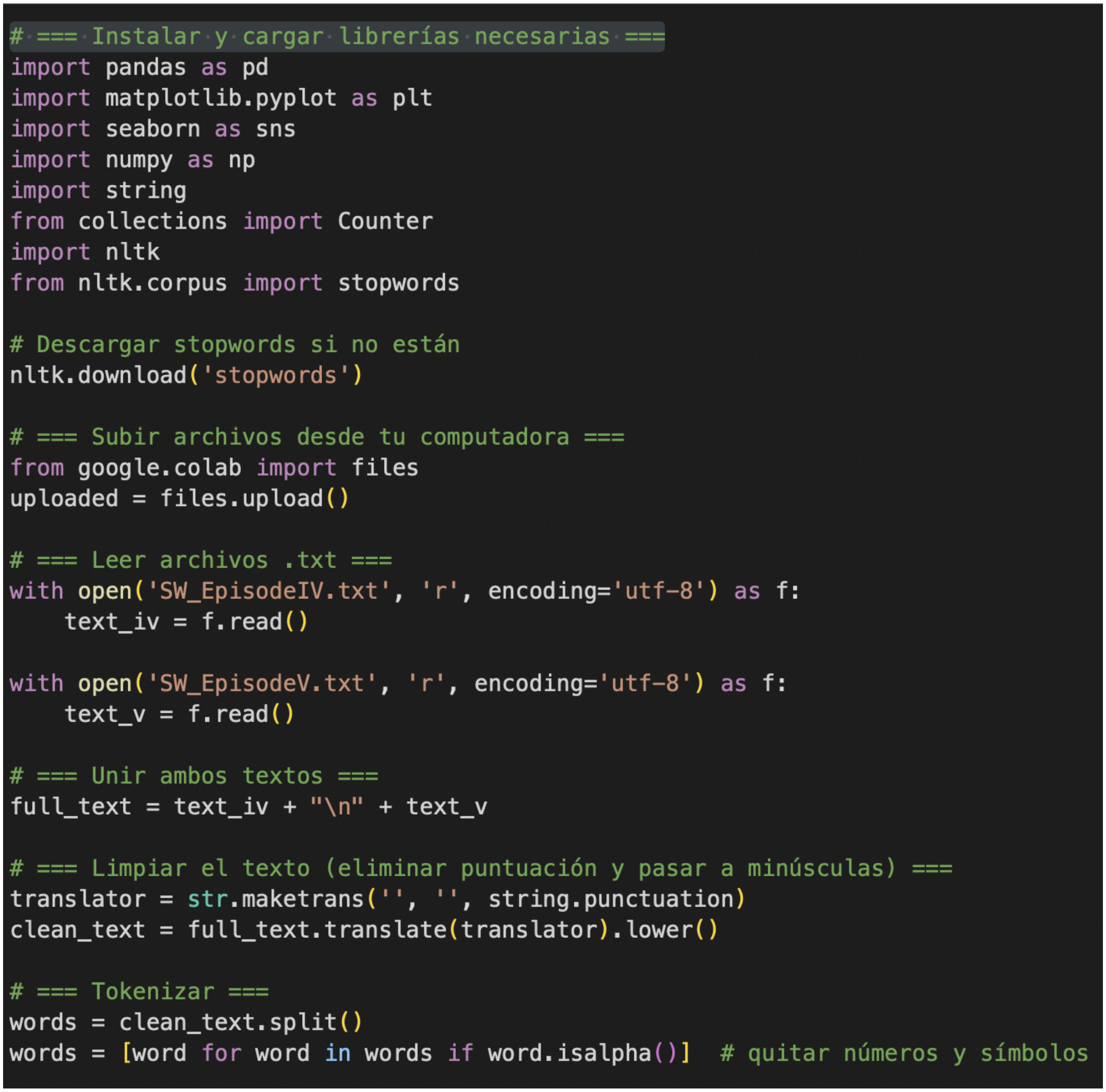
Activity 2: Descriptive statistics

Alejandro Salazar Loza A01665123

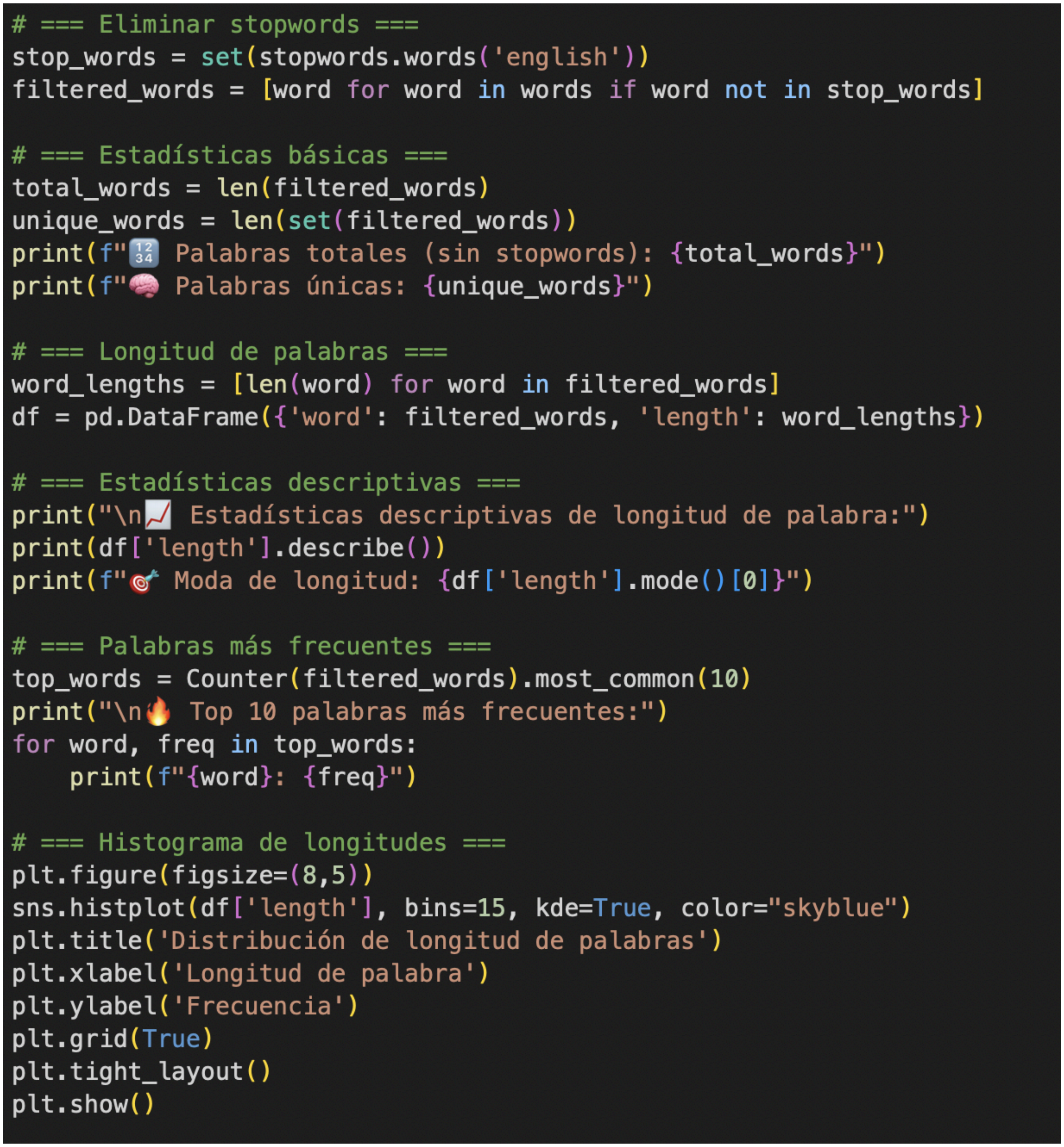
6 de mayo del 2025

CS tools

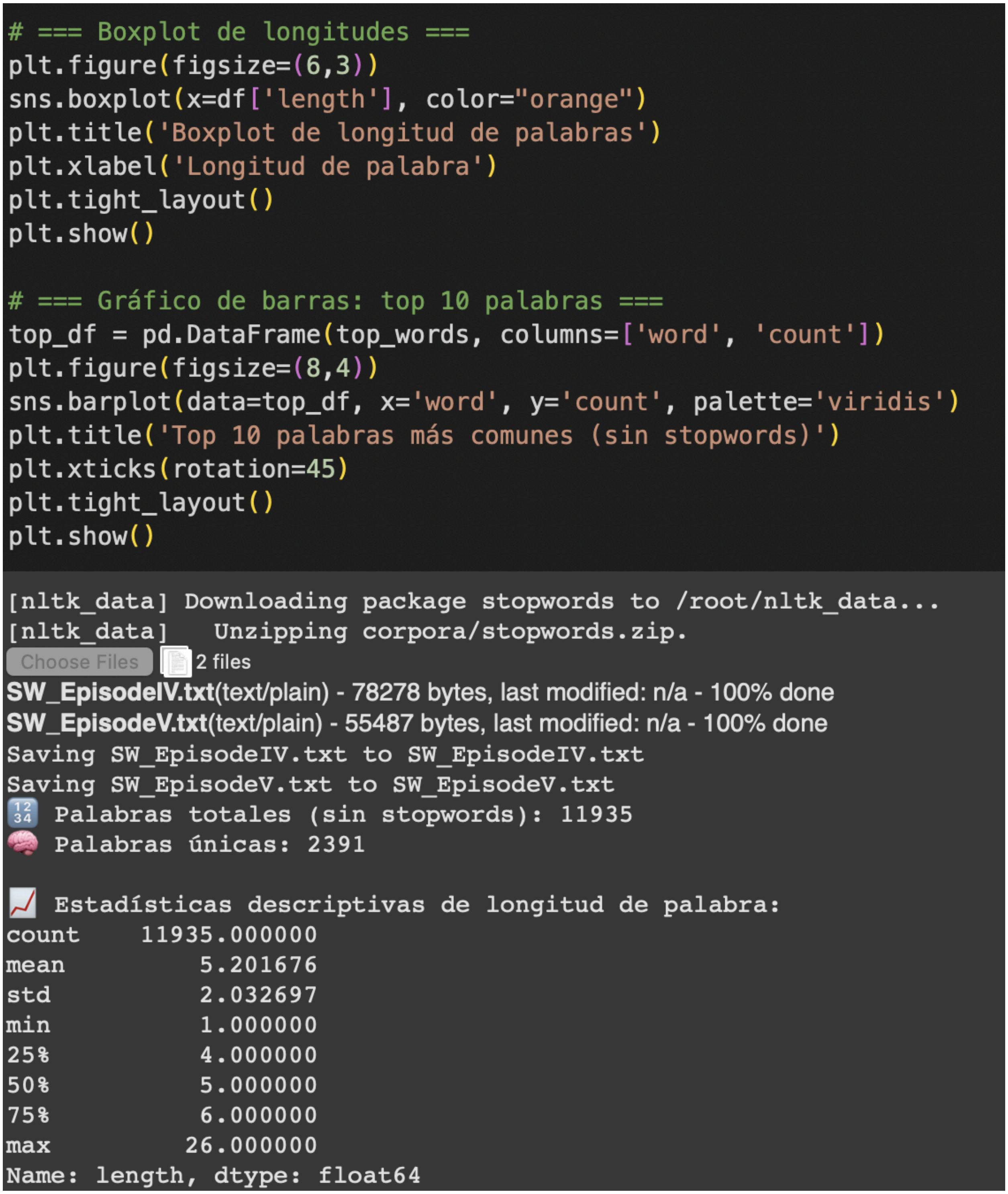
Profesor Sergio Ruiz Loza



```python
# === Instalar y cargar librerías necesarias ===
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import string
from collections import Counter
import nltk
from nltk.corpus import stopwords

# Descargar stopwords si no están
nltk.download('stopwords')

# === Subir archivos desde tu computadora ===
from google.colab import files
uploaded = files.upload()

# === Leer archivos .txt ===
with open('SW_EpisodeIV.txt', 'r', encoding='utf-8') as f:
    text_iv = f.read()

with open('SW_EpisodeV.txt', 'r', encoding='utf-8') as f:
    text_v = f.read()

# === Unir ambos textos ===
full_text = text_iv + "\n" + text_v

# === Limpiar el texto (eliminar puntuación y pasar a minúsculas) ===
translator = str.maketrans('', '', string.punctuation)
clean_text = full_text.translate(translator).lower()

# === Tokenizar ===
words = clean_text.split()
words = [word for word in words if word.isalpha()]  # quitar números y símbolos
```

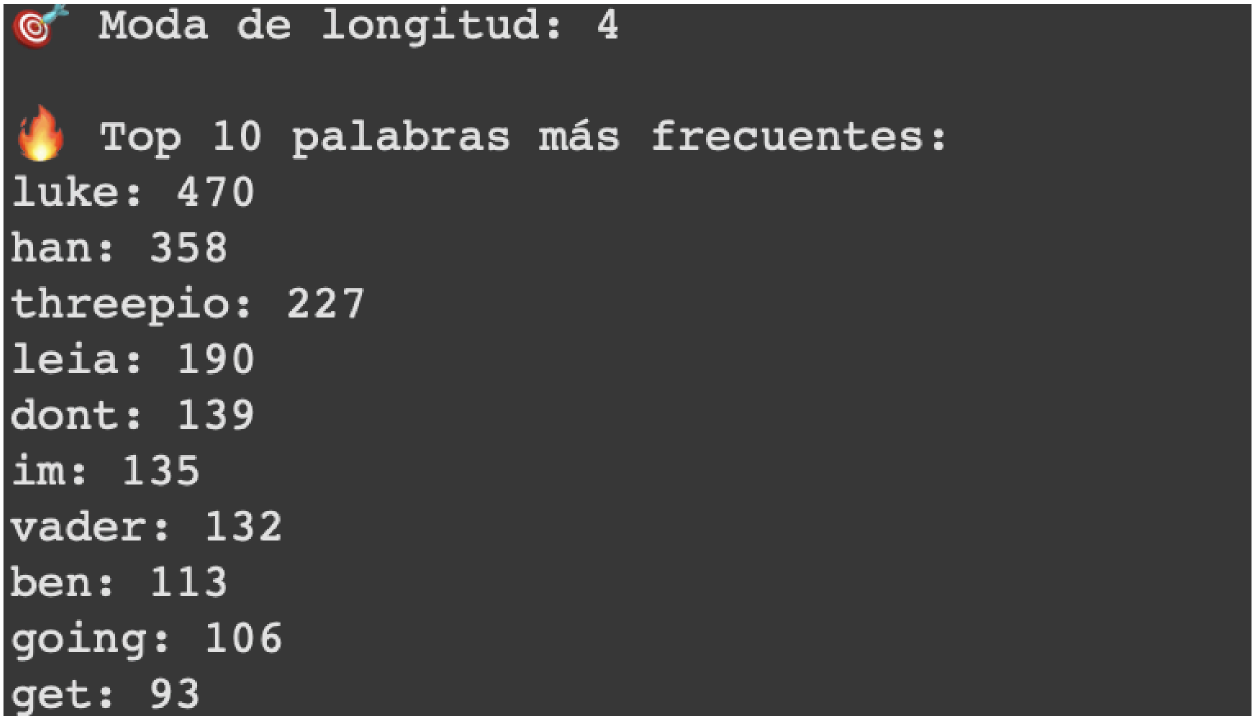


```python
# === Eliminar stopwords ===
stop_words = set(stopwords.words('english'))
filtered_words = [word for word in words if word not in stop_words]

# === Estadísticas básicas ===
total_words = len(filtered_words)
unique_words = len(set(filtered_words))
print(f"🔢 Palabras totales (sin stopwords): {total_words}")
print(f"🧠 Palabras únicas: {unique_words}")

# === Longitud de palabras ===
word_lengths = [len(word) for word in filtered_words]
df = pd.DataFrame({'word': filtered_words, 'length': word_lengths})

# === Estadísticas descriptivas ===
print("\n📈 Estadísticas descriptivas de longitud de palabra:")
print(df['length'].describe())
print(f"🎯 Moda de longitud: {df['length'].mode()[0]}")

# === Palabras más frecuentes ===
top_words = Counter(filtered_words).most_common(10)
print("\n🔥 Top 10 palabras más frecuentes:")
for word, freq in top_words:
    print(f"{word}: {freq}")

# === Histograma de longitudes ===
plt.figure(figsize=(8,5))
sns.histplot(df['length'], bins=15, kde=True, color="skyblue")
plt.title('Distribución de longitud de palabras')
plt.xlabel('Longitud de palabra')
plt.ylabel('Frecuencia')
plt.grid(True)
plt.tight_layout()
plt.show()
```

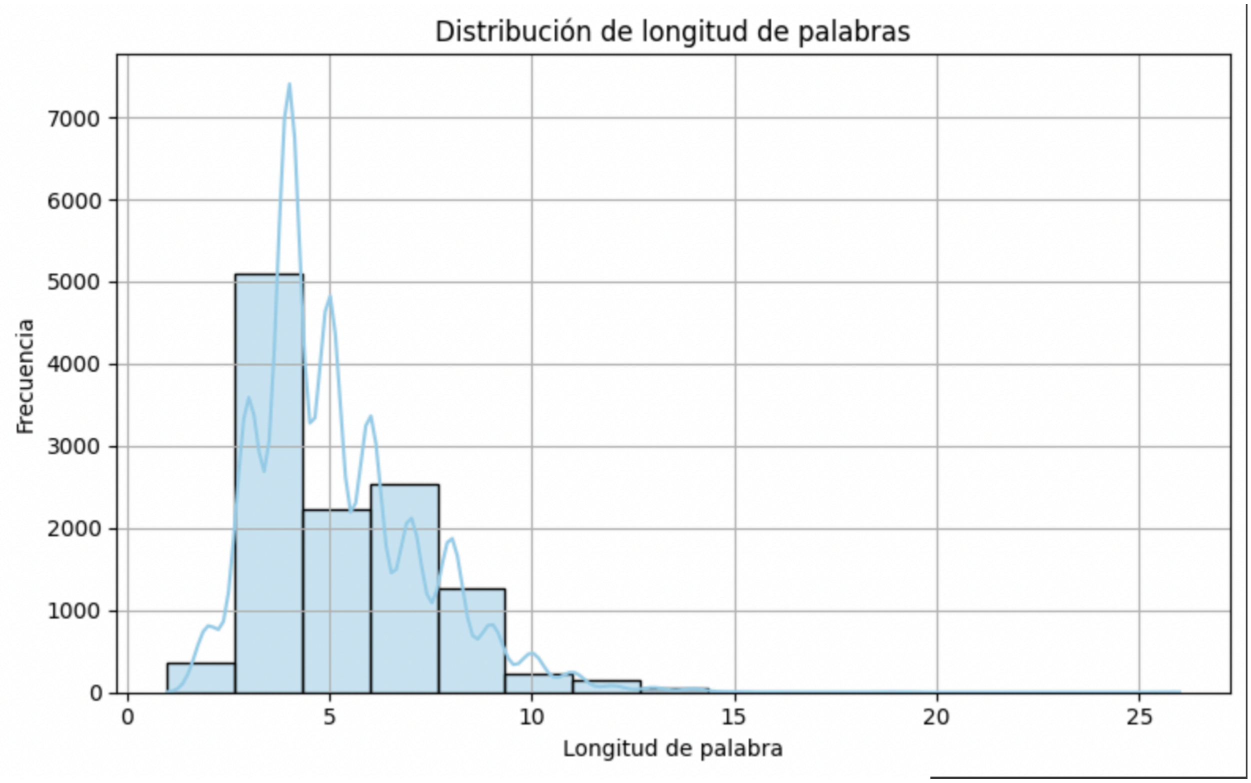


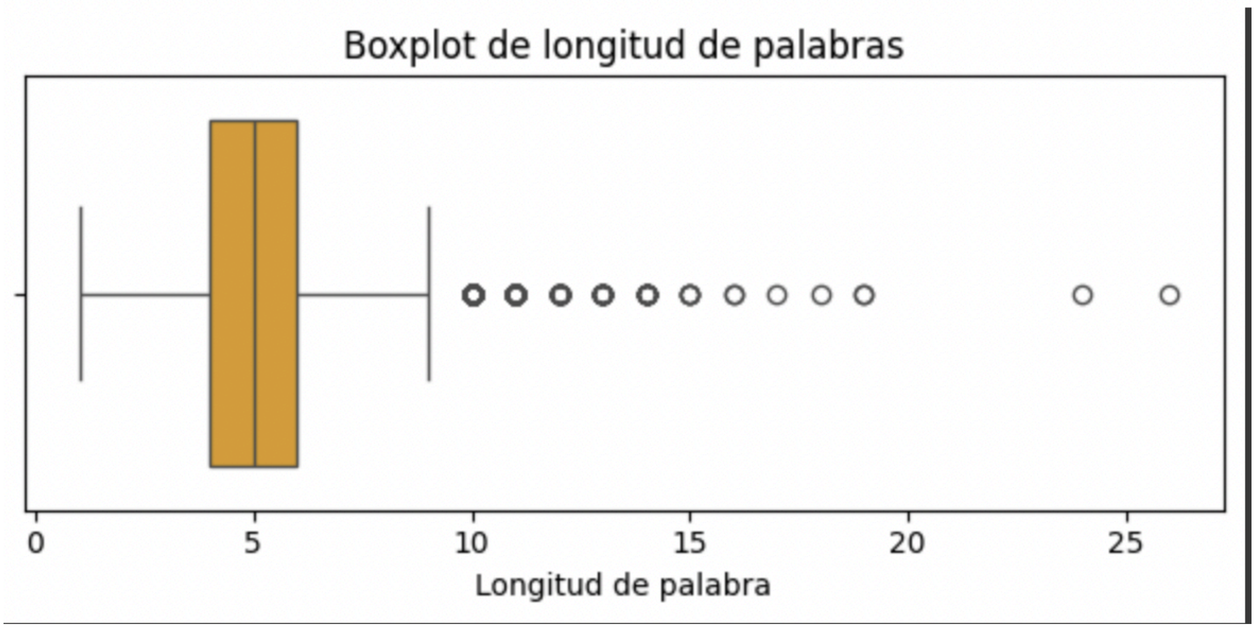
```python
# === Boxplot de longitudes ===
plt.figure(figsize=(6,3))
sns.boxplot(x=df['length'], color="orange")
plt.title('Boxplot de longitud de palabras')
plt.xlabel('Longitud de palabra')
plt.tight_layout()
plt.show()

