1B1. Práctica: Cálculo de frecuencias a nivel de un carácter

Alumno: **Luis Fernando Izquierdo Berdugo**

Materia: **Procesamiento de Información**

Fecha: **13 de Agosto de 2024**

Instrucciones:

Crear un notebook para preprocesar el texto y convertirlo a minúsculas, quitar acentos y los siguientes caracteres:

`;:,.\\-\"'/()[]¿?¡!{}~<>|«»-–’\t\n\r`

1. Calcular las frecuencias a nivel de un carácter, de los documentos text_1 al text_5 y generar los histogramas para cada archivo. Para generar las gráficas, ordenar los caracteres de acuerdo a su frecuencia; de mayor a menor frecuencia en el histograma.
2. Interpretar cada histograma en referencia a qué caracteres destacan y dar indicios del porqué ocurren esas observaciones.

```
import re

special = ";:,.\\-\"'/()[]¿?¡!{}~<>|«»-–’\t\n\r"

file = open("text_1.txt", "r")
text_1 = file.read()
file.close()
text_1 = text_1.lower()
text_1 = re.sub(r"[;:,.\-\"'/\(\)\[\]¿?¡!\{\}~<>|«»-–’\t\n\r]", "",
text_1)

file = open("text_2.txt", "r")
text_2 = file.read()
file.close()
text_2 = text_2.lower()
text_2 = re.sub(r"[;:,.\-\"'/\(\)\[\]¿?¡!\{\}~<>|«»-–’\t\n\r]", "",
text_2)

file = open("text_3.txt", "r")
text_3 = file.read()
file.close()
text_3 = text_3.lower()
text_3 = re.sub(r"[;:,.\-\"'/\(\)\[\]¿?¡!\{\}~<>|«»-–’\t\n\r]", "",
text_3)
```

```
file = open("text_4.txt", "r")
text_4 = file.read()
file.close()
text_4 = text_4.lower()
text_4 = re.sub(r"[;:,.\-\"'/\(\)\[\]¿?¡!\{\}~<>|«»-—’\t\n\r]", "",
text_4)

file = open("text_5.txt", "r")
text_5 = file.read()
file.close()
text_5 = text_5.lower()
text_5 = re.sub(r"[;:,.\-\"'/\(\)\[\]¿?¡!\{\}~<>|«»-—’\t\n\r]", "",
text_5)
```

Inciso 1

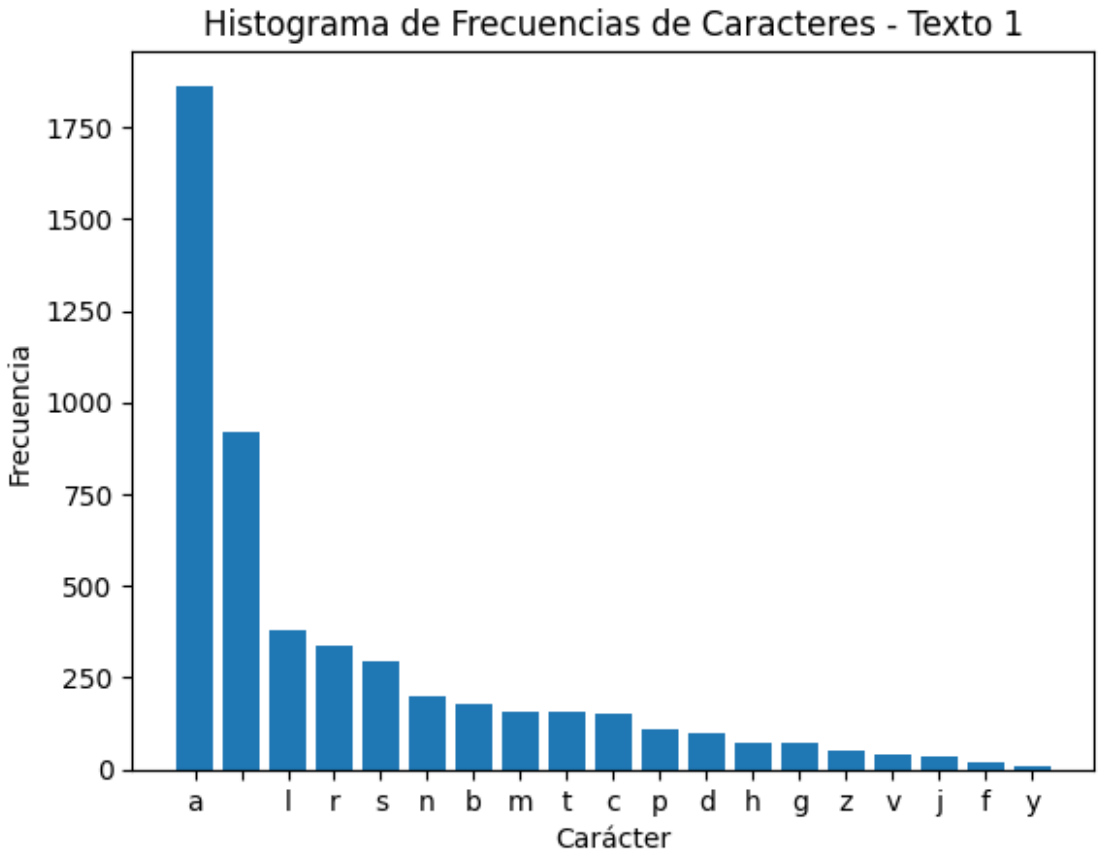
Calcular las frecuencias a nivel de un carácter, de los documentos text_1 al text_5 y generar los histogramas para cada archivo. Para generar las gráficas, ordenar los caracteres de acuerdo a su frecuencia; de mayor a menor frecuencia en el histograma.

```
def contador(texto):

    frecuencias = {}
    for caracter in texto:
        if caracter in frecuencias:
            frecuencias[caracter] += 1
        else:
            frecuencias[caracter] = 1
    return frecuencias

frecuencia_1 = contador(text_1)
print(frecuencia_1)
```

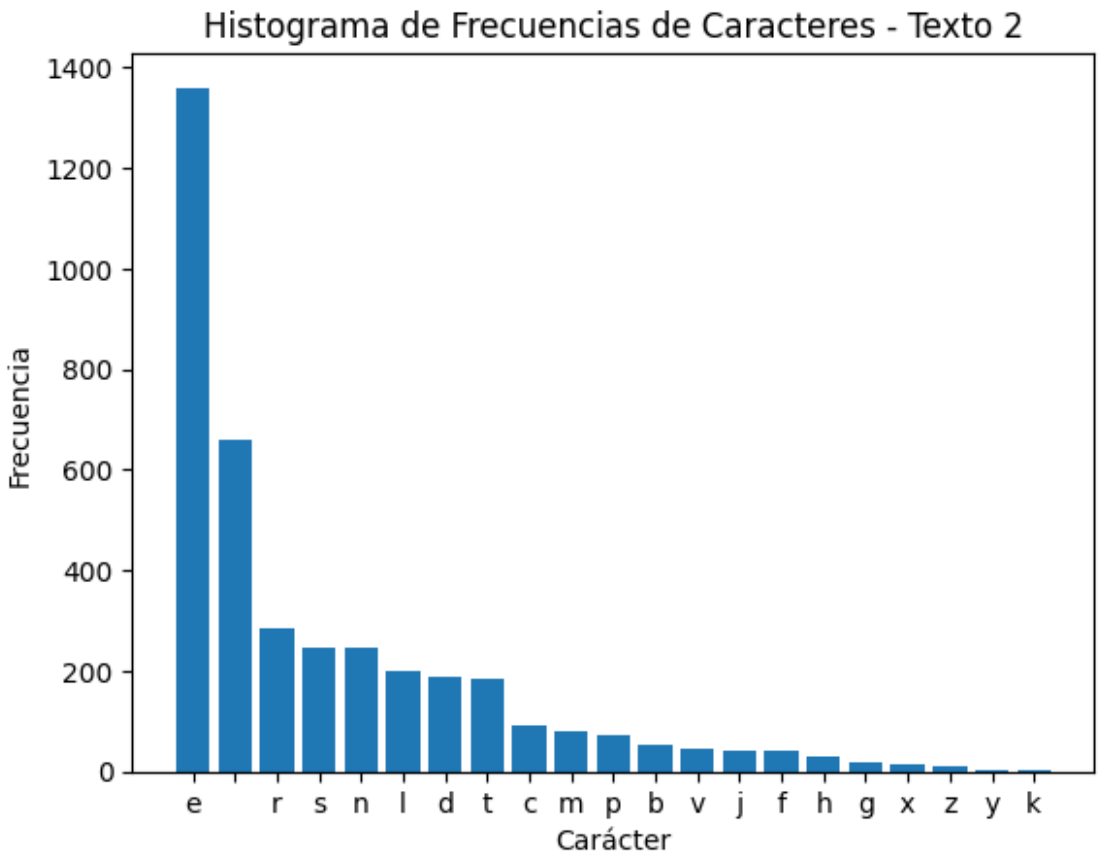
```
{'c': 153, 'a': 1863, 'n': 202, 't': 155, ' ': 920, 's': 296, 'b':
180, 'r': 338, 'h': 74, 'm': 158, 'd': 97, 'l': 381, 'p': 112, 'z':
51, 'v': 42, 'g': 71, 'y': 7, 'j': 33, 'f': 18}
```

```
import matplotlib.pyplot as plt

def grafica(frecuencias, texto):
    frecuencias = dict(sorted(frecuencias.items(), key=lambda item:
item[1], reverse=True))
    caracteres = list(frecuencias.keys())
    valores = list(frecuencias.values())
    plt.bar(caracteres, valores)
    plt.xlabel("Carácter")
    plt.ylabel("Frecuencia")
```

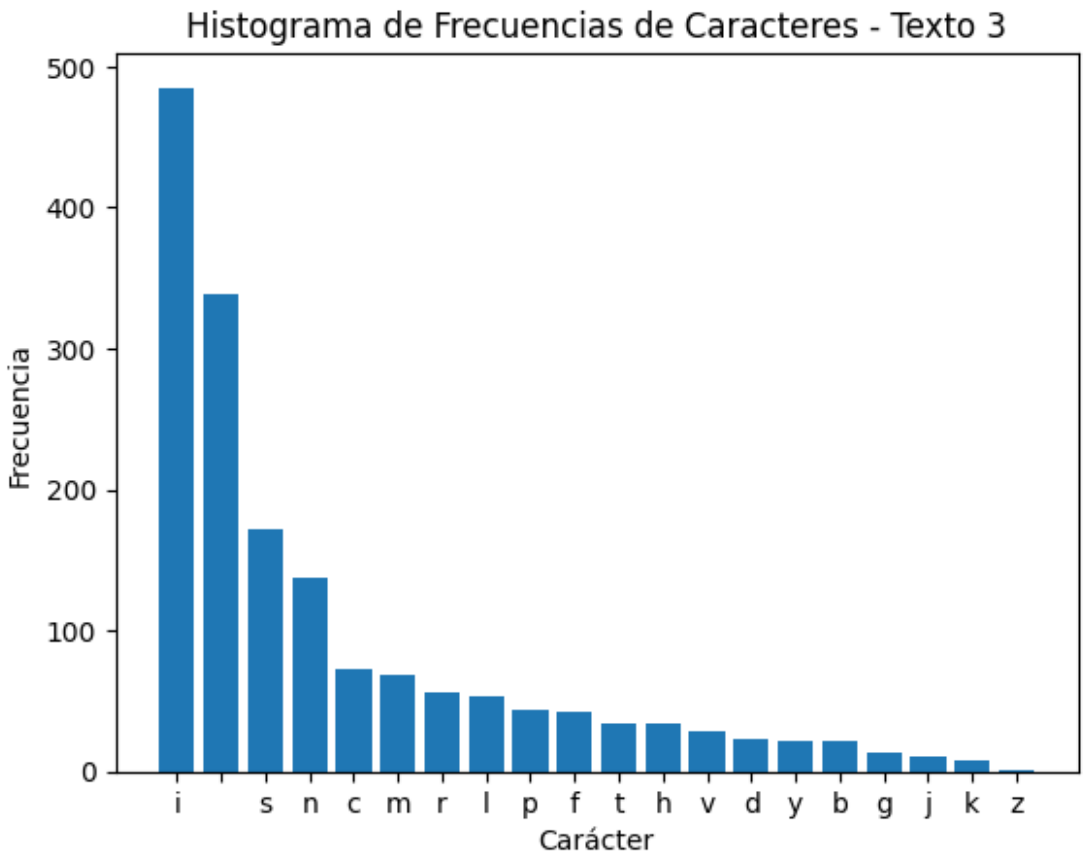
```
    plt.title(f"Histograma de Frecuencias de Caracteres {texto}")
    plt.show()

grafica(frecuencia_1, "- Texto 1")
```



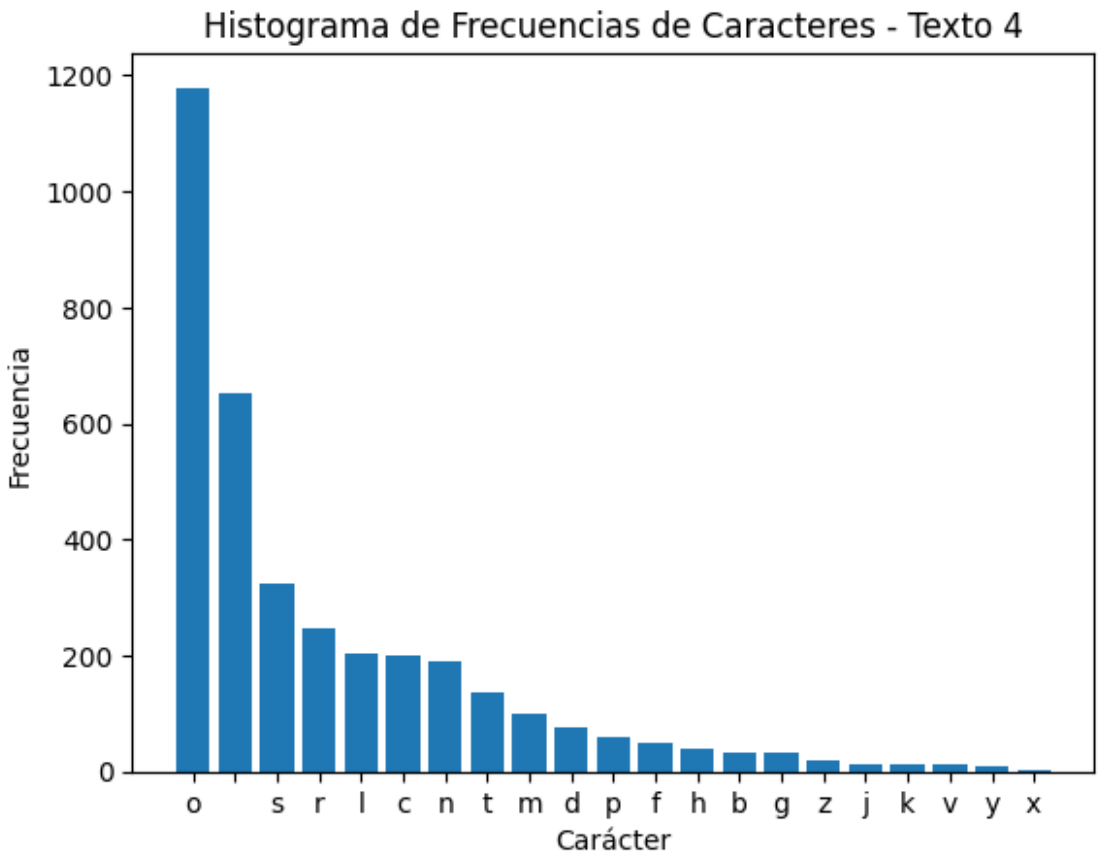

```
frecuencia_2 = contador(text_2)
grafica(frecuencia_2, "- Texto 2")
print(frecuencia_2)
```



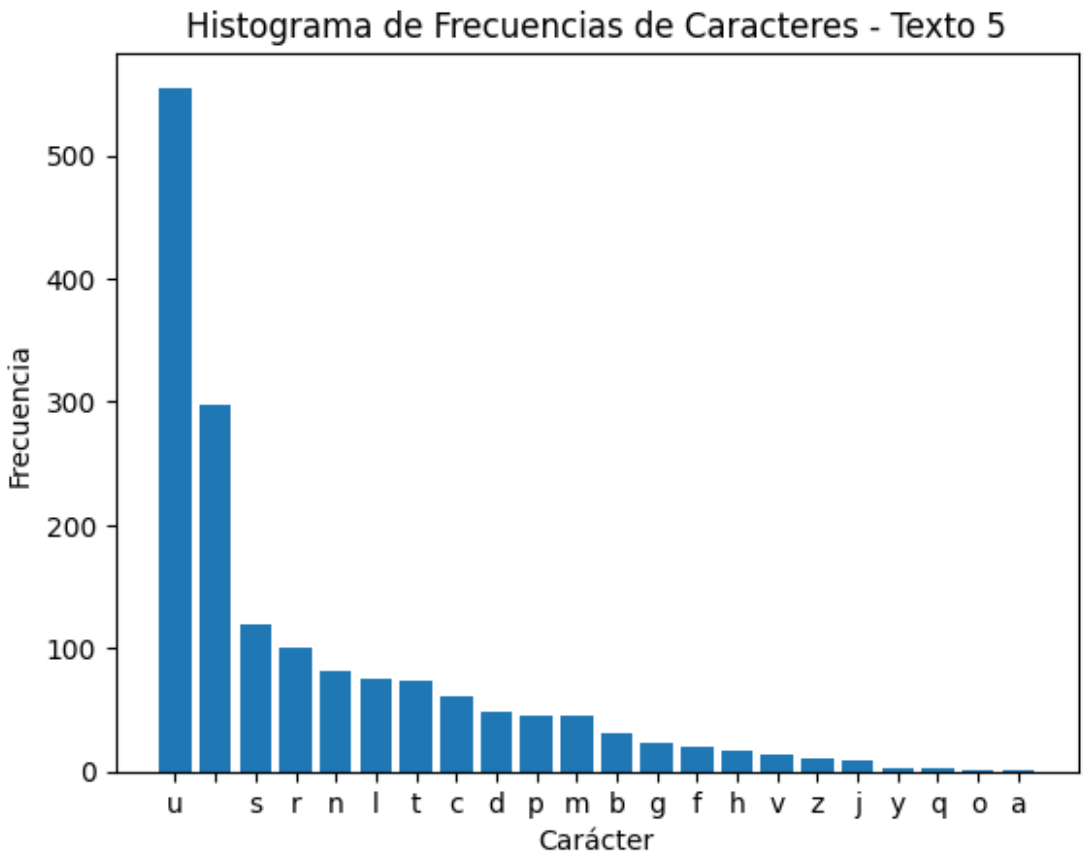

```
{'e': 1360, 'l': 199, ' ': 658, 'h': 28, 'r': 284, 'j': 42, 'b': 53,
'd': 189, 'n': 246, 'v': 44, 'c': 93, 's': 247, 'p': 71, 'y': 3, 't':
185, 'f': 40, 'm': 79, 'g': 18, 'z': 10, 'x': 13, 'k': 1}

frecuencia_3 = contador(text_3)
grafica(frecuencia_3, "- Texto 3")
print(frecuencia_3)
```




```
{'m': 69, 'i': 485, ' ': 339, 's': 172, 'n': 137, 'b': 21, 'k': 8,
't': 34, 'r': 56, 'c': 72, 'p': 44, 'v': 29, 'h': 34, 'd': 23, 'f':
42, 'l': 53, 'y': 22, 'g': 13, 'j': 10, 'z': 1}

frecuencia_4 = contador(text_4)
grafica(frecuencia_4, "- Texto 4")
print(frecuencia_4)
```




```
{'l': 204, 'o': 1179, 's': 324, ' ': 653, 'c': 201, 'm': 100, 't':
136, 'r': 247, 'h': 38, 'k': 11, 'd': 76, 'f': 50, 'j': 13, 'n': 190,
'g': 31, 'b': 32, 'v': 11, 'p': 60, 'z': 20, 'y': 10, 'x': 3}

frecuencia_5 = contador(text_5)
grafica(frecuencia_5, "- Texto 5")
print(frecuencia_5)
```

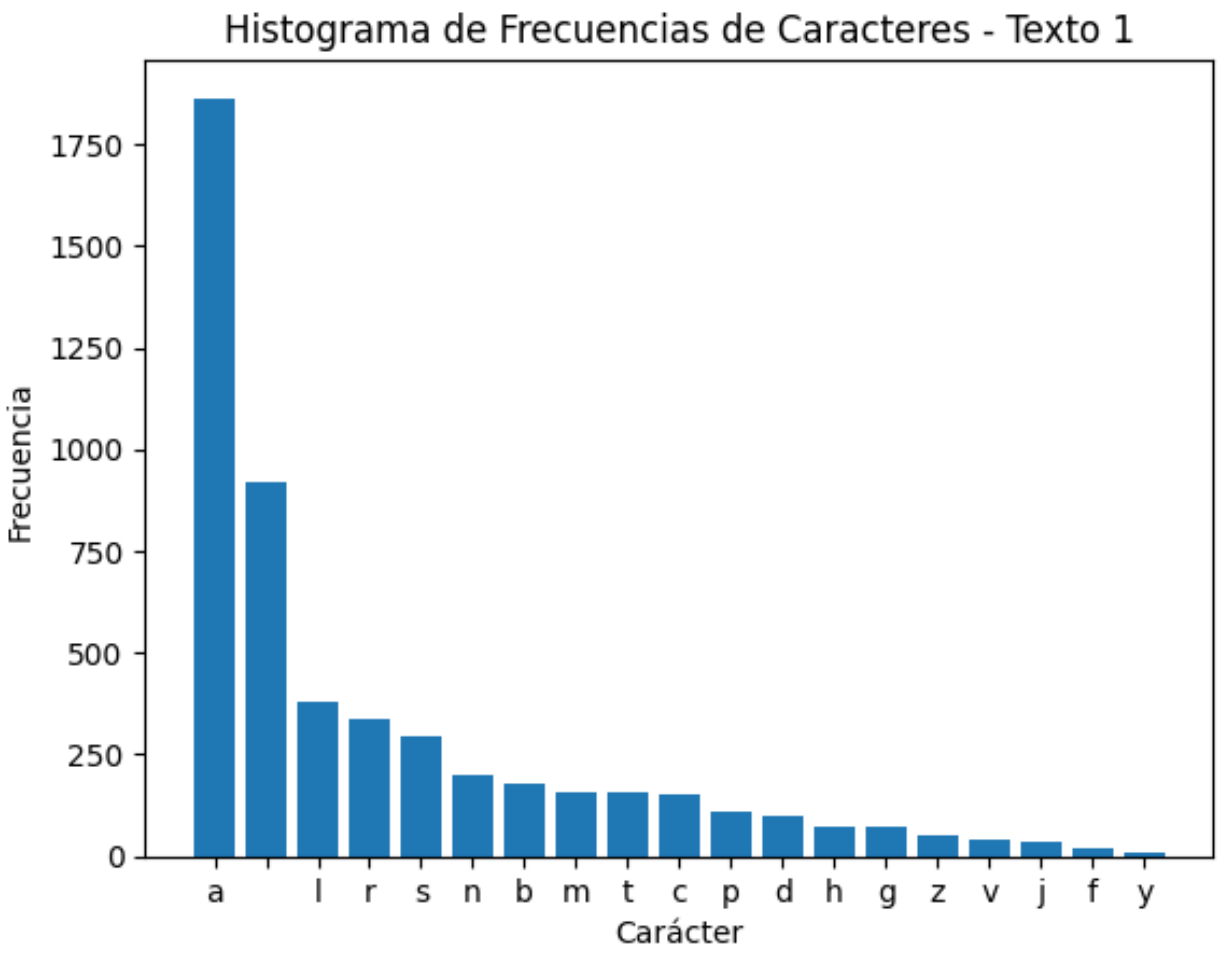



```
{'u': 555, 'n': 81, ' ': 297, 'g': 24, 'r': 100, 'v': 14, 'd': 49,
'l': 75, 's': 120, 'p': 45, 'm': 45, 'f': 20, 't': 74, 'y': 3, 'o': 1,
'c': 61, 'h': 17, 'j': 9, 'z': 10, 'b': 31, 'q': 2, 'a': 1}
```

Inciso 2

Interpretar cada histograma en referencia a qué caracteres destacan y dar indicios del porqué ocurren esas observaciones.

Para el **histograma del texto 1** podemos observar que, a lo largo del texto, predomina el uso de la letra "a", siendo incluso el doble del siguiente carácter más usado que es el espacio " ".

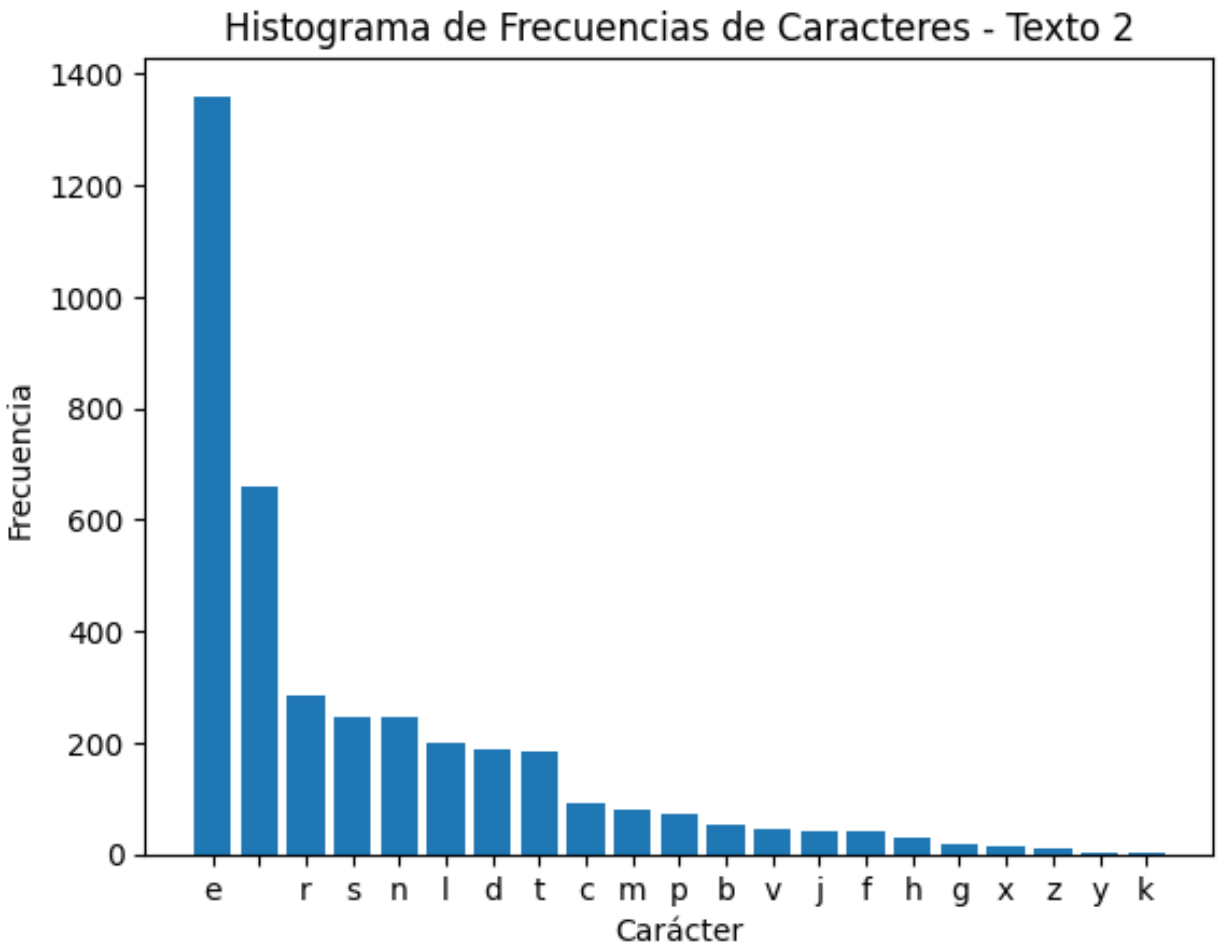


Esto nos indica que el texto está centrado en el uso de la letra A, cosa que podemos observar al leer las primeras líneas del texto:

`Cantata a Satanás Abraham amaba a Sara cada mañana clara: pasaba la manaza, arañaba la lana`

De igual manera podemos observar que el carácter menos usado fue la letra "y" Otro análisis interesante es la ausencia de los caracteres de las demás vocales "e", "i", "o," y "u", lo cual nos podría indicar que el texto es un **lipograma monovocálico**.

En el **histograma del texto 2** observamos una conclusión similar al del texto 1, teniendo esta vez el carácter de la letra "e" como predominante y la letra "k" como la menos utilizada

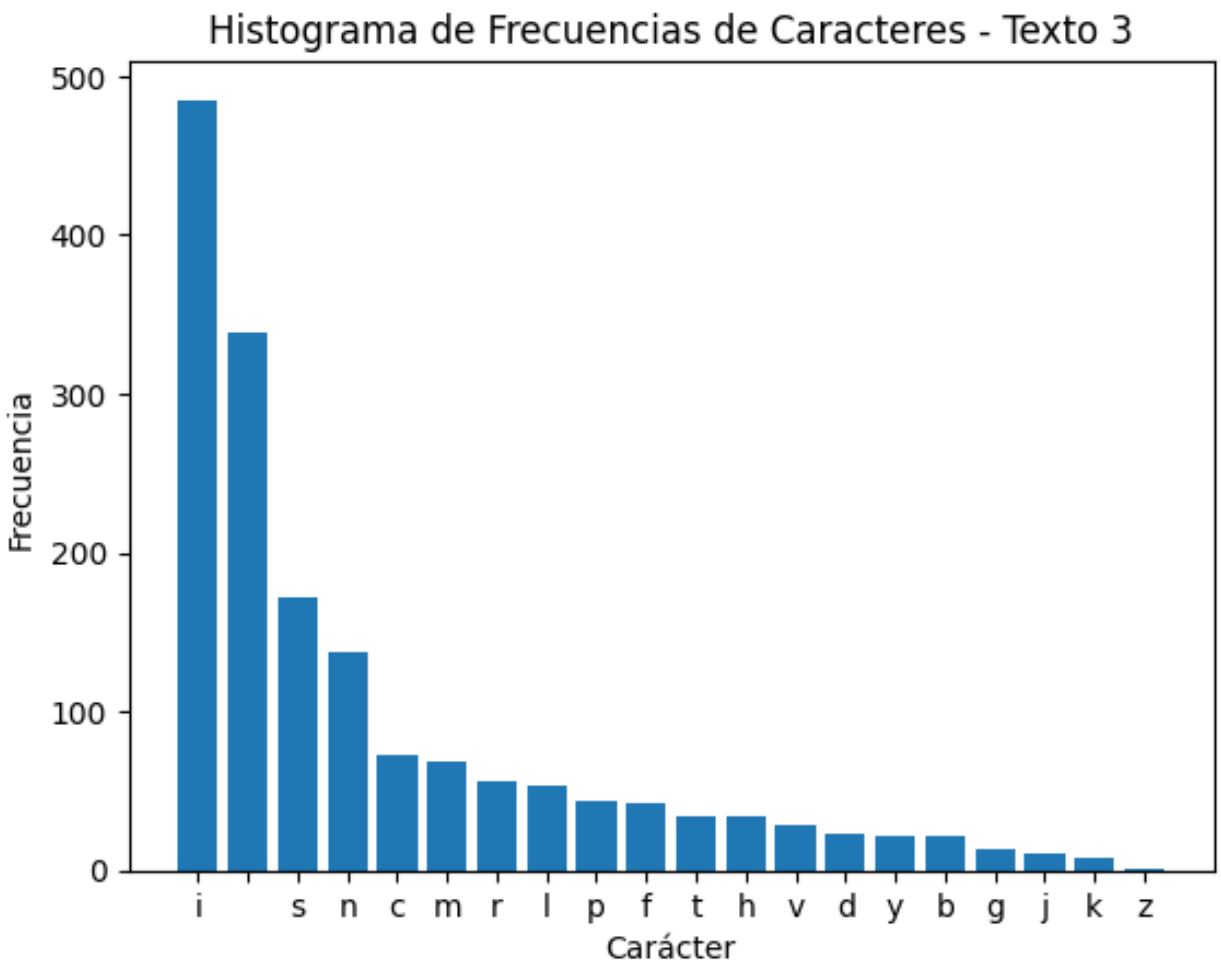


Sin embargo, esta vez podemos observar que hubo menos variación de letras en las palabras, ya que, vemos que fuera del espacio y la "e", hay más presencia de los caracteres "r", "s", "n", "l", "d" y "t", teniendo esta última el doble de la frecuencia de la siguiente más cercana que es la letra "c". De igual manera, si observamos las primeras líneas del texto, se puede observar la frecuencia de la "e":

```
El hereje rebelde En el verde césped del edén, célebre sede de creyentes, el decente Efrén se estremece. Tres deberes del mes lee en el templete del regente: «Defender el vergel del Hereje Rebelde
```

En este caso también nos encontramos con un lipograma monovocálico, ya que no están presentes las letras "a", "i", "o" y "u".

