In [1]:

```
import nltk
nltk.download('punkt')
nltk.download('stopwords')
from nltk.tokenize import sent_tokenize
import matplotlib.pyplot as plt
import pandas as pd
from nltk.corpus import stopwords
from nltk.stem.snowball import EnglishStemmer
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/alfonsoreyna/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/alfonsoreyna/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [2]:

```
sample_text = "This was not because he was cowardly and abject, quite the contrary
;; but for some time past he had been in an overstrained irritable condition, verg
ing on hypochondria. He had become so completely absorbed in himself, and isolated
from his fellows that he dreaded meeting, not only his landlady, but anyone at all
. He was crushed by poverty, but the anxieties of his position had of late ceased
to weigh upon him. He had given up attending to matters of practical importance;;
he had lost all desire to do so. Nothing that any landlady could do had a real ter
ror for him. But to be stopped on the stairs, to be forced to listen to her trivia
l, irrelevant gossip, to pestering demands for payment, threats and complaints, an
d to rack his brains for excuses, to prevaricate, to lie no, rather than that, he
would creep down the stairs like a cat and slip out unseen."
```

In [3]:

```
print(sample_text)
```

```
This was not because he was cowardly and abject, quite the contrary;; but for so
me time past he had been in an overstrained irritable condition, verging on hypo
chondria. He had become so completely absorbed in himself, and isolated from his
fellows that he dreaded meeting, not only his landlady, but anyone at all. He wa
s crushed by poverty, but the anxieties of his position had of late ceased to we
igh upon him. He had given up attending to matters of practical importance;; he
had lost all desire to do so. Nothing that any landlady could do had a real terr
or for him. But to be stopped on the stairs, to be forced to listen to her trivi
al, irrelevant gossip, to pestering demands for payment, threats and complaints,
and to rack his brains for excuses, to prevaricate, to lie no, rather than that,
he would creep down the stairs like a cat and slip out unseen.
```

In [4]:

```
# Stopwords
default_stopwords = set(nltk.corpus.stopwords.words('english'))
```

Elimino valores redundantes del texto

In [5]:

```
clean_text = sample_text.replace(',', '').replace(';', '').replace('.', '')
```

Obtengo un arreglo de palabras con el cual puedo solo hacer un count para saber cuantas palabras hay

```
In [6]: word_arr = clean_text.split(' ')
        print(len(word_arr))
        print(word_arr)
```

```
155
['This', 'was', 'not', 'because', 'he', 'was', 'cowardly', 'and', 'abject', 'qui
te', 'the', 'contrary;', 'but', 'for', 'some', 'time', 'past', 'he', 'had', 'bee
n', 'in', 'an', 'overstrained', 'irritable', 'condition', 'verging', 'on', 'hypo
chondria', 'He', 'had', 'become', 'so', 'completely', 'absorbed', 'in', 'himself
', 'and', 'isolated', 'from', 'his', 'fellows', 'that', 'he', 'dreaded', 'meetin
g', 'not', 'only', 'his', 'landlady', 'but', 'anyone', 'at', 'all', 'He', 'was',
'crushed', 'by', 'poverty', 'but', 'the', 'anxieties', 'of', 'his', 'position',
'had', 'of', 'late', 'ceased', 'to', 'weigh', 'upon', 'him', 'He', 'had', 'given
', 'up', 'attending', 'to', 'matters', 'of', 'practical', 'importance;', 'he', '
had', 'lost', 'all', 'desire', 'to', 'do', 'so', 'Nothing', 'that', 'any', 'land
lady', 'could', 'do', 'had', 'a', 'real', 'terror', 'for', 'him', 'But', 'to', '
be', 'stopped', 'on', 'the', 'stairs', 'to', 'be', 'forced', 'to', 'listen', 'to
', 'her', 'trivial', 'irrelevant', 'gossip', 'to', 'pestering', 'demands', 'for'
, 'payment', 'threats', 'and', 'complaints', 'and', 'to', 'rack', 'his', 'brains
', 'for', 'excuses', 'to', 'prevaricate', 'to', 'lie', 'no', 'rather', 'than', '
that', 'he', 'would', 'creep', 'down', 'the', 'stairs', 'like', 'a', 'cat', 'and
', 'slip', 'out', 'unseen']
```

```
In [7]: # Usando la libreria
        words = nltk.word_tokenize(clean_text)
```

Obtengo el un arreglo de palabras utilizando tokenise de nltk

```
In [8]: # Numero de palabas incluidos en el texto
        print(len(words))
```

