Tarea 1: con la librería NLTK de Python, separar un texto en palabras, así como determinar la frecuencia de estas y graficar los resultados.

In [1]:

```
%matplotlib inline
import nltk
import re
from nltk.probability import FreqDist
from urllib import request # librería que nos permite realizar peticiones web
import matplotlib.pyplot as plt # librería para realizar ploteos de figuras
from nltk.corpus import stopwords # librería que contiene las palabras auxiliares del idi
oma inglés
```

In [2]:

```
url = 'http://www.gutenberg.org/files/2554/2554-0.txt' # url de donde se extraerá el text
o
response = request.urlopen(url) # aquí se estará instanciando nuestra petición de la url
text = response.read().decode('utf-8') # aquí se guardará el texto ya decodificado por ut
f-8 obtenido por nuestr petición
print(text) # muestra del texto obtenido
```

```
oks. But to his great surprise she had not once approached the subject
and had not even offered him the Testament. He had asked her for it
himself not long before his illness and she brought him the book without
a word. Till now he had not opened it.

He did not open it now, but one thought passed through his mind: “Can
her convictions not be mine now? Her feelings, her aspirations at
least....”

She too had been greatly agitated that day, and at night she was taken
ill again. But she was so happy--and so unexpectedly happy--that she was
almost frightened of her happiness. Seven years, _only_ seven years! At
the beginning of their happiness at some moments they were both ready
to look on those seven years as though they were seven days. He did not
know that the new life would not be given him for nothing, that he would
have to pay dearly for it, that it would cost him great striving, great
suffering.

But that is the beginning of a new story--the story of the gradual
renewal of a man, the story of his gradual regeneration, of his passing
from one world into another, of his initiation into a new unknown life.
That might be the subject of a new story, but our present story is
ended.




End of Project Gutenberg’s Crime and Punishment, by Fyodor Dostoevsky

*** END OF THIS PROJECT GUTENBERG EBOOK CRIME AND PUNISHMENT ***

***** This file should be named 2554-0.txt or 2554-0.zip *****
This and all associated files of various formats will be found in:
        http://www.gutenberg.org/2/5/5/2554/

Produced by John Bickers; and Dagny and David Widger

Updated editions will replace the previous one--the old editions
will be renamed.

Creating the works from public domain print editions means that no
one owns a United States copyright in these works, so the Foundation
(and you!) can copy and distribute it in the United States without
permission and without paying copyright royalties.  Special rules,
set forth in the General Terms of Use part of this license, apply to
copying and distributing Project Gutenberg-tm electronic works to
protect the PROJECT GUTENBERG-tm concept and trademark.  Project
```

the copyright status of any work in any country outside the United States.

1.E. Unless you have removed all references to Project Gutenberg:

1.E.1. The following sentence, with active links to, or other immediate access to, the full Project Gutenberg-tm License must appear prominently whenever any copy of a Project Gutenberg-tm work (any work on which the phrase “Project Gutenberg” appears, or with which the phrase “Project Gutenberg” is associated) is accessed, displayed, performed, viewed, copied or distributed:

This eBook is for the use of anyone anywhere at no cost and with almost no restrictions whatsoever. You may copy it, give it away or re-use it under the terms of the Project Gutenberg License included with this eBook or online at www.gutenberg.org

1.E.2. If an individual Project Gutenberg-tm electronic work is derived from the public domain (does not contain a notice indicating that it is posted with permission of the copyright holder), the work can be copied and distributed to anyone in the United States without paying any fees or charges. If you are redistributing or providing access to a work with the phrase “Project Gutenberg” associated with or appearing on the work, you must comply either with the requirements of paragraphs 1.E.1 through 1.E.7 or obtain permission for the use of the work and the Project Gutenberg-tm trademark as set forth in paragraphs 1.E.8 or 1.E.9.

1.E.3. If an individual Project Gutenberg-tm electronic work is posted with the permission of the copyright holder, your use and distribution must comply with both paragraphs 1.E.1 through 1.E.7 and any additional terms imposed by the copyright holder. Additional terms will be linked to the Project Gutenberg-tm License for all works posted with the permission of the copyright holder found at the beginning of this work.

1.E.4. Do not unlink or detach or remove the full Project Gutenberg-tm License terms from this work, or any files containing a part of this work or any other work associated with Project Gutenberg-tm.