Para el **histograma del texto 3** podemos observar una tendencia marcada por los dos anteriores, ya que en este caso también observamos el predominante uso de la letra "i", seguido del carácter de "espacio". En este caso también vemos la presencia de los caracteres "s" y "n" sobre los demás.

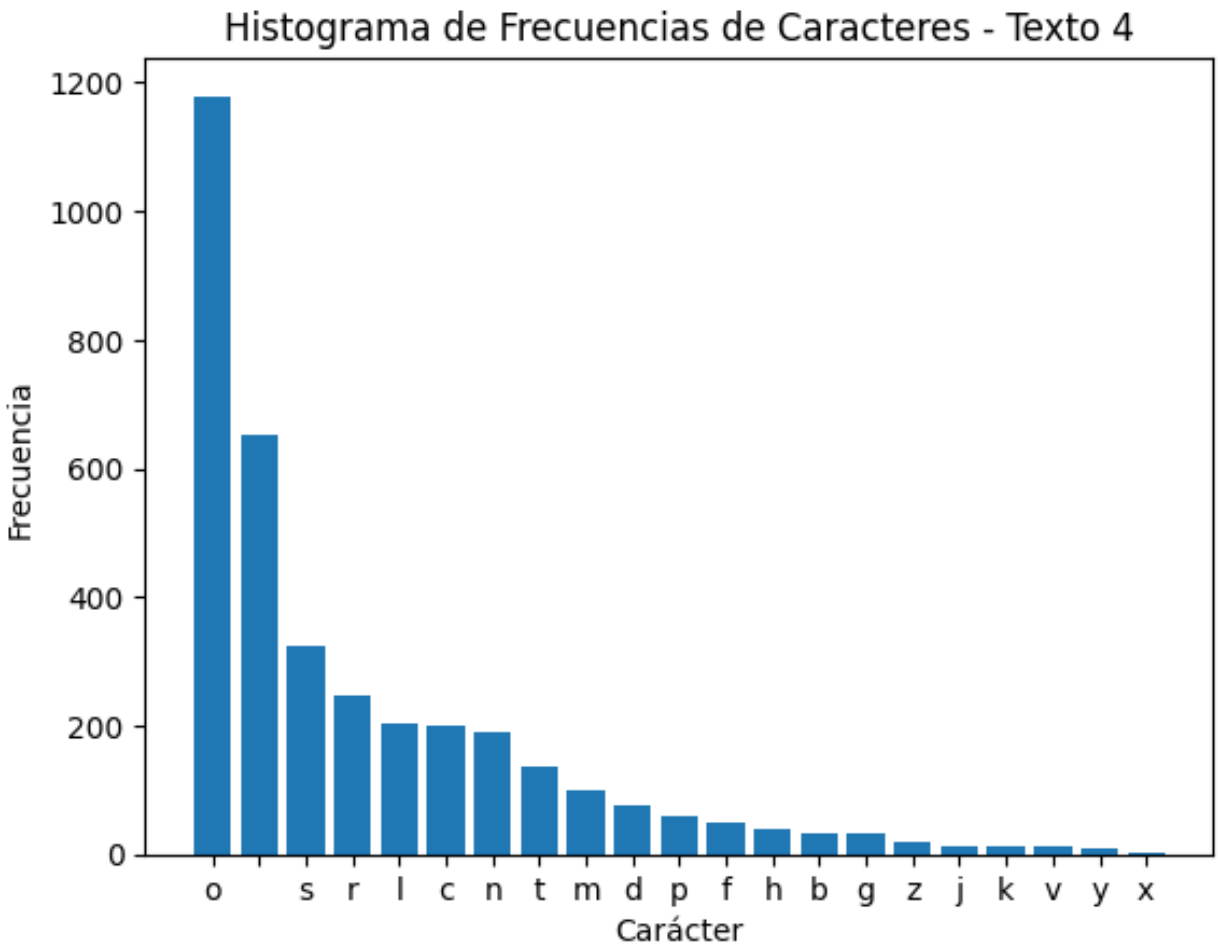


De nuevo, al analizar el texto, podemos concluir que está centrado en usar palabras únicamente con la vocal "i".

```
Mimí sin bikini Insistir, ¿Crispín?... Mi visir, mi bichín, mi cid: si sin ti viví difícil chipichipi sin fin: crisis y crisis: bilis, rinitis, tisis. (Snif, snif.)
```

También podemos concluir que es un lipograma monovocálico al notar la ausencia de las letras "a", "e", "o" y "u".

En el **histograma del texto 4** encontramos que el carácter más frecuente es la letra "o", seguido del espacio. En este caso notamos que, a pesar de que algunos caracteres se usan más que otros, no hay un salto grande entre un caracter y otro, más que los mencionados previamente que son más frecuentes.

