In [1]:
```
person='nilesh'
```

In [2]:
```
print(f"my name is {person}")
```

```
my name is nilesh
```

In [4]:
```
d={'a':123, 'b':456}
```

In [8]:
```
print(f"my numner is {d['b']}")
```

```
my numner is 456
```

In [9]:
```
my_list =[7,8,9]
```

In [10]:
```
print(f"my no is {my_list[1]}")
```

```
my no is 8
```

In [11]:
```
library =[('author','topic','pages'),('Twaig','Rafting',601),('Feynam','physics
```

In [12]:
```
library
```

Out[12]:
```
[('author', 'topic', 'pages'),
 ('Twaig', 'Rafting', 601),
 ('Feynam', 'physics', 95),
 ('hamilton', 'mythology', 144)]
```

In [13]:
```
for x in library:
    print(x)
```

```
('author', 'topic', 'pages')
('Twaig', 'Rafting', 601)
('Feynam', 'physics', 95)
('hamilton', 'mythology', 144)
```

In [14]:
```
for x in library:
    print(x[0])
```

```
author
Twaig
Feynam
hamilton
```

In [16]:
```
for author, topic, pages in library:
    print(f"{author:{10}} {topic:{30}} {pages:{10}}")
```

```
author     topic                          pages
Twaig      Rafting                              601
Feynam     physics                               95
hamilton   mythology                            144
```

In [17]:
```
from datetime import datetime
today=datetime(year=2023, month=2, day=28)
```

In [21]:
```
print(f"{today:%B %d, %Y}")
```

```
February 28, 2023
```

In [22]:
```
pwd
```

Out[22]:
```
'C:\\Users\\nilesh'
```

In [23]:
```
my_file =open('ddd.txt','a+')
```

In [25]:
```
my_file.write('my first line is a+openifffong')
```

Out[25]:
```
30
```

In [29]:
```
my_file.close()
```

In [30]:
```
my_file =open('ddd.txt')
```

In [31]:
```
my_file.read()
```

Out[31]:
```
'my first line is a+openiongmy first line is a+openifffong'
```

In [32]:
```
my_file.close()
```

In [33]:
```
my_file =open('ddd.txt',mode='a+')
```

In [34]:
```
my_file.write('this is an added line, beacuse i use a+ mode')
```

Out[34]:
```
44
```

In [35]:
```
my_file.seek(0)
```

Out[35]:
```
0
```

In [36]:
```
my_file.read()
```

Out[36]:
```
'my first line is a+openiongmy first line is a+openifffongthis is an added li
ne, beacuse i use a+ mode'
```

In [37]:
```
my_file.write('\n this is the real line, on the next line')
```

Out[37]:
```
41
```

In [39]:
```
my_file.seek(0)
```

Out[39]:
```
0
```

In [40]:
```
my_file.read()
```

Out[40]:
```
'my first line is a+openiongmy first line is a+openifffongthis is an added li
ne, beacuse i use a+ mode\n this is the real line, on the next line'
```

In [43]:
```
myfile=open(r'C:\Users\nilesh\OneDrive\Desktop\New Text Document.txt')
```

In [45]:
```
myfile.read()
```

Out[45]:
```
' getting good iq start practing mdeiation\n\n'
```

In [50]:
```
myfile.seek(0)
```

Out[50]:
```
0
```

In [51]:
```
content =myfile.read()
```

In [52]:
```
print(content)
```

```
welcome to this world ddldlandnkl
welcome to Earth 2023
for getting good iq start practing mdeiation
```

In [53]:
```
myfile.close()
```

In [4]:
```
import PyPDF2
```

In [2]:
```
myfile=open(r'F:\daily work\NLP\UPDATED_NLP_COURSE\00-Python-Text-Basics\US_Dec
```

In [8]:
```
pdf_reader=PyPDF2.PdfReader(myfile)
```

In [11]:
```
len(pdf_reader.pages)
```

Out[11]:
```
5
```

In [18]:
```
page_one =pdf_reader.pages[0]
```

In [21]:
```
myfile.close()
```

In [22]:
```
f=open(r'F:\daily work\NLP\UPDATED_NLP_COURSE\00-Python-Text-Basics\US_Declarat
```

In [23]:
```
pdf_reader=PyPDF2.PdfReader(f)
```

In [24]:
```
first_page =pdf_reader.pages[0]
```

In [26]:
```
pdf_writer=PyPDF2.PdfWriter()
```

In [29]:
```
pdf_writer.add_page(first_page)
```

Out[29]:
```
{'/Type': '/Page',
 '/Contents': {},
 '/MediaBox': [0, 0, 612, 792],
 '/Resources': {'/Font': {'/F9': {'/Type': '/Font',
    '/Subtype': '/Type1',
    '/Name': '/F9',
    '/Encoding': '/WinAnsiEncoding',
    '/FirstChar': 31,
    '/LastChar': 255,
    '/Widths': [778,
     250,
     333,
     555,
     500,
     500,
     1000,
     833,
     278,
     333,
```

In [32]:
```
pdf_output= open('F:\\daily work\\NLP\\UPDATED_NLP_COURSE\\00-Python-Text-Basic
```

In [33]:
```
pdf_writer.write(pdf_output)
```

Out[33]:
```
(False,
 <_io.BufferedWriter name='F:\\daily work\\NLP\\UPDATED_NLP_COURSE\\00-Python
-Text-Basics\\NEW_BRAND.pdf'>)
```

In [34]:
```
pdf_output.close()
```

In [35]:
```
f.close()
```

In [36]:
```
brand_new= open('F:\\daily work\\NLP\\UPDATED_NLP_COURSE\\00-Python-Text-Basics
pdf_reader=PyPDF2.PdfReader(brand_new)
```

In [37]:
```
first1_page =pdf_reader.pages[0]
```

In [38]:
```
first1_page.extract_text()
```

Out[38]:
```
"Declaration of Independence\nIN CONGRESS, July 4, 1776.  \nThe unanimous Dec
laration of the thirteen united States of America,  \nWhen in the Course of h
uman events, it becomes necessary for one people to dissolve thepolitical ban
ds which have connected them with another, and to assume among the powers of
theearth, the separate and equal station to which the Laws of Nature and of N
ature's God entitlethem, a decent respect to the opinions of mankind requires
that they should declare the causeswhich impel them to the separation. We hol
d these truths to be self-evident, that all men are created equal, that they
are endowed bytheir Creator with certain unalienable Rights, that among these
are Life, Liberty and the pursuit\nof Happiness.— \x14That to secure these ri
ghts, Governments are instituted among Men, derivingtheir just powers from th
e consent of the governed,—  \x14That whenever any Form of Government\nbecome
s destructive of these ends, it is the Right of the People to alter or to abo
lish it, and to\ninstitute new Government, laying its foundation on such prin
ciples and organizing its powers in\nsuch form, as to them shall seem most li
kely to effect their Safety and Happiness. Prudence,indeed, will dictate that
Governments long established should not be changed for light andtransient cau
ses; and accordingly all experience hath shewn, that mankind are more dispose
d to\nsuffer, while evils are sufferable, than to right themselves by abolish
ing the forms to which theyare accustomed. But when a long train of abuses an
d usurpations, pursuing invariably the same\nObject evinces a design to reduc
e them under absolute Despotism, it is their right, it is their duty,\nto thr
ow off such Government, and to provide new Guards for their future securit y.
— \x14Such has\nbeen the patient sufferance of these Colonies; and such is no
w the necessity which constrainsthem to alter their former Systems of Governm
ent. The history of the present King of GreatBritain is a history of repeated
injuries and usurpations, all having in direct object the\nestablishment of a
n absolute Tyranny over these States. To prove this, let Facts be submitted t
o a\ncandid world. \nHe has refused his Assent to Laws, the most wholesome an
d necessary for the\npublic good.He has forbidden his Governors to pass Laws
of immediate and pressingimportance, unless suspended in their operation till
his Assent should be obtained;and when so suspended, he has utterly neglected
to attend to them.He has refused to pass other Laws for the accommodation of
large districts of\npeople, unless those people would relinquish the right of
Representation in theLegislature, a right inestimable to them and formidable
to tyrants only. He has called together legislative bodies at places unusual,
uncomfortable, and distantfrom the depository of their public Records, for th
e sole purpose of fatiguing them into\ncompliance with his measures."
```

In [ ]:
```
f= open('F:\\daily work\\NLP\\UPDATED_NLP_COURSE\\00-Python-Text-Basics\\NEW_BF
pdf_text=[]
pdf_reader=PyPDF2.PdfReader(f)
count = len(pdfReader.pages)
for p in range(count):
    page=pdf_reader.pages[p]
    pdf_text.append(page.extract_text())
f.close()
```

In [64]:
```
text ='the phone number of agent is 400-555-1234. call soon'
```

In [65]:
```
"400-555-1234" in text
```

Out[65]:
```
True
```

In [66]:
```
import re
pattern="phone"
```

In [67]:
```
re.search(pattern, text)
```

Out[67]:
```
<re.Match object; span=(4, 9), match='phone'>
```

In [68]:
```
my_match=re.search(pattern, text)
```

In [69]:
```
my_match.span()
```

Out[69]:
```
(4, 9)
```

In [70]:
```
my_match.start()
```

Out[70]:
```
4
```

In [71]:
```
my_match.end()
```

Out[71]:
```
9
```

SPACY Basic

In [1]:
```
import spacy
```

In [4]:
```
import spacy
import spacy.cli
spacy.cli.download("en_core_web_lg")
nlp=spacy.load('en_core_web_lg')
```

```
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_lg')
```

In [5]:
```
nlp=spacy.load('en_core_web_lg')
```

In [8]:
```
doc=nlp(u'Tesla is looking at US start up for $6 millions')
```

In [9]:
```
for token in doc:
    print(token)
```

```
Tesla
is
looking
at
US
start
up
for
$
6
millions
```

In [11]:
```
for token in doc:
    print(token.text,token.pos_,token.dep_)
```

```
Tesla PROPN nsubj
is AUX aux
looking VERB ROOT
at ADP prep
US PROPN pobj
start VERB conj
up ADP prt
for ADP prep
$ SYM quantmod
6 NUM compound
millions NOUN pobj
```

In [13]:
```
nlp.pipeline
```

Out[13]:
```
[('tok2vec', <spacy.pipeline.tok2vec.Tok2Vec at 0x21259038d60>),
 ('tagger', <spacy.pipeline.tagger.Tagger at 0x21259038c40>),
 ('parser', <spacy.pipeline.dep_parser.DependencyParser at 0x21259030120>),
 ('attribute_ruler',
  <spacy.pipeline.attributeruler.AttributeRuler at 0x21259040500>),
 ('lemmatizer', <spacy.lang.en.lemmatizer.EnglishLemmatizer at 0x21259035280
>),
 ('ner', <spacy.pipeline.ner.EntityRecognizer at 0x21259030190>)]
```

In [16]:
```
nlp.pipe_names
```

Out[16]:
```
['tok2vec', 'tagger', 'parser', 'attribute_ruler', 'lemmatizer', 'ner']
```

In [19]:
```
doc1=nlp(u'Tesla is not looking for any startup')
```

In [20]:
```
for token in doc1:
    print(token.text,token.pos_,token.dep_)
```

```
Tesla PROPN nsubj
is AUX aux
not PART neg
looking VERB ROOT
for ADP prep
any DET det
startup NOUN pobj
```

In [21]:
```
doc1[0]
```

Out[21]:
```
Tesla
```

In [22]:
```
doc3 = nlp(u'Although commmonly attributed to John Lennon from his song "Beauti
the phrase "Life is what happens to us while we are making other plans" was wri
cartoonist Allen Saunders and published in Reader\'s Digest in 1957, when Lennc
```

In [23]:
```
life_quote = doc3[16:30]
print(life_quote)
```

```
"Life is what happens to us while we are making other plans"
```

In [24]:
```
type(life_quote)
```

Out[24]:
```
spacy.tokens.span.Span
```

In [25]:
```
type(doc3)
```

Out[25]:
```
spacy.tokens.doc.Doc
```

In [26]:
```
doc4=nlp(u'this is the first sentence. this is the second sentence. this is the
```

In [27]:
```
for sentence in doc4.sents:
    print(sentence)
```

```
this is the first sentence.
this is the second sentence.
this is the last sentence
```

In [28]:
```
doc4[6]
```

Out[28]:
```
this
```

In [29]:
```
doc4[6].is_sent_start
```

Out[29]:
```
True
```

In [30]:
```
import spacy
nlp=spacy.load('en_core_web_lg')
```

In [33]:
```
my_string= '"we\'re moving to L.A !" '
```

In [34]:
```
print(my_string)
```

```
"we're moving to L.A !"
```

In [36]:
```
doc= nlp(my_string)
```

In [37]:
```
for token in doc:
    print(token.text)
```

```
"
we
're
moving
to
L.A
!
"
```

In [43]:
```
doc1= nlp(u"we're here to help you! send email, cs@abc.com or vist or website h
```

In [44]:
```
  for t in doc1:
        print(t)
```

```
we
're
here
to
help
you
!
send
email
,
cs@abc.com
or
vist
or
website
http://www.abc.co.in (http://www.abc.co.in)
!
```

In [45]:
```
doc3=nlp(u" A 5km NYC cab ride cost $10.30")
```

In [46]:
```
for x in doc3:
    print(x)
```

```
A
5
km
NYC
cab
ride
cost
$
10.30
```

In [52]:
```
doc4= nlp(u" let's vist ST. johan in US. next week")
```

In [53]:
```
for x in doc4:
    print(x)
```

```
let
's
vist
ST
.
johan
in
US
.
next
week
```

In [54]:
```
len(doc4)
```

Out[54]: 12

In [56]:
```
len(doc4.vocab)
```

Out[56]: 795

In [59]:
```
doc5= nlp(u"it's better to give than receive")
```

In [60]:
```
doc5[0]
```

Out[60]: it

In [61]:
```
doc5[2:5]
```

Out[61]: better to give

In [62]:
```
doc6= nlp(u"Apple to build new office in India to support production in asia-pa
```

In [63]:
```
for x in doc6:
    print(x.text, end='|')
```

```
Apple|to|build|new|office|in|India|to|support|production|in|asia|-|pacific|
```

In [66]:
```
for entity in doc6.ents:
    print(entity)
    print(entity.label_)
    print(str(spacy.explain(entity.label_)))
    print('\n')
```

```
Apple
ORG
Companies, agencies, institutions, etc.


India
GPE
Countries, cities, states


asia-pacific
LOC
Non-GPE locations, mountain ranges, bodies of water
```

In [67]:
```
from spacy import displacy
```

In [69]:
```
doc=nlp(u"apple is going to build in UK factory for $6 millions.")
```

In [76]:
```
displacy.render(doc, style='dep',jupyter=True, options={'distance':90})
```

apple NOUN is AUX going VERB to PART build VERB in ADP UK PROPN factory NOUN for ADP $ SYM 6 NUM millions. NOUN nsubj aux aux xcomp prep compound pobj prep quantmod compound pobj

In [79]:
```
doc1= nlp(u" Last year Apple sold around $600 millions Iphone in profit.")
```

In [80]:
```
displacy.render(doc1, style='ent', jupyter=True)
```

Last year **DATE** Apple **ORG** sold around $600 millions **MONEY** Iphone in profit.

In [81]:
```
doc2=nlp(u" This is the sentence.")
```

In [83]:
```
displacy.serve(doc2, style='dep')
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\spacy\displacy\__init__.py:106: U
serWarning: [W011] It looks like you're calling displacy.serve from within a
Jupyter notebook or a similar environment. This likely means you're already r
unning a local web server, so there's no need to make displaCy start another
one. Instead, you should be able to replace displacy.serve with displacy.rend
er to show the visualization.
  warnings.warn(Warnings.W011)
```

displaCy

SPACE This PRON is AUX the DET sentence. NOUN dep nsubj det attr

```
Using the 'dep' visualizer
Serving on http://0.0.0.0:5000 (http://0.0.0.0:5000) ...

Shutting down server on port 5000.
```

Stemming

In [84]:
```
from nltk.stem.porter import PorterStemmer
```

In [85]:
```
p_stemmer=PorterStemmer()
```

In [86]:
```
words =['run', 'runner','ran','runs','easily','fairly']
```

In [88]:
```
for x in words:
    print((x)+'---->'+p_stemmer.stem(x))
```

```
run---->run
runner---->runner
ran---->ran
runs---->run
easily---->easili
fairly---->fairli
```

In [89]:
```
from nltk.stem.snowball import SnowballStemmer
```

In [90]:
```
s_stemmer=SnowballStemmer(language='english' )
```

In [91]:
```
for x in words:
    print((x)+'---->'+s_stemmer.stem(x))
```

```
run---->run
runner---->runner
ran---->ran
runs---->run
easily---->easili
fairly---->fair
```

In [1]:
```
import spacy
nlp=spacy.load('en_core_web_lg')
```

In [5]:
```
from spacy.matcher import PhraseMatcher
```

In [6]:
```
matcher= PhraseMatcher(nlp.vocab)
```

In [8]:
```
with open(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\reaganomics.txt') as
    doc3=nlp(f.read())
```

In [9]:
```
phrase_list=['voodoo economics','supply-side economics','trickle-down economics
```

In [10]:
```
pharase_patterns=[nlp(text) for text in phrase_list]
```

In [11]:
```
pharase_patterns
```

Out[11]:
```
[voodoo economics,
 supply-side economics,
 trickle-down economics,
 free-market economics]
```

In [13]:
```
matcher.add('EconMatcher',None, *pharase_patterns)
```

In [14]:
```
found_matches=matcher(doc3)
```

In [15]:
```
found_matches
```

Out[15]:
```
[(3680293220734633682, 41, 45),
 (3680293220734633682, 49, 53),
 (3680293220734633682, 54, 56),
 (3680293220734633682, 61, 65),
 (3680293220734633682, 673, 677),
 (3680293220734633682, 2987, 2991)]
```

In [16]:
```
for match_id, start, end in found_matches:
    string_id =nlp.vocab.strings[match_id]
    span=doc3[start:end]
    print(match_id, string_id, start,end,span.text)
```

```
3680293220734633682 EconMatcher 41 45 supply-side economics
3680293220734633682 EconMatcher 49 53 trickle-down economics
3680293220734633682 EconMatcher 54 56 voodoo economics
3680293220734633682 EconMatcher 61 65 free-market economics
3680293220734633682 EconMatcher 673 677 supply-side economics
3680293220734633682 EconMatcher 2987 2991 trickle-down economics
```

In [1]:
```
import spacy.cli
spacy.cli.download("en_core_web_lg")
nlp=spacy.load('en_core_web_lg')
```

```
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_lg')
```

In [2]:
```
import spacy
```

In [3]:
```
doc=nlp(u"the quick brown box jumped over the lazy's back dog")
```

In [4]:
```
print(doc.text)
```

```
the quick brown box jumped over the lazy's back dog
```

In [5]:
```
print(doc[4])
```

```
jumped
```

In [6]:
```
print(doc[4].pos_)
```

```
VERB
```

In [7]:
```
print(doc[4].text)
```

```
jumped
```

In [8]:
```
print(doc[4].tag_)
```

```
VBD
```

In [9]:
```
for token in doc:
    print(f"{token.text} {token.pos_} {token.tag_} {spacy.explain(token.tag_)}
```

```
the DET DT determiner
quick ADJ JJ adjective (English), other noun-modifier (Chinese)
brown PROPN NNP noun, proper singular
box NOUN NN noun, singular or mass
jumped VERB VBD verb, past tense
over ADP IN conjunction, subordinating or preposition
the DET DT determiner
lazy ADJ JJ adjective (English), other noun-modifier (Chinese)
's PART POS possessive ending
back ADJ JJ adjective (English), other noun-modifier (Chinese)
dog NOUN NN noun, singular or mass
```

In [10]:
```
doc=nlp(u"I read books on NLP")
```

In [11]:
```
word=doc[1]
```

In [12]:
```
word.text
```

Out[12]:
```
'read'
```

In [13]:
```
token=word
print(f"{token.text:{10}} {token.pos_:{10}} {token.tag_:{10}} {spacy.explain(t
```

```
  Cell In[13], line 2
    print(f"{token.text:{10}} {token.pos_:{10}} {token.tag_:{10}} {spacy.expl
ain(token.tag_:{10})}")
                                ^
SyntaxError: f-string: invalid syntax
```

In [14]:
```
doc=nlp(u"I read a book on NLP")
```

In [15]:
```
word= doc[1]
```

In [16]:
```
word.text
```

Out[16]:
```
'read'
```

In [17]:
```
token=word
print(f"{token.text} {token.pos_} {token.tag_} {spacy.explain(token.tag_)}")
```

```
read VERB VBP verb, non-3rd person singular present
```

In [18]:
```
doc=nlp(u"I slept in the night")
```

In [19]:
```
word= doc[1]
```

In [20]:
```
word.text
```

Out[20]:
```
'slept'
```

In [21]:
```
token=word
print(f"{token.text} {token.pos_} {token.tag_} {spacy.explain(token.tag_)}")
```

```
slept VERB VBD verb, past tense
```

In [26]:
```
doc1= nlp(u'I read books on NLP')
```

In [27]:
```
word= doc1[1]
```

In [28]:
```
word.text
```

Out[28]:
```
'read'
```

In [29]:
```
token=word
print(f"{token.text} {token.pos_} {token.tag_} {spacy.explain(token.tag_)}")
```

```
read VERB VBP verb, non-3rd person singular present
```

In [30]:
```
doc=nlp(u"the quick brown box jumped over the lazy's back dog")
```

In [31]:
```
pos_counts=doc.count_by(spacy.attrs.POS)
```

In [32]:
```
pos_counts
```

Out[32]:
```
{90: 2, 84: 3, 96: 1, 92: 2, 100: 1, 85: 1, 94: 1}
```

In [34]:
```
doc.vocab[84].text
```

Out[34]:
```
'ADJ'
```

In [35]:
```
for k,v in sorted (pos_counts.items()):
    print(f"{k}.{doc.vocab[k].text:{5}} {v}")
```

```
84.ADJ   3
85.ADP   1
90.DET   2
92.NOUN  2
94.PART  1
96.PROPN 1
100.VERB  1
```

In [36]:
```
len(doc.vocab)
```

Out[36]:
```
788
```

Visualzting part of speech

In [37]:
```
import spacy.cli
spacy.cli.download("en_core_web_lg")
nlp=spacy.load('en_core_web_lg')
```

```
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_lg')
```

In [38]:
```
import spacy
```

In [39]:
```
from spacy import *
```

In [40]:
```
doc= nlp(u"the quick brown fox jumped over the lazy dog")
```

In [42]:
```
displacy.render(doc, style='dep',jupyter=True)
```

the DET quick ADJ brown ADJ fox NOUN jumped VERB over ADP the DET lazy ADJ dog NOUN
det amod amod nsubj prep det amod pobj

In [44]:
```
options={'distance':110,'compact':'True','color':'yellow','bg':'#09a3d5','font
```

In [45]:
```
displacy.render(doc, style='dep',jupyter=True,options=options)
```

the DET quick ADJ brown ADJ fox NOUN jumped VERB over ADP the DET lazy ADJ dog NOUN
det amod amod nsubj prep det amod pobj

In [46]:
```
doc2=nlp(u"this is sentence. This is another sentence, possibily ,onger than ot
```

In [47]:
```
spans=list(doc2.sents)
```

In [48]:
```
displacy.serve(spans, style='dep',options={'distance':110})
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\spacy\displacy\__init__.py:106: U
serWarning: [W011] It looks like you're calling displacy.serve from within a
Jupyter notebook or a similar environment. This likely means you're already r
unning a local web server, so there's no need to make displaCy start another
one. Instead, you should be able to replace displacy.serve with displacy.rend
er to show the visualization.
  warnings.warn(Warnings.W011)
```

displaCy
this PRON is AUX sentence. NOUN nsubj attr

This PRON is AUX another DET sentence, NOUN possibily , ADV onger ADJ than ADP others NOUN nsubj det attr advmod appos prep pobj

```
Using the 'dep' visualizer
Serving on http://0.0.0.0:5000 (http://0.0.0.0:5000) ...


127.0.0.1 - - [16/Oct/2023 23:48:38] "GET / HTTP/1.1" 200 8959
127.0.0.1 - - [16/Oct/2023 23:48:40] "GET /favicon.ico HTTP/1.1" 200 8959

Shutting down server on port 5000.
```

In [ ]:
```
http://127.0.0.1:5000/
```

(NER) Name Entity recogination

In [ ]:
```
import spacy.cli
spacy.cli.download("en_core_web_lg")
nlp=spacy.load('en_core_web_lg')
```

```
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_lg')
```

In [60]:
```
def show_ents(doc):
    if doc.ents:
        for ent in doc.ents:
            print(ent.text+'-'+ent.label_+'-'+str(spacy.explain(ent.label_)))
    else:
        print('No entities found')
```

In [52]:
```
doc=nlp(u'how are you')
```

In [53]:
```
show_ents(doc)
```

```
No entities found
```

In [62]:
```
doc=nlp(u"May i go the see singapore for 1 month and japan for 2 months")
```

In [63]:
```
show_ents(doc)
```

```
singapore-GPE-Countries, cities, states
1 month-DATE-Absolute or relative dates or periods
japan-GPE-Countries, cities, states
2 months-DATE-Absolute or relative dates or periods
```

In [64]:
```
doc= nlp(u'i ant to buy 1000 quantity stock for HDFC bank and go for  USA tour
```

In [65]:
```
show_ents(doc)
```

```
1000-CARDINAL-Numerals that do not fall under another type
HDFC bank-ORG-Companies, agencies, institutions, etc.
USA-GPE-Countries, cities, states
```

In [66]:
```
doc=nlp(u"Tesla to built UK factory got $5 millions")
```

In [67]:
```
show_ents(doc)
```

```
Tesla-ORG-Companies, agencies, institutions, etc.
UK-GPE-Countries, cities, states
$5 millions-MONEY-Monetary values, including unit
```

In [92]:
```
doc = nlp(u'Our company plans to introduce a new vacuum cleaner. '
          u'If successful, the vacuum cleaner will be our first product.')

show_ents(doc)
```

```
first-ORDINAL-"first", "second", etc.
```

In [93]:
```
# Import PhraseMatcher and create a matcher object:
from spacy.matcher import PhraseMatcher
matcher = PhraseMatcher(nlp.vocab)
```

In [94]:
```
# Create the desired phrase patterns:
phrase_list = ['vacuum cleaner', 'vacuum-cleaner']
phrase_patterns = [nlp(text) for text in phrase_list]
```

In [95]:
```
# Apply the patterns to our matcher object:
matcher.add('newproduct', None, *phrase_patterns)

# Apply the matcher to our Doc object:
matches = matcher(doc)

# See what matches occur:
matches
```

Out[95]:
```
[(2689272359382549672, 7, 9), (2689272359382549672, 14, 16)]
```

In [96]:
```
# Here we create Spans from each match, and create named entities from them:
from spacy.tokens import Span

PROD = doc.vocab.strings[u'PRODUCT']

new_ents = [Span(doc, match[1],match[2],label=PROD) for match in matches]

doc.ents = list(doc.ents) + new_ents
```

In [97]:
```
show_ents(doc)
```

```
vacuum cleaner-PRODUCT-Objects, vehicles, foods, etc. (not services)
vacuum cleaner-PRODUCT-Objects, vehicles, foods, etc. (not services)
first-ORDINAL-"first", "second", etc.
```

In [12]:
```
doc=nlp(u'$28.33 is now decreasing as inflation rate is increasing so invest $8
        u'in January 2023 Sony sold only 15 millions Televison in India')
```

In [105]:
```
len([ent for ent in doc.ents if ent.label_ =='MONEY'])
```

Out[105]:
```
2
```

In [2]:
```
import spacy
```

In [3]:
```
from spacy import *
```

In [14]:
```
doc= nlp(u'As per last quarter report Apple sold 20 millions iPods for $15 mil
        u'in January 2023 Sony sold only 15 millions Televison in India')
```

In [15]:
```
displacy.render(doc, style='ent', jupyter=True)
```

As per last quarter **DATE** report Apple **ORG** sold 20 millions **CARDINAL** iPods **PRODUCT** for $ 15 **MONEY** millionsin **GPE** January 2023 **DATE** Sony **ORG** sold only 15 millions Televison **MONEY** in India **GPE**

In [16]:
```
for sent in doc.sents:
    displacy.render(nlp(sent.text),style='ent', jupyter=True)
```

As per last quarter **DATE** report Apple **ORG** sold 20 millions **CARDINAL** iPods **PRODUCT** for $ 15 **MONEY** millionsin **GPE** January 2023 **DATE** Sony **ORG** sold only 15 millions Televison **MONEY** in India **GPE**

In [17]:
```
options={'ents':['PRODUCT']}
```

In [19]:
```
displacy.render(doc,style='ent', jupyter=True, options=options)
```

As per last quarter report Apple sold 20 millions iPods **PRODUCT** for $15 millionsin January 2023 Sony sold only 15 millions Televison in India

In [20]:
```
options={'ents':['PRODUCT','ORG']}
```

In [21]:
```
displacy.render(doc,style='ent', jupyter=True, options=options)
```

As per last quarter report Apple **ORG** sold 20 millions iPods **PRODUCT** for $15 millionsin January 2023 Sony **ORG** sold only 15 millions Televison in India

In [26]:
```
colors={'ORG':'red','PRODUCT':'yellow'}
options={'ents':['PRODUCT','ORG'],'colors':colors}
```

In [27]:
```
displacy.render(doc,style='ent', jupyter=True, options=options)
```

As per last quarter report Apple **ORG** sold 20 millions iPods **PRODUCT** for $15 millionsin January 2023 Sony **ORG** sold only 15 millions Televison in India

In [28]:
```
colors={'ORG':'radial-gradient(yellow,red)','PRODUCT':'yellow'}
options={'ents':['PRODUCT','ORG'],'colors':colors}
```

In [29]:
```
displacy.render(doc,style='ent', jupyter=True, options=options)
```

As per last quarter report Apple **ORG** sold 20 millions iPods **PRODUCT** for $15 millionsin January 2023 Sony **ORG** sold only 15 millions Televison in India

In [31]:
```
displacy.serve(doc,style='ent',options=options)
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\spacy\displacy\__init__.py:106: U
serWarning: [W011] It looks like you're calling displacy.serve from within a
Jupyter notebook or a similar environment. This likely means you're already r
unning a local web server, so there's no need to make displaCy start another
one. Instead, you should be able to replace displacy.serve with displacy.rend
er to show the visualization.
  warnings.warn(Warnings.W011)
```

displaCy

As per last quarter report Apple **ORG** sold 20 millions iPods **PRODUCT** for $15 millionsin

January 2023 Sony **ORG** sold only 15 millions Televison in India

```
Using the 'ent' visualizer
Serving on http://0.0.0.0:5000 (http://0.0.0.0:5000) ...

Shutting down server on port 5000.
```

In [34]:
```
doc=nlp(u'this is first sentence. This is second sentence. This is last last se
```

In [35]:
```
for sent in doc.sents:
    print(sent)
```

```
this is first sentence.
This is second sentence.
This is last last sentence
```

In [36]:
```
doc.sents[0]
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
Cell In[36], line 1
----> 1 doc.sents[0]

TypeError: 'generator' object is not subscriptable
```

In [2]:
```
doc=nlp(u'"Management is doing right things; leadership is doing nice work"-Pet
```

In [3]:
```
for sent in doc.sents:
    print(sent)
    print('\n')
```

```
"Management is doing right things; leadership is doing nice work"-Peter kalvi
n
```

In [4]:
```
#Add a segmentation rulw
def set_custom_boundaries(doc):
    for token in doc:
        print(token)
        print(token.i)
```

In [5]:
```
set_custom_boundaries(doc)
```

```
"
0
Management
1
is
2
doing
3
right
4
things
5
;
6
leadership
7
is
8
doing
9
nice
10
work"-Peter
11
kalvin
12
```

In [6]:
```
def set_custom_coundaries(doc):
    for token in doc[:-1]:
        if token.text==';':
            doc[token.i+1].is_sent_start=True
    return doc
```

In [18]:
```
doc4=nlp(u'"Management is doing right things; Leadership is doing nice work."-P
```

In [19]:
```
for sent in doc4.sents:
    print(sent)
```

```
"Management is doing right things; Leadership is doing nice work.
"-Peter kalvin
```

In [20]:
```
mystring=u'This is sentence.This is another.\n\n this is \n third sentence'
```

In [21]:
```
print(mystring)
```

```
This is sentence.This is another.

 this is
 third sentence
```

In [22]:
```
doc=nlp(mystring)
```

In [23]:
```
for sentence in doc.sents:
    print(sentence)
```

```
This is sentence.
This is another.


this is
 third sentence
```

In [27]:
```
import spacy

nlp.enable_pipe("senter")
doc = nlp("This is sentence.This is another.\n\n this is \n third sentence")
for sent in doc.sents:
    print(sent.text)
```

```
This is sentence.
This is another.


this is
 third sentence
```

Text Calssification

In [1]:
```
import pandas as pd
import numpy as np
```

In [4]:
```
df=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\smsspamcollecti
```

In [5]:
```
df.head()
```

Out[5]:

	label	message	length	punct
0	ham	Go until jurong point, crazy.. Available only ...	111	9
1	ham	Ok lar... Joking wif u oni...	29	6
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	155	6
3	ham	U dun say so early hor... U c already then say...	49	6
4	ham	Nah I don't think he goes to usf, he lives aro...	61	2

In [6]:
```
df.describe
```

Out[6]:
```
<bound method NDFrame.describe of      label
message  length  punct
0      ham  Go until jurong point, crazy.. Available only ...     111      9
1      ham                      Ok lar... Joking wif u oni...      29      6
2     spam  Free entry in 2 a wkly comp to win FA Cup fina...     155      6
3      ham  U dun say so early hor... U c already then say...      49      6
4      ham  Nah I don't think he goes to usf, he lives aro...      61      2
...    ...                                                ...     ...    ...
5567  spam  This is the 2nd time we have tried 2 contact u...     160      8
5568   ham               Will ü b going to esplanade fr home?      36      1
5569   ham  Pity, * was in mood for that. So...any other s...      57      7
5570   ham  The guy did some bitching but I acted like i'd...     125      1
5571   ham                         Rofl. Its true to its name      26      1

[5572 rows x 4 columns]>
```

In [7]:
```
df.isnull().sum()
```

Out[7]:
```
label      0
message    0
length     0
punct      0
dtype: int64
```

In [8]:
```
len(df)
```

Out[8]:
```
5572
```

In [9]:
```
df['label'].unique()
```

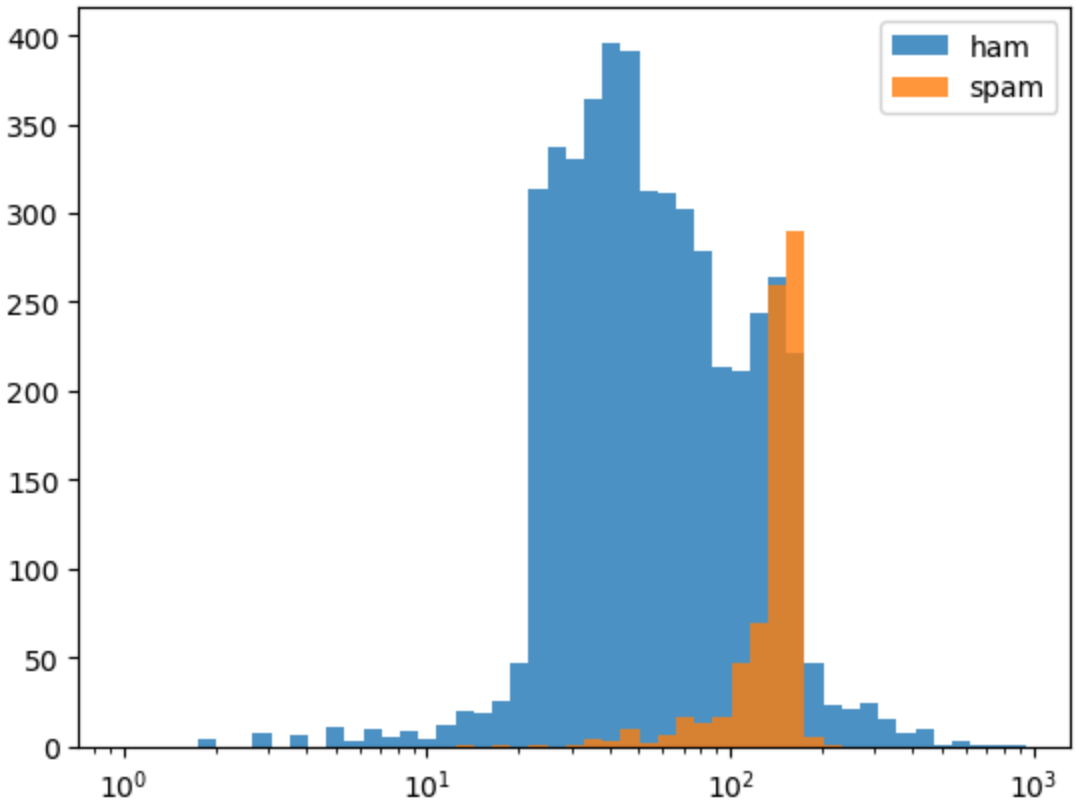
Out[9]:
```
array(['ham', 'spam'], dtype=object)
```

In [12]:
```
df['label'].value_counts()
```

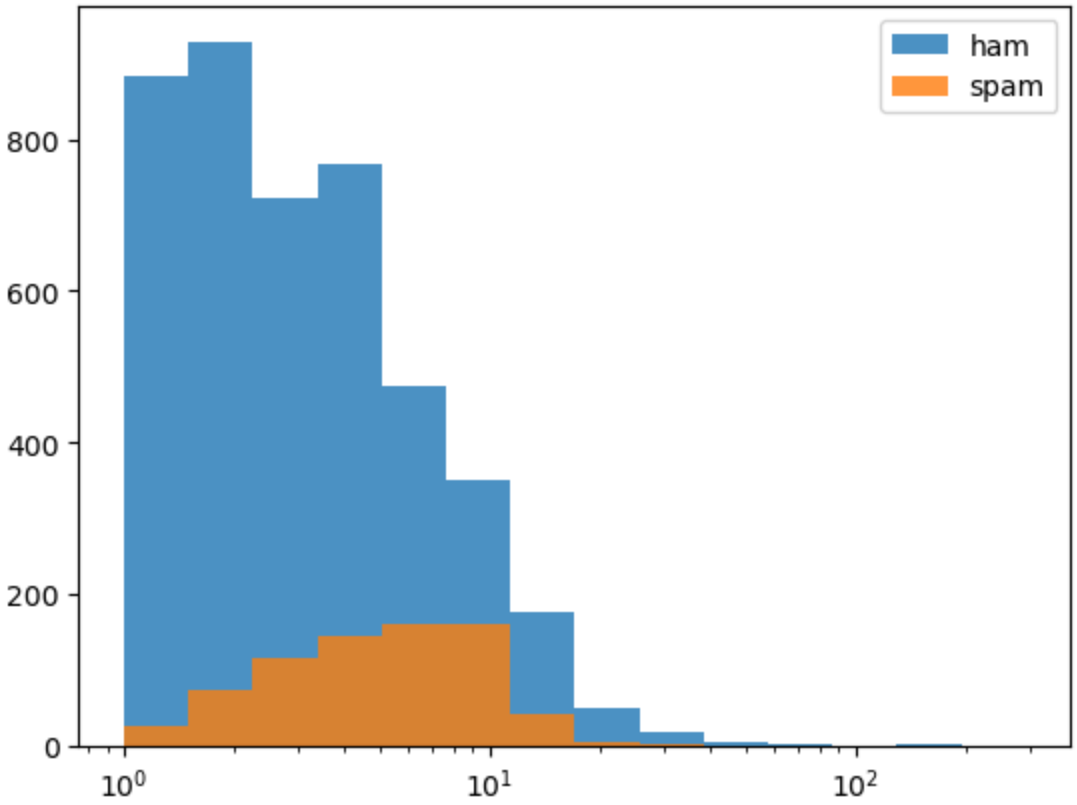
Out[12]:
```
label
ham     4825
spam     747
Name: count, dtype: int64
```

In [13]:
```
import matplotlib.pyplot as plt
%matplotlib inline

plt.xscale('log')
bins = 1.15**(np.arange(0,50))
plt.hist(df[df['label']=='ham']['length'],bins=bins,alpha=0.8)
plt.hist(df[df['label']=='spam']['length'],bins=bins,alpha=0.8)
plt.legend(('ham','spam'))
plt.show()
```



In [14]:
```
plt.xscale('log')
bins = 1.5**(np.arange(0,15))
plt.hist(df[df['label']=='ham']['punct'],bins=bins,alpha=0.8)
plt.hist(df[df['label']=='spam']['punct'],bins=bins,alpha=0.8)
plt.legend(('ham','spam'))
plt.show()
```



In [15]:
```
from sklearn.model_selection import train_test_split
```

In [16]:
```
#X features data
X=df[['length','punct']]
#y is our label
y=df['label']
```

In [17]:
```
X_train,X_test,y_train, y_test=train_test_split(X,y,test_size=0.3,random_state
```

In [18]:
```
X_train.shape
```

Out[18]: (3900, 2)

In [19]:
```
X_test.shape
```

Out[19]: (1672, 2)

In [20]:
```
from sklearn.linear_model import LogisticRegression
```

In [21]:
```
lr_model=LogisticRegression(solver='lbfgs')
```

In [22]:
```
lr_model.fit(X_train,y_train)
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:767:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if not hasattr(array, "sparse") and array.dtypes.apply(is_sparse).any():
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[22]:
```
LogisticRegression()
```
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [23]:
```
from sklearn import metrics
```

In [24]:
```
predictions=lr_model.predict(X_test)
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:767:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if not hasattr(array, "sparse") and array.dtypes.apply(is_sparse).any():
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [25]:
```
predictions
```

Out[25]:
```
array(['ham', 'ham', 'ham', ..., 'ham', 'ham', 'ham'], dtype=object)
```

In [26]:
```
y_test
```

Out[26]:
```
3245     ham
944      ham
1044     ham
2484     ham
812      ham
        ...
2505     ham
2525    spam
4975     ham
650     spam
4463     ham
Name: label, Length: 1672, dtype: object
```

In [27]:
```
print(metrics.confusion_matrix(y_test,predictions))
```

```
[[1404   44]
 [ 219    5]]

C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [28]:

```
# You can make the confusion matrix less confusing by adding labels:
df = pd.DataFrame(metrics.confusion_matrix(y_test,predictions), index=['ham','s
df
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[28]:

	ham	spam
ham	1404	44
spam	219	5

In [29]:
```
print(metrics.classification_report(y_test,predictions))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

              precision    recall  f1-score   support

         ham       0.87      0.97      0.91      1448
        spam       0.10      0.02      0.04       224

    accuracy                           0.84      1672
   macro avg       0.48      0.50      0.48      1672
weighted avg       0.76      0.84      0.80      1672
```

In [30]:
```
print(metrics.accuracy_score(y_test,predictions))
```

```
0.8427033492822966

C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [32]:
```
from sklearn.naive_bayes import MultinomialNB
nb_model=MultinomialNB()
nb_model.fit(X_train,y_train)
predictions=nb_model.predict(X_test)
print(metrics.confusion_matrix(y_test,predictions))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:767:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if not hasattr(array, "sparse") and array.dtypes.apply(is_sparse).any():
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:767:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if not hasattr(array, "sparse") and array.dtypes.apply(is_sparse).any():
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

[[1438   10]
 [ 224    0]]
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [33]:
```
print(metrics.classification_report(y_test,predictions))
```

```
              precision    recall  f1-score   support

         ham       0.87      0.99      0.92      1448
        spam       0.00      0.00      0.00       224

    accuracy                           0.86      1672
   macro avg       0.43      0.50      0.46      1672
weighted avg       0.75      0.86      0.80      1672


C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [34]:
```
print(metrics.accuracy_score(y_test,predictions))
```

```
0.8600478468899522

C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [37]:
```
from sklearn.svm import SVC
```

In [38]:
```
svc_model=SVC(gamma='auto')
svc_model.fit(X_train,y_train)
predictions=svc_model.predict(X_test)
print(metrics.confusion_matrix(y_test,predictions))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:767:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if not hasattr(array, "sparse") and array.dtypes.apply(is_sparse).any():
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:767:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if not hasattr(array, "sparse") and array.dtypes.apply(is_sparse).any():
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

[[1373   75]
 [ 121  103]]
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [39]:
```
print(metrics.classification_report(y_test,predictions))
```

```
              precision    recall  f1-score   support

         ham       0.92      0.95      0.93      1448
        spam       0.58      0.46      0.51       224

    accuracy                           0.88      1672
   macro avg       0.75      0.70      0.72      1672
weighted avg       0.87      0.88      0.88      1672


C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [40]:
```
print(metrics.accuracy_score(y_test,predictions))
```

```
0.8827751196172249

C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Features Extraction from Text

In [41]:
```
%%writefile 1.txt
This is a story about cats
our feline pets
Cats are furry animals
```

```
Writing 1.txt
```

In [42]:
```
%%writefile 2.txt
This story is about surfing
Catching waves is fun
Surfing is a popular water sport
```

```
Writing 2.txt
```

In [43]:
```
vocab = {}
i = 1

with open('1.txt') as f:
    x = f.read().lower().split()

for word in x:
    if word in vocab:
        continue
    else:
        vocab[word]=i
        i+=1

print(vocab)
```

```
{'this': 1, 'is': 2, 'a': 3, 'story': 4, 'about': 5, 'cats': 6, 'our': 7, 'fe
line': 8, 'pets': 9, 'are': 10, 'furry': 11, 'animals': 12}
```

In [44]:
```
with open('2.txt') as f:
    x = f.read().lower().split()

for word in x:
    if word in vocab:
        continue
    else:
        vocab[word]=i
        i+=1

print(vocab)
```

```
{'this': 1, 'is': 2, 'a': 3, 'story': 4, 'about': 5, 'cats': 6, 'our': 7, 'fe
line': 8, 'pets': 9, 'are': 10, 'furry': 11, 'animals': 12, 'surfing': 13, 'c
atching': 14, 'waves': 15, 'fun': 16, 'popular': 17, 'water': 18, 'sport': 1
9}
```

In [45]:
```
# Create an empty vector with space for each word in the vocabulary:
one = ['1.txt']+[0]*len(vocab)
one
```

Out[45]:
```
['1.txt', 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

In [46]:
```
# map the frequencies of each word in 1.txt to our vector:
with open('1.txt') as f:
    x = f.read().lower().split()

for word in x:
    one[vocab[word]]+=1

one
```

Out[46]:
```
['1.txt', 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0]
```

In [47]:
```
# Do the same for the second document:
two = ['2.txt']+[0]*len(vocab)

with open('2.txt') as f:
    x = f.read().lower().split()

for word in x:
    two[vocab[word]]+=1
```

In [48]:
```
# Compare the two vectors:
print(f'{one}\n{two}')
```

```
['1.txt', 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0]
['2.txt', 1, 3, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 2, 1, 1, 1, 1, 1, 1]
```

Feature Extraction from Text

In the **Scikit-learn Primer** lecture we applied a simple SVC classification model to the SMSSpamCollection dataset. We tried to predict the ham/spam label based on message length and punctuation counts. In this section we'll actually look at the text of each message and try to perform a classification based on content. We'll take advantage of some of scikit-learn's [feature extraction (https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction)](https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction) tools.

Load a dataset

In [49]:
```
# Perform imports and load the dataset:
import numpy as np
import pandas as pd
df=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\smsspamcollecti
df.head()
```

Out[49]:

	label	message	length	punct
0	ham	Go until jurong point, crazy.. Available only ...	111	9
1	ham	Ok lar... Joking wif u oni...	29	6
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	155	6
3	ham	U dun say so early hor... U c already then say...	49	6
4	ham	Nah I don't think he goes to usf, he lives aro...	61	2

In [50]:
```
df.isnull().sum()
```

Out[50]:
```
label      0
message    0
length     0
punct      0
dtype: int64
```

In [51]:
```
df['label'].value_counts()
```

Out[51]:
```
label
ham     4825
spam     747
Name: count, dtype: int64
```

In [52]:
```
from sklearn.model_selection import train_test_split
```

In [53]:
```
X=df['message']
y=df['label']
```

In [55]:
```
X_train, X_test,y_train,y_test=train_test_split(X,y, test_size=0.33, random_sta
```

In [56]:
```
from sklearn.feature_extraction.text import CountVectorizer
```

In [57]:
```
count_vect=CountVectorizer()
```

In [58]:
```
X
```

Out[58]:
```
0       Go until jurong point, crazy.. Available only ...
1                           Ok lar... Joking wif u oni...
2       Free entry in 2 a wkly comp to win FA Cup fina...
3       U dun say so early hor... U c already then say...
4       Nah I don't think he goes to usf, he lives aro...
                              ...
5567    This is the 2nd time we have tried 2 contact u...
5568                 Will ü b going to esplanade fr home?
5569    Pity, * was in mood for that. So...any other s...
5570    The guy did some bitching but I acted like i'd...
5571                           Rofl. Its true to its name
Name: message, Length: 5572, dtype: object
```

In [59]:
```
#FIT VECTORIZE TO TEH DATA(build a vocab and counts the nos of words)
#count_vect.fit(X_train)
#x_train_counts=count_vect.transform(X_train)

#Transform the original text message -->VECTOR
X_train_counts=count_vect.fit_transform(X_train)
```

In [60]:
```
X_train_counts
```

Out[60]:
```
<3733x7082 sparse matrix of type '<class 'numpy.int64'>'
        with 49992 stored elements in Compressed Sparse Row format>
```

In [61]:
```
X_train.shape
```

Out[61]:
```
(3733,)
```

In [62]:
```
X_train_counts.shape
```

Out[62]:
```
(3733, 7082)
```

In [63]:
```
from sklearn.feature_extraction.text import TfidfTransformer
```

In [64]:
```
tfidf_transformer =TfidfTransformer()
```

In [65]:
```
X_train_tfidf=tfidf_transformer.fit_transform(X_train_counts)
```

In [66]:
```
X_train_tfidf.shape
```

Out[66]:
```
(3733, 7082)
```

In [67]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [68]:
```
vectorizer=TfidfVectorizer()
```

In [69]:
```
X_train_tfidf=vectorizer.fit_transform(X_train)
```

In [70]:
```
from sklearn.svm import LinearSVC
```

In [71]:
```
clf=LinearSVC()
```

In [72]:
```
clf.fit(X_train_tfidf, y_train)
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[72]:
```
LinearSVC()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [73]:
```
from sklearn.pipeline import Pipeline
```

In [74]:
```
text_clf=Pipeline([('tfidf',TfidfVectorizer()),('clf',LinearSVC())])
```

In [75]:
```
text_clf.fit(X_train,y_train)
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[75]:
```
Pipeline(steps=[('tfidf', TfidfVectorizer()), ('clf', LinearSVC())])
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [76]:
```
predictions=text_clf.predict(X_test)
```

In [77]:
```
X_test
```

Out[77]:
```
3245    Squeeeeeze!! This is christmas hug.. If u lik ...
944     And also I've sorta blown him off a couple tim...
1044    Mmm thats better now i got a roast down me! i...
2484        Mm have some kanji dont eat anything heavy ok
812     So there's a ring that comes with the guys cos...
                              ...
4944    Check mail.i have mailed varma and kept copy t...
3313    I know you are serving. I mean what are you do...
3652         Want to send me a virtual hug?... I need one
14                    I HAVE A DATE ON SUNDAY WITH WILL!!
4758    hey, looks like I was wrong and one of the kap...
Name: message, Length: 1839, dtype: object
```

In [78]:
```
from sklearn.metrics import confusion_matrix, classification_report
```

In [80]:
```
print(confusion_matrix(y_test,predictions))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

[[1586    7]
 [  12  234]]
```

In [82]:
```
print(classification_report(y_test,predictions))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

              precision    recall  f1-score   support

         ham       0.99      1.00      0.99      1593
        spam       0.97      0.95      0.96       246

    accuracy                           0.99      1839
   macro avg       0.98      0.97      0.98      1839
weighted avg       0.99      0.99      0.99      1839
```

In [83]:
```
from sklearn import metrics
```

In [84]:
```
metrics.accuracy_score(y_test,predictions)
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[84]:
```
0.989668297988037
```

In [86]:
```
text_clf.predict(['Hi how are doing today'])
```

Out[86]:
```
array(['ham'], dtype=object)
```

In [89]:
```
text_clf.predict(['Congratulations ! You have selected as a winner. text to 544
```

Out[89]:
```
array(['spam'], dtype=object)
```

Text Classification code along Project

In [90]:
```
#Task is to predict movie review is postive or negative
import numpy as np
import pandas as pd
```

In [92]:
```
df=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\moviereviews.ts
```

In [93]:
```
df.head()
```

Out[93]:

	label	review
0	neg	how do films like mouse hunt get into theatres...
1	neg	some talented actresses are blessed with a dem...
2	pos	this has been an extraordinary year for austra...
3	pos	according to hollywood movies made in last few...
4	neg	my first press screening of 1998 and already i...

In [94]:
```
len(df)
```

Out[94]:
```
2000
```

In [95]:
```
df['review'][0]
```

Out[95]:
```
'how do films like mouse hunt get into theatres ? \r\nisn\'t there a law or s
omething ? \r\nthis diabolical load of claptrap from steven speilberg\'s drea
mworks studio is hollywood family fare at its deadly worst . \r\nmouse hunt t
akes the bare threads of a plot and tries to prop it up with overacting and f
lat-out stupid slapstick that makes comedies like jingle all the way look dec
ent by comparison . \r\nwriter adam rifkin and director gore verbinski are th
e names chiefly responsible for this swill . \r\nthe plot , for what its wort
h , concerns two brothers ( nathan lane and an appalling lee evens ) who inhe
rit a poorly run string factory and a seemingly worthless house from their ec
centric father . \r\ndeciding to check out the long-abandoned house , they so
on learn that it\'s worth a fortune and set about selling it in auction to th
e highest bidder . \r\nbut battling them at every turn is a very smart mouse
, happy with his run-down little abode and wanting it to stay that way . \r\n
the story alternates between unfunny scenes of the brothers bickering over wh
at to do with their inheritance and endless action sequences as the two take
on their increasingly determined furry foe . \r\nwhatever promise the film st
arts with soon deteriorates into boring dialogue , terrible overacting , and
increasingly uninspired slapstick that becomes all sound and fury , signifyin
g nothing . \r\nthe script becomes so unspeakably bad that the best line poor
lee evens can utter after another run in with the rodent is : " i hate that m
ouse " . \r\noh cringe ! \r\nthis is home alone all over again , and ten time
s worse . \r\none touching scene early on is worth mentioning . \r\nwe follow
the mouse through a maze of walls and pipes until he arrives at his makeshift
abode somewhere in a wall . \r\nhe jumps into a tiny bed , pulls up a makeshi
ft sheet and snuggles up to sleep , seemingly happy and just wanting to be le
ft alone . \r\nit\'s a magical little moment in an otherwise soulless film .
\r\na message to speilberg : if you want dreamworks to be associated with som
e kind of artistic credibility , then either give all concerned in mouse hunt
a swift kick up the arse or hire yourself some decent writers and directors .
\r\nthis kind of rubbish will just not do at all . \r\n'
```

In [96]:
```
df.isnull().sum()
```

Out[96]:
```
label      0
review    35
dtype: int64
```

In [97]:
```
df.dropna(inplace=True)
```

In [98]:
```
df.isnull().sum()
```

Out[98]:
```
label     0
review    0
dtype: int64
```

In [102]:
```
my_string='hello'
empty=' '
```

In [100]:
```
my_string.isspace()
```

Out[100]:
```
False
```

In [103]:
```
empty.isspace()
```

Out[103]:
```
True
```

In [104]:
```
blanks=[]
#(index,label,review text)
for i,lb,rv in df.itertuples():
    if rv.isspace():
        blanks.append(i)
```

In [105]:
```
blanks
```

Out[105]:
```
[57,
 71,
 147,
 151,
 283,
 307,
 313,
 323,
 343,
 351,
 427,
 501,
 633,
 675,
 815,
 851,
 977,
 1079,
 1299,
 1455,
 1493,
 1525,
 1531,
 1763,
 1851,
 1905,
 1993]
```

In [107]:
```
df.drop(blanks,inplace=True)
```

In [108]:
```
len(df)
```

Out[108]:
```
1938
```

In [109]:
```
from sklearn.model_selection import train_test_split
```

In [110]:
```
X=df['review']
```

In [111]:
```
y=df['label']
```

In [112]:
```
X_train,X_test, y_train,y_test=train_test_split(X,y,test_size=0.3,random_state
```

In [113]:
```
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import LinearSVC
```

In [114]:
```
text_clf=Pipeline([('tfidf',TfidfVectorizer()),('clf',LinearSVC())])
```

In [115]:
```
text_clf.fit(X_train, y_train)
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[115]:
```
Pipeline(steps=[('tfidf', TfidfVectorizer()), ('clf', LinearSVC())])
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [116]:
```
predictions=text_clf.predict(X_test)
```

In [117]:
```
from sklearn.metrics import confusion_matrix, classification_report
```

In [118]:
```
print(confusion_matrix(y_test, predictions))
```

```
[[235  47]
 [ 41 259]]

C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [119]:
```
print(classification_report(y_test, predictions))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

              precision    recall  f1-score   support

         neg       0.85      0.83      0.84       282
         pos       0.85      0.86      0.85       300

    accuracy                           0.85       582
   macro avg       0.85      0.85      0.85       582
weighted avg       0.85      0.85      0.85       582
```

In [121]:
```
from sklearn import metrics
```

In [122]:
```
metrics.accuracy_score(y_test,predictions)
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[122]: 0.8487972508591065

Second Project

In [123]:
```
#Task is to predict movie review is postive or negative
import numpy as np
import pandas as pd
```

In [125]:
```
df1=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\moviereviews2
```

In [126]:
```
df1.head()
```

Out[126]:

	label	review
0	pos	I loved this movie and will watch it again. Or...
1	pos	A warm, touching movie that has a fantasy-like...
2	pos	I was not expecting the powerful filmmaking ex...
3	neg	This so-called "documentary" tries to tell tha...
4	pos	This show has been my escape from reality for ...

In [127]:
```
len(df1)
```

Out[127]: 6000

In [130]:
```
df1.isnull().sum()
```

Out[130]:
```
label      0
review    20
dtype: int64
```

In [132]:
```
df1.dropna(inplace=True)
```

In [133]:
```
blanks=[]
#(index,label,review text)
for i,lb,rv in df1.itertuples():
    if rv.isspace():
        blanks.append(i)
```

In [134]:
```
blanks
```

Out[134]:
```
[]
```

In [135]:
```
len(df1)
```

Out[135]:
```
5980
```

In [136]:
```
from sklearn.model_selection import train_test_split
```

In [139]:
```
X=df1['review']
```

In [140]:
```
y=df1['label']
```

In [141]:
```
X_train,X_test, y_train,y_test=train_test_split(X,y,test_size=0.3,random_state
```

In [142]:
```
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import LinearSVC
```

In [143]:
```
text_clf=Pipeline([('tfidf1',TfidfVectorizer()),('clf',LinearSVC())])
```

In [144]:
```
text_clf.fit(X_train, y_train)
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[144]:
```
Pipeline(steps=[('tfidf1', TfidfVectorizer()), ('clf', LinearSVC())])
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [145]:
```
predictions=text_clf.predict(X_test)
```

In [146]:
```
from sklearn.metrics import confusion_matrix, classification_report
```

In [147]:
```
print(confusion_matrix(y_test, predictions))
```

```
[[821  78]
 [ 58 837]]

C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [148]:
```
print(classification_report(y_test, predictions))
```

```
              precision    recall  f1-score   support

         neg       0.93      0.91      0.92       899
         pos       0.91      0.94      0.92       895

    accuracy                           0.92      1794
   macro avg       0.92      0.92      0.92      1794
weighted avg       0.92      0.92      0.92      1794


C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [149]:
```
from sklearn import metrics
```

In [150]:
```
metrics.accuracy_score(y_test,predictions)
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[150]:
```
0.9241917502787068
```

Spacy word vector

In [1]:
```
import spacy.cli
spacy.cli.download("en_core_web_lg")
nlp=spacy.load('en_core_web_lg')
```

```
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_lg')
```

In [2]:
```
nlp(u'lion').vector
```

Out[2]:
```
array([  1.2746  ,   0.46242 ,  -1.1829  ,  -5.2661  ,  -2.7128  ,
         1.8521  ,  -0.94273 ,   2.1865  ,   6.503   ,   0.6704  ,
         1.5361  ,   2.5992  ,  -0.36233 ,   4.3965  ,  -6.5644  ,
         1.6141  ,  -1.2897  ,   2.1184  ,  -0.63654 ,  -3.4572  ,
        -4.3771  ,   4.2074  ,  -3.6411  ,  -0.97214 ,   1.3253  ,
        -2.3125  ,  -3.6531  ,  -2.8398  ,   2.7913  ,  -1.53    ,
        -2.9984  ,  -2.6357  ,   0.50615 ,  -2.6925  ,   4.3401  ,
        -5.6017  ,   0.045691,   4.3832  ,  -0.19535 ,  -1.0751  ,
         0.32172 ,   2.4395  ,   4.6638  ,   3.4471  ,  -3.3847  ,
        -1.8238  ,   0.70212 ,   0.58557 ,   5.0032  ,  -3.1072  ,
         1.2364  ,   7.4595  ,   0.057368,   1.0111  ,  -1.0827  ,
         0.69113 ,   2.8009  ,  -3.4383  ,  -1.0599  ,  -2.2627  ,
        -5.149   ,  -5.0636  ,   3.1405  ,   1.0793  ,  -0.72892 ,
        -3.9939  ,  -0.69551 ,  -0.55767 ,   3.2555  ,  -2.9449  ,
         4.7114  ,   1.6388  ,   1.3828  ,   1.4255  ,  -3.2334  ,
        -2.274   ,  -1.8136  ,   2.2966  ,   2.5462  ,   1.0722  ,
        -0.73447 ,   1.2148  ,  -0.9196  ,  -0.065012,   2.088   ,
         0.57002 ,   3.5746  ,   1.7192  ,  -8.335   ,   0.71079 ,
         0.91314 ,  -5.0107  ,   1.899   ,  -4.4658  ,   4.7993  ,
        -0.39899 ,  -2.673   ,  -2.9354  ,   4.304   ,   1.4336  ,
         3.7121  ,   0.34882 ,   4.6512  ,  -4.5731  ,  -4.5665  ,
         1.5988  ,  -0.50383 ,   0.95857 ,   0.68728 ,  -0.39976 ,
        -3.1922  ,   4.4363  ,  -0.69479 ,  -1.9528  ,   4.9376  ,
         2.7259  ,   2.2485  ,   5.5734  ,   2.5842  ,   4.7836  ,
        -1.0274  ,   2.2703  ,  -2.0696  ,  -1.0642  ,  -4.932   ,
        -2.274   ,   4.1409  ,   0.73313 ,   2.1889  ,  -0.098888,
         1.6472  ,  -2.3985  ,   2.5911  ,   3.6026  ,   1.885   ,
         5.7822  ,  -1.4481  ,   1.8914  , -10.044   ,  -5.7452  ,
        -4.3224  ,  -3.854   ,   2.3084  ,  -0.84018 ,  -0.40526 ,
         4.7741  ,  -2.3271  ,   7.064   ,   0.95753 ,  -2.356   ,
         0.83953 ,   0.40004 ,   0.33743 ,   0.8376  ,   3.9285  ,
         0.05955 ,   2.4422  ,   4.3492  ,   3.9861  ,   2.1043  ,
        -1.0197  ,  -0.61752 ,  -0.42999 ,  -0.1014  ,  -5.9571  ,
        -0.53818 ,  -1.7797  ,   1.7446  ,   2.3934  ,  -0.50263 ,
        -1.6222  ,  -0.37372 ,  -6.8938  ,   0.55018 ,  -2.267   ,
         0.64912 ,   3.1525  ,  -2.2541  ,  -4.0384  ,   3.206   ,
         0.14962 ,  -2.6662  ,   0.18167 ,   5.0028  ,   2.1521  ,
         0.92419 ,   5.4163  ,  -2.2408  ,   1.6585  ,  -5.1625  ,
         5.029   ,   0.1026  ,  -0.44542 ,   2.0557  ,   3.7778  ,
         3.8679  ,  -2.7135  ,   5.3242  ,  -3.2916  ,   5.6421  ,
         5.0466  ,   1.6072  ,  -1.3206  ,   4.2044  ,  -0.33793 ,
        -3.1139  ,   2.8841  ,  -3.1565  ,  -2.9832  ,  -0.23235 ,
         2.3259  ,   3.5477  ,  -2.1299  ,  -1.8344  ,   2.7271  ,
         1.5568  ,   5.6865  ,   0.9412  ,  -2.6412  ,  -5.3254  ,
         1.3494  ,  -0.47159 ,   2.4979  ,  -1.5568  ,  -1.6911  ,
        -2.1842  ,   6.0319  ,   0.022573,   2.3824  ,  -1.1002  ,
         0.90216 ,  -1.9113  ,   1.5527  ,   5.7413  ,  -3.1956  ,
         0.68655 ,  -1.6068  ,   1.7404  ,  -3.2142  ,   6.4783  ,
         1.7548  ,  -2.9795  ,   0.97631 ,  -0.018354,  -0.6379  ,
         0.80559 ,   3.1923  ,   3.3335  ,   4.3068  ,  -1.0819  ,
        -1.3839  ,  -4.7626  ,  -4.6637  ,  -1.2201  ,  -3.2741  ,
         1.5204  ,   0.78119 ,   8.7339  ,   1.6009  ,  -0.79332 ,
         5.8416  ,  -1.485   ,   1.5978  ,   2.9746  ,  -0.30759 ,
        -1.8023  ,  -4.8344  ,   1.2817  ,  -2.5469  ,   2.6517  ,
         1.4881  ,   2.1952  ,  -0.12652 ,   1.2223  ,   0.44763 ,
        -3.1445  ,  -2.2051  ,  -4.1785  ,  -3.6539  ,   5.1929  ,
         0.78457 ,  -1.2312  ,   5.5624  ,  -1.8462  ,   6.1262  ,
```

```
  -1.6653  ,  -2.7557  ,  -0.066465,  -3.6362  ,   5.2005  ,
  -1.2865  ,   2.8855  ,   6.1219  ,   1.7824  ,   1.4264  ,
  10.628   ,  -0.36028 ,   1.9268  ,  -7.835   ,   0.57865 ],
dtype=float32)
```

In [4]:
```
nlp(u'the quick brown fox jumped').vector
```

Out[4]:
```
array([-1.18224001e+00,  1.37251186e+00, -9.51321959e-01,  1.90348044e-01,
        2.23220205e+00, -1.80298805e+00, -5.02611995e-01,  3.87434816e+00,
        1.20122206e+00,  1.17870200e+00,  5.13008022e+00,  2.30294418e+00,
       -3.43085170e+00,  8.87973964e-01,  1.43500805e+00, -2.52099991e-01,
        1.64447999e+00,  5.38545966e-01,  1.38572001e+00, -1.45570815e+00,
       -4.49979961e-01, -2.67203957e-01,  1.12585390e+00, -2.43520021e+00,
        1.16168082e+00,  1.02620004e-02, -3.85214376e+00, -2.39539409e+00,
        1.19893003e+00,  1.58175802e+00, -1.14021003e+00, -4.03898001e-01,
       -1.11064994e+00, -1.65323222e+00, -1.39682198e+00, -2.35803604e+00,
        6.77866042e-01,  1.38088202e+00,  2.82931995e+00, -1.20117211e+00,
        1.05320811e+00,  4.73229170e-01,  1.89112020e+00, -1.00325203e+00,
       -1.31561995e-01,  2.56538808e-01,  1.71212006e+00, -8.41005981e-01,
       -1.82731986e+00,  4.83720005e-01, -4.97079380e-02,  3.82681990e+00,
       -4.06570017e-01, -1.86495930e-01, -1.02143991e+00, -3.80337983e-01,
        1.71682000e-01,  9.62980092e-01,  1.47140396e+00,  6.69384450e-02,
        8.05018067e-01, -2.97483993e+00, -1.34168029e-01,  6.26000047e-01,
       -7.70565987e-01, -6.39220059e-01, -3.13495779e+00, -1.06219399e+00,
        1.14220798e+00,  2.85654008e-01,  7.77800083e-01,  2.59539969e-02,
        7.14900076e-01, -4.59673971e-01,  3.37721974e-01, -1.79241985e-01,
       -1.50710213e+00,  1.19176006e+00, -6.42620265e-01, -1.39475632e+00,
       -3.42332983e+00, -1.77089405e+00,  5.29478081e-02,  2.42739987e+00,
        2.05983996e+00, -6.70646071e-01,  1.04680002e+00,  2.75402074e-03,
       -1.89217794e+00,  4.82790053e-01, -2.82747805e-01,  5.51725864e-01,
        2.66382027e+00, -1.27522445e+00,  6.28995001e-01,  1.07828999e+00,
        2.35402584e+00,  2.05999941e-01,  2.93441796e+00, -2.24882036e-01,
        1.45308399e+00, -9.23099741e-02,  2.74249387e+00,  1.27958405e+00,
       -7.46407807e-01,  2.21395993e+00, -1.81273997e+00,  9.30260003e-01,
       -1.87481189e+00, -7.58661985e-01, -8.02040696e-01, -9.21366096e-01,
       -1.95599914e-01,  9.79211986e-01,  9.90561962e-01,  2.50984013e-01,
        2.31348419e+00, -7.98141003e-01,  1.33613396e+00, -2.29319222e-02,
       -2.32175916e-01,  1.52176023e-01, -2.17740011e+00,  1.94582009e+00,
       -1.95228386e+00, -1.50523400e+00,  1.62459183e+00, -8.71561825e-01,
        4.69488049e+00,  1.21837997e+00, -2.30739784e+00, -1.45342803e+00,
        2.13762021e+00, -1.52957976e-01, -2.13279966e-02,  5.75737953e-01,
       -4.80286032e-01, -3.02331996e+00,  9.43019986e-01, -1.04377997e+00,
       -4.57494020e+00,  4.89240177e-02, -1.69130003e+00,  6.37752056e-01,
       -7.89700031e-01,  1.25091195e+00, -2.19885969e+00,  8.78104031e-01,
        1.71543372e+00,  8.11782002e-01, -8.98997962e-01,  3.59739995e+00,
        7.00516105e-01, -9.37424064e-01,  4.38241184e-01,  9.93759990e-01,
        3.40121222e+00, -1.44444001e+00,  1.15776992e+00,  6.18332744e-01,
        7.11503983e-01, -2.18797779e+00,  1.81680810e+00, -1.69167590e+00,
       -2.66606402e+00, -1.77599899e-02, -2.86205006e+00,  1.18827987e+00,
       -2.71388054e-01,  3.47555965e-01,  1.06615946e-01,  2.34087992e+00,
        2.35564995e+00, -1.12218022e-01, -1.07647002e+00, -1.57532021e-01,
       -4.44869995e-01, -1.99166012e+00, -1.33363998e+00, -3.25271189e-01,
       -1.64782596e+00, -5.63333988e-01, -8.87355357e-02,  1.35562003e+00,
       -4.68459994e-01, -7.51655936e-01, -9.87359881e-01, -1.77089095e+00,
        8.87893975e-01, -2.39797211e+00,  1.15864003e+00,  2.08151007e+00,
       -2.33528063e-01, -1.56540000e+00,  1.09336996e+00, -8.69010091e-01,
       -1.60074008e+00, -1.18240472e-02, -1.70169199e+00,  8.91156018e-01,
       -1.57864064e-01, -9.28445935e-01, -1.04275191e+00, -2.20953989e+00,
        1.07129979e+00,  4.11802381e-01, -3.19541168e+00,  4.25209999e-01,
        4.57892001e-01,  9.05926824e-01,  9.55353916e-01,  1.76145196e+00,
       -2.09932017e+00,  6.11521184e-01,  1.14155555e+00,  8.64779949e-01,
        3.06578588e+00, -9.32515919e-01, -2.23505211e+00,  5.51597103e-02,
       -2.48792791e+00,  1.23801589e+00, -8.07563961e-01,  1.85346007e+00,
       -1.75209200e+00,  3.61921877e-01,  5.31363010e-01,  2.04426193e+00,
```

```
 2.58960843e+00,  8.87745976e-01,  1.52296901e+00, -3.47569013e+00,
 1.53510594e+00,  1.78362012e+00, -6.50260091e-01,  1.51403844e+00,
-1.21742392e+00,  8.38418007e-01,  1.39599796e-02,  1.95636010e+00,
 5.98335981e-01, -1.55487192e+00,  1.00260007e+00, -5.04159816e-02,
-1.00793841e-03,  1.43645990e+00, -1.34034181e+00,  9.29039717e-02,
-3.09747994e-01, -4.44676012e-01, -1.69196391e+00, -1.97914410e+00,
-3.76150370e+00, -8.31771195e-01,  1.75440311e-02, -1.33436882e+00,
 8.00455928e-01,  4.29681587e+00, -9.22177970e-01,  3.25948030e-01,
 2.48811388e+00,  1.94026399e+00,  9.14101243e-01,  1.31440103e+00,
-4.59141910e-01, -1.68114603e+00, -2.33204389e+00, -2.92747885e-01,
-2.69432592e+00,  5.43431997e-01,  1.90584981e+00, -3.94615978e-01,
 1.07552016e+00, -1.54083407e+00, -7.39986002e-01, -2.44169784e+00,
 5.72397947e-01, -1.63752007e+00, -4.27471966e-01,  1.89340782e+00,
-9.70826030e-01,  1.96122020e-01,  9.99798000e-01, -4.63580042e-01,
 1.97484016e+00, -3.29501957e-01, -8.72795939e-01, -9.23940063e-01,
-6.44327998e-01,  5.83415985e-01, -8.48168552e-01,  7.95794010e-01,
 2.25744200e+00,  4.86402035e-01, -8.06441963e-01,  2.12743402e+00,
 6.97450042e-01,  2.01827958e-01, -3.40572023e+00, -2.99616009e-01],
dtype=float32)
```

In [5]:
```
nlp(u'the quick brown fox jumped').vector.shape
```

Out[5]:
```
(300,)
```

In [6]:
```
nlp(u'fox').vector.shape
```

Out[6]:
```
(300,)
```

In [7]:
```
tokens=nlp(u'lion cat pet')
```

In [9]:
```
for token1 in tokens:
    for tokens2 in tokens:
        print(token1.text,tokens2.text,token1.similarity(tokens2))
```

```
lion lion 1.0
lion cat 0.3854507803916931
lion pet 0.20031583309173584
cat lion 0.3854507803916931
cat cat 1.0
cat pet 0.732966423034668
pet lion 0.20031583309173584
pet cat 0.732966423034668
pet pet 1.0
```

In [10]:
```
tokens=nlp(u'like love hate')
```

In [11]:
```
for token1 in tokens:
    for tokens2 in tokens:
        print(token1.text,tokens2.text,token1.similarity(tokens2))
```

```
like like 1.0
like love 0.5212638974189758
like hate 0.5065140724182129
love like 0.5212638974189758
love love 1.0
love hate 0.5708349943161011
hate like 0.5065140724182129
hate love 0.5708349943161011
hate hate 1.0
```

In [13]:
```
len(nlp.vocab.vectors)
```

Out[13]: 514157

In [14]:
```
tokens=nlp(u'dog cat nargle')
```

In [15]:
```
for token in tokens:
    print(token.text, token.has_vector,token.vector_norm,token.is_oov)
```

```
dog True 75.254234 False
cat True 63.188496 False
nargle False 0.0 True
```

Sentiment analysis

In [1]:
```
import nltk
```

In [2]:
```
nltk.download('vader_lexicon')
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     C:\Users\nilesh\AppData\Roaming\nltk_data...
```

Out[2]: True

In [3]:
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

In [4]:
```
sid=SentimentIntensityAnalyzer()
```

In [5]:
```
a='This is a good movie'
```

In [6]:
```
sid.polarity_scores(a)
```

Out[6]: {'neg': 0.0, 'neu': 0.508, 'pos': 0.492, 'compound': 0.4404}

In [7]:
```
a='This is the best movie i have EVER MADE!!!'
```

In [8]:
```
sid.polarity_scores(a)
```

Out[8]:
```
{'neg': 0.0, 'neu': 0.58, 'pos': 0.42, 'compound': 0.7249}
```

In [9]:
```
a='This is the worst movie in the history of cinema'
```

In [10]:
```
sid.polarity_scores(a)
```

Out[10]:
```
{'neg': 0.313, 'neu': 0.687, 'pos': 0.0, 'compound': -0.6249}
```

In [11]:
```
import pandas as pd
```

In [13]:
```
df=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\amazonreviews.t
```

In [14]:
```
df.head()
```

Out[14]:

	label	review
0	pos	Stuning even for the non-gamer: This sound tra...
1	pos	The best soundtrack ever to anything.: I'm rea...
2	pos	Amazing!: This soundtrack is my favorite music...
3	pos	Excellent Soundtrack: I truly like this soundt...
4	pos	Remember, Pull Your Jaw Off The Floor After He...

In [15]:
```
df['label'].value_counts()
```

Out[15]:
```
label
neg    5097
pos    4903
Name: count, dtype: int64
```

In [16]:
```
df.dropna(inplace=True)
```

In [18]:
```
blanks=[]
for i,lb,rv in df.itertuples():
    #(index,label,review)
    if type(rv)==str:
        if rv.isspace():
            blanks.append(i)
```

In [19]:
```
blanks
```

Out[19]:
```
[]
```

In [20]:
```
#df.drop(blanks, inplace=True)
```

In [21]:
```
df.iloc[0]['review']
```

Out[21]:
```
'Stuning even for the non-gamer: This sound track was beautiful! It paints th
e senery in your mind so well I would recomend it even to people who hate vi
d. game music! I have played the game Chrono Cross but out of all of the game
s I have ever played it has the best music! It backs away from crude keyboard
ing and takes a fresher step with grate guitars and soulful orchestras. It wo
uld impress anyone who cares to listen! ^_^'
```

In [22]:
```
sid.polarity_scores(df.iloc[0]['review'])
```

Out[22]:
```
{'neg': 0.088, 'neu': 0.669, 'pos': 0.243, 'compound': 0.9454}
```

In [23]:
```
df['scores']=df['review'].apply(lambda review:sid.polarity_scores(review))
```

In [24]:
```
df.head()
```

Out[24]:

	label	review	scores
0	pos	Stuning even for the non-gamer: This sound tra...	{'neg': 0.088, 'neu': 0.669, 'pos': 0.243, 'co...
1	pos	The best soundtrack ever to anything.: I'm rea...	{'neg': 0.018, 'neu': 0.837, 'pos': 0.145, 'co...
2	pos	Amazing!: This soundtrack is my favorite music...	{'neg': 0.04, 'neu': 0.692, 'pos': 0.268, 'com...
3	pos	Excellent Soundtrack: I truly like this soundt...	{'neg': 0.09, 'neu': 0.615, 'pos': 0.295, 'com...
4	pos	Remember, Pull Your Jaw Off The Floor After He...	{'neg': 0.0, 'neu': 0.746, 'pos': 0.254, 'comp...

In [25]:
```
df['compound']=df['scores'].apply(lambda d:d['compound'])
```

In [26]:
```
df.head()
```

Out[26]:

	label	review	scores	compound
0	pos	Stuning even for the non-gamer: This sound tra...	{'neg': 0.088, 'neu': 0.669, 'pos': 0.243, 'co...	0.9454
1	pos	The best soundtrack ever to anything.: I'm rea...	{'neg': 0.018, 'neu': 0.837, 'pos': 0.145, 'co...	0.8957
2	pos	Amazing!: This soundtrack is my favorite music...	{'neg': 0.04, 'neu': 0.692, 'pos': 0.268, 'com...	0.9858
3	pos	Excellent Soundtrack: I truly like this soundt...	{'neg': 0.09, 'neu': 0.615, 'pos': 0.295, 'com...	0.9814
4	pos	Remember, Pull Your Jaw Off The Floor After He...	{'neg': 0.0, 'neu': 0.746, 'pos': 0.254, 'comp...	0.9781

In [27]:
```
df.tail()
```

Out[27]:

	label	review	scores	compound
9995	pos	A revelation of life in small town America in ...	{'neg': 0.017, 'neu': 0.846, 'pos': 0.136, 'co...	0.9610
9996	pos	Great biography of a very interesting journali...	{'neg': 0.0, 'neu': 0.868, 'pos': 0.132, 'comp...	0.9544
9997	neg	Interesting Subject; Poor Presentation: You'd ...	{'neg': 0.084, 'neu': 0.754, 'pos': 0.162, 'co...	0.9102
9998	neg	Don't buy: The box looked used and it is obvio...	{'neg': 0.091, 'neu': 0.909, 'pos': 0.0, 'comp...	-0.3595
9999	pos	Beautiful Pen and Fast Delivery.: The pen was ...	{'neg': 0.028, 'neu': 0.811, 'pos': 0.161, 'co...	0.9107

In [28]:
```
df['comp_score']=df['compound'].apply(lambda score:'pos' if score >=0 else 'neg
```

In [29]:
```
df.head()
```

Out[29]:

	label	review	scores	compound	comp_score
0	pos	Stuning even for the non-gamer: This sound tra...	{'neg': 0.088, 'neu': 0.669, 'pos': 0.243, 'co...	0.9454	pos
1	pos	The best soundtrack ever to anything.: I'm rea...	{'neg': 0.018, 'neu': 0.837, 'pos': 0.145, 'co...	0.8957	pos
2	pos	Amazing!: This soundtrack is my favorite music...	{'neg': 0.04, 'neu': 0.692, 'pos': 0.268, 'com...	0.9858	pos
3	pos	Excellent Soundtrack: I truly like this soundt...	{'neg': 0.09, 'neu': 0.615, 'pos': 0.295, 'com...	0.9814	pos
4	pos	Remember, Pull Your Jaw Off The Floor After He...	{'neg': 0.0, 'neu': 0.746, 'pos': 0.254, 'comp...	0.9781	pos

In [30]:
```
df.tail()
```

Out[30]:

	label	review	scores	compound	comp_score
9995	pos	A revelation of life in small town America in ...	{'neg': 0.017, 'neu': 0.846, 'pos': 0.136, 'co...	0.9610	pos
9996	pos	Great biography of a very interesting journali...	{'neg': 0.0, 'neu': 0.868, 'pos': 0.132, 'comp...	0.9544	pos
9997	neg	Interesting Subject; Poor Presentation: You'd ...	{'neg': 0.084, 'neu': 0.754, 'pos': 0.162, 'co...	0.9102	pos
9998	neg	Don't buy: The box looked used and it is obvio...	{'neg': 0.091, 'neu': 0.909, 'pos': 0.0, 'comp...	-0.3595	neg
9999	pos	Beautiful Pen and Fast Delivery.: The pen was ...	{'neg': 0.028, 'neu': 0.811, 'pos': 0.161, 'co...	0.9107	pos

In [31]:
```
from sklearn.metrics import accuracy_score, classification_report, confusion_ma
```

In [32]:
```
accuracy_score(df['label'],df['comp_score'])
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[32]:
```
0.7097
```

In [33]:
```
print(classification_report(df['label'],df['comp_score']))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

	precision	recall	f1-score	support
neg	0.86	0.52	0.64	5097
pos	0.64	0.91	0.75	4903
accuracy			0.71	10000
macro avg	0.75	0.71	0.70	10000
weighted avg	0.75	0.71	0.70	10000

In [35]:
```
print(confusion_matrix(df['label'],df['comp_score']))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

[[2629 2468]
 [ 435 4468]]
```

In [1]:
```
import numpy as np
import pandas as pd
```

In [2]:
```
df=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\moviereviews.ts
```

In [3]:
```
df.head()
```

Out[3]:

	label	review
0	neg	how do films like mouse hunt get into theatres...
1	neg	some talented actresses are blessed with a dem...
2	pos	this has been an extraordinary year for austra...
3	pos	according to hollywood movies made in last few...
4	neg	my first press screening of 1998 and already i...

In [4]:
```
df.dropna(inplace=True)
```

In [5]:
```
blanks=[]
for i, lb, rv in df.itertuples():
    if type(rv)==str:
        if rv.isspace():
            blanks.append(i)
```

In [6]:
```
blanks
```

Out[6]:
```
[57,
 71,
 147,
 151,
 283,
 307,
 313,
 323,
 343,
 351,
 427,
 501,
 633,
 675,
 815,
 851,
 977,
 1079,
 1299,
 1455,
 1493,
 1525,
 1531,
 1763,
 1851,
 1905,
 1993]
```

In [8]:
```
df.iloc[0]['review']
```

Out[8]:
```
'how do films like mouse hunt get into theatres ? \r\nisn\'t there a law or s
omething ? \r\nthis diabolical load of claptrap from steven speilberg\'s drea
mworks studio is hollywood family fare at its deadly worst . \r\nmouse hunt t
akes the bare threads of a plot and tries to prop it up with overacting and f
lat-out stupid slapstick that makes comedies like jingle all the way look dec
ent by comparison . \r\nwriter adam rifkin and director gore verbinski are th
e names chiefly responsible for this swill . \r\nthe plot , for what its wort
h , concerns two brothers ( nathan lane and an appalling lee evens ) who inhe
rit a poorly run string factory and a seemingly worthless house from their ec
centric father . \r\ndeciding to check out the long-abandoned house , they so
on learn that it\'s worth a fortune and set about selling it in auction to th
e highest bidder . \r\nbut battling them at every turn is a very smart mouse
, happy with his run-down little abode and wanting it to stay that way . \r\n
the story alternates between unfunny scenes of the brothers bickering over wh
at to do with their inheritance and endless action sequences as the two take
on their increasingly determined furry foe . \r\nwhatever promise the film st
arts with soon deteriorates into boring dialogue , terrible overacting , and
increasingly uninspired slapstick that becomes all sound and fury , signifyin
g nothing . \r\nthe script becomes so unspeakably bad that the best line poor
lee evens can utter after another run in with the rodent is : " i hate that m
ouse " . \r\noh cringe ! \r\nthis is home alone all over again , and ten time
s worse . \r\none touching scene early on is worth mentioning . \r\nwe follow
the mouse through a maze of walls and pipes until he arrives at his makeshift
abode somewhere in a wall . \r\nhe jumps into a tiny bed , pulls up a makeshi
ft sheet and snuggles up to sleep , seemingly happy and just wanting to be le
ft alone . \r\nit\'s a magical little moment in an otherwise soulless film .
\r\na message to speilberg : if you want dreamworks to be associated with som
e kind of artistic credibility , then either give all concerned in mouse hunt
a swift kick up the arse or hire yourself some decent writers and directors .
\r\nthis kind of rubbish will just not do at all . \r\n'
```

In [9]:
```
df.iloc[57]['review']
```

Out[9]:
```
'  '
```

In [10]:
```
df.drop(blanks,inplace=True)
```

In [11]:
```
df['label'].value_counts()
```

Out[11]:
```
label
neg    969
pos    969
Name: count, dtype: int64
```

In [12]:
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

In [13]:
```
sid=SentimentIntensityAnalyzer()
```

In [14]:
```
df['scores']=df['review'].apply(lambda review:sid.polarity_scores(review))
```

In [15]:
```
df.head()
```

Out[15]:

	label	review	scores
0	neg	how do films like mouse hunt get into theatres...	{'neg': 0.121, 'neu': 0.778, 'pos': 0.101, 'co...
1	neg	some talented actresses are blessed with a dem...	{'neg': 0.12, 'neu': 0.775, 'pos': 0.105, 'com...
2	pos	this has been an extraordinary year for austra...	{'neg': 0.068, 'neu': 0.781, 'pos': 0.15, 'com...
3	pos	according to hollywood movies made in last few...	{'neg': 0.071, 'neu': 0.782, 'pos': 0.147, 'co...
4	neg	my first press screening of 1998 and already i...	{'neg': 0.091, 'neu': 0.817, 'pos': 0.093, 'co...

In [16]:
```
df['compound']=df['scores'].apply(lambda d:d['compound'])
```

In [17]:
```
df.head()
```

Out[17]:

	label	review	scores	compound
0	neg	how do films like mouse hunt get into theatres...	{'neg': 0.121, 'neu': 0.778, 'pos': 0.101, 'co...	-0.9125
1	neg	some talented actresses are blessed with a dem...	{'neg': 0.12, 'neu': 0.775, 'pos': 0.105, 'com...	-0.8618
2	pos	this has been an extraordinary year for austra...	{'neg': 0.068, 'neu': 0.781, 'pos': 0.15, 'com...	0.9951
3	pos	according to hollywood movies made in last few...	{'neg': 0.071, 'neu': 0.782, 'pos': 0.147, 'co...	0.9972
4	neg	my first press screening of 1998 and already i...	{'neg': 0.091, 'neu': 0.817, 'pos': 0.093, 'co...	-0.2484

In [18]:
```
df['comp_score']=df['compound'].apply(lambda score:'pos' if score >=0 else 'neg
```

In [19]:
```
df.head()
```

Out[19]:

	label	review	scores	compound	comp_score
0	neg	how do films like mouse hunt get into theatres...	{'neg': 0.121, 'neu': 0.778, 'pos': 0.101, 'co...	-0.9125	neg
1	neg	some talented actresses are blessed with a dem...	{'neg': 0.12, 'neu': 0.775, 'pos': 0.105, 'com...	-0.8618	neg
2	pos	this has been an extraordinary year for austra...	{'neg': 0.068, 'neu': 0.781, 'pos': 0.15, 'com...	0.9951	pos
3	pos	according to hollywood movies made in last few...	{'neg': 0.071, 'neu': 0.782, 'pos': 0.147, 'co...	0.9972	pos
4	neg	my first press screening of 1998 and already i...	{'neg': 0.091, 'neu': 0.817, 'pos': 0.093, 'co...	-0.2484	neg

In [20]:
```
from sklearn.metrics import accuracy_score, classification_report, confusion_ma
```

In [21]:
```python
accuracy_score(df['label'],df['comp_score'])
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[21]: 0.6357069143446853

In [22]:
```
print(classification_report(df['label'],df['comp_score']))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

	precision	recall	f1-score	support
neg	0.72	0.44	0.55	969
pos	0.60	0.83	0.70	969
accuracy			0.64	1938
macro avg	0.66	0.64	0.62	1938
weighted avg	0.66	0.64	0.62	1938

In [23]:
```
print(confusion_matrix(df['label'],df['comp_score']))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

[[427 542]
 [164 805]]
```

Kaggle projects

In [26]:
```
df1=pd.read_csv(r'C:\Users\nilesh\Downloads\IMDB Dataset.csv')
```

In [27]:
```
df1.head()
```

Out[27]:

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

In [28]:
```
df1.dropna(inplace=True)
```

In [29]:
```
blanks=[]
for i, lb, rv in df1.itertuples():
    if type(rv)==str:
        if rv.isspace():
            blanks.append(i)
```

In [30]:
```
blanks
```

Out[30]:
```
[]
```

In [31]:
```
df1.drop(blanks,inplace=True)
```

In [32]:
```
df1['sentiment'].value_counts()
```

Out[32]:
```
sentiment
positive    25000
negative    25000
Name: count, dtype: int64
```

In [33]:
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

In [34]:
```
sid=SentimentIntensityAnalyzer()
```

In [38]:
```
df1['scores']=df1['review'].apply(lambda review:sid.polarity_scores(review))
```

In [39]:
```
df1.head()
```

Out[39]:

	review	sentiment	scores
0	One of the other reviewers has mentioned that ...	positive	{'neg': 0.203, 'neu': 0.748, 'pos': 0.048, 'co...
1	A wonderful little production. The...	positive	{'neg': 0.053, 'neu': 0.776, 'pos': 0.172, 'co...
2	I thought this was a wonderful way to spend ti...	positive	{'neg': 0.094, 'neu': 0.714, 'pos': 0.192, 'co...
3	Basically there's a family where a little boy ...	negative	{'neg': 0.138, 'neu': 0.797, 'pos': 0.065, 'co...
4	Petter Mattei's "Love in the Time of Money" is...	positive	{'neg': 0.052, 'neu': 0.801, 'pos': 0.147, 'co...

In [40]:
```
df1['compound']=df1['scores'].apply(lambda d:d['compound'])
```

In [41]:
```
df1.head()
```

Out[41]:

	review	sentiment	scores	compound
0	One of the other reviewers has mentioned that ...	positive	{'neg': 0.203, 'neu': 0.748, 'pos': 0.048, 'co...	-0.9951
1	A wonderful little production. The...	positive	{'neg': 0.053, 'neu': 0.776, 'pos': 0.172, 'co...	0.9641
2	I thought this was a wonderful way to spend ti...	positive	{'neg': 0.094, 'neu': 0.714, 'pos': 0.192, 'co...	0.9605
3	Basically there's a family where a little boy ...	negative	{'neg': 0.138, 'neu': 0.797, 'pos': 0.065, 'co...	-0.9213
4	Petter Mattei's "Love in the Time of Money" is...	positive	{'neg': 0.052, 'neu': 0.801, 'pos': 0.147, 'co...	0.9744

In [47]:
```
df1['comp_score']=df1['compound'].apply(lambda score:'positive' if score >=0 e
```

In [48]:
```
df1.head()
```

Out[48]:

	review	sentiment	scores	compound	comp_score
0	One of the other reviewers has mentioned that ...	positive	{'neg': 0.203, 'neu': 0.748, 'pos': 0.048, 'co...	-0.9951	negative
1	A wonderful little production. The...	positive	{'neg': 0.053, 'neu': 0.776, 'pos': 0.172, 'co...	0.9641	positive
2	I thought this was a wonderful way to spend ti...	positive	{'neg': 0.094, 'neu': 0.714, 'pos': 0.192, 'co...	0.9605	positive
3	Basically there's a family where a little boy ...	negative	{'neg': 0.138, 'neu': 0.797, 'pos': 0.065, 'co...	-0.9213	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive	{'neg': 0.052, 'neu': 0.801, 'pos': 0.147, 'co...	0.9744	positive

In [49]:
```
from sklearn.metrics import accuracy_score, classification_report, confusion_ma
```

In [50]:
```
accuracy_score(df1['sentiment'],df1['comp_score'])
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[50]: 0.69626

In [51]:
```
print(classification_report(df1['sentiment'],df1['comp_score']))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

	precision	recall	f1-score	support
negative	0.79	0.54	0.64	25000
positive	0.65	0.86	0.74	25000
accuracy			0.70	50000
macro avg	0.72	0.70	0.69	50000
weighted avg	0.72	0.70	0.69	50000

In [52]:
```
print(confusion_matrix(df1['sentiment'],df1['comp_score']))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

[[13410 11590]
 [ 3597 21403]]
```

LDA-- Latent dirichlet allocation

In [1]:
```
import pandas as pd
```

In [2]:
```
npr=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\05-Topic-Modeling\npr.cs
```

In [3]:
```
npr.head()
```

Out[3]:

	Article
0	In the Washington of 2016, even when the polic...
1	Donald Trump has used Twitter — his prefe...
2	Donald Trump is unabashedly praising Russian...
3	Updated at 2:50 p. m. ET, Russian President Vl...
4	From photography, illustration and video, to d...

In [54]:
```
npr['Article'][0]
```

Out[54]: 'In the Washington of 2016, even when the policy can be bipartisan, the politics cannot. And in that sense, this year shows little sign of ending on Dec. 31. When President Obama moved to sanction Russia over its alleged interference in the U. S. election just concluded, some Republicans who had long called for similar or more severe measures could scarcely bring themselves to approve. House Speaker Paul Ryan called the Obama measures ”appropriate” but also ”overdue” and ”a prime example of this administration’s ineffective foreign policy that has left America weaker in the eyes of the world.” Other GOP leaders sounded much the same theme. ”[We have] been urging President Obama for years to take strong action to deter Russia’s worldwide aggression, including its operations,” wrote Rep. Devin Nunes, . chairman of the House Intelligence Committee. ”Now with just a few weeks left in office, the president has suddenly decided that some stronger measures are indeed warranted.” Appearing on CNN, frequent Obama critic Trent Franks, . called for ”much tougher” actions and said three times that Obama had ”finally found his tongue.” Meanwhile, at and on Fox News, various spokesmen for Trump said Obama’s real target was not the Russians at all but the man poised to take over the White House in less than three weeks. They spoke of Obama trying to ”tie Trump’s hands” or ”box him in,” meaning the would be forced either to keep the sanctions or be at odds with Republicans who want to be tougher still on Moscow. Throughout 2016, Trump has repeatedly called not for sanctions but for closer ties with Russia, including cooperation in the fight against ISIS. Russia has battled ISIS in Syria on behalf of that country’s embattled dictator, Bashar Assad, bombing the besieged city of Aleppo that fell to Assad’s forces this week. During the campaign, Trump even urged Russia to ”find” missing emails from the private server of his opponent, Hillary Clinton. He has exchanged public enco miums with Russian President Vladimir Putin on several occasions and added his doubts about the current U. S. levels of support for NATO — Putin’s longtime nemesis. There have also been suggestions that Trump’s extensive business dealings with various Russians are the reason he refuses to release his tax returns. All those issues have been disquieting to some Republicans for many months. Sens. John McCain, . and Lindsay Graham, . C. prominent senior members of the Armed Services Committee, have accepted the assessment of 17 U. S. intelligence agencies regarding the role of Russia in the hacking of various Democratic committees last year. That includes the FBI and CIA consensus that the Russian goal was not just to discredit American democracy but to defeat Clinton and elect Trump. They say the great majority of their Senate colleagues agree with them, and McCain has slated an Armed Services hearing on cyberthreats for Jan. 5. But the politicizing of the Russian actions — the idea that they helped Trump win — has also made the issue difficult for Republican leaders. It has allowed Trump supporters to push back on the intelligence agencies and say the entire issue is designed to undermine Trump’s legitimacy. Senate Majority Leader Mitch McConnell has so far resisted calls for a select committee to look into the Russian interference in the 2016 campaign. He has said it is enough for Sen. Richard Burr, . C. to look into it as chairman of the Senate Intelligence Committee. Typically, Republican leaders and spokesmen say there is no evidence that the actual voting or tallying on Nov. 8 was compromised, and that is true. But it is also a red herring, as interference in those functions has not been alleged and is not the focus of the U. S. intelligence agencies’ concern. For his part, Trump has shown little interest in delving into what happened. He has cast doubt on the U. S. intelligence reports to date and suggested ”no one really knows what happened.” He also has suggested that computers make it very difficult to know who is using them. This week, Trump said it was time to ”get on with our lives and do more important things.” However, at week’s end he did agree to have an intelligence briefing on the subject next week. The has not wanted the daily intelligence briefings available to him in recent weeks, preferring that they be given to the

```
men he has chosen as his vice president (Mike Pence) and national security ad
viser (Mike Flynn) with Trump taking them only occasionally. The irony of thi
s controversy arising at the eleventh hour of the Obama presidency can scarce
ly be overstated, and it defines the dilemma facing both the outgoing preside
nt and the incoming party in control. Obama appears to have been reluctant to
retaliate against the Russian hacking before the election for fear of seeming
to interfere with the election himself. The Republicans, meanwhile, have for
years called for greater confrontation with the Russians, with Obama usually
resisting. Obama did join with NATO in punishing the Russians with economic s
anctions over the annexation of Crimea. Those sanctions may have been painfu
l, coming as they did alongside falling prices for oil  —  the commodity tha
t keeps the Russian economy afloat. On other occasions, despite Russian provo
cations through surrogates in Syria and elsewhere, Obama did not make overt m
oves to force Russia's hand. That includes occasions when Russia was believed
to be hacking critical computer systems in neighboring Ukraine, Estonia and P
oland. But this week, following a chorus of confirmation from the U. S. intel
ligence community regarding the Russian role in computer hacking in the polit
ical campaign, Obama acted. He imposed a set of mostly diplomatic actions suc
h as sanctioning some Russian officials, closing two diplomatic compounds and
expelling 35 Russian diplomats. There may have been more damaging measures ta
ken covertly, and some Russophobes in Washington held out hope for that. But
the visible portion of the program scarcely amounted to major retribution. An
d Putin saw fit to diminish the Obama sanctions further by declining to respo
nd. Although his government has steadfastly denied any interference in the U.
S. election, Putin rejected his own foreign minister's recommended package of
responses. (He even sent an invitation for U. S. diplomats to send their chil
dren to a holiday party in Moscow.) That allowed Putin to appear for the mome
nt to be "the bigger man," even as he spurned Obama and kept up what has look
ed like a public bromance with Trump, who tweeted: "Great move on delay (by
V. Putin)  I always knew he was very smart!" At the moment it may seem that
the overall Russia question amounts to the first crisis facing the Trump pres
idency. Whether forced by this campaign interference issue or not, Trump must
grasp the nettle of a relationship Mitt Romney once called the greatest threa
t to U. S. security in the world. To be sure, Trump needs to dispel doubts ab
out his ability to stand up to Putin, who has bullied and cajoled his way to
center stage in recent world affairs. But Trump also seems determined to turn
the page on past U. S. commitments, from free trade philosophy to funding of
NATO and the United Nations. And if his Twitter account is any guide, Trump s
hows little concern about the conundrum others perceive to be facing him. Abo
ve all, Trump has shown himself determined to play by his own rules. A year a
go, many were confident that would not work for him in the world of president
ial politics. We are about to find out whether it works for him in the Oval O
ffice.'
```

In [5]:
```
from sklearn.feature_extraction.text import CountVectorizer
```

In [6]:
```
cv=CountVectorizer(max_df=0.9,min_df=2,stop_words='english')
```

In [7]:
```
dtm=cv.fit_transform(npr['Article'])
```

In [8]:
```
dtm
```

Out[8]:
```
<11992x54777 sparse matrix of type '<class 'numpy.int64'>'
        with 3033388 stored elements in Compressed Sparse Row format>
```

In [9]:
```
from sklearn.decomposition import LatentDirichletAllocation
```

In [10]:
```
LDA=LatentDirichletAllocation(n_components=7, random_state=42)
```

In [11]:
```
LDA.fit(dtm)
```

Out[11]:
```
LatentDirichletAllocation(n_components=7, random_state=42)
```
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [12]:
```
#Grab the Vocabulary of words
```

In [22]:
```
len(cv.get_feature_names_out())
```

Out[22]:
```
54777
```

In [23]:
```
type(cv.get_feature_names_out())
```

Out[23]:
```
numpy.ndarray
```

In [35]:
```
import random
id=random.randint(0,54777)
cv.get_feature_names_out()[id]
```

Out[35]:
```
'complimentary'
```

In [36]:
```
#Grab the topic
```

In [37]:
```
len(LDA.components_)
```

Out[37]:
```
7
```

In [39]:
```
type(LDA.components_)
```

Out[39]:
```
numpy.ndarray
```

In [40]:
```
LDA.components_.shape
```

Out[40]:
```
(7, 54777)
```

In [41]:
```
LDA.components_
```

Out[41]:
```
array([[8.64332806e+00, 2.38014333e+03, 1.42900522e-01, ...,
        1.43006821e-01, 1.42902042e-01, 1.42861626e-01],
       [2.76191749e+01, 5.36394437e+02, 1.42857148e-01, ...,
        1.42861973e-01, 1.42857147e-01, 1.42906875e-01],
       [7.22783888e+00, 8.24033986e+02, 1.42857148e-01, ...,
        6.14236247e+00, 2.14061364e+00, 1.42923753e-01],
       ...,
       [3.11488651e+00, 3.50409655e+02, 1.42857147e-01, ...,
        1.42859912e-01, 1.42857146e-01, 1.42866614e-01],
       [4.61486388e+01, 5.14408600e+01, 3.14281373e+00, ...,
        1.43107628e-01, 1.43902481e-01, 2.14271779e+00],
       [4.93991422e-01, 4.18841042e+02, 1.42857151e-01, ...,
        1.42857146e-01, 1.43760101e-01, 1.42866201e-01]])
```

In [42]:
```
#Grab the highest probability word per topic
```

In [43]:
```
single_topic=LDA.components_[0]
```

In [44]:
```
single_topic.argsort()
```

Out[44]:
```
array([ 2475, 18302, 35285, ..., 22673, 42561, 42993], dtype=int64)
```

In [45]:
```
import numpy as np
arr=np.array([10,200,1])
```

In [46]:
```
arr
```

Out[46]:
```
array([ 10, 200,   1])
```

In [47]:
```
arr.argsort()
```

Out[47]:
```
array([2, 0, 1], dtype=int64)
```

In [48]:
```
#ARGSORT --INDEX POSITION SORTED FROM LEAST--->GREATEST
#Top 10 values (10 Greatest values)
#Last 10 values of ARGSORT
single_topic.argsort()[-10:]#grab teh last 10 values of argsort()
```

Out[48]:
```
array([33390, 36310, 21228, 10425, 31464,  8149, 36283, 22673, 42561,
       42993], dtype=int64)
```

In [51]:
```
top_ten_words=single_topic.argsort()[-10:]
```

In [52]:
```
for index in top_ten_words:
    print(cv.get_feature_names_out()[index])
```

```
new
percent
government
company
million
care
people
health
said
says
```

In [55]:
```
top_twenty_words=single_topic.argsort()[-20:]
```

In [57]:
```
for index in top_twenty_words:
    print(cv.get_feature_names_out()[index])
```

```
president
state
tax
insurance
trump
companies
money
year
federal
000
new
percent
government
company
million
care
people
health
said
says
```

In [58]:
```
#GRAB the highest probability words per topic
```

In [60]:
```
for i, topic in enumerate(LDA.components_):
    print(f"THE TOP 15 WORDS FOR TOPIC #{i}")
    print([cv.get_feature_names_out()[index] for index in topic.argsort()[-15:]
    print('\n')
    print('\n')
```

```
THE TOP 15 WORDS FOR TOPIC #0
['companies', 'money', 'year', 'federal', '000', 'new', 'percent', 'governmen
t', 'company', 'million', 'care', 'people', 'health', 'said', 'says']



THE TOP 15 WORDS FOR TOPIC #1
['military', 'house', 'security', 'russia', 'government', 'npr', 'reports',
'says', 'news', 'people', 'told', 'police', 'president', 'trump', 'said']



THE TOP 15 WORDS FOR TOPIC #2
['way', 'world', 'family', 'home', 'day', 'time', 'water', 'city', 'new', 'ye
ars', 'food', 'just', 'people', 'like', 'says']



THE TOP 15 WORDS FOR TOPIC #3
['time', 'new', 'don', 'years', 'medical', 'disease', 'patients', 'just', 'ch
ildren', 'study', 'like', 'women', 'health', 'people', 'says']



THE TOP 15 WORDS FOR TOPIC #4
['voters', 'vote', 'election', 'party', 'new', 'obama', 'court', 'republica
n', 'campaign', 'people', 'state', 'president', 'clinton', 'said', 'trump']



THE TOP 15 WORDS FOR TOPIC #5
['years', 'going', 've', 'life', 'don', 'new', 'way', 'music', 'really', 'tim
e', 'know', 'think', 'people', 'just', 'like']



THE TOP 15 WORDS FOR TOPIC #6
['student', 'years', 'data', 'science', 'university', 'people', 'time', 'scho
ols', 'just', 'education', 'new', 'like', 'students', 'school', 'says']
```

In [61]:
```
dtm
```

Out[61]:
```
<11992x54777 sparse matrix of type '<class 'numpy.int64'>'
        with 3033388 stored elements in Compressed Sparse Row format>
```

In [62]:
```
npr
```

Out[62]:

	Article
0	In the Washington of 2016, even when the polic...
1	Donald Trump has used Twitter — his prefe...
2	Donald Trump is unabashedly praising Russian...
3	Updated at 2:50 p. m. ET, Russian President Vl...
4	From photography, illustration and video, to d...
...	...
11987	The number of law enforcement officers shot an...
11988	Trump is busy these days with victory tours,...
11989	It's always interesting for the Goats and Soda...
11990	The election of Donald Trump was a surprise to...
11991	Voters in the English city of Sunderland did s...

11992 rows × 1 columns

In [63]:
```
topic_results=LDA.transform(dtm)
```

In [64]:
```
topic_results[0].round(2)
```

Out[64]:
```
array([0.02, 0.68, 0.  , 0.  , 0.3 , 0.  , 0.  ])
```

In [65]:
```
topic_results[0].argmax()
```

Out[65]:
```
1
```

In [66]:
```
npr['TOPIC']=topic_results.argmax(axis=1)
```

In [70]:
```
npr
```

Out[70]:

	Article	TOPIC
0	In the Washington of 2016, even when the polic...	1
1	Donald Trump has used Twitter — his prefe...	1
2	Donald Trump is unabashedly praising Russian...	1
3	Updated at 2:50 p. m. ET, Russian President Vl...	1
4	From photography, illustration and video, to d...	2
...	...	...
11987	The number of law enforcement officers shot an...	1
11988	Trump is busy these days with victory tours,...	4
11989	It's always interesting for the Goats and Soda...	3
11990	The election of Donald Trump was a surprise to...	4
11991	Voters in the English city of Sunderland did s...	0

11992 rows × 2 columns

Non-Negative Matrix Factorization

In [71]:
```
import pandas as pd
import numpy as np
npr=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\05-Topic-Modeling\npr.cs
```

In [72]:
```
npr
```

Out[72]:

	Article
0	In the Washington of 2016, even when the polic...
1	Donald Trump has used Twitter — his prefe...
2	Donald Trump is unabashedly praising Russian...
3	Updated at 2:50 p. m. ET, Russian President Vl...
4	From photography, illustration and video, to d...
...	...
11987	The number of law enforcement officers shot an...
11988	Trump is busy these days with victory tours,...
11989	It's always interesting for the Goats and Soda...
11990	The election of Donald Trump was a surprise to...
11991	Voters in the English city of Sunderland did s...

11992 rows × 1 columns

In [74]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [75]:
```
tfidf=TfidfVectorizer(max_df=0.95, min_df=2, stop_words='english')
```

In [76]:
```
dtm=tfidf.fit_transform(npr['Article'])
```

In [77]:
```
dtm
```

Out[77]:
```
<11992x54777 sparse matrix of type '<class 'numpy.float64'>'
        with 3033388 stored elements in Compressed Sparse Row format>
```

In [78]:
```
from sklearn.decomposition import NMF
```

In [79]:
```
nmf_model=NMF(n_components=7, random_state=42)
```

In [80]:
```
nmf_model.fit(dtm)
```

Out[80]:
```
NMF(n_components=7, random_state=42)
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [82]:
```
tfidf.get_feature_names_out()[2300]
```

Out[82]:
```
'albala'
```

In [88]:
```
for index,topic in enumerate(nmf_model.components_):
    print(f'THE TOP 15 WORDS FOR TOPIC #{index}')
    print([tfidf.get_feature_names_out()[i] for i in topic.argsort()[-15:]])
    print('\n')
```

```
THE TOP 15 WORDS FOR TOPIC #0
['new', 'research', 'like', 'patients', 'health', 'disease', 'percent', 'wome
n', 'virus', 'study', 'water', 'food', 'people', 'zika', 'says']


THE TOP 15 WORDS FOR TOPIC #1
['gop', 'pence', 'presidential', 'russia', 'administration', 'election', 'rep
ublican', 'obama', 'white', 'house', 'donald', 'campaign', 'said', 'presiden
t', 'trump']


THE TOP 15 WORDS FOR TOPIC #2
['senate', 'house', 'people', 'act', 'law', 'tax', 'plan', 'republicans', 'af
fordable', 'obamacare', 'coverage', 'medicaid', 'insurance', 'care', 'healt
h']


THE TOP 15 WORDS FOR TOPIC #3
['officers', 'syria', 'security', 'department', 'law', 'isis', 'russia', 'gov
ernment', 'state', 'attack', 'president', 'reports', 'court', 'said', 'polic
e']


THE TOP 15 WORDS FOR TOPIC #4
['primary', 'cruz', 'election', 'democrats', 'percent', 'party', 'delegates',
'vote', 'state', 'democratic', 'hillary', 'campaign', 'voters', 'sanders', 'c
linton']


THE TOP 15 WORDS FOR TOPIC #5
['love', 've', 'don', 'album', 'way', 'time', 'song', 'life', 'really', 'kno
w', 'people', 'think', 'just', 'music', 'like']


THE TOP 15 WORDS FOR TOPIC #6
['teacher', 'state', 'high', 'says', 'parents', 'devos', 'children', 'colleg
e', 'kids', 'teachers', 'student', 'education', 'schools', 'school', 'student
s']
```

In [89]:
```
topic_results=nmf_model.transform(dtm)
```

In [90]:
```
topic_results.argmax(axis=1)
```

Out[90]:
```
array([1, 1, 1, ..., 0, 4, 3], dtype=int64)
```

In [91]:
```
npr['Topic']=topic_results.argmax(axis=1)
```

In [92]:
```
npr.head()
```

Out[92]:

	Article	Topic
0	In the Washington of 2016, even when the polic...	1
1	Donald Trump has used Twitter — his prefe...	1
2	Donald Trump is unabashedly praising Russian...	1
3	Updated at 2:50 p. m. ET, Russian President Vl...	3
4	From photography, illustration and video, to d...	6

In [93]:
```
mytopic_dict={0:'health',1:'election',2:'legis',3:'ploicts',4:'elections',5:'mu
npr['Topic Label']=npr['Topic'].map(mytopic_dict)
```

In [94]:
```
npr.head()
```

Out[94]:

	Article	Topic	Topic Label
0	In the Washington of 2016, even when the polic...	1	election
1	Donald Trump has used Twitter — his prefe...	1	election
2	Donald Trump is unabashedly praising Russian...	1	election
3	Updated at 2:50 p. m. ET, Russian President Vl...	3	ploicts
4	From photography, illustration and video, to d...	6	educa

Assignments--Projects

In [95]:
```
import numpy as np
import pandas as pd
```

In [96]:
```
quora=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\05-Topic-Modeling\quor
```

In [97]:
```
quora
```

Out[97]:

	Question
0	What is the step by step guide to invest in sh...
1	What is the story of Kohinoor (Koh-i-Noor) Dia...
2	How can I increase the speed of my internet co...
3	Why am I mentally very lonely? How can I solve...
4	Which one dissolve in water quikly sugar, salt...
...	...
404284	How many keywords are there in the Racket prog...
404285	Do you believe there is life after death?
404286	What is one coin?
404287	What is the approx annual cost of living while...
404288	What is like to have sex with cousin?

404289 rows × 1 columns

In [98]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [99]:
```
tfidf=TfidfVectorizer(max_df=0.95, min_df=2, stop_words='english')
```

In [100]:
```
dtm=tfidf.fit_transform(quora['Question'])
```

In [102]:
```
dtm
```

Out[102]:
```
<404289x38669 sparse matrix of type '<class 'numpy.float64'>'
        with 2002912 stored elements in Compressed Sparse Row format>
```

In [103]:
```
from sklearn.decomposition import NMF
```

In [104]:
```
nmf_model=NMF(n_components=7, random_state=42)
```

In [105]:
```
nmf_model.fit(dtm)
```

Out[105]:
```
NMF(n_components=7, random_state=42)
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [107]:
```
tfidf.get_feature_names_out()[2300]
```

Out[107]:
```
'algeria'
```

In [108]:
```
for index,topic in enumerate(nmf_model.components_):
    print(f'THE TOP 15 WORDS FOR TOPIC #{index}')
    print([tfidf.get_feature_names_out()[i] for i in topic.argsort()[-15:]])
    print('\n')
```

```
THE TOP 15 WORDS FOR TOPIC #0
['phone', 'india', 'lose', 'buy', 'laptop', 'time', 'movie', 'ways', '2016',
'weight', 'books', 'book', 'movies', 'way', 'best']


THE TOP 15 WORDS FOR TOPIC #1
['new', 'compare', 'look', 'cost', 'really', 'girl', 'love', 'long', 'sex',
'time', 'work', 'feel', 'like', 'mean', 'does']


THE TOP 15 WORDS FOR TOPIC #2
['post', 'answered', 'use', 'improvement', 'delete', 'easily', 'asked', 'goog
le', 'answer', 'answers', 'ask', 'question', 'questions', 'people', 'quora']


THE TOP 15 WORDS FOR TOPIC #3
['easiest', 'rupee', 'home', 'easy', 'notes', '1000', '500', 'black', 'youtub
e', 'ways', 'way', 'earn', 'online', 'make', 'money']


THE TOP 15 WORDS FOR TOPIC #4
['moment', 'live', 'employees', 'like', 'want', 'real', 'love', 'things', 'da
y', 'important', 'thing', 'know', 'meaning', 'purpose', 'life']


THE TOP 15 WORDS FOR TOPIC #5
['election', 'war', '1000', 'people', 'notes', '500', 'win', 'think', 'did',
'hillary', 'clinton', 'president', 'donald', 'trump', 'india']


THE TOP 15 WORDS FOR TOPIC #6
['speaking', 'languages', 'writing', 'java', 'speak', 'learning', 'skills',
'start', 'way', 'good', 'improve', 'programming', 'language', 'english', 'lea
rn']
```

In [109]:
```
topic_results=nmf_model.transform(dtm)
```

In [110]:
```
topic_results.argmax(axis=1)
```

Out[110]:
```
array([5, 4, 3, ..., 5, 5, 1], dtype=int64)
```

In [112]:
```
quora['Topic']=topic_results.argmax(axis=1)
```

In [113]:
```
quora
```

Out[113]:

	Question	Topic
0	What is the step by step guide to invest in sh...	5
1	What is the story of Kohinoor (Koh-i-Noor) Dia...	4
2	How can I increase the speed of my internet co...	3
3	Why am I mentally very lonely? How can I solve...	1
4	Which one dissolve in water quikly sugar, salt...	1
...	...	...
404284	How many keywords are there in the Racket prog...	6
404285	Do you believe there is life after death?	4
404286	What is one coin?	5
404287	What is the approx annual cost of living while...	5
404288	What is like to have sex with cousin?	1

404289 rows × 2 columns

In [114]:
```
mytopic_dict={0:'lifestyle',1:'romance',2:'ques and ans',3:'earning online mone
quora['Topic Label']=quora['Topic'].map(mytopic_dict)
```

In [115]:
```
quora
```

Out[115]:

	Question	Topic	Topic Label
0	What is the step by step guide to invest in sh...	5	politics
1	What is the story of Kohinoor (Koh-i-Noor) Dia...	4	routine
2	How can I increase the speed of my internet co...	3	earning online money
3	Why am I mentally very lonely? How can I solve...	1	romance
4	Which one dissolve in water quikly sugar, salt...	1	romance
...	...	...	...
404284	How many keywords are there in the Racket prog...	6	educa
404285	Do you believe there is life after death?	4	routine
404286	What is one coin?	5	politics
404287	What is the approx annual cost of living while...	5	politics
404288	What is like to have sex with cousin?	1	romance

404289 rows × 3 columns

LDA assumes that the documents are generated using a statistical generative process, such that each document is a mixture of topics, and each topics are a mixture of words.

Deep Learning in NLP

In [3]:
```
import numpy as np
```

In [4]:
```
from sklearn.datasets import load_iris
```

In [5]:
```
iris=load_iris()
```

In [6]:
```
type(iris)
```

Out[6]:
```
sklearn.utils._bunch.Bunch
```

In [7]:
```
print(iris.DESCR)
```

```
.. _iris_dataset:

Iris plants dataset
--------------------

**Data Set Characteristics:**

    :Number of Instances: 150 (50 in each of three classes)
    :Number of Attributes: 4 numeric, predictive attributes and the class
    :Attribute Information:
        - sepal length in cm
        - sepal width in cm
        - petal length in cm
        - petal width in cm
        - class:
                - Iris-Setosa
                - Iris-Versicolour
                - Iris-Virginica

    :Summary Statistics:

    ============== ==== ==== ======= ===== ====================
                    Min  Max   Mean    SD   Class Correlation
    ============== ==== ==== ======= ===== ====================
    sepal length:   4.3  7.9   5.84   0.83    0.7826
    sepal width:    2.0  4.4   3.05   0.43   -0.4194
    petal length:   1.0  6.9   3.76   1.76    0.9490  (high!)
    petal width:    0.1  2.5   1.20   0.76    0.9565  (high!)
    ============== ==== ==== ======= ===== ====================

    :Missing Attribute Values: None
    :Class Distribution: 33.3% for each of 3 classes.
    :Creator: R.A. Fisher
    :Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
    :Date: July, 1988

The famous Iris database, first used by Sir R.A. Fisher. The dataset is taken
from Fisher's paper. Note that it's the same as in R, but not as in the UCI
Machine Learning Repository, which has two wrong data points.

This is perhaps the best known database to be found in the
pattern recognition literature.  Fisher's paper is a classic in the field and
is referenced frequently to this day.  (See Duda & Hart, for example.)  The
data set contains 3 classes of 50 instances each, where each class refers to
a
type of iris plant.  One class is linearly separable from the other 2; the
latter are NOT linearly separable from each other.

.. topic:: References

   - Fisher, R.A. "The use of multiple measurements in taxonomic problems"
     Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to
     Mathematical Statistics" (John Wiley, NY, 1950).
   - Duda, R.O., & Hart, P.E. (1973) Pattern Classification and Scene Analysi
s.
     (Q327.D83) John Wiley & Sons.  ISBN 0-471-22361-1.  See page 218.
   - Dasarathy, B.V. (1980) "Nosing Around the Neighborhood: A New System
```

```
     Structure and Classification Rule for Recognition in Partially Exposed
     Environments".  IEEE Transactions on Pattern Analysis and Machine
     Intelligence, Vol. PAMI-2, No. 1, 67-71.
   - Gates, G.W. (1972) "The Reduced Nearest Neighbor Rule".  IEEE Transactio
ns
     on Information Theory, May 1972, 431-433.
   - See also: 1988 MLC Proceedings, 54-64.  Cheeseman et al"s AUTOCLASS II
     conceptual clustering system finds 3 classes in the data.
   - Many, many more ...
```

In [8]:
```
X=iris.data
```

In [9]:
```
X
```

Out[9]:
```
array([[5.1, 3.5, 1.4, 0.2],
       [4.9, 3. , 1.4, 0.2],
       [4.7, 3.2, 1.3, 0.2],
       [4.6, 3.1, 1.5, 0.2],
       [5. , 3.6, 1.4, 0.2],
       [5.4, 3.9, 1.7, 0.4],
       [4.6, 3.4, 1.4, 0.3],
       [5. , 3.4, 1.5, 0.2],
       [4.4, 2.9, 1.4, 0.2],
       [4.9, 3.1, 1.5, 0.1],
       [5.4, 3.7, 1.5, 0.2],
       [4.8, 3.4, 1.6, 0.2],
       [4.8, 3. , 1.4, 0.1],
       [4.3, 3. , 1.1, 0.1],
       [5.8, 4. , 1.2, 0.2],
       [5.7, 4.4, 1.5, 0.4],
       [5.4, 3.9, 1.3, 0.4],
       [5.1, 3.5, 1.4, 0.3],
       [5.7, 3.8, 1.7, 0.3],
```

In [10]:
```
y=iris.target
```

In [11]:
```
y
```

Out[11]:
```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2])
```

In [12]:
```
#class 0----->[1,0,0]
#class1 ----->[0,1,0]
#class 2---->[0,0,1]
```

In [13]:
```
from keras.utils import to_categorical
```

In [14]:
```
y=to_categorical(y)
```

In [16]:
```
y.shape
```

Out[16]:
```
(150, 3)
```

In [17]:
```
from sklearn.model_selection import train_test_split
```

In [18]:
```
X_train, X_test,y_train, y_test=train_test_split(X,y, test_size=0.33, random_st
```

In [19]:
```
X_train
```

Out[19]:
```
array([[5.7, 2.9, 4.2, 1.3],
       [7.6, 3. , 6.6, 2.1],
       [5.6, 3. , 4.5, 1.5],
       [5.1, 3.5, 1.4, 0.2],
       [7.7, 2.8, 6.7, 2. ],
       [5.8, 2.7, 4.1, 1. ],
       [5.2, 3.4, 1.4, 0.2],
       [5. , 3.5, 1.3, 0.3],
       [5.1, 3.8, 1.9, 0.4],
       [5. , 2. , 3.5, 1. ],
       [6.3, 2.7, 4.9, 1.8],
       [4.8, 3.4, 1.9, 0.2],
       [5. , 3. , 1.6, 0.2],
       [5.1, 3.3, 1.7, 0.5],
       [5.6, 2.7, 4.2, 1.3],
       [5.1, 3.4, 1.5, 0.2],
       [5.7, 3. , 4.2, 1.2],
       [7.7, 3.8, 6.7, 2.2],
       [4.6, 3.2, 1.4, 0.2],
       [6.2, 2.9, 4.3, 1.3],
       [5.7, 2.5, 5. , 2. ],
       [5.5, 4.2, 1.4, 0.2],
       [6. , 3. , 4.8, 1.8],
       [5.8, 2.7, 5.1, 1.9],
       [6. , 2.2, 4. , 1. ],
       [5.4, 3. , 4.5, 1.5],
       [6.2, 3.4, 5.4, 2.3],
       [5.5, 2.3, 4. , 1.3],
       [5.4, 3.9, 1.7, 0.4],
       [5. , 2.3, 3.3, 1. ],
       [6.4, 2.7, 5.3, 1.9],
       [5. , 3.3, 1.4, 0.2],
       [5. , 3.2, 1.2, 0.2],
       [5.5, 2.4, 3.8, 1.1],
       [6.7, 3. , 5. , 1.7],
       [4.9, 3.1, 1.5, 0.2],
       [5.8, 2.8, 5.1, 2.4],
       [5. , 3.4, 1.5, 0.2],
       [5. , 3.5, 1.6, 0.6],
       [5.9, 3.2, 4.8, 1.8],
       [5.1, 2.5, 3. , 1.1],
       [6.9, 3.2, 5.7, 2.3],
       [6. , 2.7, 5.1, 1.6],
       [6.1, 2.6, 5.6, 1.4],
       [7.7, 3. , 6.1, 2.3],
       [5.5, 2.5, 4. , 1.3],
       [4.4, 2.9, 1.4, 0.2],
       [4.3, 3. , 1.1, 0.1],
       [6. , 2.2, 5. , 1.5],
       [7.2, 3.2, 6. , 1.8],
       [4.6, 3.1, 1.5, 0.2],
       [5.1, 3.5, 1.4, 0.3],
       [4.4, 3. , 1.3, 0.2],
       [6.3, 2.5, 4.9, 1.5],
       [6.3, 3.4, 5.6, 2.4],
       [4.6, 3.4, 1.4, 0.3],
       [6.8, 3. , 5.5, 2.1],
```

```
[6.3, 3.3, 6. , 2.5],
[4.7, 3.2, 1.3, 0.2],
[6.1, 2.9, 4.7, 1.4],
[6.5, 2.8, 4.6, 1.5],
[6.2, 2.8, 4.8, 1.8],
[7. , 3.2, 4.7, 1.4],
[6.4, 3.2, 5.3, 2.3],
[5.1, 3.8, 1.6, 0.2],
[6.9, 3.1, 5.4, 2.1],
[5.9, 3. , 4.2, 1.5],
[6.5, 3. , 5.2, 2. ],
[5.7, 2.6, 3.5, 1. ],
[5.2, 2.7, 3.9, 1.4],
[6.1, 3. , 4.6, 1.4],
[4.5, 2.3, 1.3, 0.3],
[6.6, 2.9, 4.6, 1.3],
[5.5, 2.6, 4.4, 1.2],
[5.3, 3.7, 1.5, 0.2],
[5.6, 3. , 4.1, 1.3],
[7.3, 2.9, 6.3, 1.8],
[6.7, 3.3, 5.7, 2.1],
[5.1, 3.7, 1.5, 0.4],
[4.9, 2.4, 3.3, 1. ],
[6.7, 3.3, 5.7, 2.5],
[7.2, 3. , 5.8, 1.6],
[4.9, 3.6, 1.4, 0.1],
[6.7, 3.1, 5.6, 2.4],
[4.9, 3. , 1.4, 0.2],
[6.9, 3.1, 4.9, 1.5],
[7.4, 2.8, 6.1, 1.9],
[6.3, 2.9, 5.6, 1.8],
[5.7, 2.8, 4.1, 1.3],
[6.5, 3. , 5.5, 1.8],
[6.3, 2.3, 4.4, 1.3],
[6.4, 2.9, 4.3, 1.3],
[5.6, 2.8, 4.9, 2. ],
[5.9, 3. , 5.1, 1.8],
[5.4, 3.4, 1.7, 0.2],
[6.1, 2.8, 4. , 1.3],
[4.9, 2.5, 4.5, 1.7],
[5.8, 4. , 1.2, 0.2],
[5.8, 2.6, 4. , 1.2],
[7.1, 3. , 5.9, 2.1]])
```

In [20]:
```
y_train
```

Out[20]:
```
array([[0., 1., 0.],
       [0., 0., 1.],
       [0., 1., 0.],
       [1., 0., 0.],
       [0., 0., 1.],
       [0., 1., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [1., 0., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [0., 1., 0.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [1., 0., 0.],
       [0., 0., 1.],
       [0., 0., 1.],
       [0., 1., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [0., 1., 0.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [1., 0., 0.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 1., 0.],
       [1., 0., 0.],
       [0., 0., 1.],
       [1., 0., 0.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [0., 1., 0.],
       [0., 0., 1.],
       [0., 0., 1.],
       [0., 1., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [0., 0., 1.],
       [0., 0., 1.],
       [1., 0., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [1., 0., 0.],
       [0., 0., 1.],
```

```
       [0., 0., 1.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [0., 1., 0.],
       [0., 0., 1.],
       [1., 0., 0.],
       [0., 0., 1.],
       [0., 1., 0.],
       [0., 0., 1.],
       [0., 1., 0.],
       [0., 1., 0.],
       [0., 1., 0.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 1., 0.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [0., 0., 1.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [0., 0., 1.],
       [1., 0., 0.],
       [0., 0., 1.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [0., 0., 1.],
       [0., 1., 0.],
       [0., 0., 1.],
       [0., 1., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [0., 0., 1.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 0., 1.]], dtype=float32)
```

In [21]:
```
from sklearn.preprocessing import MinMaxScaler
```

In [22]:
```
np.array([5,10,15,20])/20
```

Out[22]:
```
array([0.25, 0.5 , 0.75, 1.  ])
```

In [23]:
```
scaler_object=MinMaxScaler()
```

In [24]:
```
scaler_object.fit(X_train)
```

Out[24]:
```
MinMaxScaler()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [25]:
```
scaled_X_train=scaler_object.transform(X_train)
```

In [26]:
```
scaled_X_test=scaler_object.transform(X_test)
```

In [27]:
```
scaled_X_train
```

Out[27]:
```
array([[0.41176471, 0.40909091, 0.55357143, 0.5       ],
       [0.97058824, 0.45454545, 0.98214286, 0.83333333],
       [0.38235294, 0.45454545, 0.60714286, 0.58333333],
       [0.23529412, 0.68181818, 0.05357143, 0.04166667],
       [1.        , 0.36363636, 1.        , 0.79166667],
       [0.44117647, 0.31818182, 0.53571429, 0.375     ],
       [0.26470588, 0.63636364, 0.05357143, 0.04166667],
       [0.20588235, 0.68181818, 0.03571429, 0.08333333],
       [0.23529412, 0.81818182, 0.14285714, 0.125     ],
       [0.20588235, 0.        , 0.42857143, 0.375     ],
       [0.58823529, 0.31818182, 0.67857143, 0.70833333],
       [0.14705882, 0.63636364, 0.14285714, 0.04166667],
       [0.20588235, 0.45454545, 0.08928571, 0.04166667],
       [0.23529412, 0.59090909, 0.10714286, 0.16666667],
       [0.38235294, 0.31818182, 0.55357143, 0.5       ],
       [0.23529412, 0.63636364, 0.07142857, 0.04166667],
       [0.41176471, 0.45454545, 0.55357143, 0.45833333],
       [1.        , 0.81818182, 1.        , 0.875     ],
       [0.08823529, 0.54545455, 0.05357143, 0.04166667],
       [0.55882353, 0.40909091, 0.57142857, 0.5       ],
       [0.41176471, 0.22727273, 0.69642857, 0.79166667],
       [0.35294118, 1.        , 0.05357143, 0.04166667],
       [0.5       , 0.45454545, 0.66071429, 0.70833333],
       [0.44117647, 0.31818182, 0.71428571, 0.75      ],
       [0.5       , 0.09090909, 0.51785714, 0.375     ],
       [0.32352941, 0.45454545, 0.60714286, 0.58333333],
       [0.55882353, 0.63636364, 0.76785714, 0.91666667],
       [0.35294118, 0.13636364, 0.51785714, 0.5       ],
       [0.32352941, 0.86363636, 0.10714286, 0.125     ],
       [0.20588235, 0.13636364, 0.39285714, 0.375     ],
       [0.61764706, 0.31818182, 0.75      , 0.75      ],
       [0.20588235, 0.59090909, 0.05357143, 0.04166667],
       [0.20588235, 0.54545455, 0.01785714, 0.04166667],
       [0.35294118, 0.18181818, 0.48214286, 0.41666667],
       [0.70588235, 0.45454545, 0.69642857, 0.66666667],
       [0.17647059, 0.5       , 0.07142857, 0.04166667],
       [0.44117647, 0.36363636, 0.71428571, 0.95833333],
       [0.20588235, 0.63636364, 0.07142857, 0.04166667],
       [0.20588235, 0.68181818, 0.08928571, 0.20833333],
       [0.47058824, 0.54545455, 0.66071429, 0.70833333],
       [0.23529412, 0.22727273, 0.33928571, 0.41666667],
       [0.76470588, 0.54545455, 0.82142857, 0.91666667],
       [0.5       , 0.31818182, 0.71428571, 0.625     ],
       [0.52941176, 0.27272727, 0.80357143, 0.54166667],
       [1.        , 0.45454545, 0.89285714, 0.91666667],
       [0.35294118, 0.22727273, 0.51785714, 0.5       ],
       [0.02941176, 0.40909091, 0.05357143, 0.04166667],
       [0.        , 0.45454545, 0.        , 0.        ],
       [0.5       , 0.09090909, 0.69642857, 0.58333333],
       [0.85294118, 0.54545455, 0.875     , 0.70833333],
       [0.08823529, 0.5       , 0.07142857, 0.04166667],
       [0.23529412, 0.68181818, 0.05357143, 0.08333333],
       [0.02941176, 0.45454545, 0.03571429, 0.04166667],
       [0.58823529, 0.22727273, 0.67857143, 0.58333333],
       [0.58823529, 0.63636364, 0.80357143, 0.95833333],
       [0.08823529, 0.63636364, 0.05357143, 0.08333333],
       [0.73529412, 0.45454545, 0.78571429, 0.83333333],
```

```
[0.58823529, 0.59090909, 0.875     , 1.        ],
[0.11764706, 0.54545455, 0.03571429, 0.04166667],
[0.52941176, 0.40909091, 0.64285714, 0.54166667],
[0.64705882, 0.36363636, 0.625     , 0.58333333],
[0.55882353, 0.36363636, 0.66071429, 0.70833333],
[0.79411765, 0.54545455, 0.64285714, 0.54166667],
[0.61764706, 0.54545455, 0.75      , 0.91666667],
[0.23529412, 0.81818182, 0.08928571, 0.04166667],
[0.76470588, 0.5       , 0.76785714, 0.83333333],
[0.47058824, 0.45454545, 0.55357143, 0.58333333],
[0.64705882, 0.45454545, 0.73214286, 0.79166667],
[0.41176471, 0.27272727, 0.42857143, 0.375     ],
[0.26470588, 0.31818182, 0.5       , 0.54166667],
[0.52941176, 0.45454545, 0.625     , 0.54166667],
[0.05882353, 0.13636364, 0.03571429, 0.08333333],
[0.67647059, 0.40909091, 0.625     , 0.5       ],
[0.35294118, 0.27272727, 0.58928571, 0.45833333],
[0.29411765, 0.77272727, 0.07142857, 0.04166667],
[0.38235294, 0.45454545, 0.53571429, 0.5       ],
[0.88235294, 0.40909091, 0.92857143, 0.70833333],
[0.70588235, 0.59090909, 0.82142857, 0.83333333],
[0.23529412, 0.77272727, 0.07142857, 0.125     ],
[0.17647059, 0.18181818, 0.39285714, 0.375     ],
[0.70588235, 0.59090909, 0.82142857, 1.        ],
[0.85294118, 0.45454545, 0.83928571, 0.625     ],
[0.17647059, 0.72727273, 0.05357143, 0.        ],
[0.70588235, 0.5       , 0.80357143, 0.95833333],
[0.17647059, 0.45454545, 0.05357143, 0.04166667],
[0.76470588, 0.5       , 0.67857143, 0.58333333],
[0.91176471, 0.36363636, 0.89285714, 0.75      ],
[0.58823529, 0.40909091, 0.80357143, 0.70833333],
[0.41176471, 0.36363636, 0.53571429, 0.5       ],
[0.64705882, 0.45454545, 0.78571429, 0.70833333],
[0.58823529, 0.13636364, 0.58928571, 0.5       ],
[0.61764706, 0.40909091, 0.57142857, 0.5       ],
[0.38235294, 0.36363636, 0.67857143, 0.79166667],
[0.47058824, 0.45454545, 0.71428571, 0.70833333],
[0.32352941, 0.63636364, 0.10714286, 0.04166667],
[0.52941176, 0.36363636, 0.51785714, 0.5       ],
[0.17647059, 0.22727273, 0.60714286, 0.66666667],
[0.44117647, 0.90909091, 0.01785714, 0.04166667],
[0.44117647, 0.27272727, 0.51785714, 0.45833333],
[0.82352941, 0.45454545, 0.85714286, 0.83333333]])
```

In [29]:
```
from keras.models import Sequential
from keras.layers import Dense
```

In [30]:
```
model=Sequential()
model.add(Dense(8,input_dim=4,activation='relu'))
model.add(Dense(8,input_dim=4,activation='relu'))
model.add(Dense(3,activation='softmax'))##[0.2,0.3,0.5]
model.compile(loss='categorical_crossentropy',optimizer='adam',metrics=['accura
```

In [31]:
```
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 8)                 40

 dense_1 (Dense)             (None, 8)                 72

 dense_2 (Dense)             (None, 3)                 27

=================================================================
Total params: 139 (556.00 Byte)
Trainable params: 139 (556.00 Byte)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

In [32]:
```
model.fit(scaled_X_train, y_train, epochs=150, verbose=2)
```

```
Epoch 1/150
4/4 - 8s - loss: 1.0905 - accuracy: 0.3800 - 8s/epoch - 2s/step
Epoch 2/150
4/4 - 0s - loss: 1.0878 - accuracy: 0.3800 - 25ms/epoch - 6ms/step
Epoch 3/150
4/4 - 0s - loss: 1.0856 - accuracy: 0.3800 - 27ms/epoch - 7ms/step
Epoch 4/150
4/4 - 0s - loss: 1.0837 - accuracy: 0.3800 - 29ms/epoch - 7ms/step
Epoch 5/150
4/4 - 0s - loss: 1.0819 - accuracy: 0.3800 - 29ms/epoch - 7ms/step
Epoch 6/150
4/4 - 0s - loss: 1.0800 - accuracy: 0.3800 - 25ms/epoch - 6ms/step
Epoch 7/150
4/4 - 0s - loss: 1.0780 - accuracy: 0.3800 - 22ms/epoch - 5ms/step
Epoch 8/150
4/4 - 0s - loss: 1.0762 - accuracy: 0.3600 - 24ms/epoch - 6ms/step
Epoch 9/150
4/4 - 0s - loss: 1.0741 - accuracy: 0.3600 - 24ms/epoch - 6ms/step
Epoch 10/150
```

In [33]:
```
scaled_X_test
```

Out[33]:
```
array([[ 0.52941176,  0.36363636,  0.64285714,  0.45833333],
       [ 0.41176471,  0.81818182,  0.10714286,  0.08333333],
       [ 1.        ,  0.27272727,  1.03571429,  0.91666667],
       [ 0.5       ,  0.40909091,  0.60714286,  0.58333333],
       [ 0.73529412,  0.36363636,  0.66071429,  0.54166667],
       [ 0.32352941,  0.63636364,  0.07142857,  0.125     ],
       [ 0.38235294,  0.40909091,  0.44642857,  0.5       ],
       [ 0.76470588,  0.5       ,  0.71428571,  0.91666667],
       [ 0.55882353,  0.09090909,  0.60714286,  0.58333333],
       [ 0.44117647,  0.31818182,  0.5       ,  0.45833333],
       [ 0.64705882,  0.54545455,  0.71428571,  0.79166667],
       [ 0.14705882,  0.45454545,  0.05357143,  0.        ],
       [ 0.35294118,  0.68181818,  0.03571429,  0.04166667],
       [ 0.17647059,  0.5       ,  0.07142857,  0.        ],
       [ 0.23529412,  0.81818182,  0.07142857,  0.08333333],
       [ 0.58823529,  0.59090909,  0.64285714,  0.625     ],
       [ 0.64705882,  0.45454545,  0.83928571,  0.875     ],
       [ 0.38235294,  0.22727273,  0.5       ,  0.41666667],
       [ 0.41176471,  0.36363636,  0.60714286,  0.5       ],
       [ 0.61764706,  0.36363636,  0.80357143,  0.875     ],
       [ 0.11764706,  0.54545455,  0.08928571,  0.04166667],
       [ 0.52941176,  0.45454545,  0.67857143,  0.70833333],
       [ 0.20588235,  0.63636364,  0.08928571,  0.125     ],
       [ 0.61764706,  0.36363636,  0.80357143,  0.83333333],
       [ 1.05882353,  0.81818182,  0.94642857,  0.79166667],
       [ 0.70588235,  0.45454545,  0.73214286,  0.91666667],
       [ 0.70588235,  0.22727273,  0.83928571,  0.70833333],
       [ 0.73529412,  0.54545455,  0.85714286,  0.91666667],
       [ 0.14705882,  0.45454545,  0.05357143,  0.08333333],
       [ 0.14705882,  0.5       ,  0.08928571,  0.04166667],
       [ 0.08823529,  0.72727273, -0.01785714,  0.04166667],
       [ 0.41176471,  1.09090909,  0.07142857,  0.125     ],
       [ 0.70588235,  0.5       ,  0.58928571,  0.54166667],
       [ 0.14705882,  0.63636364,  0.08928571,  0.04166667],
       [ 0.02941176,  0.54545455,  0.03571429,  0.04166667],
       [ 0.58823529,  0.22727273,  0.69642857,  0.75      ],
       [ 0.61764706,  0.54545455,  0.60714286,  0.58333333],
       [ 0.26470588,  0.68181818,  0.07142857,  0.04166667],
       [ 0.20588235,  0.72727273,  0.05357143,  0.04166667],
       [ 0.26470588,  0.95454545,  0.07142857,  0.        ],
       [ 0.44117647,  0.31818182,  0.71428571,  0.75      ],
       [ 0.5       ,  0.63636364,  0.60714286,  0.625     ],
       [ 0.70588235,  0.5       ,  0.64285714,  0.58333333],
       [ 0.32352941,  0.86363636,  0.03571429,  0.125     ],
       [ 0.32352941,  0.77272727,  0.07142857,  0.04166667],
       [ 0.35294118,  0.18181818,  0.46428571,  0.375     ],
       [ 0.58823529,  0.36363636,  0.71428571,  0.58333333],
       [ 0.61764706,  0.5       ,  0.78571429,  0.70833333],
       [ 0.67647059,  0.45454545,  0.58928571,  0.54166667],
       [ 0.85294118,  0.72727273,  0.89285714,  1.        ]])
```

In [35]:
```
predict_x=model.predict(scaled_X_test)
```

```
2/2 [==============================] - 0s 25ms/step
```

In [36]:
```
classes_x=np.argmax(predict_x,axis=1)
```

In [39]:
```
classes_x
```

Out[39]:
```
array([2, 0, 2, 2, 2, 0, 1, 2, 2, 1, 2, 0, 0, 0, 0, 2, 2, 1, 1, 2, 0, 2,
       0, 2, 2, 2, 2, 2, 0, 0, 0, 0, 2, 0, 0, 2, 2, 0, 0, 0, 2, 2, 2, 0,
       0, 1, 2, 2, 2, 2], dtype=int64)
```

In [40]:
```
y_test
```

Out[40]:
```
array([[0., 1., 0.],
       [1., 0., 0.],
       [0., 0., 1.],
       [0., 1., 0.],
       [0., 1., 0.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [0., 1., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [1., 0., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [0., 1., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [1., 0., 0.],
       [0., 0., 1.],
       [1., 0., 0.],
       [0., 0., 1.],
       [0., 0., 1.],
       [0., 0., 1.],
       [0., 0., 1.],
       [0., 0., 1.],
       [1., 0., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [0., 1., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [0., 0., 1.],
       [0., 1., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [0., 0., 1.],
       [0., 1., 0.],
       [0., 1., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [0., 1., 0.],
       [0., 0., 1.],
       [0., 0., 1.],
       [0., 1., 0.],
       [0., 0., 1.]], dtype=float32)
```

In [41]:
```
y_test.argmax(axis=1)
```

Out[41]:
```
array([1, 0, 2, 1, 1, 0, 1, 2, 1, 1, 2, 0, 0, 0, 0, 1, 2, 1, 1, 2, 0, 2,
       0, 2, 2, 2, 2, 2, 0, 0, 0, 0, 1, 0, 0, 2, 1, 0, 0, 0, 2, 1, 1, 0,
       0, 1, 2, 2, 1, 2], dtype=int64)
```

In [42]:
```
from sklearn.metrics import confusion_matrix, classification_report, accuracy_s
```

In [44]:
```
confusion_matrix(y_test.argmax(axis=1),classes_x)
```

Out[44]:
```
array([[19,  0,  0],
       [ 0,  5, 10],
       [ 0,  0, 16]], dtype=int64)
```

In [45]:
```
print(classification_report(y_test.argmax(axis=1),classes_x))
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        19
           1       1.00      0.33      0.50        15
           2       0.62      1.00      0.76        16

    accuracy                           0.80        50
   macro avg       0.87      0.78      0.75        50
weighted avg       0.88      0.80      0.77        50
```

In [46]:
```
accuracy_score(y_test.argmax(axis=1),classes_x)
```

Out[46]:
```
0.8
```

In [47]:
```
model.save('myfirstmodel.h5')
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\keras\src\engine\training.py:307
9: UserWarning: You are saving your model as an HDF5 file via `model.save()`.
This file format is considered legacy. We recommend using instead the native
Keras format, e.g. `model.save('my_model.keras')`.
  saving_api.save_model(
```

In [48]:
```
from keras.models import load_model
```

In [49]:
```
new_model=load_model('myfirstmodel.h5')
```

In [52]:
```
new_model.predict(scaled_X_test)
```

```
2/2 [==============================] - 12s 110ms/step
```

Out[52]:
```
array([[3.72950286e-02, 4.77029473e-01, 4.85675514e-01],
       [9.57485199e-01, 3.26164514e-02, 9.89839528e-03],
       [4.69889725e-04, 2.21053183e-01, 7.78477013e-01],
       [2.07423847e-02, 4.57594246e-01, 5.21663368e-01],
       [1.04048010e-02, 3.94045413e-01, 5.95549881e-01],
       [9.15957987e-01, 6.43589497e-02, 1.96830090e-02],
       [7.07160309e-02, 5.20747662e-01, 4.08536285e-01],
       [1.36550947e-03, 2.88264364e-01, 7.10370183e-01],
       [8.71101208e-03, 4.13282305e-01, 5.78006625e-01],
       [5.21281883e-02, 5.09663582e-01, 4.38208193e-01],
       [3.89115536e-03, 3.46420676e-01, 6.49688184e-01],
       [9.44849253e-01, 4.21767756e-02, 1.29739465e-02],
       [9.53903377e-01, 3.54445949e-02, 1.06520448e-02],
       [9.49557245e-01, 3.87575366e-02, 1.16852513e-02],
       [9.68420565e-01, 2.45877281e-02, 6.99162111e-03],
       [1.93525180e-02, 4.39726561e-01, 5.40920913e-01],
       [1.60067074e-03, 2.97571450e-01, 7.00827897e-01],
       [6.33995980e-02, 5.21492481e-01, 4.15107995e-01],
       [4.24933545e-02, 4.97849643e-01, 4.59657013e-01],
       [1.72494480e-03, 3.09142798e-01, 6.89132273e-01],
       [9.51612055e-01, 3.73286270e-02, 1.10592339e-02],
       [8.23729206e-03, 4.00967866e-01, 5.90794802e-01],
       [9.34379101e-01, 5.06824143e-02, 1.49385557e-02],
       [2.07051379e-03, 3.16506952e-01, 6.81422472e-01],
       [1.55514234e-03, 2.50328064e-01, 7.48116791e-01],
       [1.43415632e-03, 2.95325965e-01, 7.03239858e-01],
       [2.70402874e-03, 3.21981102e-01, 6.75314844e-01],
       [1.14161568e-03, 2.72467613e-01, 7.26390779e-01],
       [9.20837343e-01, 6.14083894e-02, 1.77542847e-02],
       [9.43354130e-01, 4.37369645e-02, 1.29089355e-02],
       [9.66618717e-01, 2.60281004e-02, 7.35318428e-03],
       [9.78227735e-01, 1.69255994e-02, 4.84671770e-03],
       [1.93068329e-02, 4.31969941e-01, 5.48723221e-01],
       [9.59436357e-01, 3.15483399e-02, 9.01533104e-03],
       [9.53291118e-01, 3.58581655e-02, 1.08506884e-02],
       [3.58221075e-03, 3.56418997e-01, 6.39998853e-01],
       [2.15098187e-02, 4.44904894e-01, 5.33585310e-01],
       [9.60251570e-01, 3.06268055e-02, 9.12159309e-03],
       [9.65278149e-01, 2.69955415e-02, 7.72629073e-03],
       [9.80350375e-01, 1.54410442e-02, 4.20857640e-03],
       [5.11426991e-03, 3.84048849e-01, 6.10836983e-01],
       [3.22732404e-02, 4.71489608e-01, 4.96237129e-01],
       [1.33845322e-02, 4.09667224e-01, 5.76948225e-01],
       [9.62660611e-01, 2.89545525e-02, 8.38495977e-03],
       [9.66209769e-01, 2.60396283e-02, 7.75068579e-03],
       [8.68299380e-02, 5.28946102e-01, 3.84223938e-01],
       [1.16176773e-02, 4.11438972e-01, 5.76943338e-01],
       [6.01350563e-03, 3.67488235e-01, 6.26498282e-01],
       [1.84905343e-02, 4.33547288e-01, 5.47962248e-01],
       [6.44849322e-04, 2.34623030e-01, 7.64732122e-01]], dtype=float32)
```

Text Generation with Python and Keras

In [1]:
```
def read_file(filepath):
    with open(filepath) as f:
        str_text=f.read()
    return str_text
```

In [3]:
```
#read_file(r'F:\daily work\NLP\UPDATED_NLP_COURSE\06-Deep-Learning\moby_dick_fo
```

In [1]:
```
import spacy
```

In [4]:
```
try:
  nlp_lg = spacy.load("en_core_web_lg")
except ModuleNotFoundError:
  download(model="en_core_web_lg")
  nlp_lg = spacy.load("en_core_web_lg")
```

In [5]:
```
nlp_lg.max_length=1198623
```

In [7]:
```
try:
  nlp = spacy.load("en_core_web_lg",disable=['parser','tagger','ner'])
except ModuleNotFoundError:
  download(model="en_core_web_lg")
  nlp = spacy.load("en_core_web_lg",disable=['parser','tagger','ner'])
```

In [8]:
```
nlp.max_length=1198623
```

Revision

In [1]:
```
# working with Text file
name='nilesh'
print ("my name is {}".format(name))
```

```
my name is nilesh
```

In [2]:
```
name='nilesh'
print(f"my name is {name}")
```

```
my name is nilesh
```

In [3]:
```
d={'a':123, 'b':456}
```

In [5]:
```
print(f"my number is {d['b']}")
```

```
my number is 456
```

In [6]:
```
my_list=[100,200, 300]
```

In [7]:
```
print(f'my number belongs to {my_list[1]}')
```

```
my number belongs to 200
```

In [8]:
```
library=[('Author','Topic','Pages'),('Twain','Rafting',601),('Feynman','physics
```

In [9]:
```
library
```

Out[9]:
```
[('Author', 'Topic', 'Pages'),
 ('Twain', 'Rafting', 601),
 ('Feynman', 'physics', 95),
 ('Hamilton', 'mythology', 144)]
```

In [10]:
```
for book in library:
    print(book)
```

```
('Author', 'Topic', 'Pages')
('Twain', 'Rafting', 601)
('Feynman', 'physics', 95)
('Hamilton', 'mythology', 144)
```

In [11]:
```
for book in library:
    print(book[0])
```

```
Author
Twain
Feynman
Hamilton
```

In [15]:
```
for author, topic, pages in library:
    print(f'{author:{10}} {topic:{20}} {pages:>{10}}')
```

```
Author     Topic                      Pages
Twain      Rafting                      601
Feynman    physics                       95
Hamilton   mythology                    144
```

In [1]:
```
import spacy
import spacy.cli
spacy.cli.download("en_core_web_lg")
nlp=spacy.load('en_core_web_lg')
```

```
An exception has occurred, use %tb to see the full traceback.

SystemExit: 1



C:\Users\nilesh\AppData\Roaming\Python\Python39\site-packages\IPython\core\in
teractiveshell.py:3516: UserWarning: To exit: use 'exit', 'quit', or Ctrl-D.
  warn("To exit: use 'exit', 'quit', or Ctrl-D.", stacklevel=1)
```

In [2]:
```
spacy.cli.download("en_core_web_sm")
nlp=spacy.load('en_core_web_sm')
```

```
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')

C:\Users\nilesh\anaconda3\lib\site-packages\spacy\util.py:910: UserWarning:
[W095] Model 'en_core_web_sm' (3.7.1) was trained with spaCy v3.7.2 and may n
ot be 100% compatible with the current version (3.7.0). If you see errors or
degraded performance, download a newer compatible model or retrain your custo
m model with the current spaCy version. For more details and available update
s, run: python -m spacy validate
  warnings.warn(warn_msg)
```

In [3]:
```
doc=nlp(u'Tesla is lookig to buy US start up for $6 million')
```

In [6]:
```
for token in doc:
    print(token.text,token.pos_,token.dep_)
```

```
Tesla PROPN nsubj
is AUX ROOT
lookig ADJ attr
to PART aux
buy VERB xcomp
US PROPN dobj
start VERB dep
up ADP prt
for ADP prep
$ SYM quantmod
6 NUM compound
million NUM pobj
```

In [7]:
```
nlp.pipeline
```

Out[7]:
```
[('tok2vec', <spacy.pipeline.tok2vec.Tok2Vec at 0x22c03fe3e20>),
 ('tagger', <spacy.pipeline.tagger.Tagger at 0x22c03fe3f40>),
 ('parser', <spacy.pipeline.dep_parser.DependencyParser at 0x22c03eb76d0>),
 ('attribute_ruler',
  <spacy.pipeline.attributeruler.AttributeRuler at 0x22c041f4c40>),
 ('lemmatizer', <spacy.lang.en.lemmatizer.EnglishLemmatizer at 0x22c04204640
>),
 ('ner', <spacy.pipeline.ner.EntityRecognizer at 0x22c03eb7510>)]
```

In [11]:
```
doc2= nlp(u"Tesla isn't   looking for start-up anymore." )
```

In [12]:
```
for token in doc2:
    print(token.text, token.pos_, token.dep_)
```

```
Tesla PROPN nsubj
is AUX aux
n't PART neg
   SPACE dep
looking VERB ROOT
for ADP prep
start NOUN compound
- PUNCT punct
up NOUN pobj
anymore ADV advmod
. PUNCT punct
```

In [13]:
```
doc2[0].pos_
```

Out[13]: 'PROPN'

In [14]:
```
doc2[0].lemma_
```

Out[14]: 'Tesla'

In [15]:
```
doc2[1].lemma_
```

Out[15]: 'be'

In [17]:
```
doc3=nlp(u"although commonly used to john lennane as his song 'beautiful boy'
```

In [18]:
```
life_q=doc3[0:5]
```

In [19]:
```
print(life_q)
```

```
although commonly used to john
```

In [20]:
```
type(life_q)
```

Out[20]: spacy.tokens.span.Span

In [21]:
```
type(doc3)
```

Out[21]: spacy.tokens.doc.Doc

In [26]:
```
doc4=nlp(u'This is the first sentence. This is the the second sentence. This is
```

In [27]:
```
for token in doc4.sents:
    print(token)
```

```
This is the first sentence.
This is the the second sentence.
This is the last sentence
```

In [1]:
```
#Tokenization
import spacy
nlp=spacy.load('en_core_web_sm')
```

In [29]:
```
mystring= ' "We\'re moving to L.A. ! "'
```

In [30]:
```
print(mystring)
```

```
 "We're moving to L.A. ! "
```

In [31]:
```
doc=nlp(mystring)
```

In [32]:
```
for token in doc:
    print(token)
```

```
"
We
're
moving
to
L.A.
!
"
```

In [33]:
```
doc2=nlp(u" we're here to help. Please send email to abc@gmail.com or vist to o
```

In [34]:
```
for token in doc2:
    print(token)
```

```
we
're
here
to
help
.
Please
send
email
to
abc@gmail.com
or
vist
to
our
websitte
:
https://www.linkedin.com/nhome/?trk (https://www.linkedin.com/nhome/?trk)
```

In [2]:
```
doc3=nlp(u"a 5 km NYC cab ride cost $10.30")
```

In [3]:
```
for t in doc3:
    print(t)
```

```
a
5
km
NYC
cab
ride
cost
$
10.30
```

In [4]:
```
doc4= nlp(u"Let's visit St. Louis in the U.S. next year")
```

In [5]:
```
for x in doc4:
    print(x)
```

```
Let
's
visit
St.
Louis
in
the
U.S.
next
year
```

In [6]:
```
len(doc4)
```

Out[6]: 10

In [7]:
```
len(doc4.vocab)
```

Out[7]: 778

In [1]:
```
import spacy.cli
spacy.cli.download("en_core_web_lg")
nlp=spacy.load('en_core_web_lg')
```

```
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_lg')
```

In [10]:
```
doc5= nlp(u"Let's visit St. Louis in the U.S. next year")
```

In [11]:
```
for x in doc5:
    print(x)
```

```
Let
's
visit
St.
Louis
in
the
U.S.
next
year
```

In [12]:
```
len(doc5.vocab)
```

Out[12]: 771

In [13]:
```
doc6= nlp(u"It is better than to give")
```

In [15]:
```
doc6[0]
```

Out[15]: It

In [16]:
```
doc6[0:4]
```

Out[16]: It is better than

In [17]:
```
doc6[0]="final"
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
Cell In[17], line 1
----> 1 doc6[0]="final"

TypeError: 'spacy.tokens.doc.Doc' object does not support item assignment
```

In [18]:
```
doc8= nlp(u"India economy is excepted to grow with 7% in this year")
```

In [19]:
```
for token in doc8:
    print(token.text, end='|')
```

```
India|economy|is|excepted|to|grow|with|7|%|in|this|year|
```

In [23]:
```
for entity in doc8.ents:
    print(entity)
    print(entity.label_)
    print(str(spacy.explain(entity.label_)))
    print('\n')
```

```
India
GPE
Countries, cities, states


7%
PERCENT
Percentage, including "%"


this year
DATE
Absolute or relative dates or periods
```

In [24]:
```
doc9= nlp(u'Autonomous cars shift insurance liability toward manufactures')
```

In [25]:
```
for junk in doc9.noun_chunks:
    print(junk)
```

```
Autonomous cars
insurance liability
manufactures
```

In [26]:
```
from spacy import displacy
```

In [28]:
```
doc9= nlp(u'Apple is going to built factorty in India for $12 millions.')
```

In [30]:
```
displacy.render(doc9, style='dep', jupyter=True, options={'distance':80})
```

Apple PROPN is AUX going VERB to PART built VERB factorty NOUN in ADP India PROPN for ADP $ SYM 12 NUM millions. NOUN nsubj aux aux xcomp dobj prep pobj prep quantmod compound pobj

In [35]:
```
doc10= nlp(u'This year Apple sold 120 millions iPods and earns $200 millions wh
```

In [36]:
```
displacy.render(doc10, style='ent', jupyter=True)
```

This year **DATE** Apple **ORG** sold 120 millions **CARDINAL** iPods **PRODUCT** and earns $200 millions **MONEY** which is greater than last year **DATE**

In [40]:
```
doc10=nlp(u"this is the sentence.")
displacy.serve(doc10, style='dep')
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\spacy\displacy\__init__.py:106: U
serWarning: [W011] It looks like you're calling displacy.serve from within a
Jupyter notebook or a similar environment. This likely means you're already r
unning a local web server, so there's no need to make displaCy start another
one. Instead, you should be able to replace displacy.serve with displacy.rend
er to show the visualization.
  warnings.warn(Warnings.W011)
```

displaCy

this PRON is AUX the DET sentence. NOUN nsubj det attr

```
Using the 'dep' visualizer
Serving on http://0.0.0.0:5000 (http://0.0.0.0:5000) ...


127.0.0.1 - - [07/Jan/2024 16:14:42] "GET / HTTP/1.1" 200 3397
127.0.0.1 - - [07/Jan/2024 16:14:44] "GET /favicon.ico HTTP/1.1" 200 3397

Shutting down server on port 5000.
```

In [41]:
```
##Stemming in NLTK
from nltk.stem.porter import PorterStemmer
```

In [42]:
```
p_stemmer=PorterStemmer()
```

In [43]:
```
words=['run','runner','running','boat','boating','boats','easily','fairly']
```

In [44]:
```
for word in words:
    print(word +'------>'+ p_stemmer.stem(word))
```

```
run------>run
runner------>runner
running------>run
boat------>boat
boating------>boat
boats------>boat
easily------>easili
fairly------>fairli
```

In [45]:
```
from nltk.stem.snowball import SnowballStemmer
```

In [46]:
```
s_stemmer=SnowballStemmer(language='english')
```

In [47]:
```
for word in words:
    print(word +'------>'+ s_stemmer.stem(word))
```

```
run------>run
runner------>runner
running------>run
boat------>boat
boating------>boat
boats------>boat
easily------>easili
fairly------>fair
```

In [48]:
```
#Lemmatization

doc1= nlp(u"I am runner running in the race because i love to run since i am ru
```

In [49]:
```
for token in doc1:
    print(token.text, '\t', token.pos_,'\t',token.lemma,'\t',token.lemma_)
```

```
I        PRON    4690420944186131903     I
am       AUX     10382539506755952630    be
runner   NOUN    12640964157389618806    runner
running          VERB    12767647472892411841    run
in       ADP     3002984154512732771     in
the      DET     7425985699627899538     the
race     NOUN    8048469955494714898     race
because          SCONJ   16950148841647037698    because
i        PRON    4690420944186131903     I
love     VERB    3702023516439754181     love
to       PART    3791531372978436496     to
run      VERB    12767647472892411841    run
since    SCONJ   10066841407251338481    since
i        PRON    4690420944186131903     I
am       AUX     10382539506755952630    be
running          VERB    12767647472892411841    run
too      ADV     12286903790479710773    too
fast     ADV     1826119438242743099     fast
```

In [2]:
```
#STOP WORDS
print(nlp.Defaults.stop_words)
```

```
{''re', 'done', 'any', 'whom', 'against', 'three', 'become', 'no', 'though',
'fifty', 'whither', 'am', 'nothing', 'full', 'via', 'say', 'do', ''d', 'among
st', 'have', 'sometime', 'due', 'thence', 'please', 'thru', 'yourself', 'hers
elf', 'could', 'thereafter', 'for', 'part', 'then', 'mostly', 'twelve', 'furt
her', 'ours', 'had', 'bottom', 'onto', 'well', 'her', 'they', 'hers', 'althou
gh', 'rather', 'last', 'most', 'still', 'never', 'among', 'call', 'eight', 's
omething', 'becomes', 'him', 'give', 'off', 'under', 'former', 'would', 'afte
r', 'five', 'from', 'that', 'without', 'anyway', 'too', 'everywhere', 'becam
e', 'nine', 'hereby', 'our', 'already', 'anyhow', 'my', 'move', 'does', 'arou
nd', 'one', 'this', 'almost', 'least', 'than', 'ten', 'used', 'he', 'top', 'w
henever', 'what', 'within', 'both', 'using', 're', 'even', 'anything', 'someh
ow', 'to', 'per', 'amount', 'doing', 'why', 'may', 'did', 'therein', 'own',
'such', 'nevertheless', 'whereby', 'into', 'go', 'your', 'themselves', 'kee
p', 'with', 'being', 'ca', 'fifteen', "'s", 'all', 'whereupon', 'below', 'thr
ough', 'before', 'it', 'be', 'here', 'between', 'towards', 'everything', 'mor
e', 'ever', 'whatever', 'several', 'regarding', 'anyone', 'throughout', 'no
t', 'i', 'elsewhere', 'cannot', 'me', 'must', 'serious', 'if', ''s', 'm', 'b
ack', 'next', 'seems', 'wherein', 'is', 'd', 'unless', 'few', 'everyone', 'i
n', 'wherever', 'much', 'therefore', 'often', 'hundred', 'these', 'else', 'la
tter', 'thereupon', 'can', 'eleven', 'along', 'out', 'them', 'mine', "'m", 'w
here', 'thus', 'while', 'across', 'beforehand', 'made', 'might', 'was', 'mak
e', 'because', 'however', 'the', 'put', 'yours', 'beyond', 'side', 'as', 'hen
ce', 'toward', 'since', 'hereupon', 'seeming', 'other', 'thereby', 'forty',
'name', 'their', 'except', 'until', 'hereafter', 'very', 'how', 'who', 'dow
n', 'sixty', 'enough', 've', 'about', 'namely', 'really', 'sometimes', 'no
w', 'so', 'were', 'whether', 'its', 'same', 'his', 'less', 'take', 'either',
'otherwise', 'yet', 'or', 'many', 'formerly', 'a', 'alone', 'are', 'nobody',
'whence', "n't", 'latterly', 'six', 'seemed', 'himself', 'beside', 'togethe
r', 'quite', 'll', 'whereas', 'll', 'seem', 'various', 'indeed', 'during',
'over', 'two', 'and', 'an', 'noone', 'empty', 'twenty', 'ourselves', 'when',
'some', 'up', 'becoming', 'moreover', 'only', 'besides', 'at', 'will', 'nt',
'also', 'which', 'perhaps', 'yourselves', 'neither', 'third', 'always',
'nt', 'itself', 'meanwhile', 're', 'get', 'of', 'another', "'re", 'myself',
'us', 'there', 'see', "'d", 'front', 'first', 'we', 'whoever', 'has', 'you',
'whose', 'those', 'others', 'been', "'ve", 'someone', 'behind', 'on', 'none',
'herein', 'but', 'she', 'anywhere', 'whereafter', 'four', "'ll", 'show', 'som
ewhere', 'each', 'again', 'by', 'should', 'just', 'whole', 'every', 've', 'a
bove', 'nowhere', 'nor', 'once', 'm', 'upon', 's', 'afterwards'}
```

In [3]:
```
len(nlp.Defaults.stop_words)
```

Out[3]: 326

In [4]:
```
nlp.vocab['is'].is_stop
```

Out[4]: True

In [6]:
```
nlp.Defaults.stop_words.add('btw')
```

In [7]:
```
nlp.vocab['btw'].is_stop=True
```

In [8]:
```
len(nlp.Defaults.stop_words)
```

Out[8]: 327

In [9]:
```
nlp.Defaults.stop_words.remove('btw')
```

In [10]:
```
nlp.vocab['btw'].is_stop=False
```

In [11]:
```
nlp.vocab['btw'].is_stop
```

Out[11]: False

In [1]:
```
#Pharase matching vocabulary
import spacy
nlp=spacy.load('en_core_web_sm')
```

In [13]:
```
from spacy.matcher import Matcher
```

In [14]:
```
matcher=Matcher(nlp.vocab)
```

In [26]:
```
#SolarPower
#Solar-power
#Solar power
pattern1 = [{'LOWER':'solarpower'}]
pattern2 = [{'LOWER':'solar'},{'IS_PUNCT':True},{'LOWER':'power'}]
pattern3 = [{'LOWER':'solar'},{'LOWER':'power'}]
```

In [27]:
```
matcher.add('SolarPower',[pattern1,pattern2,pattern3])
```

In [28]:
```
doc=nlp(u'The Solar power industry continues to grow as solarpower increases. S
```

In [29]:
```
found_matches=matcher(doc)
```

In [30]:
```
print(found_matches)
```

```
[(8656102463236116519, 1, 3), (8656102463236116519, 8, 9), (86561024632361165
19, 11, 14)]
```

In [3]:
```
from spacy.matcher import PhraseMatcher
```

In [4]:
```
matcher=PhraseMatcher(nlp.vocab)
```

In [7]:
```
with open(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\reaganomics.txt') as
    doc3=nlp(f.read())
```

In [8]:
```
phrase_list=['voodo economic','supply-side economic','trickle-down economics',
```

In [9]:
```
phrase_patterns=[nlp(text) for text in phrase_list]
```

In [10]:
```
phrase_patterns
```

Out[10]:
```
[voodo economic,
 supply-side economic,
 trickle-down economics,
 free-market economics]
```

In [11]:
```
type(phrase_patterns[0])
```

Out[11]:
```
spacy.tokens.doc.Doc
```

In [12]:
```
matcher.add('EcoMatcher',None,*phrase_patterns)
```

In [13]:
```
found_matches=matcher(doc3)
```

In [14]:
```
found_matches
```

Out[14]:
```
[(2351661100535932681, 49, 53),
 (2351661100535932681, 61, 65),
 (2351661100535932681, 2987, 2991)]
```

In [1]:
```
#Part of speech
import spacy
nlp=spacy.load('en_core_web_sm')
```

In [16]:
```
doc=nlp(u"the quick brown fox jumped over teh lazy dog's back")
```

In [17]:
```
doc.text
```

Out[17]:
```
"the quick brown fox jumped over teh lazy dog's back"
```

In [19]:
```
print(doc[4].pos_)
```

```
VERB
```

In [20]:
```
print(doc[4].tag_)
```

```
VBD
```

In [26]:
```
for token in doc:
    print(f"{token.text} {token.pos_} {token.tag_}, {spacy.explain(token.tag_)}
```

```
the DET DT, determiner
quick ADJ JJ, adjective (English), other noun-modifier (Chinese)
brown ADJ JJ, adjective (English), other noun-modifier (Chinese)
fox NOUN NN, noun, singular or mass
jumped VERB VBD, verb, past tense
over ADP IN, conjunction, subordinating or preposition
teh NOUN NN, noun, singular or mass
lazy ADJ JJ, adjective (English), other noun-modifier (Chinese)
dog NOUN NN, noun, singular or mass
's PART POS, possessive ending
back NOUN NN, noun, singular or mass
```

In [30]:
```
for token in doc:
    print(f"{token.text:{10}} {token.pos_:{10}} {token.tag_:{10}} {spacy.expla
```

```
the        DET        DT         determiner
quick      ADJ        JJ         adjective (English), other noun-modifier (Ch
inese)
brown      ADJ        JJ         adjective (English), other noun-modifier (Ch
inese)
fox        NOUN       NN         noun, singular or mass
jumped     VERB       VBD        verb, past tense
over       ADP        IN         conjunction, subordinating or preposition
teh        NOUN       NN         noun, singular or mass
lazy       ADJ        JJ         adjective (English), other noun-modifier (Ch
inese)
dog        NOUN       NN         noun, singular or mass
's         PART       POS        possessive ending
back       NOUN       NN         noun, singular or mass
```

In [31]:
```
doc1= nlp(u"I read books in NLP")
```

In [32]:
```
for token in doc1:
    print(f"{token.text:{10}} {token.pos_:{10}} {token.tag_:{10}} {spacy.expla
```

```
I          PRON       PRP        pronoun, personal
read       VERB       VBP        verb, non-3rd person singular present
books      NOUN       NNS        noun, plural
in         ADP        IN         conjunction, subordinating or preposition
NLP        PROPN      NNP        noun, proper singular
```

In [33]:
```
doc2= nlp(u"I read a book on NLP")
```

In [34]:
```
for token in doc2:
    print(f"{token.text:{10}} {token.pos_:{10}} {token.tag_:{10}} {spacy.expla
```

```
I          PRON       PRP        pronoun, personal
read       VERB       VBD        verb, past tense
a          DET        DT         determiner
book       NOUN       NN         noun, singular or mass
on         ADP        IN         conjunction, subordinating or preposition
NLP        PROPN      NNP        noun, proper singular
```

In [2]:
```
doc=nlp(u"the quick brown fox jumped over teh lazy dog's back")
```

In [3]:
```
POS_counts=doc.count_by(spacy.attrs.POS)
```

In [4]:
```
POS_counts
```

Out[4]:
```
{90: 1, 84: 3, 92: 4, 100: 1, 85: 1, 94: 1}
```

In [6]:
```
doc.vocab[84].text
```

Out[6]:
```
'ADJ'
```

In [7]:
```
doc[2]
```

Out[7]:
```
brown
```

In [9]:
```
doc[2].text
```

Out[9]:
```
'brown'
```

In [10]:
```
doc[2].pos
```

Out[10]:
```
84
```

In [11]:
```
for k, v in sorted(POS_counts.items()):
    print(f"{k}.{doc.vocab[k].text:{5}} {v}")
```

```
84.ADJ   3
85.ADP   1
90.DET   1
92.NOUN  4
94.PART  1
100.VERB  1
```

In [12]:
```
len(doc.vocab)
```

Out[12]:
```
781
```

In [14]:
```
#Visulize the part of speech
import spacy
nlp=spacy.load('en_core_web_sm')
```

In [15]:
```
doc=nlp(u"The quick brown fox jumped over the lazy dog in delhi India last year
```

In [16]:
```
from spacy import displacy
```

In [18]:
```
displacy.render(doc, style='dep',jupyter=True)
```

The DET quick ADJ brown ADJ fox NOUN jumped VERB over ADP the DET lazy ADJ dog NOUN in ADP delhi PROPN India PROPN last ADJ year NOUN while SCONJ going VERB to ADP the DET market NOUN det amod amod nsubj prep det amod pobj prep amod pobj amod npadvmod mark advcl prep det pobj

In [20]:
```
options={'distance':110,'compact':'True','color':'yellow','bg':'#09a3d5','font
```

In [21]:
```
displacy.render(doc, style='dep',jupyter=True,options=options)
```

The DET quick ADJ brown ADJ fox NOUN jumped VERB over ADP the DET lazy ADJ dog NOUN in ADP delhi PROPN India PROPN last ADJ year NOUN while SCONJ going VERB to ADP the DET market NOUN det amod amod nsubj prep det amod pobj prep amod pobj amod npadvmod mark advcl prep det pobj

In [22]:
```
doc2=nlp(u"This is the sentence. This is another sentence, possibiliy longer th
```

In [23]:
```
spans=list(doc2.sents)
```

In [24]:
```
displacy.serve(spans, style='dep', options={'distance':110})
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\spacy\displacy\__init__.py:106: U
serWarning: [W011] It looks like you're calling displacy.serve from within a
Jupyter notebook or a similar environment. This likely means you're already r
unning a local web server, so there's no need to make displaCy start another
one. Instead, you should be able to replace displacy.serve with displacy.rend
er to show the visualization.
  warnings.warn(Warnings.W011)
```

displaCy
This PRON is AUX the DET sentence. NOUN nsubj det attr

This PRON is AUX another DET sentence, NOUN possibiliy ADV longer ADV than ADP others NOUN nsubj det attr npadvmod advmod prep pobj

```
Using the 'dep' visualizer
Serving on http://0.0.0.0:5000 (http://0.0.0.0:5000) ...


127.0.0.1 - - [13/Jan/2024 17:20:32] "GET / HTTP/1.1" 200 9805
127.0.0.1 - - [13/Jan/2024 17:20:33] "GET /favicon.ico HTTP/1.1" 200 9805

Shutting down server on port 5000.
```

In [1]:
```
##NER
import spacy
nlp=spacy.load('en_core_web_sm')
```

In [26]:
```
def show_ents(doc):
    if doc.ents:
        for ent in doc.ents:
            print(ent.text +' -'+ ent.label_+'-' + str(spacy.explain(ent.label_
    else:
        print ('No entity found')
```

In [27]:
```
doc= nlp(u"How are you")
```

In [28]:
```
show_ents(doc)
```

```
No entity found
```

In [29]:
```
doc= nlp(u" Next year i am planning to visit U.S.A in march and after that i wi
```

In [30]:
```
show_ents(doc)
```

```
Next year -DATE-Absolute or relative dates or periods
U.S.A -GPE-Countries, cities, states
TCS -ORG-Companies, agencies, institutions, etc.
India -GPE-Countries, cities, states
```

In [31]:
```
doc= nlp(u"Can i get Microsoft shares in 500 millions dollar")
```

In [32]:
```
show_ents(doc)
```

```
Microsoft -ORG-Companies, agencies, institutions, etc.
500 millions dollar -MONEY-Monetary values, including unit
```

In [33]:
```
doc= nlp(u"Tesla is to built $5 millions factory in India this year")
```

In [34]:
```
show_ents(doc)
```

```
Tesla -ORG-Companies, agencies, institutions, etc.
$5 millions -MONEY-Monetary values, including unit
India -GPE-Countries, cities, states
this year -DATE-Absolute or relative dates or periods
```

In [35]:
```
doc= nlp(u"Levadata is to built $5 millions factory in India this year")
```

In [36]:
```
show_ents(doc)
```

```
Levadata -ORG-Companies, agencies, institutions, etc.
$5 millions -MONEY-Monetary values, including unit
India -GPE-Countries, cities, states
this year -DATE-Absolute or relative dates or periods
```

In [45]:
```
doc= nlp(u"Nilesh is to built $5 millions factory in India this year")
```

In [46]:
```
show_ents(doc)
```

```
$5 millions -MONEY-Monetary values, including unit
India -GPE-Countries, cities, states
this year -DATE-Absolute or relative dates or periods
```

In [38]:
```
#to add label in doc

from spacy.tokens import Span
```

In [39]:
```
ORG=doc.vocab.strings[u"ORG"]
```

In [40]:
```
ORG
```

Out[40]: 383

In [47]:
```
new_ent=Span(doc,0,1,label=ORG)
```

In [49]:
```
doc.ents=list(doc.ents)+[new_ent]
```

In [50]:
```
show_ents(doc)
```

```
Nilesh -ORG-Companies, agencies, institutions, etc.
$5 millions -MONEY-Monetary values, including unit
India -GPE-Countries, cities, states
this year -DATE-Absolute or relative dates or periods
```

In [90]:
```
#How to labelmuliple data in String

doc=nlp(u'Our compnay created brand new vacuum cleaner.'
        u'This new vacuum-cleaner is best in the market')
```

In [52]:
```
show_ents(doc)
```

```
No entity found
```

In [53]:
```
from spacy.matcher import PhraseMatcher
```

In [56]:
```
matcher=PhraseMatcher(nlp.vocab)
```

In [54]:
```
phrase_list=['vacuum cleaner','vacuum-cleaner']
```

In [55]:
```
phrase_patterns=[nlp(text) for text in phrase_list]
```

In [57]:
```
matcher.add('newproduct',None,*phrase_patterns)
```

In [70]:
```
found_matches=matcher(doc)
```

In [71]:
```
found_matches
```

Out[71]:
```
[(2689272359382549672, 5, 7), (2689272359382549672, 10, 13)]
```

In [80]:
```
from spacy.tokens import Span
```

In [91]:
```
PROD=doc.vocab.strings[u"PRODUCT"]
```

In [82]:
```
found_matches
```

Out[82]:
```
[(2689272359382549672, 5, 7), (2689272359382549672, 10, 13)]
```

In [83]:
```
new_ents=[Span(doc,match[1],match[2],label=PROD) for match in found_matches]
```

In [84]:
```
doc_ents=list(doc.ents)+new_ents
```

In [89]:
```
show_ents(doc)
```

```
No entity found
```

In [92]:
```
doc=nlp("In USA $5 dollar is not enough you need atleast $100 dollar in L.A and
```

In [95]:
```
len([x for x in doc.ents if x.label_=='MONEY'])
```

Out[95]: 3

In [21]:
```
from spacy import displacy
```

In [3]:
```
doc=nlp(u"In the last quater Apple sold 500 units iPods for $5 millions and ger
```

In [4]:
```
displacy.render(doc, style='ent', jupyter=True)
```

In the last quater **DATE** Apple **ORG** sold 500 **CARDINAL** units iPods **PRODUCT** for $5 millions **MONEY** and generate 15% **PERCENT** of the profit

In [5]:
```
for sent in doc.sents:
    displacy.render(nlp(sent.text), style='ent', jupyter=True)
```

In the last quater **DATE** Apple **ORG** sold 500 **CARDINAL** units iPods **PRODUCT** for $5 millions **MONEY** and generate 15% **PERCENT** of the profit

In [14]:
```
colors={'ORG':'red'}
options={'ents':['PRODUCT','ORG'],'colors':colors}
```

In [15]:
```
displacy.render(doc, style='ent', jupyter=True, options=options)
```

In the last quater Apple **ORG** sold 500 units iPods **PRODUCT** for $5 millions and generate 15% of the profit

In [1]:
```
##SENTENCE SEGMENTATION
import spacy
nlp=spacy.load('en_core_web_sm')
```

In [2]:
```
doc= nlp(u"This is the first sentence. This is the second senetnce. This is the
```

In [3]:
```
for sent in doc.sents:
    print(sent)
```

```
This is the first sentence.
This is the second senetnce.
This is the last sentence
```

In [4]:
```
doc.sents[0]
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
Cell In[4], line 1
----> 1 doc.sents[0]

TypeError: 'generator' object is not subscriptable
```

In [5]:
```
doc[0]
```

Out[5]:
```
This
```

In [11]:
```
doc1= nlp(u'"managemnt is doing right thing; leadership is also doing right thi
```

In [12]:
```
for sent in doc1.sents:
    print(sent)
    print('\n')
```

```
"managemnt is doing right thing; leadership is also doing right thing."


-peter
```

In [13]:
```
#Add segmentation rule-

def set_custom_boundaries(doc):
    for token in doc1:
        print(token)
        print(token.i)
```

In [14]:
```
set_custom_boundaries(doc1)
```

```
"
0
managemnt
1
is
2
doing
3
right
4
thing
5
;
6
leadership
7
is
8
also
9
doing
10
right
11
thing
12
.
13
"
14
-peter
15
```

In [15]:
```
mystring= nlp(u'This is the first sentence. This is another.\n\n This is a \n t
```

In [16]:
```
print(mystring)
```

```
This is the first sentence. This is another.

 This is a
 third sentence
```

In [17]:
```
for doc in mystring.sents:
    print(doc)
```

```
This is the first sentence.
This is another.


This is a
 third sentence
```

In [19]:
```
with open('F:/daily work/NLP/UPDATED_NLP_COURSE/TextFiles/peterrabbit.txt') as
    doc=nlp(f.read())
```

In [20]:
```
doc
```

Out[20]:
```
The Tale of Peter Rabbit, by Beatrix Potter (1902).

Once upon a time there were four little Rabbits, and their names
were--

          Flopsy,
       Mopsy,
   Cotton-tail,
and Peter.

They lived with their Mother in a sand-bank, underneath the root of a
very big fir-tree.

'Now my dears,' said old Mrs. Rabbit one morning, 'you may go into
the fields or down the lane, but don't go into Mr. McGregor's garden:
your Father had an accident there; he was put in a pie by Mrs.
McGregor.'

'Now run along, and don't get into mischief. I am going out.'
```

In [23]:
```
for token in list(doc.sents)[2]:
    print(f'{token.text:{10}} {token.pos_:{10}} {token.tag_:{10}} {str(spacy.e
```

```
They       PRON       PRP        pronoun, personal
lived      VERB       VBD        verb, past tense
with       ADP        IN         conjunction, subordinating or preposition
their      PRON       PRP$       pronoun, possessive
Mother     NOUN       NN         noun, singular or mass
in         ADP        IN         conjunction, subordinating or preposition
a          DET        DT         determiner
sand       NOUN       NN         noun, singular or mass
-          PUNCT      HYPH       punctuation mark, hyphen
bank       NOUN       NN         noun, singular or mass
,          PUNCT      ,          punctuation mark, comma
underneath ADP        IN         conjunction, subordinating or preposition
the        DET        DT         determiner
root       NOUN       NN         noun, singular or mass
of         ADP        IN         conjunction, subordinating or preposition
a          DET        DT         determiner

          SPACE      _SP        whitespace
very       ADV        RB         adverb
big        ADJ        JJ         adjective (English), other noun-modifier (Ch
inese)
fir        NOUN       NN         noun, singular or mass
-          PUNCT      HYPH       punctuation mark, hyphen
tree       NOUN       NN         noun, singular or mass
.          PUNCT      .          punctuation mark, sentence closer


         SPACE      _SP        whitespace
```

In [1]:
```
import schedule
import time

# Functions setup
def sudo_placement():
    print("Get ready for Sudo Placement at Geeksforgeeks")

def good_luck():
    print("Good Luck for Test")

def work():
    print("Study and work hard")

def bedtime():
    print("It is bed time go rest")

def geeks():
    print("Shaurya says Geeksforgeeks")
```

In [2]:
```
# Task scheduling
# After every 10mins geeks() is called.
schedule.every(2).minutes.do(geeks)
```

Out[2]:
```
Every 2 minutes do geeks() (last run: [never], next run: 2024-01-18 00:34:36)
```

In [3]:
```
# Loop so that the scheduling task
# keeps on running all time.
while True:

    # Checks whether a scheduled task
    # is pending to run or not
    schedule.run_pending()
    time.sleep(1)
```

```
Shaurya says Geeksforgeeks
Shaurya says Geeksforgeeks

---------------------------------------------------------------------------
KeyboardInterrupt                         Traceback (most recent call last)
Cell In[3], line 8
      3 while True:
      4
      5     # Checks whether a scheduled task
      6     # is pending to run or not
      7     schedule.run_pending()
----> 8     time.sleep(1)

KeyboardInterrupt:
```

In [2]:
```
with open('F:/daily work/NLP/UPDATED_NLP_COURSE/TextFiles/peterrabbit.txt') as
    doc=nlp(f.read())
```

In [6]:
```
doc
```

Out[6]:
```
The Tale of Peter Rabbit, by Beatrix Potter (1902).

Once upon a time there were four little Rabbits, and their names
were--

          Flopsy,
       Mopsy,
   Cotton-tail,
and Peter.

They lived with their Mother in a sand-bank, underneath the root of a
very big fir-tree.

'Now my dears,' said old Mrs. Rabbit one morning, 'you may go into
the fields or down the lane, but don't go into Mr. McGregor's garden:
your Father had an accident there; he was put in a pie by Mrs.
McGregor.'

'Now run along, and don't get into mischief. I am going out.'
```

In [5]:
```
for token in list(doc.sents)[2]:
    print(f"{token.text:{10}} {token.pos_:{10}} {str(spacy.explain(token.tag_)
```

```
They       PRON       pronoun, personal
lived      VERB       verb, past tense
with       ADP        conjunction, subordinating or preposition
their      PRON       pronoun, possessive
Mother     NOUN       noun, singular or mass
in         ADP        conjunction, subordinating or preposition
a          DET        determiner
sand       NOUN       noun, singular or mass
-          PUNCT      punctuation mark, hyphen
bank       NOUN       noun, singular or mass
,          PUNCT      punctuation mark, comma
underneath ADP        conjunction, subordinating or preposition
the        DET        determiner
root       NOUN       noun, singular or mass
of         ADP        conjunction, subordinating or preposition
a          DET        determiner

          SPACE      whitespace
very       ADV        adverb
big        ADJ        adjective (English), other noun-modifier (Chinese)
fir        NOUN       noun, singular or mass
-          PUNCT      punctuation mark, hyphen
tree       NOUN       noun, singular or mass
.          PUNCT      punctuation mark, sentence closer


         SPACE      whitespace
```

In [10]:
```
pos_counts=doc.count_by(spacy.attrs.POS)
for k,v in sorted (pos_counts.items()):
    print(f"id:{k} POS:{doc.vocab[k].text} {v} counts")
```

```
id:84 POS:ADJ 53 counts
id:85 POS:ADP 125 counts
id:86 POS:ADV 63 counts
id:87 POS:AUX 49 counts
id:89 POS:CCONJ 61 counts
id:90 POS:DET 90 counts
id:92 POS:NOUN 172 counts
id:93 POS:NUM 9 counts
id:94 POS:PART 28 counts
id:95 POS:PRON 110 counts
id:96 POS:PROPN 74 counts
id:97 POS:PUNCT 171 counts
id:98 POS:SCONJ 19 counts
id:100 POS:VERB 135 counts
id:103 POS:SPACE 99 counts
```

In [11]:
```
len(doc)
```

Out[11]: 1258

In [14]:
```
pos_counts[12]/len(doc)
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
Cell In[14], line 1
----> 1 pos_counts[12]/len(doc)

KeyError: 12
```

In [18]:
```
for ent in doc.ents[:2]:
    print(ent.text + ' '+ ent.label_ + ' '+ str(spacy.explain(ent.label_)))
```

```
The Tale of Peter Rabbit WORK_OF_ART Titles of books, songs, etc.
Beatrix Potter PERSON People, including fictional
```

In [19]:
```
len(list(doc.sents))
```

Out[19]:
```
55
```

In [24]:
```
list_of_sents=[nlp(sent.text) for sent in  doc.sents]
```

In [28]:
```
displacy.render(list_of_sents[0], style="ent", jupyter=True)
```

The Tale of Peter Rabbit **WORK_OF_ART** , by Beatrix Potter **PERSON** (1902 **DATE**).

In [1]:
```
import spacy
nlp=spacy.load('en_core_web_sm')
```

In [2]:
```
##PROJECT

import pandas as pd
import numpy as np
```

In [3]:
```
df=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\smsspamcollecti
```

In [4]:
```
df.head()
```

Out[4]:

	label	message	length	punct
0	ham	Go until jurong point, crazy.. Available only ...	111	9
1	ham	Ok lar... Joking wif u oni...	29	6
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	155	6
3	ham	U dun say so early hor... U c already then say...	49	6
4	ham	Nah I don't think he goes to usf, he lives aro...	61	2

In [5]:
```
df.isnull().sum()
```

Out[5]:
```
label      0
message    0
length     0
punct      0
dtype: int64
```

In [6]:
```
len(df)
```

Out[6]:
```
5572
```

In [7]:
```
df['label'].unique()
```

Out[7]:
```
array(['ham', 'spam'], dtype=object)
```

In [9]:
```
df['label'].value_counts()
```

Out[9]:
```
label
ham     4825
spam     747
Name: count, dtype: int64
```

In [29]:
```
from sklearn.model_selection import train_test_split
```

In [28]:
```
#X features data
X=df[['length','punct']]
#y is our label
y=df['label']

X_train, X_test, y_train, y_test=train_test_split(X,y, test_size=0.3, random_st
```

In [30]:
```
X_train.shape
```

Out[30]:
```
(3900, 2)
```

In [31]:
```
X_test.head()
```

Out[31]:

	length	punct
3245	147	14
944	116	1
1044	102	3
2484	45	0
812	112	4

In [32]:
```
X_test.shape
```

Out[32]:
```
(1672, 2)
```

In [33]:
```
y_train.shape
```

Out[33]:
```
(3900,)
```

In [16]:
```
from sklearn.linear_model import LogisticRegression
```

In [17]:
```
lr_model=LogisticRegression(solver='lbfgs')
```

In [36]:
```
lr_model.fit(X_train,y_train)
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:767:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if not hasattr(array, "sparse") and array.dtypes.apply(is_sparse).any():
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[36]:
```
LogisticRegression()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [37]:
```
from sklearn import metrics
```

In [38]:
```
predictions=lr_model.predict(X_test)
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:767:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if not hasattr(array, "sparse") and array.dtypes.apply(is_sparse).any():
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [39]:
```
predictions
```

Out[39]:
```
array(['ham', 'ham', 'ham', ..., 'ham', 'ham', 'ham'], dtype=object)
```

In [40]:
```
y_test
```

Out[40]:
```
3245     ham
944      ham
1044     ham
2484     ham
812      ham
        ...
2505     ham
2525    spam
4975     ham
650     spam
4463     ham
Name: label, Length: 1672, dtype: object
```

In [23]:
```
print(metrics.confusion_matrix(y_test, predictions))
```

```
[[1404   44]
 [ 219    5]]

C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [41]:
```
print(metrics.classification_report(y_test, predictions))
```

```
              precision    recall  f1-score   support

         ham       0.87      0.97      0.91      1448
        spam       0.10      0.02      0.04       224

    accuracy                           0.84      1672
   macro avg       0.48      0.50      0.48      1672
weighted avg       0.76      0.84      0.80      1672


C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [42]:
```
print(metrics.accuracy_score(y_test, predictions))
```

```
0.8427033492822966

C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [44]:
```
from sklearn.naive_bayes import MultinomialNB
nb_model=MultinomialNB()
nb_model.fit(X_train,y_train)
predictions=nb_model.predict(X_test)
print(metrics.confusion_matrix(y_test,predictions))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:767:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if not hasattr(array, "sparse") and array.dtypes.apply(is_sparse).any():
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:767:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if not hasattr(array, "sparse") and array.dtypes.apply(is_sparse).any():
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

[[1438   10]
 [ 224    0]]
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [45]:
```
print(metrics.classification_report(y_test,predictions))
```

```
              precision    recall  f1-score   support

         ham       0.87      0.99      0.92      1448
        spam       0.00      0.00      0.00       224

    accuracy                           0.86      1672
   macro avg       0.43      0.50      0.46      1672
weighted avg       0.75      0.86      0.80      1672


C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [47]:
```
from sklearn.svm import SVC
```

In [49]:
```
svc_model=SVC(gamma='auto')
svc_model.fit(X_train, y_train)
predictions=svc_model.predict(X_test)
print(metrics.confusion_matrix(y_test, predictions))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:767:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if not hasattr(array, "sparse") and array.dtypes.apply(is_sparse).any():
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:767:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if not hasattr(array, "sparse") and array.dtypes.apply(is_sparse).any():
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

[[1373   75]
 [ 121  103]]
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [50]:
```
print(metrics.classification_report(y_test,predictions))
```

```
              precision    recall  f1-score   support

         ham       0.92      0.95      0.93      1448
        spam       0.58      0.46      0.51       224

    accuracy                           0.88      1672
   macro avg       0.75      0.70      0.72      1672
weighted avg       0.87      0.88      0.88      1672


C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [51]:
```
##Text features Extractions
#Count Vectorization

message=["Hey, let's go to the game today!",
        "Call your sisters.",
        "Want to go walk your dogs"]
```

In [55]:
```
import pandas as pd
import numpy as np
```

In [56]:
```
df=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\smsspamcollecti
```

In [57]:
```
df
```

Out[57]:

	label	message	length	punct
0	ham	Go until jurong point, crazy.. Available only ...	111	9
1	ham	Ok lar... Joking wif u oni...	29	6
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	155	6
3	ham	U dun say so early hor... U c already then say...	49	6
4	ham	Nah I don't think he goes to usf, he lives aro...	61	2
...	...	...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...	160	8
5568	ham	Will ü b going to esplanade fr home?	36	1
5569	ham	Pity, * was in mood for that. So...any other s...	57	7
5570	ham	The guy did some bitching but I acted like i'd...	125	1
5571	ham	Rofl. Its true to its name	26	1

5572 rows × 4 columns

In [58]:
```
df.isnull().sum()
```

Out[58]:
```
label      0
message    0
length     0
punct      0
dtype: int64
```

In [59]:
```
df['label'].value_counts()
```

Out[59]:
```
label
ham     4825
spam     747
Name: count, dtype: int64
```

In [60]:
```
from sklearn.model_selection import train_test_split
```

In [61]:
```
X=df['message']
```

In [62]:
```
y=df['label']
```

In [63]:
```
X_train, X_test, y_train, y_test=train_test_split(X,y, test_size=0.3, random_s
```

In [64]:
```
from sklearn.feature_extraction.text import CountVectorizer
```

In [65]:
```
count_vect=CountVectorizer()
```

In [66]:
```
X
```

Out[66]:
```
0       Go until jurong point, crazy.. Available only ...
1                           Ok lar... Joking wif u oni...
2       Free entry in 2 a wkly comp to win FA Cup fina...
3       U dun say so early hor... U c already then say...
4       Nah I don't think he goes to usf, he lives aro...
                              ...
5567    This is the 2nd time we have tried 2 contact u...
5568                 Will ü b going to esplanade fr home?
5569    Pity, * was in mood for that. So...any other s...
5570    The guy did some bitching but I acted like i'd...
5571                           Rofl. Its true to its name
Name: message, Length: 5572, dtype: object
```

In [68]:
```
X_train_counts=count_vect.fit_transform(X_train)
```

In [69]:
```
X_train_counts
```

Out[69]:
```
<3900x7263 sparse matrix of type '<class 'numpy.int64'>'
        with 52150 stored elements in Compressed Sparse Row format>
```

In [70]:
```
X_train.shape
```

Out[70]:
```
(3900,)
```

In [71]:
```
X_train_counts.shape
```

Out[71]:
```
(3900, 7263)
```

In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
df=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\moviereviews.ts
```

In [4]:
```
df.head()
```

Out[4]:

	label	review
0	neg	how do films like mouse hunt get into theatres...
1	neg	some talented actresses are blessed with a dem...
2	pos	this has been an extraordinary year for austra...
3	pos	according to hollywood movies made in last few...
4	neg	my first press screening of 1998 and already i...

In [5]:
```
len(df)
```

Out[5]:
```
2000
```

In [6]:
```
df['review'][0]
```

Out[6]:
```
'how do films like mouse hunt get into theatres ? \r\nisn\'t there a law or s
omething ? \r\nthis diabolical load of claptrap from steven speilberg\'s drea
mworks studio is hollywood family fare at its deadly worst . \r\nmouse hunt t
akes the bare threads of a plot and tries to prop it up with overacting and f
lat-out stupid slapstick that makes comedies like jingle all the way look dec
ent by comparison . \r\nwriter adam rifkin and director gore verbinski are th
e names chiefly responsible for this swill . \r\nthe plot , for what its wort
h , concerns two brothers ( nathan lane and an appalling lee evens ) who inhe
rit a poorly run string factory and a seemingly worthless house from their ec
centric father . \r\ndeciding to check out the long-abandoned house , they so
on learn that it\'s worth a fortune and set about selling it in auction to th
e highest bidder . \r\nbut battling them at every turn is a very smart mouse
, happy with his run-down little abode and wanting it to stay that way . \r\n
the story alternates between unfunny scenes of the brothers bickering over wh
at to do with their inheritance and endless action sequences as the two take
on their increasingly determined furry foe . \r\nwhatever promise the film st
arts with soon deteriorates into boring dialogue , terrible overacting , and
increasingly uninspired slapstick that becomes all sound and fury , signifyin
g nothing . \r\nthe script becomes so unspeakably bad that the best line poor
lee evens can utter after another run in with the rodent is : " i hate that m
ouse " . \r\noh cringe ! \r\nthis is home alone all over again , and ten time
s worse . \r\none touching scene early on is worth mentioning . \r\nwe follow
the mouse through a maze of walls and pipes until he arrives at his makeshift
abode somewhere in a wall . \r\nhe jumps into a tiny bed , pulls up a makeshi
ft sheet and snuggles up to sleep , seemingly happy and just wanting to be le
ft alone . \r\nit\'s a magical little moment in an otherwise soulless film .
\r\na message to speilberg : if you want dreamworks to be associated with som
e kind of artistic credibility , then either give all concerned in mouse hunt
a swift kick up the arse or hire yourself some decent writers and directors .
\r\nthis kind of rubbish will just not do at all . \r\n'
```

In [7]:
```
df.isnull().sum()
```

Out[7]:
```
label      0
review    35
dtype: int64
```

In [12]:
```
df.dropna(inplace=True)
```

In [13]:
```
df.isnull().sum()
```

Out[13]:
```
label     0
review    0
dtype: int64
```

In [7]:
```
mystring='hello'
empty=' '
```

In [5]:
```
mystring.isspace()
```

Out[5]:
```
False
```

In [8]:
```
empty.isspace()
```

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

In [14]:
```
blank=[]
for i, lb, rv in df.itertuples():
    if rv.isspace():
        blank.append(i)
```

In [15]:
```
blank
```

Out[15]:
```
[57,
 71,
 147,
 151,
 283,
 307,
 313,
 323,
 343,
 351,
 427,
 501,
 633,
 675,
 815,
 851,
 977,
 1079,
 1299,
 1455,
 1493,
 1525,
 1531,
 1763,
 1851,
 1905,
 1993]
```

In [16]:
```
df.drop(blank, inplace=True)
```

In [17]:
```
len(df)
```

Out[17]:
```
1938
```

In [18]:
```
from sklearn.model_selection import train_test_split
```

In [19]:
```
X=df['review']
```

In [20]:
```
y=df['label']
```

In [21]:
```
X_train, X_test, y_train, y_test=train_test_split(X,y, test_size=0.3, random_st
```

In [26]:
```
 from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import LinearSVC
```

In [28]:
```
text_clf=Pipeline([('tfidf',TfidfVectorizer()),('clf',LinearSVC())])
```

In [29]:
```
text_clf.fit(X_train, y_train)
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[29]:
```
Pipeline(steps=[('tfidf', TfidfVectorizer()), ('clf', LinearSVC())])
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [30]:
```
predictions=text_clf.predict(X_test)
```

In [31]:
```
from sklearn.metrics import confusion_matrix, classification_report, accuracy_s
```

In [32]:
```
print(confusion_matrix(y_test, predictions))
```

```
[[235  47]
 [ 41 259]]

C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605: FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614: FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605: FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614: FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605: FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614: FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [33]:
```
print(classification_report(y_test,predictions))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

              precision    recall  f1-score   support

         neg       0.85      0.83      0.84       282
         pos       0.85      0.86      0.85       300

    accuracy                           0.85       582
   macro avg       0.85      0.85      0.85       582
weighted avg       0.85      0.85      0.85       582
```

In [35]:
```
print(accuracy_score(y_test, predictions))
```

```
0.8487972508591065

C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [36]:
```
##Task
```

In [37]:
```
import pandas as pd
import numpy as np
```

In [38]:
```
df=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\moviereviews2.t
```

In [39]:
```
df.head()
```

Out[39]:

	label	review
0	pos	I loved this movie and will watch it again. Or...
1	pos	A warm, touching movie that has a fantasy-like...
2	pos	I was not expecting the powerful filmmaking ex...
3	neg	This so-called "documentary" tries to tell tha...
4	pos	This show has been my escape from reality for ...

In [40]:
```
df.isnull().sum()
```

Out[40]:
```
label      0
review    20
dtype: int64
```

In [44]:
```
df.dropna(inplace=True)
```

In [45]:
```
df.isnull().sum()
```

Out[45]:
```
label     0
review    0
dtype: int64
```

In [46]:
```
len(df)
```

Out[46]:
```
5980
```

In [48]:
```
blank=[]
for i, lb, rv in df.itertuples():
    if rv.isspace():
        blank.append(i)
```

In [49]:
```
blank
```

Out[49]:
```
[]
```

In [51]:
```
df.drop(blank, inplace=True)
```

In [52]:
```
from sklearn.model_selection import train_test_split
```

In [53]:
```
X=df['review']
```

In [54]:
```
y=df['label']
```

In [55]:
```
X_train, X_test, y_train, y_test=train_test_split(X,y, test_size=0.3, random_st
```

In [56]:
```
 from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import LinearSVC
```

In [57]:
```
text_clf=Pipeline([('tfidf',TfidfVectorizer()),('clf',LinearSVC())])
```

In [58]:
```
text_clf.fit(X_train, y_train)
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[58]:
```
Pipeline(steps=[('tfidf', TfidfVectorizer()), ('clf', LinearSVC())])
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [59]:
```
predictions=text_clf.predict(X_test)
```

In [60]:
```
from sklearn.metrics import confusion_matrix, classification_report, accuracy_s
```

In [61]:
```
print(confusion_matrix(y_test, predictions))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

[[821  78]
 [ 58 837]]
```

In [62]:
```
print(classification_report(y_test,predictions))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

              precision    recall  f1-score   support

         neg       0.93      0.91      0.92       899
         pos       0.91      0.94      0.92       895

    accuracy                           0.92      1794
   macro avg       0.92      0.92      0.92      1794
weighted avg       0.92      0.92      0.92      1794
```

In [63]:
```
print(accuracy_score(y_test, predictions))
```

```
0.9241917502787068

C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [1]:
```
#Sentiment Analysis

import spacy
import spacy.cli
spacy.cli.download("en_core_web_lg")
nlp=spacy.load('en_core_web_lg')
```

```
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_lg')
```