1.E.5. Do not copy, display, perform, distribute or redistribute this electronic work, or any part of this electronic work, without prominently displaying the sentence set forth in paragraph 1.E.1 with active links or immediate access to the full terms of the Project Gutenberg-tm License.

1.E.6. You may convert to and distribute this work in any binary, compressed, marked up, nonproprietary or proprietary form, including any word processing or hypertext form. However, if you provide access to or distribute copies of a Project Gutenberg-tm work in a format other than “Plain Vanilla ASCII” or other format used in the official version posted on the official Project Gutenberg-tm web site (www.gutenberg.org), you must, at no additional cost, fee or expense to the user, provide a copy, a means of exporting a copy, or a means of obtaining a copy upon request, of the work in its original “Plain Vanilla ASCII” or other form. Any alternate format must include the full Project Gutenberg-tm License as specified in paragraph 1.E.1.

1.E.7. Do not charge a fee for access to, viewing, displaying, performing, copying or distributing any Project Gutenberg-tm works unless you comply with paragraph 1.E.8 or 1.E.9.

1.E.8. You may charge a reasonable fee for copies of or providing access to or distributing Project Gutenberg-tm electronic works provided that

- You pay a royalty fee of 20% of the gross profits you derive from the use of Project Gutenberg-tm works calculated using the method you already use to calculate your applicable taxes. The fee is owed to the owner of the Project Gutenberg-tm trademark, but he has agreed to donate royalties under this paragraph to the Project Gutenberg Literary Archive Foundation. Royalty payments must be paid within 60 days following each date on which you

prepare (or are legally required to prepare) your periodic tax returns. Royalty payments should be clearly marked as such and sent to the Project Gutenberg Literary Archive Foundation at the address specified in Section 4, "Information about donations to the Project Gutenberg Literary Archive Foundation."

- You provide a full refund of any money paid by a user who notifies you in writing (or by e-mail) within 30 days of receipt that s/he does not agree to the terms of the full Project Gutenberg-tm License. You must require such a user to return or destroy all copies of the works possessed in a physical medium and discontinue all use of and all access to other copies of Project Gutenberg-tm works.

- You provide, in accordance with paragraph 1.F.3, a full refund of any money paid for a work or a replacement copy, if a defect in the electronic work is discovered and reported to you within 90 days of receipt of the work.

- You comply with all other terms of this agreement for free distribution of Project Gutenberg-tm works.

1.E.9. If you wish to charge a fee or distribute a Project Gutenberg-tm electronic work or group of works on different terms than are set forth in this agreement, you must obtain permission in writing from both the Project Gutenberg Literary Archive Foundation and Michael Hart, the owner of the Project Gutenberg-tm trademark. Contact the Foundation as set forth in Section 3 below.

1.F.

1.F.1. Project Gutenberg volunteers and employees expend considerable effort to identify, do copyright research on, transcribe and proofread public domain works in creating the Project Gutenberg-tm collection. Despite these efforts, Project Gutenberg-tm electronic works, and the medium on which they may be stored, may contain "Defects," such as, but not limited to, incomplete, inaccurate or corrupt data, transcription errors, a copyright or other intellectual property infringement, a defective or damaged disk or other medium, a computer virus, or computer codes that damage or cannot be read by your equipment.

1.F.2. LIMITED WARRANTY, DISCLAIMER OF DAMAGES - Except for the "Right of Replacement or Refund" described in paragraph 1.F.3, the Project Gutenberg Literary Archive Foundation, the owner of the Project Gutenberg-tm trademark, and any other party distributing a Project Gutenberg-tm electronic work under this agreement, disclaim all liability to you for damages, costs and expenses, including legal fees. YOU AGREE THAT YOU HAVE NO REMEDIES FOR NEGLIGENCE, STRICT LIABILITY, BREACH OF WARRANTY OR BREACH OF CONTRACT EXCEPT THOSE PROVIDED IN PARAGRAPH F3. YOU AGREE THAT THE FOUNDATION, THE TRADEMARK OWNER, AND ANY DISTRIBUTOR UNDER THIS AGREEMENT WILL NOT BE LIABLE TO YOU FOR ACTUAL, DIRECT, INDIRECT, CONSEQUENTIAL, PUNITIVE OR INCIDENTAL DAMAGES EVEN IF YOU GIVE NOTICE OF THE POSSIBILITY OF SUCH DAMAGE.

1.F.3. LIMITED RIGHT OF REPLACEMENT OR REFUND - If you discover a defect in this electronic work within 90 days of receiving it, you can receive a refund of the money (if any) you paid for it by sending a written explanation to the person you received the work from. If you received the work on a physical medium, you must return the medium with your written explanation. The person or entity that provided you with the defective work may elect to provide a replacement copy in lieu of a refund. If you received the work electronically, the person or entity providing it to you may choose to give you a second opportunity to receive the work electronically in lieu of a refund. If the second copy is also defective, you may demand a refund in writing without further opportunities to fix the problem.

1.F.4. Except for the limited right of replacement or refund set forth in paragraph 1.F.3, this work is provided to you 'AS-IS' WITH NO OTHER WARRANTIES OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO

WARRANTIES OF MERCHANTIBILITY OR FITNESS FOR ANY PURPOSE.

1.F.5. Some states do not allow disclaimers of certain implied warranties or the exclusion or limitation of certain types of damages. If any disclaimer or limitation set forth in this agreement violates the law of the state applicable to this agreement, the agreement shall be interpreted to make the maximum disclaimer or limitation permitted by the applicable state law. The invalidity or unenforceability of any provision of this agreement shall not void the remaining provisions.

1.F.6. INDEMNITY - You agree to indemnify and hold the Foundation, the trademark owner, any agent or employee of the Foundation, anyone providing copies of Project Gutenberg-tm electronic works in accordance with this agreement, and any volunteers associated with the production, promotion and distribution of Project Gutenberg-tm electronic works, harmless from all liability, costs and expenses, including legal fees, that arise directly or indirectly from any of the following which you do or cause to occur: (a) distribution of this or any Project Gutenberg-tm work, (b) alteration, modification, or additions or deletions to any Project Gutenberg-tm work, and (c) any Defect you cause.

Section 2. Information about the Mission of Project Gutenberg-tm

Project Gutenberg-tm is synonymous with the free distribution of electronic works in formats readable by the widest variety of computers including obsolete, old, middle-aged and new computers. It exists because of the efforts of hundreds of volunteers and donations from people in all walks of life.

Volunteers and financial support to provide volunteers with the assistance they need, is critical to reaching Project Gutenberg-tm's goals and ensuring that the Project Gutenberg-tm collection will remain freely available for generations to come. In 2001, the Project Gutenberg Literary Archive Foundation was created to provide a secure and permanent future for Project Gutenberg-tm and future generations. To learn more about the Project Gutenberg Literary Archive Foundation and how your efforts and donations can help, see Sections 3 and 4 and the Foundation web page at http://www.pglaf.org.

Section 3. Information about the Project Gutenberg Literary Archive Foundation

The Project Gutenberg Literary Archive Foundation is a non profit 501(c)(3) educational corporation organized under the laws of the state of Mississippi and granted tax exempt status by the Internal Revenue Service. The Foundation's EIN or federal tax identification number is 64-6221541. Its 501(c)(3) letter is posted at http://pglaf.org/fundraising. Contributions to the Project Gutenberg Literary Archive Foundation are tax deductible to the full extent permitted by U.S. federal laws and your state's laws.

The Foundation's principal office is located at 4557 Melan Dr. S. Fairbanks, AK, 99712., but its volunteers and employees are scattered throughout numerous locations. Its business office is located at 809 North 1500 West, Salt Lake City, UT 84116, (801) 596-1887, email business@pglaf.org. Email contact links and up to date contact information can be found at the Foundation's web site and official page at http://pglaf.org

For additional contact information:
Dr. Gregory B. Newby
Chief Executive and Director
gbnewby@pglaf.org

Section 4. Information about Donations to the Project Gutenberg Literary Archive Foundation

Project Gutenberg-tm depends upon and cannot survive without wide spread public support and donations to carry out its mission of

increasing the number of public domain and licensed works that can be
freely distributed in machine readable form accessible by the widest
array of equipment including outdated equipment. Many small donations
($1 to $5,000) are particularly important to maintaining tax exempt
status with the IRS.

The Foundation is committed to complying with the laws regulating
charities and charitable donations in all 50 states of the United
States. Compliance requirements are not uniform and it takes a
considerable effort, much paperwork and many fees to meet and keep up
with these requirements. We do not solicit donations in locations
where we have not received written confirmation of compliance. To
SEND DONATIONS or determine the status of compliance for any
particular state visit http://pglaf.org

While we cannot and do not solicit contributions from states where we
have not met the solicitation requirements, we know of no prohibition
against accepting unsolicited donations from donors in such states who
approach us with offers to donate.

International donations are gratefully accepted, but we cannot make
any statements concerning tax treatment of donations received from
outside the United States. U.S. laws alone swamp our small staff.

Please check the Project Gutenberg Web pages for current donation
methods and addresses. Donations are accepted in a number of other
ways including checks, online payments and credit card donations.
To donate, please visit: http://pglaf.org/donate

Section 5. General Information About Project Gutenberg-tm electronic
works.

Professor Michael S. Hart is the originator of the Project Gutenberg-tm
concept of a library of electronic works that could be freely shared
with anyone. For thirty years, he produced and distributed Project
Gutenberg-tm eBooks with only a loose network of volunteer support.

Project Gutenberg-tm eBooks are often created from several printed
editions, all of which are confirmed as Public Domain in the U.S.
unless a copyright notice is included. Thus, we do not necessarily
keep eBooks in compliance with any particular paper edition.

Most people start at our Web site which has the main PG search facility:

     http://www.gutenberg.org

This Web site includes information about Project Gutenberg-tm,
including how to make donations to the Project Gutenberg Literary
Archive Foundation, how to help produce our new eBooks, and how to
subscribe to our email newsletter to hear about new eBooks.

In [3]:

```
# tamaño del texto en letras
print(len(text))
```

1176967

In [4]:

```
palabras_auxiliares = set(stopwords.words('english')) # se crea una variable que contenga
 todas las palabras auxiliares
print(palabras_auxiliares) # mostramos las palabras auxiliares
```

{"wasn't", 'after', 'do', "she's", 'hadn', 'most', 'it', 'not', 'about', 'ourselves', 'ot
her', 'who', 'myself', 'few', 'too', 'very', 'and', 'mightn', 'does', 'own', 'be', 'itsel
f', 'hers', 'here', "it's", 'from', "that'll", 'am', 't', "you're", "don't", 'we', 'until
', 'mustn', "shouldn't", 'some', 'hasn', 'doing', 'her', 'yourself', 'wouldn', 'at', 'bot

```
h', 'the', 'didn', 'what', 'below', 'he', "hasn't", 'before', 'over', 'yours', "couldn't"
, "shan't", 'did', 'between', 'against', "weren't", 've', 'how', "doesn't", 'all', 'once'
, 'won', 'off', 'don', 'them', 'those', "didn't", 'me', 'i', 'just', "should've", 'same',
'these', 'so', "aren't", 'its', 'than', 'wasn', 'having', 'again', 'our', "you've", 'any'
, 'ours', 'she', 'where', 'theirs', 'themselves', 'are', 'were', 'isn', 'as', 'of', 'd',
'in', 'ma', 'had', 'should', 'a', "hadn't", "you'll", 'into', 'his', 'now', 'has', 'durin
g', "haven't", 'll', 'an', 'when', 'is', 'o', 'him', 'under', 'herself', 'to', 's', 'bein
g', 'couldn', 'with', 'further', 'but', 'no', 'haven', 'will', 'shouldn', 'your', "you'd"
, 'have', 'my', 'whom', 'yourselves', 'there', 'such', "won't", 'himself', 'because', "ne
edn't", 're', 'for', 'was', 'or', 'this', 'down', 'been', 'you', 'while', 'then', 'why',
'their', 'y', 'more', 'by', 'only', 'each', "mustn't", 'weren', 'nor', 'that', 'can', 'ne
edn', 'doesn', "isn't", 'above', 'which', 'they', 'on', 'aren', 'through', 'if', 'out', '
up', "wouldn't", "mightn't", 'shan', 'm', 'ain'}
```

In [5]:

```
#  \w+ acepta cualquier caracter que se encuentre en el rango de [a-zA-Z0-9_]
texto_en_tokens = nltk.regexp_tokenize(text, '\w+') # aplicamos una expresión regular que
 haga match con nuestro texto

texto_en_tokens_sin_palabras_auxiliares = [palabra.lower() for palabra in texto_en_token
s] # transformar palabras a minúsculas para que hagan match con las palabras auxiliares
texto_en_tokens_sin_palabras_auxiliares = [palabra for palabra in texto_en_tokens_sin_pal
abras_auxiliares if not palabra in palabras_auxiliares] # quitamos las palabras auxiliare
s mediante un bucle que las busque en el texto

print(f'Longitud del texto en tokens: {len(texto_en_tokens)}') # cantidad de palabras en
el texto
print(f'Longitud del texto en tokens sin palabras auxiliares: {len(texto_en_tokens_sin_pa
labras_auxiliares)}') # cantidad de palabras en el texto sin palabras auxiliares
```

```
Longitud del texto en tokens: 212001
Longitud del texto en tokens sin palabras auxiliares: 94137
```

In [6]:

```
# realizamos la construcción de un vocabulario
vocabulario = sorted(set(texto_en_tokens)) # set para identificar palabras únicas en el t
exto; sorted se utiliza acomodar por órden alfabético nuestro arreglo
vocabulario_sin_palabras_auxiliares = sorted(set(texto_en_tokens_sin_palabras_auxiliares)
)

print(f'Vocabulario: {len(vocabulario)}') # cantidad de palabras únicas en el texto
print(f'Vocabulario sin palabras auxiliares: {len(vocabulario_sin_palabras_auxiliares)}')
# cantidad de palabras únicas en el texto sin palabras auxiliares
```

```
Vocabulario: 10729
Vocabulario sin palabras auxiliares: 9742
```

In [7]:

```
# definimos una función para obtener la riqueza lexica
def riqueza_lexica(texto):
    vocabulario = sorted(set(texto)) # set para identificar palabras únicas en el texto;
sorted se utiliza acomodar por órden alfabético nuestro arreglo
    return 100 * len(vocabulario) / len(texto) # retorna el valor de la cantidad de palab
ras únicas entre la cantidad de palabras del texto

print(f'Riqueza léxica: {riqueza_lexica(texto_en_tokens)}') # llamar la función
print(f'Riqueza léxica sin palabras auxiliares: {riqueza_lexica(texto_en_tokens_sin_palab
ras_auxiliares)}') # llamar la función
```

```
Riqueza léxica: 5.060825184786864
Riqueza léxica sin palabras auxiliares: 10.348747038890128
```

In [8]:

```
# porcentaje de aparición de una palabra dentro de un texto
def porcentaje_palabra(palabra, texto):
    return 100 * texto.count(palabra) / len(texto) # retorna el valor de las veces que s
e encontró la palabra buscada entre la longitud del texto multiplicado por 100
```

```
palabra = 'reason'
print(f'Porcentaje de la palabra {palabra} en el texto: {porcentaje_palabra(palabra, texto_en_tokens)}') # llamar la función para buscar la palabra "reason" en nuestro texto
```

```
Porcentaje de la palabra reason en el texto: 0.02405649029957406
```

In [9]:

```
# añadiendo la función FreqDist a nuestros tokens, esta función nos permite saber la cantidad de veces de aparición de una palabra en el texto, y mostrar el resultado en tuplas
fdist = FreqDist(texto_en_tokens)
fdist_sin_palabras_auxiliares = FreqDist(texto_en_tokens_sin_palabras_auxiliares)
```

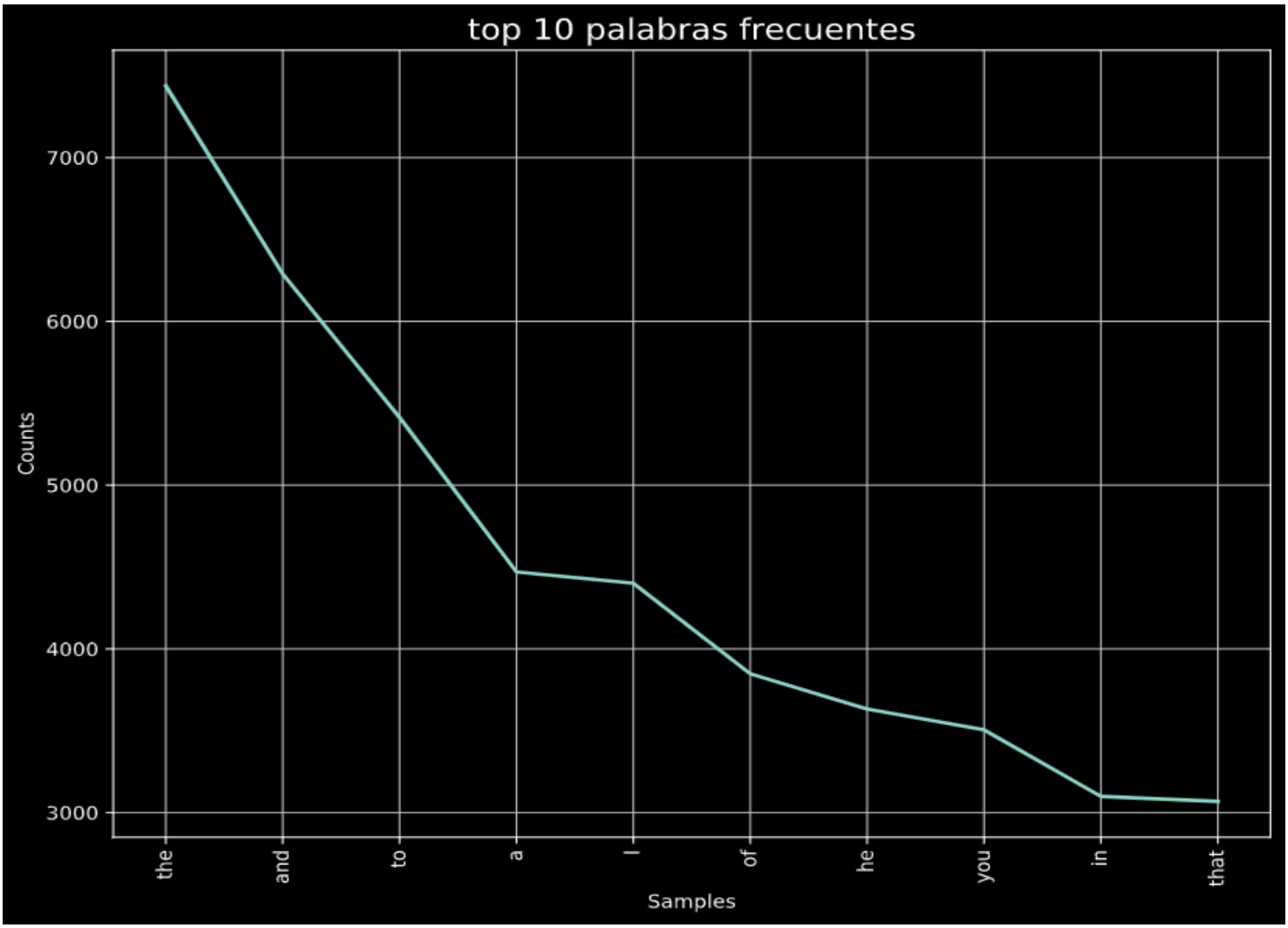
In [10]:

```
# Vocabulario mas comun
print(f'10 palabras mas comunes: {fdist.most_common(10)}')
print(f'10 palabras mas comunes sin palabras auxiliares: {fdist_sin_palabras_auxiliares.most_common(10)}')
```

```
10 palabras mas comunes: [('the', 7437), ('and', 6290), ('to', 5413), ('a', 4470), ('I', 4401), ('of', 3848), ('he', 3633), ('you', 3506), ('in', 3099), ('that', 3068)]
10 palabras mas comunes sin palabras auxiliares: [('raskolnikov', 785), ('one', 645), ('would', 573), ('know', 530), ('said', 519), ('could', 496), ('come', 480), ('man', 479), ('like', 453), ('though', 445)]
```

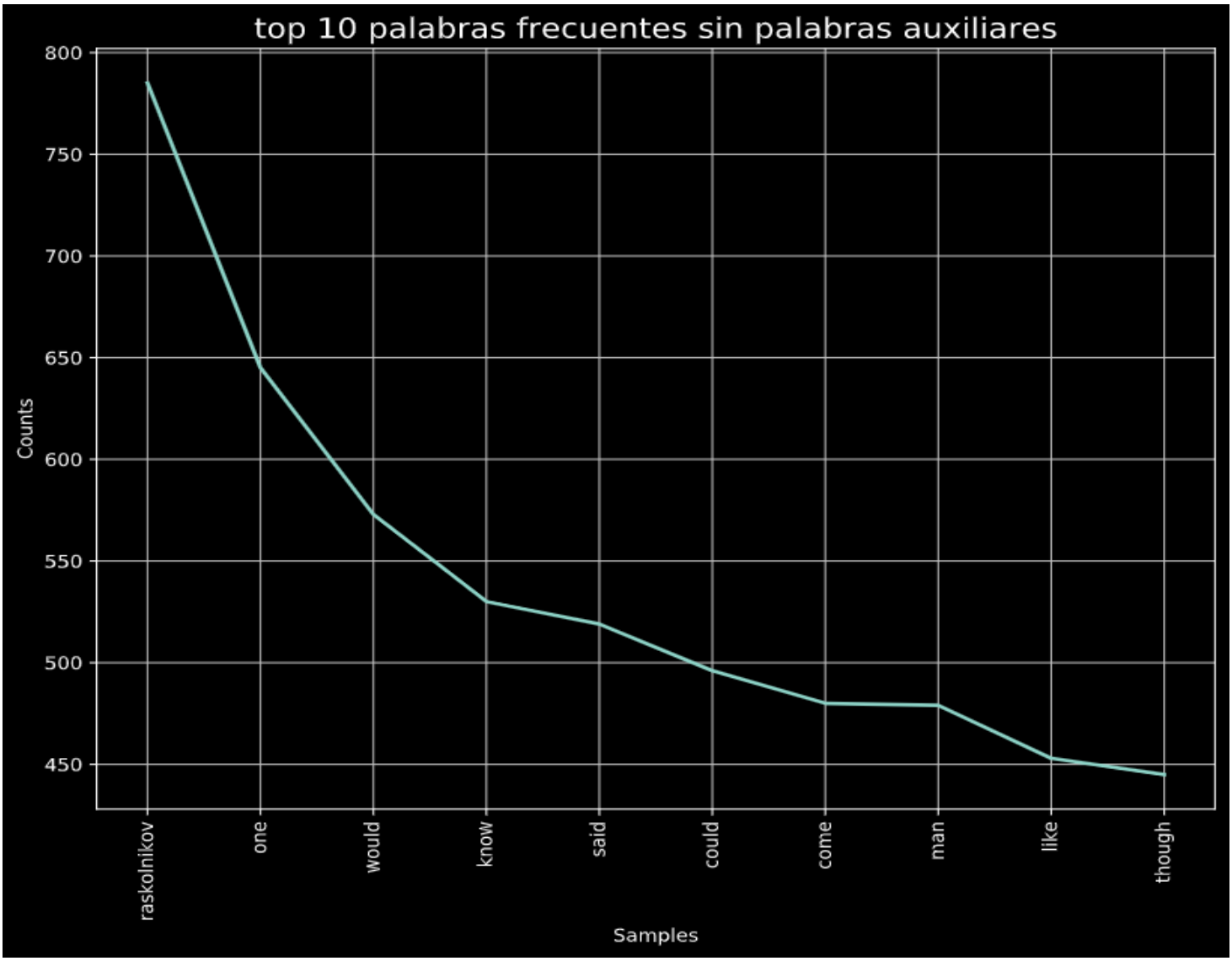
In [11]:

```
# graficando el top 10 de palabras repetidas
plt.style.use('dark_background') # usar fondo negro para nuestro plot
plt.figure(figsize=(10, 8)) # definir el tamaño de nuestra figura que se mostrará
plt.title('top 10 palabras frecuentes', fontsize=16) # título y tamaño de la letra de nuestro plot
fdist.plot(10) # plot de nuestra función
```



In [12]:

```
# graficando el top 10 de palabras repetidas
plt.style.use('dark_background') # usar fondo negro para nuestro plot
plt.figure(figsize=(10, 8)) # definir el tamaño de nuestra figura que se mostrará
plt.title('top 10 palabras frecuentes sin palabras auxiliares', fontsize=16) # título y t
amaño de la letra de nuestro plot
fdist_sin_palabras_auxiliares.plot(10) # plot de nuestra función
```




```
<matplotlib.axes._subplots.AxesSubplot at 0x21c73d7e978>
```