## 1. Write a Python code using NLP to Pre-Process the text data and convert Text-Numeric vectors.

I. Use Tokenization, Stopword removal, Stemming/Lemmatization , text preprocess logic using NLTK II. Use SKLearn for converting Text-Numeric vectors using TF-IDF model consider novel.txt as text document for implementing question 1.

In [1]:
```
import numpy as np
import pandas as pd
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import TfidfVectorizer
```

# Loading the Data to Read

In [2]:
```
with open(r"E:\!DataScience_DSPP_JNTU\Assignments\NLP\novel.txt", 'r') as f:
    text = f.read(1000)
    f.close()
print(text)
```

```
/:;<=@
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
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.

"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 textile samples
lay spread out on the table - Samsa was a travelling salesman - and
above it there hung a picture that he had recently cut out of an
illustrated magazine and housed in a nice, gilded frame.  It showed
a lady fitted out with a fur hat and fur boa who sat upright,
raising a heavy fur muff that covered the whole of her lowe
```

# Split by Whitespace

In [3]:
```
# split into words by white space
words = text.split()
print(words[:150])
```

```
['/:;<=@', 'One', 'morning,', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams,', 'he', 'found', 'himself', 'trans
formed', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin.', 'He', 'lay', 'on', 'his', 'armour-like', 'back,', 'and', 'if',
'he', 'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly,', 'slightly', 'domed', 'and', 'divid
ed', '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', 'si
ze', 'of', 'the', 'rest', 'of', 'him,', 'waved', 'about', 'helplessly', 'as', 'he', 'looked.', '"What\'s', 'happened', 'to', 'm
e?"', 'he', 'thought.', 'It', "wasn't", 'a', 'dream.', 'His', 'room,', 'a', 'proper', 'human', 'room', 'although', 'a', 'littl
e', 'too', 'small,', 'lay', 'peacefully', 'between', 'its', 'four', 'familiar', 'walls.', 'A', 'collection', 'of', 'textile', 's
amples', 'lay', 'spread', 'out', 'on', 'the', 'table', '-', 'Samsa', 'was', 'a', 'travelling', 'salesman', '-', 'and', 'above',
'it', 'there', 'hung', 'a', 'picture', 'that', 'he', 'had', 'recently', 'cut', 'out', 'of', 'an', 'illustrated', 'magazine', 'an
d']
```

# Selecting Words

In [4]:
```
# split based on words only
import re
words = re.split(r'\W+', text)
print(words[:150])
```

```
['', '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', 'l
ifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly', 'slightly', 'domed', 'and', 'divided', 'b
y', 'arches', 'into', 'stiff', 'sections', 'The', 'bedding', 'was', 'hardly', 'able', 'to', 'cover', 'it', 'and', 'seemed', 'rea
dy', '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', 'What', 's', 'happened', 'to', 'me', 'he', 'th
ought', 'It', 'wasn', 't', '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', 's
pread', 'out', 'on', 'the', 'table', 'Samsa', 'was', 'a', 'travelling', 'salesman', 'and', 'above', 'it', 'there', 'hung', 'a',
'picture', 'that', 'he', 'had', 'recently', 'cut', 'out', 'of', 'an', 'illustrated', 'magazine']
```

# Split by Whitespace and Remove Punctuation

In [5]:
```
# split into words by white space
words = text.split()
# remove punctuation from each word
import string
table = str.maketrans('', '',string.punctuation)
stripped = [w.translate(table) for w in words]
print(stripped[:150])
```

```
['', 'One', 'morning', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams', 'he', 'found', 'himself', 'transformed',
'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin', 'He', 'lay', 'on', 'his', 'armourlike', 'back', 'and', 'if', 'he', 'lifte
d', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly', 'slightly', 'domed', 'and', 'divided', 'by', 'a
rches', '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', 'th
e', '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', 'small', 'lay', 'p
eacefully', 'between', 'its', 'four', 'familiar', 'walls', 'A', 'collection', 'of', 'textile', 'samples', 'lay', 'spread', 'ou
t', 'on', 'the', 'table', '', 'Samsa', 'was', 'a', 'travelling', 'salesman', '', 'and', 'above', 'it', 'there', 'hung', 'a', 'pi
cture', 'that', 'he', 'had', 'recently', 'cut', 'out', 'of', 'an', 'illustrated', 'magazine', 'and']
```

# Normalizing Case

In [6]:
```
# split into words by white space
words = text.split()
# convert to lower case
words = [word.lower() for word in words]
print(words[:150])
```

```
['/:;<=@', 'one', 'morning,', 'when', 'gregor', 'samsa', 'woke', 'from', 'troubled', 'dreams,', 'he', 'found', 'himself', 'trans
formed', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin.', 'he', 'lay', 'on', 'his', 'armour-like', 'back,', 'and', 'if',
'he', 'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly,', 'slightly', 'domed', 'and', 'divid
ed', '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', 'si
ze', 'of', 'the', 'rest', 'of', 'him,', 'waved', 'about', 'helplessly', 'as', 'he', 'looked.', '"what\'s', 'happened', 'to', 'm
e?"', 'he', 'thought.', 'it', "wasn't", 'a', 'dream.', 'his', 'room,', 'a', 'proper', 'human', 'room', 'although', 'a', 'littl
e', 'too', 'small,', 'lay', 'peacefully', 'between', 'its', 'four', 'familiar', 'walls.', 'a', 'collection', 'of', 'textile', 's
amples', 'lay', 'spread', 'out', 'on', 'the', 'table', '-', 'samsa', 'was', 'a', 'travelling', 'salesman', '-', 'and', 'above',
'it', 'there', 'hung', 'a', 'picture', 'that', 'he', 'had', 'recently', 'cut', 'out', 'of', 'an', 'illustrated', 'magazine', 'an
d']
```

# Tokenization and Cleaning with NLTK

In [7]:
```
# split into sentences
from nltk import sent_tokenize
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.
```

# Split into Words

In [8]:
```
nltk.download('wordnet')
```

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

Out[8]:
```
True
```

In [9]:
```
nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\raju\AppData\Roaming\nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
```

Out[9]:
```
True
```

In [10]:
```
# split into words
from nltk.tokenize import word_tokenize
tokens = word_tokenize(text)
print(tokens[:150])
```

```
['/', ':', ';', '<', '=', '@', 'One', 'morning', ',', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams', ',', 'h
e', 'found', 'himself', 'transformed', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin', '.', 'He', 'lay', 'on', 'his', 'ar
mour-like', 'back', ',', 'and', 'if', 'he', 'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'bell
y', ',', 'slightly', 'domed', 'and', 'divided', 'by', 'arches', 'into', 'stiff', 'sections', '.', 'The', 'bedding', 'was', 'hard
ly', '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', 'helplessl
y', 'as', 'he', 'looked', '.', '``', 'What', "'s", 'happened', 'to', 'me', '?', "''", 'he', 'thought', '.', 'It', 'was', "n't",
'a', 'dream', '.', 'His', 'room', ',', 'a', 'proper', 'human', 'room', 'although', 'a', 'little', 'too', 'small', ',', 'lay', 'p
eacefully', 'between', 'its', 'four', 'familiar', 'walls', '.', 'A', 'collection', 'of', 'textile', 'samples', 'lay', 'spread',
'out', 'on', 'the', 'table']
```

# Filter Out Punctuation

In [11]:
```
from nltk.tokenize import word_tokenize
tokens = word_tokenize(text)
# remove all tokens that are not alphabetic
words = [word for word in tokens if word.isalpha()]
print(words[:150])
```

```
['One', 'morning', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams', 'he', 'found', 'himself', 'transformed', 'i
n', 'his', 'bed', 'into', 'a', 'horrible', 'vermin', 'He', 'lay', 'on', 'his', 'back', 'and', 'if', 'he', 'lifted', 'his', 'hea
d', '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', 'o
ff', 'any', 'moment', 'His', 'many', 'legs', 'pitifully', 'thin', 'compared', 'with', 'the', 'size', 'of', 'the', 'rest', 'of',
'him', 'waved', 'about', 'helplessly', 'as', 'he', 'looked', 'What', 'happened', 'to', 'me', 'he', 'thought', 'It', 'was', 'a',
'dream', 'His', 'room', 'a', 'proper', 'human', 'room', 'although', 'a', 'little', 'too', 'small', 'lay', 'peacefully', 'betwee
n', 'its', 'four', 'familiar', 'walls', 'A', 'collection', 'of', 'textile', 'samples', 'lay', 'spread', 'out', 'on', 'the', 'tab
le', 'Samsa', 'was', 'a', 'travelling', 'salesman', 'and', 'above', 'it', 'there', 'hung', 'a', 'picture', 'that', 'he', 'had',
'recently', 'cut', 'out', 'of', 'an', 'illustrated', 'magazine', 'and', 'housed', 'in', 'a', 'nice']
```

# Filter out Stop Words (and Pipeline)

In [12]:
```
from nltk.corpus import stopwords
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', 'yo
urself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself',
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'a
m', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an',
'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'betwee
n', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'ove
r', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each',
'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's',
't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren',
"aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'i
sn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'was
n', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

In [13]:
```
# split into words
from nltk.tokenize import word_tokenize
tokens = word_tokenize(text)

# convert to lower case
tokens = [w.lower() for w in tokens]

# remove punctuation from each word
import string
table = str.maketrans('', '', string.punctuation)
stripped = [w.translate(table) for w in tokens]
```