```
155
```

```
In [9]: print(words)
```

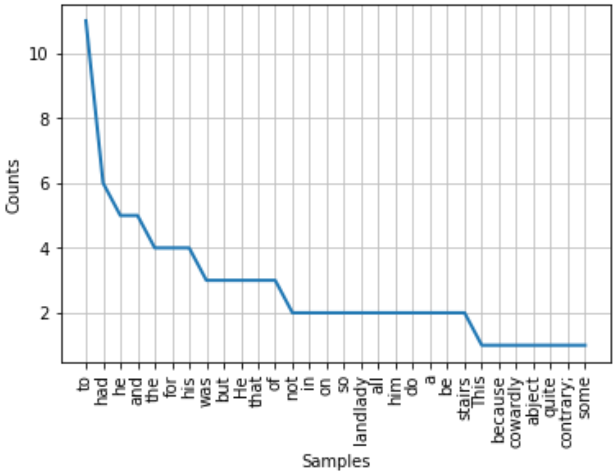
```
['This', 'was', 'not', 'because', 'he', 'was', 'cowardly', 'and', 'abject', 'qui
te', 'the', 'contrary;', 'but', 'for', 'some', 'time', 'past', 'he', 'had', 'bee
n', 'in', 'an', 'overstrained', 'irritable', 'condition', 'verging', 'on', 'hypo
chondria', 'He', 'had', 'become', 'so', 'completely', 'absorbed', 'in', 'himself
', 'and', 'isolated', 'from', 'his', 'fellows', 'that', 'he', 'dreaded', 'meetin
g', 'not', 'only', 'his', 'landlady', 'but', 'anyone', 'at', 'all', 'He', 'was',
'crushed', 'by', 'poverty', 'but', 'the', 'anxieties', 'of', 'his', 'position',
'had', 'of', 'late', 'ceased', 'to', 'weigh', 'upon', 'him', 'He', 'had', 'given
', 'up', 'attending', 'to', 'matters', 'of', 'practical', 'importance;', 'he', '
had', 'lost', 'all', 'desire', 'to', 'do', 'so', 'Nothing', 'that', 'any', 'land
lady', 'could', 'do', 'had', 'a', 'real', 'terror', 'for', 'him', 'But', 'to', '
be', 'stopped', 'on', 'the', 'stairs', 'to', 'be', 'forced', 'to', 'listen', 'to
', 'her', 'trivial', 'irrelevant', 'gossip', 'to', 'pestering', 'demands', 'for'
, 'payment', 'threats', 'and', 'complaints', 'and', 'to', 'rack', 'his', 'brains
', 'for', 'excuses', 'to', 'prevaricate', 'to', 'lie', 'no', 'rather', 'than', '
that', 'he', 'would', 'creep', 'down', 'the', 'stairs', 'like', 'a', 'cat', 'and
', 'slip', 'out', 'unseen']
```

```
In [10]: sentences = sent_tokenize(sample_text)
In [11]: # Numero de oraciónes en el texto
         print(len(sentences))

         6
```

```
In [12]: freq_words_dist = nltk.FreqDist(words)
         freq_words_dist_arr = freq_words_dist.items()
```

```
In [19]: freq_words_dist.most_common(5)

Out[19]: [('to', 11), ('had', 6), ('he', 5), ('and', 5), ('the', 4)]
```

```
In [13]: freq_words_dist.plot(30)
```



```
In [14]: # Quitar los stopwords del lenguaje
         words = [word for word in words if word not in default_stopwords]
```

```
In [15]: print(len(words))

         70
```

In [18]:

```python
# Iterando palabras para obtener su versión normalizada
est = EnglishStemmer()

normalized_words = []

for word in words:
    normalized_words.append(est.stem(word))

# Validar palabras unicas
normalized_words = list(normalized_words)
print(normalized_words)
print(len(normalized_words))
```

```
['this', 'coward', 'abject', 'quit', 'contrary;', 'time', 'past', 'overstrain',
'irrit', 'condit', 'verg', 'hypochondria', 'he', 'becom', 'complet', 'absorb', '
isol', 'fellow', 'dread', 'meet', 'landladi', 'anyon', 'he', 'crush', 'poverti',
'anxieti', 'posit', 'late', 'ceas', 'weigh', 'upon', 'he', 'given', 'attend', 'm
atter', 'practic', 'importance;', 'lost', 'desir', 'noth', 'landladi', 'could',
'real', 'terror', 'but', 'stop', 'stair', 'forc', 'listen', 'trivial', 'irrelev'
, 'gossip', 'pester', 'demand', 'payment', 'threat', 'complaint', 'rack', 'brain
', 'excus', 'prevar', 'lie', 'rather', 'would', 'creep', 'stair', 'like', 'cat',
'slip', 'unseen']
70
```

In [ ]: