In [1]:
```
text_file = open("sentences.txt")
text = text_file.read()
```

In [2]:
```
print(type(text))
```

```
<class 'str'>
```

In [3]:
```
print(text)
print("\n")
```

```
QVC Network Inc. said it completed its acquisition of CVN Cos. for about $ 423 mil
lion .
The spirits , of course , could hardly care less whether people do or do n't belie
ve in them .
The debt ceiling is scheduled to fall to $ 2.8 trillion from $ 2.87 trillion at mi
dnight tonight .
```

In [4]:
```
print(len(text))
```

```
283
```

In [5]:
```
import nltk
from nltk import sent_tokenize
from nltk import word_tokenize
```

In [6]:
```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\admine\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[6]:
```
True
```

In [7]:
```
sentences = sent_tokenize(text)
print(len(sentences))
```

```
3
```

In [8]:
```
words = word_tokenize(text)
print(len(words))
print(words)
```

```
56
['QVC', 'Network', 'Inc.', 'said', 'it', 'completed', 'its', 'acquisition', 'of',
'CVN', 'Cos.', 'for', 'about', '$', '423', 'million', '.', 'The', 'spirits', ',',
'of', 'course', ',', 'could', 'hardly', 'care', 'less', 'whether', 'people', 'do',
'or', 'do', "n't", 'believe', 'in', 'them', '.', 'The', 'debt', 'ceiling', 'is',
'scheduled', 'to', 'fall', 'to', '$', '2.8', 'trillion', 'from', '$', '2.87', 'tri
llion', 'at', 'midnight', 'tonight', '.']
```

In [9]:
```
from nltk.probability import FreqDist
fdist = FreqDist(words)
fdist.most_common(10)
```

Out[9]:
```
[('$', 3),
 ('.', 3),
 ('of', 2),
 ('The', 2),
 (',', 2),
 ('do', 2),
 ('to', 2),
 ('trillion', 2),
 ('QVC', 1),
 ('Network', 1)]
```

In [10]:
```
from nltk.corpus import stopwords
nltk.download('stopwords')
stopwords = stopwords.words('english')
print(stopwords)
```

```
['a', 'about', 'above', 'after', 'again', 'against', 'ain', 'all', 'am', 'an', 'an
d', 'any', 'are', 'aren', "aren't", 'as', 'at', 'be', 'because', 'been', 'before',
'being', 'below', 'between', 'both', 'but', 'by', 'can', 'couldn', "couldn't",
'd', 'did', 'didn', "didn't", 'do', 'does', 'doesn', "doesn't", 'doing', 'don', "d
on't", 'down', 'during', 'each', 'few', 'for', 'from', 'further', 'had', 'hadn',
"hadn't", 'has', 'hasn', "hasn't", 'have', 'haven', "haven't", 'having', 'he', "h
e'd", "he'll", 'her', 'here', 'hers', 'herself', "he's", 'him', 'himself', 'his',
'how', 'i', "i'd", 'if', "i'll", "i'm", 'in', 'into', 'is', 'isn', "isn't", 'it',
"it'd", "it'll", "it's", 'its', 'itself', "i've", 'just', 'll', 'm', 'ma', 'me',
'mightn', "mightn't", 'more', 'most', 'mustn', "mustn't", 'my', 'myself', 'needn',
"needn't", 'no', 'nor', 'not', 'now', 'o', 'of', 'off', 'on', 'once', 'only', 'o
r', 'other', 'our', 'ours', 'ourselves', 'out', 'over', 'own', 're', 's', 'same',
'shan', "shan't", 'she', "she'd", "she'll", "she's", 'should', 'shouldn', "should
n't", "should've", 'so', 'some', 'such', 't', 'than', 'that', "that'll", 'the', 't
heir', 'theirs', 'them', 'themselves', 'then', 'there', 'these', 'they', "they'd",
"they'll", "they're", "they've", 'this', 'those', 'through', 'to', 'too', 'under',
'until', 'up', 've', 'very', 'was', 'wasn', "wasn't", 'we', "we'd", "we'll", "we'r
e", 'were', 'weren', "weren't", "we've", 'what', 'when', 'where', 'which', 'whil
e', 'who', 'whom', 'why', 'will', 'with', won', "won't", 'wouldn', "wouldn't",
'y', 'you', "you'd", "you'll", 'your', "you're", 'yours', 'yourself', 'yourselve
s', "you've"]
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\admine\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [11]:
```
import string
```

In [12]:
```
words_no_punc = [w for w in words if w not in string.punctuation]
```

In [13]:
```
clean_words = []

for w in words_no_punc:
    if w not in stopwords:
        clean_words.append(w)

print(clean_words)
print("\n")
print(len(clean_words))
```

```
['QVC', 'Network', 'Inc.', 'said', 'completed', 'acquisition', 'CVN', 'Cos.', '42
3', 'million', 'The', 'spirits', 'course', 'could', 'hardly', 'care', 'less', 'whe
ther', 'people', "n't", 'believe', 'The', 'debt', 'ceiling', 'scheduled', 'fall',
'2.8', 'trillion', '2.87', 'trillion', 'midnight', 'tonight']


32
```

In [14]:
```
fdist = FreqDist(clean_words)
fdist.most_common(10)
```

Out[14]:
```
[('The', 2),
 ('trillion', 2),
 ('QVC', 1),
 ('Network', 1),
 ('Inc.', 1),
 ('said', 1),
 ('completed', 1),
 ('acquisition', 1),
 ('CVN', 1),
 ('Cos.', 1)]
```