```
#!/usr/bin/env python
# coding: utf-8

#
In[66]:


file1=open(r'C:\Users\SATHWIK\Desktop\novel.txt','r')
d=file1.readlines()
file1.close
()


# In[67]:


pip install nltk


# In[68]:


import nltk
from nltk.corpus import
stopwords
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize
from
sklearn.feature_extraction.text import
CountVectorizer
nltk.download('punkt')
nltk.download('stopwords')


# In[69]:


# Tokenize the
text data into individual words
tokenized_data = [word_tokenize(text) for text in d]


#
In[70]:


# Remove stop words from the tokenized data
stop_words =
set(stopwords.words('english'))
filtered_data = [[word for word in words if not word.lower() in
stop_words] for words in tokenized_data]


# In[71]:


#filtered_data


# In[72]:


# Stem the
filtered data
ps = PorterStemmer()
stemmed_data = [[ps.stem(word) for word in words] for words
in filtered_data]


# In[73]:


# Convert the pre-processed text data into numeric vectors
using the Bag-of-Words model
```

```
vectorizer = CountVectorizer()
bow_data =
vectorizer.fit_transform([' '.join(words) for words in stemmed_data])


# In[74]:


# Print the
pre-processed text data and the numeric vectors
print("Pre-processed text data:\n",
stemmed_data)
print("\nNumeric vectors:\n", bow_data.toarray())


# In[75]:


#2
from
sklearn.feature_extraction.text import TfidfVectorizer

# Example corpus
corpus = [

"This is the first document.",
    "This document is the second
document.",
    "And this is the third one.",
    "Is this the first
document?",
]

# Create TfidfVectorizer object
vectorizer = TfidfVectorizer()

# Fit the
vectorizer to the corpus and transform the corpus
tfidf_matrix =
vectorizer.fit_transform(corpus)

# Get the feature names (words) from the
vectorizer
feature_names = vectorizer.get_feature_names()

# Print the feature names and tf-idf
matrix
print("Feature names:", feature_names)
print("TF-IDF matrix:\n",
tfidf_matrix.toarray())


# In[ ]:
```