In [29]:
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import re
import string
import  nltk
from nltk import sent_tokenize
from nltk.tokenize import word_tokenize

from nltk.corpus import stopwords
nltk.download('stopwords')

from nltk.stem.porter import PorterStemmer
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/wsyed2/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [30]:
```
porter = PorterStemmer()
stemmed = [porter.stem(word) for word in tokens]
print(stemmed[:100])
```

```
['/', ':', ';', '<', '=', '@', 'one', 'morn', ',', 'when', 'gregor', 'samsa',
'woke', 'from', 'troubl', 'dream', ',', 'he', 'found', 'himself', 'transform',
'in', 'hi', 'bed', 'into', 'a', 'horribl', 'vermin', '.', 'he', 'lay', 'on',
'hi', 'armour-lik', 'back', ',', 'and', 'if', 'he', 'lift', 'hi', 'head', 'a',
'littl', 'he', 'could', 'see', 'hi', 'brown', 'belli', ',', 'slightli', 'dom
e', 'and', 'divid', 'by', 'arch', 'into', 'stiff', 'section', '.', 'the', 'be
d', 'wa', 'hardli', 'abl', 'to', 'cover', 'it', 'and', 'seem', 'readi', 'to',
'slide', 'off', 'ani', 'moment', '.', 'hi', 'mani', 'leg', ',', 'piti', 'thi
n', 'compar', 'with', 'the', 'size', 'of', 'the', 'rest', 'of', 'him', ',', 'w
ave', 'about', 'helplessli', 'as', 'he', 'look']
```

In [11]:
```
words = re.split(r'\W+', text)
print(words[:100])
```

```
['', 'One', 'morning', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled',
'dreams', 'he', 'found', 'himself', 'transformed', 'in', 'his', 'bed', 'into',
'a', 'horrible', 'vermin', 'He', 'lay', 'on', 'his', 'armour', 'like', 'back',
'and', 'if', 'he', 'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'se
e', 'his', 'brown', 'belly', 'slightly', 'domed', 'and', 'divided', 'by', 'arc
hes', 'into', 'stiff', 'sections', 'The', 'bedding', 'was', 'hardly', 'able',
'to', 'cover', 'it', 'and', 'seemed', 'ready', 'to', 'slide', 'off', 'any', 'm
oment', 'His', 'many', 'legs', 'pitifully', 'thin', 'compared', 'with', 'the',
'size', 'of', 'the', 'rest', 'of', 'him', 'waved', 'about', 'helplessly', 'a
s', 'he', 'looked', 'What', 's', 'happened', 'to', 'me', 'he', 'thought', 'I
t', 'wasn', 't', 'a', 'dream', 'His']
```

In [14]:
```
table = str.maketrans('', '', string.punctuation)
stripped = [w.translate(table) for w in words]
print(stripped[:100])
```

```
['', 'One', 'morning', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled',
'dreams', 'he', 'found', 'himself', 'transformed', 'in', 'his', 'bed', 'into',
'a', 'horrible', 'vermin', 'He', 'lay', 'on', 'his', 'armour', 'like', 'back',
'and', 'if', 'he', 'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'se
e', 'his', 'brown', 'belly', 'slightly', 'domed', 'and', 'divided', 'by', 'arc
hes', 'into', 'stiff', 'sections', 'The', 'bedding', 'was', 'hardly', 'able',
'to', 'cover', 'it', 'and', 'seemed', 'ready', 'to', 'slide', 'off', 'any', 'm
oment', 'His', 'many', 'legs', 'pitifully', 'thin', 'compared', 'with', 'the',
'size', 'of', 'the', 'rest', 'of', 'him', 'waved', 'about', 'helplessly', 'a
s', 'he', 'looked', 'What', 's', 'happened', 'to', 'me', 'he', 'thought', 'I
t', 'wasn', 't', 'a', 'dream', 'His']
```

In [20]:

```python
sentences = sent_tokenize(text)
print(sentences[0])
```

```
/:;<=@
One morning, when Gregor Samsa woke from troubled dreams, he found
himself transformed in his bed into a horrible vermin.
```

In [22]:

```python
tokens = word_tokenize(text)
print(tokens[:100])
```

```
['/', ':', ';', '<', '=', '@', 'One', 'morning', ',', 'when', 'Gregor', 'Sams
a', 'woke', 'from', 'troubled', 'dreams', ',', 'he', 'found', 'himself', 'tran
sformed', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin', '.', 'He', 'l
ay', 'on', 'his', 'armour-like', 'back', ',', 'and', 'if', 'he', 'lifted', 'hi
s', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly', ',',
'slightly', 'domed', 'and', 'divided', 'by', 'arches', 'into', 'stiff', 'secti
ons', '.', 'The', 'bedding', 'was', 'hardly', 'able', 'to', 'cover', 'it', 'an
d', 'seemed', 'ready', 'to', 'slide', 'off', 'any', 'moment', '.', 'His', 'man
y', 'legs', ',', 'pitifully', 'thin', 'compared', 'with', 'the', 'size', 'of',
'the', 'rest', 'of', 'him', ',', 'waved', 'about', 'helplessly', 'as', 'he',
'looked']
```

In [33]:

```python
# remove all tokens that are not alphabetic
words = [word for word in tokens if word.isalpha()]
print(words[:100])
```

```
['One', 'morning', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dre
ams', 'he', 'found', 'himself', 'transformed', 'in', 'his', 'bed', 'into',
'a', 'horrible', 'vermin', 'He', 'lay', 'on', 'his', 'back', 'and', 'if', 'h
e', 'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brow
n', 'belly', 'slightly', 'domed', 'and', 'divided', 'by', 'arches', 'into', 's
tiff', 'sections', 'The', 'bedding', 'was', 'hardly', 'able', 'to', 'cover',
'it', 'and', 'seemed', 'ready', 'to', 'slide', 'off', 'any', 'moment', 'His',
'many', 'legs', 'pitifully', 'thin', 'compared', 'with', 'the', 'size', 'of',
'the', 'rest', 'of', 'him', 'waved', 'about', 'helplessly', 'as', 'he', 'looke
d', 'What', 'happened', 'to', 'me', 'he', 'thought', 'It', 'was', 'a', 'drea
m', 'His', 'room', 'a', 'proper', 'human', 'room']
```

In [34]:

```python
# Filter out Stop Words (and Pipeline)
stop_words = stopwords.words('english')
print(stop_words)
```

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

In [35]:

```python
words = [w for w in words if not w in stop_words]
print(words[:100])
```

```
['One', 'morning', 'Gregor', 'Samsa', 'woke', 'troubled', 'dreams', 'found',
'transformed', 'bed', 'horrible', 'vermin', 'He', 'lay', 'back', 'lifted', 'he
ad', 'little', 'could', 'see', 'brown', 'belly', 'slightly', 'domed', 'divide
d', 'arches', 'stiff', 'sections', 'The', 'bedding', 'hardly', 'able', 'cove
r', 'seemed', 'ready', 'slide', 'moment', 'His', 'many', 'legs', 'pitifully',
'thin', 'compared', 'size', 'rest', 'waved', 'helplessly', 'looked', 'What',
'happened', 'thought', 'It', 'dream', 'His', 'room', 'proper', 'human', 'roo
m', 'although', 'little', 'small', 'lay', 'peacefully', 'four', 'familiar', 'w
alls', 'A', 'collection', 'textile', 'samples', 'lay', 'spread', 'table', 'Sam
sa', 'travelling', 'salesman', 'hung', 'picture', 'recently', 'cut', 'illustra
ted', 'magazine', 'housed', 'nice', 'gilded', 'frame', 'It', 'showed', 'lady',
'fitted', 'fur', 'hat', 'fur', 'boa', 'sat', 'upright', 'raising', 'heavy', 'f
ur', 'muff']
```

In [37]:

```python
porter = PorterStemmer()
stemmed = [porter.stem(word) for word in tokens]
print(stemmed[:100])
```

```
['/', ':', ';', '<', '=', '@', 'one', 'morn', ',', 'when', 'gregor', 'samsa',
'woke', 'from', 'troubl', 'dream', ',', 'he', 'found', 'himself', 'transform',
'in', 'hi', 'bed', 'into', 'a', 'horribl', 'vermin', '.', 'he', 'lay', 'on',
'hi', 'armour-lik', 'back', ',', 'and', 'if', 'he', 'lift', 'hi', 'head', 'a',
'littl', 'he', 'could', 'see', 'hi', 'brown', 'belli', ',', 'slightli', 'dom
e', 'and', 'divid', 'by', 'arch', 'into', 'stiff', 'section', '.', 'the', 'be
d', 'wa', 'hardli', 'abl', 'to', 'cover', 'it', 'and', 'seem', 'readi', 'to',
'slide', 'off', 'ani', 'moment', '.', 'hi', 'mani', 'leg', ',', 'piti', 'thi
n', 'compar', 'with', 'the', 'size', 'of', 'the', 'rest', 'of', 'him', ',', 'w
ave', 'about', 'helplessli', 'as', 'he', 'look']
```

In [ ]:

```python
```