```
# remove remaining tokens that are not alphabetic
words = [word for word in stripped if word.isalpha()]

# filter out stop words
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))
words = [w for w in words if not w in stop_words]

print(words[:150])
```

```
['one', 'morning', 'gregor', 'samsa', 'woke', 'troubled', 'dreams', 'found', 'transformed', 'bed', 'horrible', 'vermin', 'lay',
'armourlike', 'back', 'lifted', 'head', 'little', 'could', 'see', 'brown', 'belly', 'slightly', 'domed', 'divided', 'arches', 's
tiff', 'sections', 'bedding', 'hardly', 'able', 'cover', 'seemed', 'ready', 'slide', 'moment', 'many', 'legs', 'pitifully', 'thi
n', 'compared', 'size', 'rest', 'waved', 'helplessly', 'looked', 'happened', 'thought', 'nt', 'dream', 'room', 'proper', 'huma
n', 'room', 'although', 'little', 'small', 'lay', 'peacefully', 'four', 'familiar', 'walls', 'collection', 'textile', 'samples',
'lay', 'spread', 'table', 'samsa', 'travelling', 'salesman', 'hung', 'picture', 'recently', 'cut', 'illustrated', 'magazine', 'h
oused', 'nice', 'gilded', 'frame', 'showed', 'lady', 'fitted', 'fur', 'hat', 'fur', 'boa', 'sat', 'upright', 'raising', 'heavy',
'fur', 'muff', 'covered', 'whole', 'lowe']
```

# Stem Words

In [14]:
```
# stemming of words
from nltk.stem.porter import PorterStemmer
porter = PorterStemmer()
stemmed = [porter.stem(word) for word in tokens]
print(stemmed[:150])
```

```
['/', ':', ';', '<', '=', '@', 'one', 'morn', ',', 'when', 'gregor', 'samsa', 'woke', 'from', 'troubl', 'dream', ',', 'he', 'fou
nd', 'himself', 'transform', 'in', 'hi', 'bed', 'into', 'a', 'horribl', 'vermin', '.', 'he', 'lay', 'on', 'hi', 'armour-lik', 'b
ack', ',', 'and', 'if', 'he', 'lift', 'hi', 'head', 'a', 'littl', 'he', 'could', 'see', 'hi', 'brown', 'belli', ',', 'slightli',
'dome', 'and', 'divid', 'by', 'arch', 'into', 'stiff', 'section', '.', 'the', 'bed', 'wa', 'hardli', 'abl', 'to', 'cover', 'it',
'and', 'seem', 'readi', 'to', 'slide', 'off', 'ani', 'moment', '.', 'hi', 'mani', 'leg', ',', 'piti', 'thin', 'compar', 'with',
'the', 'size', 'of', 'the', 'rest', 'of', 'him', ',', 'wave', 'about', 'helplessli', 'as', 'he', 'look', '.', '``', 'what',
"'s", 'happen', 'to', 'me', '?', "''", 'he', 'thought', '.', 'it', 'wa', "n't", 'a', 'dream', '.', 'hi', 'room', ',', 'a', 'prop
er', 'human', 'room', 'although', 'a', 'littl', 'too', 'small', ',', 'lay', 'peac', 'between', 'it', 'four', 'familiar', 'wall',
'.', 'a', 'collect', 'of', 'textil', 'sampl', 'lay', 'spread', 'out', 'on', 'the', 'tabl']
```

# Encoding (word vectorizing) With SciKit-Learn:

In [15]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
# list of text documents
text = ["One morning, when Gregor Samsa woke from troubled dreams, he found himself transformed in his bed into a horrible vermi
# create the transform
vectorizer = TfidfVectorizer()
# tokenize and build vocab
vectorizer.fit(text)
# summarize
print(vectorizer.vocabulary_)
print(vectorizer.idf_)
# encode document
vector = vectorizer.transform([text[0]])
# summarize encoded vector
print(vector.shape)
print(vector.toarray())
```

```
{'one': 12, 'morning': 11, 'when': 17, 'gregor': 4, 'samsa': 13, 'woke': 18, 'from': 3, 'troubled': 15, 'dreams': 1, 'he': 5, 'f
ound': 2, 'himself': 6, 'transformed': 14, 'in': 9, 'his': 7, 'bed': 0, 'into': 10, 'horrible': 8, 'vermin': 16}
[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
(1, 19)
[[0.22941573 0.22941573 0.22941573 0.22941573 0.22941573 0.22941573
  0.22941573 0.22941573 0.22941573 0.22941573 0.22941573 0.22941573
  0.22941573 0.22941573 0.22941573 0.22941573 0.22941573 0.22941573
  0.22941573]]
```

In [ ]: