Downloads/assignment11/ | Untitled - Jupyter Notebook | Untitled - Jupyter Notebook | School of Continuing and Distan

localhost:8888/notebooks/Downloads/assignment11/Untitled.ipynb?kernel_name=python3

jupyter Untitled Last Checkpoint: 15 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code

In [11]:
```
import re,string
filename='novel (1).txt'
file=open(filename,'rt',encoding='utf-8')
text=file.read()
file.close()
```

In [12]:
```
words=text.split()
print(words[:120])
```

```
['/:;<=@', 'One', 'morning,', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams,', 'he', 'found', 'himself', 'tran
sformed', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin.', 'He', 'lay', 'on', 'his', 'armour-like', 'back,', 'and', 'i
f', 'he', 'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly,', 'slightly', 'domed', 'and',
'divided', 'by', 'arches', 'into', 'stiff', 'sections.', 'The', 'bedding', 'was', 'hardly', 'able', 'to', 'cover', 'it', 'and',
'seemed', 'ready', 'to', 'slide', 'off', 'any', 'moment.', 'His', 'many', 'legs,', 'pitifully', 'thin', 'compared', 'with', 'th
e', 'size', 'of', 'the', 'rest', 'of', 'him,', 'waved', 'about', 'helplessly', 'as', 'he', 'looked.', '"What\'s', 'happened',
'to', 'me?"', 'he', 'thought.', 'It', "wasn't", 'a', 'dream.', 'His', 'room,', 'a', 'proper', 'human', 'room', 'although', 'a',
'little', 'too', 'small,', 'lay', 'peacefully', 'between', 'its', 'four', 'familiar', 'walls.', 'A', 'collection', 'of', 'textil
e', 'samples', 'lay']
```

In [13]:
```
#splitting the words at non word character
words=re.split(r'\W+',text)
print(words[:120])
```

```
['', 'One', 'morning', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams', 'he', 'found', 'himself', 'transforme
d', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin', 'He', 'lay', 'on', 'his', 'armour', 'like', 'back', 'and', 'if', 'h
e', 'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly', 'slightly', 'domed', 'and', 'divide
d', 'by', 'arches', 'into', 'stiff', 'sections', 'The', 'bedding', 'was', 'hardly', 'able', 'to', 'cover', 'it', 'and', 'seeme
d', 'ready', 'to', 'slide', 'off', 'any', 'moment', 'His', 'many', 'legs', 'pitifully', 'thin', 'compared', 'with', 'the', 'siz
e', 'of', 'the', 'rest', 'of', 'him', 'waved', 'about', 'helplessly', 'as', 'he', 'looked', 'What', 's', 'happened', 'to', 'm
```

30° Search ENG IN 7:52 PM 3/27/2023

localhost:8888/notebooks/Downloads/assignment11/Untitled.ipynb?kernel_name=python3

jupyter Untitled Last Checkpoint: 16 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted Python 3 (ipykernel)

Run Code

In [15]:
```
#split the words and remove the punctuations
words=text.split()
re_punc=re.compile('[%s]' % re.escape(string.punctuation))
stripped=[re_punc.sub('',word)for word in words]
print(stripped[:120])
```

```
['', 'One', 'morning', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams', 'he', 'found', 'himself', 'transforme
d', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin', 'He', 'lay', 'on', 'his', 'armourlike', 'back', 'and', 'if', 'he',
'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly', 'slightly', 'domed', 'and', 'divided',
'by', 'arches', 'into', 'stiff', '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', 'looked', 'Whats', 'happened', 'to', 'me', 'he',
'thought', 'It', 'wasnt', 'a', 'dream', 'His', 'room', 'a', 'proper', 'human', 'room', 'although', 'a', 'little', 'too', 'smal
l', 'lay', 'peacefully', 'between', 'its', 'four', 'familiar', 'walls', 'A', 'collection', 'of', 'textile', 'samples', 'lay']
```

In [18]:
```
#some times text contains the characters that cannot be printed . so filtering is to be done
re_print=re.compile('[^%s]' %re.escape(string.printable))
result=[re_print.sub('',word) for word in stripped]
print(result[:120])
```

```
['', 'One', 'morning', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams', 'he', 'found', 'himself', 'transforme
d', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin', 'He', 'lay', 'on', 'his', 'armourlike', 'back', 'and', 'if', 'he',
'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly', 'slightly', 'domed', 'and', 'divided',
'by', 'arches', 'into', 'stiff', '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', 'looked', 'Whats', 'happened', 'to', 'me', 'he',
'thought', 'It', 'wasnt', 'a', 'dream', 'His', 'room', 'a', 'proper', 'human', 'room', 'although', 'a', 'little', 'too', 'smal
l', 'lay', 'peacefully', 'between', 'its', 'four', 'familiar', 'walls', 'A', 'collection', 'of', 'textile', 'samples', 'lay']
```

In [19]:
```
#conversion of entire text into lowercase
result=[word.lower() for word in result]
print(result[:120])
```

Downloads/assignment11/ | Untitled - Jupyter Notebook | Untitled - Jupyter Notebook | School of Continuing and Distan

localhost:8888/notebooks/Downloads/assignment11/Untitled.ipynb?kernel_name=python3

jupyter Untitled Last Checkpoint: 16 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code

In [19]:
```
#conversion of entire text into lowercase
result=[word.lower() for word in result]
print(result[:120])
```

```
['', 'one', 'morning', 'when', 'gregor', 'samsa', 'woke', 'from', 'troubled', 'dreams', 'he', 'found', 'himself', 'transforme
d', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin', 'he', 'lay', 'on', 'his', 'armourlike', 'back', 'and', 'if', 'he',
'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly', 'slightly', 'domed', 'and', 'divided',
'by', 'arches', 'into', 'stiff', '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', 'looked', 'whats', 'happened', 'to', 'me', 'he',
'thought', 'it', 'wasnt', 'a', 'dream', 'his', 'room', 'a', 'proper', 'human', 'room', 'although', 'a', 'little', 'too', 'smal
l', 'lay', 'peacefully', 'between', 'its', 'four', 'familiar', 'walls', 'a', 'collection', 'of', 'textile', 'samples', 'lay']
```

In [20]:
```
#1 character word won't contribute to NLP tasks so removing them
result=[word for word in result if len(word) >1]
print(result[:120])
```

```
['one', 'morning', 'when', 'gregor', 'samsa', 'woke', 'from', 'troubled', 'dreams', 'he', 'found', 'himself', 'transformed', 'i
n', 'his', 'bed', 'into', 'horrible', 'vermin', 'he', 'lay', 'on', 'his', 'armourlike', 'back', 'and', 'if', 'he', 'lifted', 'h
is', 'head', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly', 'slightly', 'domed', 'and', 'divided', 'by', 'arches', 'i
nto', 'stiff', 'sections', 'the', 'bedding', 'was', 'hardly', 'able', 'to', 'cover', 'it', 'and', 'seemed', 'ready', 'to', 'sli
de', 'off', 'any', 'moment', 'his', 'many', 'legs', 'pitifully', 'thin', 'compared', 'with', 'the', 'size', 'of', 'the', 'res
t', 'of', 'him', 'waved', 'about', 'helplessly', 'as', 'he', 'looked', 'whats', 'happened', 'to', 'me', 'he', 'thought', 'it',
'wasnt', 'dream', 'his', 'room', 'proper', 'human', 'room', 'although', 'little', 'too', 'small', 'lay', 'peacefully', 'betwee
n', 'its', 'four', 'familiar', 'walls', 'collection', 'of', 'textile', 'samples', 'lay', 'spread', 'out', 'on', 'the', 'table',
'samsa', 'was']
```

In [ ]:

30° Search ENG IN 7:53 PM 3/27/2023