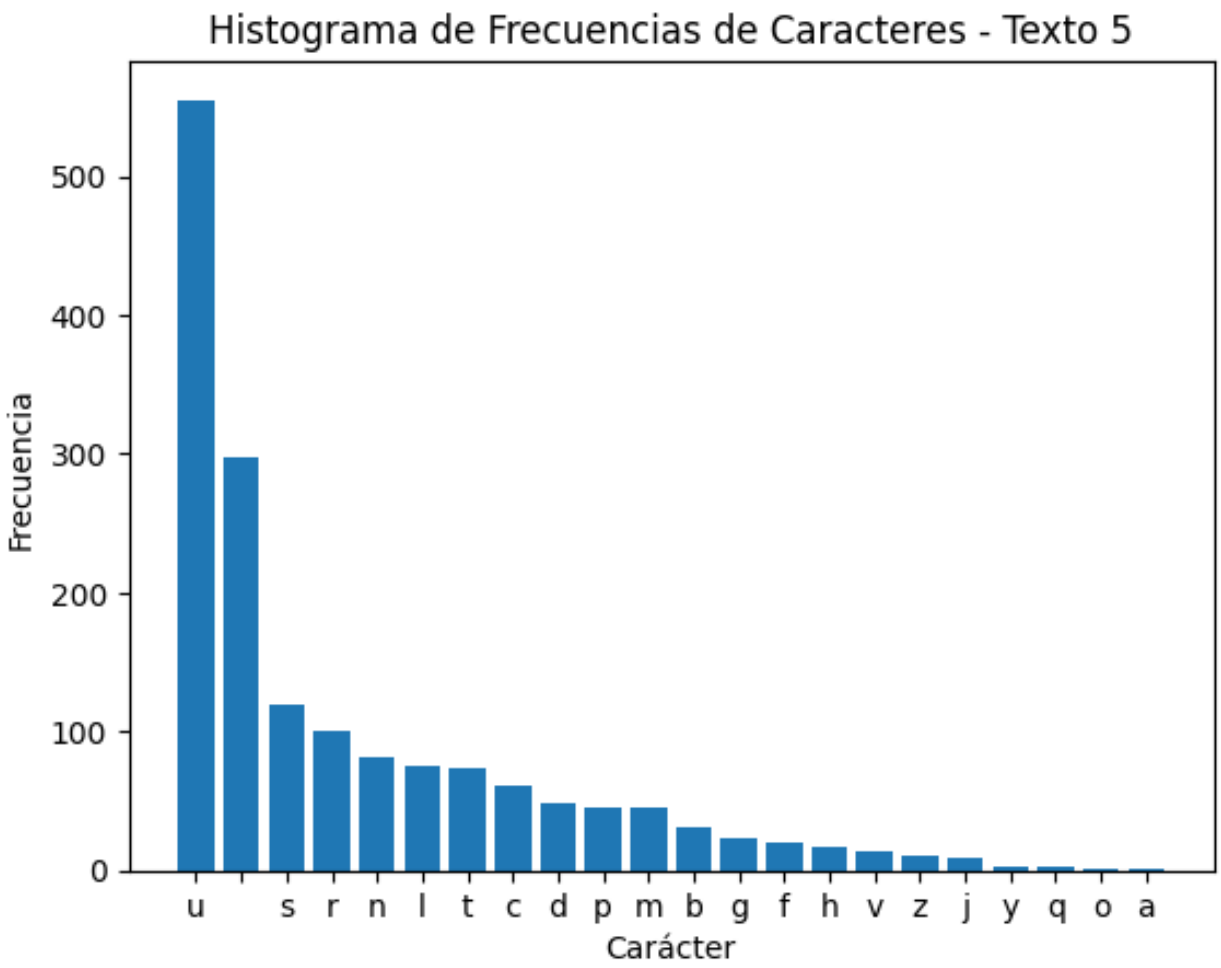


Analizando unas líneas del texto, observamos que si se presenta de manera predominante el carácter "o" y el " ".

```
Los locos somos otro cosmosOtto colocó los shocks. Rodolfo mostró los ojos con horror: dos globos rojos, torvos, con poco fósforo como bolsos fofos; combó los hombros, sollozó: «No doctor, no... loco no...» Sor Socorro lo frotó con yodo: «Pon flojos los codos — rogó—, ponlos como yo.
```

Como en los textos anteriores, nos encontramos con un lipograma monovocálico centrado en solamente usar la "o".

Finalmente, al analizar el **histograma del texto 5** observamos que el caracter con mayor frecuencia es el de la vocal "u" y el de espacio " ".



Si analizamos el texto, observamos que realmente está la presencia de la letra "u".

```
Un gurú vudúUn gurú vudú, un Duvulur, supusu un mundu futuru mu suyu;
un mundo cuyu multutud frustrudu pur sus Tuntuns Mucutus nuncu
luchuru, nuncu junturu sus músculus puru hundur su curul. Su tutur,
Pupú Duc, un sultún mu crul, un furúnculu du Luzbul, fundú su brutul
club cun un grupúsculu du brujus du truculuntus trucus cun sustu vudú.
Muchus uñus ul publu sufrú pústulus, sudú jugus púrpuru, tuvu tumurus
du pus, susurrú su runcur, su humbru, su murtu, su cruz.
```

A diferencia de los demás textos, en este si encontramos la presencia de otra vocal, la cual es la "a" en la palabra "brutal". Esto podría deberse a un error de transcripción.

En conclusión, cada una de los textos trata de utilizar una de las 5 vocales únicamente. Lo cual logran con éxito a excepción del de la letra "u". Algo que se puede observar en todos los textos es la fuerte presencia del carácter de espacio " ", esto se debe a que cada una de las palabras está separada por un espacio, por lo cual aumenta de manera constante a lo largo de todos los textos.