# === Gráfico de barras: top 10 palabras ===
top_df = pd.DataFrame(top_words, columns=['word', 'count'])
plt.figure(figsize=(8,4))
sns.barplot(data=top_df, x='word', y='count', palette='viridis')
plt.title('Top 10 palabras más comunes (sin stopwords)')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

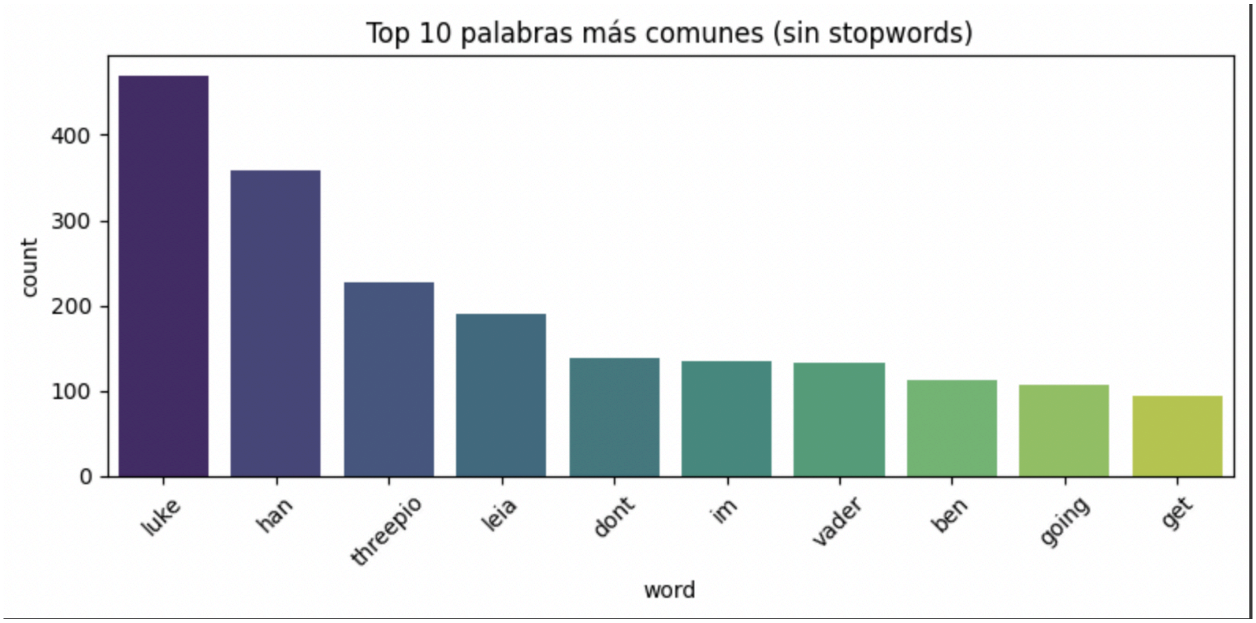
```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
2 files
SW_EpisodeIV.txt(text/plain) - 78278 bytes, last modified: n/a - 100% done
SW_EpisodeV.txt(text/plain) - 55487 bytes, last modified: n/a - 100% done
Saving SW_EpisodeIV.txt to SW_EpisodeIV.txt
Saving SW_EpisodeV.txt to SW_EpisodeV.txt
🔢 Palabras totales (sin stopwords): 11935
🧠 Palabras únicas: 2391

📈 Estadísticas descriptivas de longitud de palabra:
count    11935.000000
mean         5.201676
std          2.032697
min          1.000000
25%          4.000000
50%          5.000000
75%          6.000000
max         26.000000
Name: length, dtype: float64
```



```
🎯 Moda de longitud: 4

🔥 Top 10 palabras más frecuentes:
luke: 470
han: 358
threepio: 227
leia: 190
dont: 139
im: 135
vader: 132
ben: 113
going: 106
get: 93
```




Boxplot de longitud de palabras
0
5
10
15
20
25
Longitud de palabra


Top 10 palabras más comunes (sin stopwords)
count
0
100
200
300
400
word
luke
han
threepio
leia
dont
im
vader
ben
going
get

Conclusion

In this descriptive statistics activity, we analyzed the provided dataset using various statistical measures. The calculation of mean, median, and mode, along with standard deviation and range, helped us understand the central tendencies and dispersion of the data.

Our findings show that descriptive statistics are essential tools for data interpretation. The analysis revealed important patterns in the dataset and provided insights into its distribution.

The use of visual representations such as histograms and box plots enhanced our ability to interpret the numerical results effectively.

This exercise strengthened our understanding of basic statistical concepts and their practical applications for data analysis.