


```
import nltk
from nltk.stem import PorterStemmer
from collections import Counter
import re

# Asegúrate de tener los recursos necesarios de NLTK
nltk.download('punkt')

# Leer el archivo de texto
with open('/content/Amor y Llanto.txt', 'r', encoding='utf-8') as file:
    text = file.read()

# Preprocesamiento del texto: quitar caracteres no deseados
text = re.sub(r'[^\w\s]', '', text.lower()) # Eliminar puntuación y convertir a minúsculas
```


 [nltk\_data] Downloading package punkt to /root/nltk\_data...  
[nltk\_data] Unzipping tokenizers/punkt.zip.

```
# Tokenización: dividir el texto en palabras
words = text.split()

# Aplicar stemming con el Porter Stemmer de NLTK
stemmer = PorterStemmer()
stemmed_words = [stemmer.stem(word) for word in words]

# Contar la frecuencia de las palabras
word_counts = Counter(stemmed_words)

# Mostrar las palabras más comunes
print("Frecuencia de las palabras (con stem):")
for word, count in word_counts.most_common(10): # Muestra las 10 palabras más comunes
    print(f"{word}: {count}")
```

 Frecuencia de las palabras (con stem):  
de: 4734  
la: 3736  
y: 2936  
que: 2691  
a: 2470  
su: 2290  
el: 2281  
en: 1918  
lo: 1212  
con: 1016



si quieres no tomar en cuenta los conectores:

si quieres no tomar en cuenta los conectores:

```
import nltk
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from collections import Counter
import re

# Asegúrate de tener los recursos necesarios de NLTK
nltk.download('punkt')
nltk.download('stopwords')

# Leer el archivo de texto
with open('/content/Amor y Llanto.txt', 'r', encoding='utf-8') as file:
    text = file.read()

# Preprocesamiento del texto: quitar caracteres no deseados
text = re.sub(r'[^\w\s]', '', text.lower()) # Eliminar puntuación y convertir a minúsculas

# Tokenización: dividir el texto en palabras
words = text.split()

# Aplicar stemming con el Porter Stemmer de NLTK
stemmer = PorterStemmer()
stemmed_words = [stemmer.stem(word) for word in words]
```

```
# Obtener las stopwords en español
stop_words = set(stopwords.words('spanish'))

# Filtrar las stopwords
filtered_words = [word for word in stemmed_words if word not in stop_words]

# Contar la frecuencia de las palabras filtradas
word_counts = Counter(filtered_words)

# Mostrar las palabras más comunes
print("Frecuencia de las palabras (con stem y sin conectores):")
for word, count in word_counts.most_common(10): # Muestra las 10 palabras más comunes
    print(f"{word}: {count}")
```

```
↗ [nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Unzipping corpora/stopwords.zip.
Frecuencia de las palabras (con stem y sin conectores):
rey: 482
don: 437
má: 325
reina: 302
mano: 263
ojo: 262
dijo: 243
joven: 205
the: 189
cond: 170
```