In [2]:
```
nlp(u'lion').vector
```

Out[2]:
```
array([  1.2746  ,   0.46242 ,  -1.1829  ,  -5.2661  ,  -2.7128  ,
         1.8521  ,  -0.94273 ,   2.1865  ,   6.503   ,   0.6704  ,
         1.5361  ,   2.5992  ,  -0.36233 ,   4.3965  ,  -6.5644  ,
         1.6141  ,  -1.2897  ,   2.1184  ,  -0.63654 ,  -3.4572  ,
        -4.3771  ,   4.2074  ,  -3.6411  ,  -0.97214 ,   1.3253  ,
        -2.3125  ,  -3.6531  ,  -2.8398  ,   2.7913  ,  -1.53    ,
        -2.9984  ,  -2.6357  ,   0.50615 ,  -2.6925  ,   4.3401  ,
        -5.6017  ,   0.045691,   4.3832  ,  -0.19535 ,  -1.0751  ,
         0.32172 ,   2.4395  ,   4.6638  ,   3.4471  ,  -3.3847  ,
        -1.8238  ,   0.70212 ,   0.58557 ,   5.0032  ,  -3.1072  ,
         1.2364  ,   7.4595  ,   0.057368,   1.0111  ,  -1.0827  ,
         0.69113 ,   2.8009  ,  -3.4383  ,  -1.0599  ,  -2.2627  ,
        -5.149   ,  -5.0636  ,   3.1405  ,   1.0793  ,  -0.72892 ,
        -3.9939  ,  -0.69551 ,  -0.55767 ,   3.2555  ,  -2.9449  ,
         4.7114  ,   1.6388  ,   1.3828  ,   1.4255  ,  -3.2334  ,
        -2.274   ,  -1.8136  ,   2.2966  ,   2.5462  ,   1.0722  ,
        -0.73447 ,   1.2148  ,  -0.9196  ,  -0.065012,   2.088   ,
         0.57002 ,   3.5746  ,   1.7192  ,  -8.335   ,   0.71079 ,
         0.91314 ,  -5.0107  ,   1.899   ,  -4.4658  ,   4.7993  ,
        -0.39899 ,  -2.673   ,  -2.9354  ,   4.304   ,   1.4336  ,
         3.7121  ,   0.34882 ,   4.6512  ,  -4.5731  ,  -4.5665  ,
         1.5988  ,  -0.50383 ,   0.95857 ,   0.68728 ,  -0.39976 ,
        -3.1922  ,   4.4363  ,  -0.69479 ,  -1.9528  ,   4.9376  ,
         2.7259  ,   2.2485  ,   5.5734  ,   2.5842  ,   4.7836  ,
        -1.0274  ,   2.2703  ,  -2.0696  ,  -1.0642  ,  -4.932   ,
        -2.274   ,   4.1409  ,   0.73313 ,   2.1889  ,  -0.098888,
         1.6472  ,  -2.3985  ,   2.5911  ,   3.6026  ,   1.885   ,
         5.7822  ,  -1.4481  ,   1.8914  , -10.044   ,  -5.7452  ,
        -4.3224  ,  -3.854   ,   2.3084  ,  -0.84018 ,  -0.40526 ,
         4.7741  ,  -2.3271  ,   7.064   ,   0.95753 ,  -2.356   ,
         0.83953 ,   0.40004 ,   0.33743 ,   0.8376  ,   3.9285  ,
         0.05955 ,   2.4422  ,   4.3492  ,   3.9861  ,   2.1043  ,
        -1.0197  ,  -0.61752 ,  -0.42999 ,  -0.1014  ,  -5.9571  ,
        -0.53818 ,  -1.7797  ,   1.7446  ,   2.3934  ,  -0.50263 ,
        -1.6222  ,  -0.37372 ,  -6.8938  ,   0.55018 ,  -2.267   ,
         0.64912 ,   3.1525  ,  -2.2541  ,  -4.0384  ,   3.206   ,
         0.14962 ,  -2.6662  ,   0.18167 ,   5.0028  ,   2.1521  ,
         0.92419 ,   5.4163  ,  -2.2408  ,   1.6585  ,  -5.1625  ,
         5.029   ,   0.1026  ,  -0.44542 ,   2.0557  ,   3.7778  ,
         3.8679  ,  -2.7135  ,   5.3242  ,  -3.2916  ,   5.6421  ,
         5.0466  ,   1.6072  ,  -1.3206  ,   4.2044  ,  -0.33793 ,
        -3.1139  ,   2.8841  ,  -3.1565  ,  -2.9832  ,  -0.23235 ,
         2.3259  ,   3.5477  ,  -2.1299  ,  -1.8344  ,   2.7271  ,
         1.5568  ,   5.6865  ,   0.9412  ,  -2.6412  ,  -5.3254  ,
         1.3494  ,  -0.47159 ,   2.4979  ,  -1.5568  ,  -1.6911  ,
        -2.1842  ,   6.0319  ,   0.022573,   2.3824  ,  -1.1002  ,
         0.90216 ,  -1.9113  ,   1.5527  ,   5.7413  ,  -3.1956  ,
         0.68655 ,  -1.6068  ,   1.7404  ,  -3.2142  ,   6.4783  ,
         1.7548  ,  -2.9795  ,   0.97631 ,  -0.018354,  -0.6379  ,
         0.80559 ,   3.1923  ,   3.3335  ,   4.3068  ,  -1.0819  ,
        -1.3839  ,  -4.7626  ,  -4.6637  ,  -1.2201  ,  -3.2741  ,
         1.5204  ,   0.78119 ,   8.7339  ,   1.6009  ,  -0.79332 ,
         5.8416  ,  -1.485   ,   1.5978  ,   2.9746  ,  -0.30759 ,
        -1.8023  ,  -4.8344  ,   1.2817  ,  -2.5469  ,   2.6517  ,
         1.4881  ,   2.1952  ,  -0.12652 ,   1.2223  ,   0.44763 ,
        -3.1445  ,  -2.2051  ,  -4.1785  ,  -3.6539  ,   5.1929  ,
         0.78457 ,  -1.2312  ,   5.5624  ,  -1.8462  ,   6.1262  ,
```

```
  -1.6653  ,  -2.7557  ,  -0.066465,  -3.6362  ,   5.2005  ,
  -1.2865  ,   2.8855  ,   6.1219  ,   1.7824  ,   1.4264  ,
  10.628   ,  -0.36028 ,   1.9268  ,  -7.835   ,   0.57865 ],
dtype=float32)
```

In [3]:
```
nlp(u'The quick brown fox jumped').vector
```

Out[3]:
```
array([-1.615     ,  0.73115796, -0.43866196,  0.017288  ,  1.7033421 ,
       -2.4918282 ,  0.13813403,  3.185108  ,  1.431782  ,  0.16901802,
        2.7527    ,  0.64076406, -2.342192  ,  0.41953403,  0.797184  ,
       -1.1299    ,  1.4179399 ,  1.630098  ,  2.56572   , -1.1786883 ,
       -1.8828003 ,  0.10497598,  1.5746659 , -2.4389    ,  1.6264408 ,
       -0.195288  , -3.663484  , -2.367334  ,  1.120072  ,  1.517998  ,
       -1.5409119 , -0.74236   , -1.70981   ,  0.21122456, -2.4972022 ,
       -2.9451559 ,  0.741366  ,  0.69130594,  2.56254   , -2.1797023 ,
        1.455228  ,  0.19872916,  0.96076   , -0.1895    , -0.50636196,
       -0.3994352 ,  1.72168   , -0.43335405, -2.79734   ,  0.04104001,
        0.870872  ,  3.07908   , -0.85011005,  0.15068403, -0.992664  ,
       -0.5621961 , -0.08378398,  1.3055401 ,  2.0414014 ,  0.24415842,
        0.33115202, -2.2055602 ,  0.27012798,  1.9971001 , -0.91020405,
       -1.43928   , -3.3562179 , -2.326374  ,  1.410608  ,  0.12847403,
        0.52536   ,  0.494448  ,  0.47294015, -0.5851339 , -0.22306204,
       -0.48378196,  0.15025802,  0.48946   ,  0.12913978, -0.73069644,
       -2.379094  , -2.0692801 , -1.0096788 ,  1.25272   ,  1.7908599 ,
       -0.9036261 ,  2.79906   , -0.04914599, -2.5143478 ,  0.61309   ,
       -0.6750078 , -0.32888407,  2.54184   , -0.17071861,  0.26411504,
       -0.08619805,  2.3224058 ,  0.89006007,  2.165062  ,  0.25915796,
        0.8285859 , -1.261076  ,  2.086974  ,  2.1299443 , -0.21456781,
        0.5569401 , -1.60378   ,  1.9442778 , -0.913892  , -0.95242196,
       -0.7128006 ,  0.98087406, -0.5353999 ,  1.931312  ,  0.894812  ,
       -1.6521759 ,  0.886004  , -0.6169411 ,  1.6394199 , -0.11687199,
       -0.30550203, -0.190044  , -0.70105994,  2.47268   , -1.375986  ,
       -0.648144  ,  0.97559243, -0.231462  ,  4.2825003 ,  0.31813997,
       -2.8985977 , -1.7611881 ,  1.7779    , -0.580056  ,  0.39031202,
        0.85677797, -0.60733604, -3.11234   ,  0.97764003, -0.15887599,
       -3.6071002 ,  0.13601403, -1.73926   , -0.18580799, -0.70062   ,
        0.685252  , -0.43827993,  0.99436396,  1.3498939 ,  0.65000397,
        0.7475419 ,  2.8786798 ,  0.20917603, -0.77544403, -0.52630484,
        1.10618   ,  2.88104   , -1.4148799 ,  1.196578  ,  1.7139801 ,
        1.236796  , -1.0417379 ,  1.1953919 , -2.6903958 , -3.408744  ,
       -0.42112   , -1.429724  ,  0.6638    , -0.313608  ,  0.878424  ,
       -0.41962606,  3.11922   ,  1.7436501 ,  0.22487998, -1.51121   ,
       -0.20153204,  0.22349992, -2.05798   , -1.58918   , -0.7827472 ,
       -1.330986  ,  0.09964222,  0.06207442,  0.9869    , -0.30682   ,
       -1.50693   , -1.2281196 , -2.191202  ,  0.47935557, -2.3384001 ,
        1.20602   ,  1.31761   , -0.8693681 , -0.65864   ,  1.2964699 ,
       -0.38857   , -1.44396   , -0.16076404, -1.309934  , -0.21346001,
        0.34005594, -0.535046  , -1.1445577 , -0.7100479 ,  0.93315995,
        1.0069644 , -3.7653117 ,  0.14853014,  0.191566  ,  0.23523481,
        0.435994  ,  1.7084318 , -1.03388   , -0.2700088 ,  1.4422156 ,
        0.40605992,  2.61922   ,  0.139244  , -2.405738  ,  0.10005765,
       -1.648928  ,  0.67755586, -1.522344  ,  1.3883001 , -1.922932  ,
        0.67168695,  0.47529   ,  1.632602  ,  3.0956483 ,  0.7767579 ,
        0.802909  , -3.05359   ,  1.635946  ,  0.578938  , -2.0028203 ,
        1.5184625 , -1.128546  ,  1.2410339 , -0.45437998,  3.326482  ,
        1.8903358 , -2.4961722 ,  1.77654   , -0.15610322, -0.94260806,
        1.4371201 , -0.50928175, -0.59553605, -1.221542  , -0.415412  ,
       -0.624964  , -1.675564  , -2.5494435 , -1.323242  , -0.666676  ,
       -1.729809  ,  1.1727359 ,  4.390082  , -0.28563803, -1.214412  ,
        1.775902  , -0.414936  ,  1.1300812 ,  0.77161705, -1.810062  ,
       -1.393386  , -2.9972858 , -0.30180797, -2.399366  ,  0.261624  ,
        0.6725    , -0.27217802,  0.69558007, -1.088974  , -0.264166  ,
       -2.4113898 ,  0.23572204, -0.84566003, -0.16156402,  1.9245478 ,
       -0.08114004, -1.1815579 ,  0.18797809, -0.64504004,  1.6769199 ,
```

```
       -1.097682  ,  0.07828407, -2.18926   ,  1.10807   ,  0.80761206,
       -1.1212146 ,  0.84139407,  3.769678  ,  1.2326021 , -0.30580598,
        3.210394  ,  0.63234997,  0.676008  , -2.833096  , -1.0153    ],
      dtype=float32)
```

In [4]:
```
nlp(u'The quick brown fox jumped').vector.shape
```

Out[4]:
```
(300,)
```

In [5]:
```
tokens=nlp(u'lion cat pet')
```

In [7]:
```
for token1 in tokens:
    for token2 in tokens:
        print(token1.text, token2.text, token1.similarity(token2))
```

```
lion lion 1.0
lion cat 0.3854507803916931
lion pet 0.20031583309173584
cat lion 0.3854507803916931
cat cat 1.0
cat pet 0.732966423034668
pet lion 0.20031583309173584
pet cat 0.732966423034668
pet pet 1.0
```

In [8]:
```
tokens=nlp(u'love hate like')
```

In [9]:
```
for token1 in tokens:
    for token2 in tokens:
        print(token1.text, token2.text, token1.similarity(token2))
```

```
love love 1.0
love hate 0.5708349943161011
love like 0.5212638974189758
hate love 0.5708349943161011
hate hate 1.0
hate like 0.5065140724182129
like love 0.5212638974189758
like hate 0.5065140724182129
like like 1.0
```

In [10]:
```
len(nlp.vocab.vectors)
```

Out[10]:
```
514157
```

In [11]:
```
nlp.vocab.vectors.shape
```

Out[11]:
```
(514157, 300)
```

In [12]:
```
tokens=nlp(u"dog cat nargle")
```

In [13]:
```
for token in tokens:
    print(token.text, token.has_vector,token.vector_norm, token.is_oov)
```

```
dog True 75.254234 False
cat True 63.188496 False
nargle False 0.0 True
```

In [14]:
```
import nltk
```

In [15]:
```
nltk.download('vader_lexicon')
```

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

Out[15]:
```
True
```

In [16]:
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

In [17]:
```
sid=SentimentIntensityAnalyzer()
```

In [18]:
```
a='This is a good movie'
```

In [19]:
```
sid.polarity_scores(a)
```

Out[19]:
```
{'neg': 0.0, 'neu': 0.508, 'pos': 0.492, 'compound': 0.4404}
```

In [20]:
```
a='This was the best movie ever mode in this History'
```

In [21]:
```
sid.polarity_scores(a)
```

Out[21]:
```
{'neg': 0.0, 'neu': 0.682, 'pos': 0.318, 'compound': 0.6369}
```

In [22]:
```
a='This was the worst movie for last year. Please ignore this movie'
```

In [23]:
```
sid.polarity_scores(a)
```

Out[23]:
```
{'neg': 0.369, 'neu': 0.503, 'pos': 0.128, 'compound': -0.6486}
```

In [24]:
```
import pandas as pd
```

In [25]:
```
df=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\amazonreviews.t
```

In [26]:
```
df
```

Out[26]:

	label	review
0	pos	Stuning even for the non-gamer: This sound tra...
1	pos	The best soundtrack ever to anything.: I'm rea...
2	pos	Amazing!: This soundtrack is my favorite music...
3	pos	Excellent Soundtrack: I truly like this soundt...
4	pos	Remember, Pull Your Jaw Off The Floor After He...
...	...	...
9995	pos	A revelation of life in small town America in ...
9996	pos	Great biography of a very interesting journali...
9997	neg	Interesting Subject; Poor Presentation: You'd ...
9998	neg	Don't buy: The box looked used and it is obvio...
9999	pos	Beautiful Pen and Fast Delivery.: The pen was ...

10000 rows × 2 columns

In [27]:
```
df['label'].value_counts()
```

Out[27]:
```
label
neg    5097
pos    4903
Name: count, dtype: int64
```

In [28]:
```
df.dropna(inplace=True)
```

In [29]:
```
blanks=[]

for i, lb, rv in df.itertuples():
    if type(rv)== str:
        if rv.isspace():
            blanks.append(i)
```

In [30]:
```
blanks
```

Out[30]:
```
[]
```

In [33]:
```
df.drop(blanks, inplace=True)
```

In [34]:
```
df.iloc[0]['review']
```

Out[34]:
```
'Stuning even for the non-gamer: This sound track was beautiful! It paints th
e senery in your mind so well I would recomend it even to people who hate vi
d. game music! I have played the game Chrono Cross but out of all of the game
s I have ever played it has the best music! It backs away from crude keyboard
ing and takes a fresher step with grate guitars and soulful orchestras. It wo
uld impress anyone who cares to listen! ^_^'
```

In [35]:
```
sid.polarity_scores(df.iloc[0]['review'])
```

Out[35]:
```
{'neg': 0.088, 'neu': 0.669, 'pos': 0.243, 'compound': 0.9454}
```

In [36]:
```
df['scores']=df['review'].apply(lambda review: sid.polarity_scores(review))
```

In [38]:
```
df.head()
```

Out[38]:

	label	review	scores
0	pos	Stuning even for the non-gamer: This sound tra...	{'neg': 0.088, 'neu': 0.669, 'pos': 0.243, 'co...
1	pos	The best soundtrack ever to anything.: I'm rea...	{'neg': 0.018, 'neu': 0.837, 'pos': 0.145, 'co...
2	pos	Amazing!: This soundtrack is my favorite music...	{'neg': 0.04, 'neu': 0.692, 'pos': 0.268, 'com...
3	pos	Excellent Soundtrack: I truly like this soundt...	{'neg': 0.09, 'neu': 0.615, 'pos': 0.295, 'com...
4	pos	Remember, Pull Your Jaw Off The Floor After He...	{'neg': 0.0, 'neu': 0.746, 'pos': 0.254, 'comp...

In [39]:
```
df['compound']=df['scores'].apply(lambda d:d['compound'])
```

In [40]:
```
df.head()
```

Out[40]:

	label	review	scores	compound
0	pos	Stuning even for the non-gamer: This sound tra...	{'neg': 0.088, 'neu': 0.669, 'pos': 0.243, 'co...	0.9454
1	pos	The best soundtrack ever to anything.: I'm rea...	{'neg': 0.018, 'neu': 0.837, 'pos': 0.145, 'co...	0.8957
2	pos	Amazing!: This soundtrack is my favorite music...	{'neg': 0.04, 'neu': 0.692, 'pos': 0.268, 'com...	0.9858
3	pos	Excellent Soundtrack: I truly like this soundt...	{'neg': 0.09, 'neu': 0.615, 'pos': 0.295, 'com...	0.9814
4	pos	Remember, Pull Your Jaw Off The Floor After He...	{'neg': 0.0, 'neu': 0.746, 'pos': 0.254, 'comp...	0.9781

In [41]:
```
df['comp_score']=df['compound'].apply(lambda score: 'pos' if score>=0 else 'neg
```

In [42]:
```
df.head()
```

Out[42]:

	label	review	scores	compound	comp_score
0	pos	Stuning even for the non-gamer: This sound tra...	{'neg': 0.088, 'neu': 0.669, 'pos': 0.243, 'co...	0.9454	pos
1	pos	The best soundtrack ever to anything.: I'm rea...	{'neg': 0.018, 'neu': 0.837, 'pos': 0.145, 'co...	0.8957	pos
2	pos	Amazing!: This soundtrack is my favorite music...	{'neg': 0.04, 'neu': 0.692, 'pos': 0.268, 'com...	0.9858	pos
3	pos	Excellent Soundtrack: I truly like this soundt...	{'neg': 0.09, 'neu': 0.615, 'pos': 0.295, 'com...	0.9814	pos
4	pos	Remember, Pull Your Jaw Off The Floor After He...	{'neg': 0.0, 'neu': 0.746, 'pos': 0.254, 'comp...	0.9781	pos

In [43]:
```
from sklearn.metrics import accuracy_score,classification_report, confusion_mat
```

In [44]:
```
accuracy_score(df['label'], df['comp_score'])
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[44]:
```
0.7097
```

In [45]:
```
print(classification_report(df['label'], df['comp_score']))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

	precision	recall	f1-score	support
neg	0.86	0.52	0.64	5097
pos	0.64	0.91	0.75	4903
accuracy			0.71	10000
macro avg	0.75	0.71	0.70	10000
weighted avg	0.75	0.71	0.70	10000

In [46]:
```
print(confusion_matrix(df['label'],df['comp_score']))
```

```
[[2629 2468]
 [ 435 4468]]

C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614: FutureWarning: is_sparse is deprecated and will be removed in a future version. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [73]:
```
#Sentiment analysis

%time
import numpy as np
import pandas as pd
df=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\moviereviews.ts
```

```
CPU times: total: 0 ns
Wall time: 1 ms
```

In [49]:
```
df=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\moviereviews.ts
```

In [50]:
```
df.head()
```

Out[50]:

	label	review
0	neg	how do films like mouse hunt get into theatres...
1	neg	some talented actresses are blessed with a dem...
2	pos	this has been an extraordinary year for austra...
3	pos	according to hollywood movies made in last few...
4	neg	my first press screening of 1998 and already i...

In [51]:
```
df.dropna(inplace=True)
```

In [52]:
```
df.count()
```

Out[52]:
```
label     1965
review    1965
dtype: int64
```

In [53]:
```
blanks=[]

for i, lb, rv in df.itertuples():
    if type(rv)== str:
        if rv.isspace():
            blanks.append(i)
```

In [54]:
```
blanks
```

Out[54]:
```
[57,
 71,
 147,
 151,
 283,
 307,
 313,
 323,
 343,
 351,
 427,
 501,
 633,
 675,
 815,
 851,
 977,
 1079,
 1299,
 1455,
 1493,
 1525,
 1531,
 1763,
 1851,
 1905,
 1993]
```

In [56]:
```
df.drop(blanks,inplace=True)
```

In [59]:
```
df['label'].value_counts()
```

Out[59]:
```
label
neg    969
pos    969
Name: count, dtype: int64
```

In [60]:
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

In [61]:
```
sid=SentimentIntensityAnalyzer()
```

In [62]:
```
df['scores']=df['review'].apply(lambda review:sid.polarity_scores(review))
```

In [63]:
```
df['compound']=df['scores'].apply(lambda d:d['compound'])
```

In [64]:
```
df.head()
```

Out[64]:

	label	review	scores	compound
0	neg	how do films like mouse hunt get into theatres...	{'neg': 0.121, 'neu': 0.778, 'pos': 0.101, 'co...	-0.9125
1	neg	some talented actresses are blessed with a dem...	{'neg': 0.12, 'neu': 0.775, 'pos': 0.105, 'com...	-0.8618
2	pos	this has been an extraordinary year for austra...	{'neg': 0.068, 'neu': 0.781, 'pos': 0.15, 'com...	0.9951
3	pos	according to hollywood movies made in last few...	{'neg': 0.071, 'neu': 0.782, 'pos': 0.147, 'co...	0.9972
4	neg	my first press screening of 1998 and already i...	{'neg': 0.091, 'neu': 0.817, 'pos': 0.093, 'co...	-0.2484

In [65]:
```
df['comp_score']=df['compound'].apply(lambda score:'pos' if score >=0 else 'neg
```

In [66]:
```
df.head()
```

Out[66]:

	label	review	scores	compound	comp_score
0	neg	how do films like mouse hunt get into theatres...	{'neg': 0.121, 'neu': 0.778, 'pos': 0.101, 'co...	-0.9125	neg
1	neg	some talented actresses are blessed with a dem...	{'neg': 0.12, 'neu': 0.775, 'pos': 0.105, 'com...	-0.8618	neg
2	pos	this has been an extraordinary year for austra...	{'neg': 0.068, 'neu': 0.781, 'pos': 0.15, 'com...	0.9951	pos
3	pos	according to hollywood movies made in last few...	{'neg': 0.071, 'neu': 0.782, 'pos': 0.147, 'co...	0.9972	pos
4	neg	my first press screening of 1998 and already i...	{'neg': 0.091, 'neu': 0.817, 'pos': 0.093, 'co...	-0.2484	neg

In [67]:
```
from sklearn.metrics import accuracy_score, classification_report, confusion_ma
```

In [68]:
```
accuracy_score(df['label'],df['comp_score'])
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[68]:
```
0.6357069143446853
```

In [69]:
```python
print(classification_report(df['label'], df['comp_score']))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

	precision	recall	f1-score	support
neg	0.72	0.44	0.55	969
pos	0.60	0.83	0.70	969
accuracy			0.64	1938
macro avg	0.66	0.64	0.62	1938
weighted avg	0.66	0.64	0.62	1938

In [70]:
```
print(confusion_matrix(df['label'],df['comp_score']))
```

```
[[427 542]
 [164 805]]

C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [2]:
```
#Sentiment Analysis

import spacy
import spacy.cli
spacy.cli.download("en_core_web_lg")
nlp=spacy.load('en_core_web_lg')
```

```
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_lg')
```

In [3]:
```
nlp(u'lion').vector
```

Out[3]:
```
array([  1.2746  ,   0.46242 ,  -1.1829  ,  -5.2661  ,  -2.7128  ,
         1.8521  ,  -0.94273 ,   2.1865  ,   6.503   ,   0.6704  ,
         1.5361  ,   2.5992  ,  -0.36233 ,   4.3965  ,  -6.5644  ,
         1.6141  ,  -1.2897  ,   2.1184  ,  -0.63654 ,  -3.4572  ,
        -4.3771  ,   4.2074  ,  -3.6411  ,  -0.97214 ,   1.3253  ,
        -2.3125  ,  -3.6531  ,  -2.8398  ,   2.7913  ,  -1.53    ,
        -2.9984  ,  -2.6357  ,   0.50615 ,  -2.6925  ,   4.3401  ,
        -5.6017  ,   0.045691,   4.3832  ,  -0.19535 ,  -1.0751  ,
         0.32172 ,   2.4395  ,   4.6638  ,   3.4471  ,  -3.3847  ,
        -1.8238  ,   0.70212 ,   0.58557 ,   5.0032  ,  -3.1072  ,
         1.2364  ,   7.4595  ,   0.057368,   1.0111  ,  -1.0827  ,
         0.69113 ,   2.8009  ,  -3.4383  ,  -1.0599  ,  -2.2627  ,
        -5.149   ,  -5.0636  ,   3.1405  ,   1.0793  ,  -0.72892 ,
        -3.9939  ,  -0.69551 ,  -0.55767 ,   3.2555  ,  -2.9449  ,
         4.7114  ,   1.6388  ,   1.3828  ,   1.4255  ,  -3.2334  ,
        -2.274   ,  -1.8136  ,   2.2966  ,   2.5462  ,   1.0722  ,
        -0.73447 ,   1.2148  ,  -0.9196  ,  -0.065012,   2.088   ,
         0.57002 ,   3.5746  ,   1.7192  ,  -8.335   ,   0.71079 ,
         0.91314 ,  -5.0107  ,   1.899   ,  -4.4658  ,   4.7993  ,
        -0.39899 ,  -2.673   ,  -2.9354  ,   4.304   ,   1.4336  ,
         3.7121  ,   0.34882 ,   4.6512  ,  -4.5731  ,  -4.5665  ,
         1.5988  ,  -0.50383 ,   0.95857 ,   0.68728 ,  -0.39976 ,
        -3.1922  ,   4.4363  ,  -0.69479 ,  -1.9528  ,   4.9376  ,
         2.7259  ,   2.2485  ,   5.5734  ,   2.5842  ,   4.7836  ,
        -1.0274  ,   2.2703  ,  -2.0696  ,  -1.0642  ,  -4.932   ,
        -2.274   ,   4.1409  ,   0.73313 ,   2.1889  ,  -0.098888,
         1.6472  ,  -2.3985  ,   2.5911  ,   3.6026  ,   1.885   ,
         5.7822  ,  -1.4481  ,   1.8914  , -10.044   ,  -5.7452  ,
        -4.3224  ,  -3.854   ,   2.3084  ,  -0.84018 ,  -0.40526 ,
         4.7741  ,  -2.3271  ,   7.064   ,   0.95753 ,  -2.356   ,
         0.83953 ,   0.40004 ,   0.33743 ,   0.8376  ,   3.9285  ,
         0.05955 ,   2.4422  ,   4.3492  ,   3.9861  ,   2.1043  ,
        -1.0197  ,  -0.61752 ,  -0.42999 ,  -0.1014  ,  -5.9571  ,
        -0.53818 ,  -1.7797  ,   1.7446  ,   2.3934  ,  -0.50263 ,
        -1.6222  ,  -0.37372 ,  -6.8938  ,   0.55018 ,  -2.267   ,
         0.64912 ,   3.1525  ,  -2.2541  ,  -4.0384  ,   3.206   ,
         0.14962 ,  -2.6662  ,   0.18167 ,   5.0028  ,   2.1521  ,
         0.92419 ,   5.4163  ,  -2.2408  ,   1.6585  ,  -5.1625  ,
         5.029   ,   0.1026  ,  -0.44542 ,   2.0557  ,   3.7778  ,
         3.8679  ,  -2.7135  ,   5.3242  ,  -3.2916  ,   5.6421  ,
         5.0466  ,   1.6072  ,  -1.3206  ,   4.2044  ,  -0.33793 ,
        -3.1139  ,   2.8841  ,  -3.1565  ,  -2.9832  ,  -0.23235 ,
         2.3259  ,   3.5477  ,  -2.1299  ,  -1.8344  ,   2.7271  ,
         1.5568  ,   5.6865  ,   0.9412  ,  -2.6412  ,  -5.3254  ,
         1.3494  ,  -0.47159 ,   2.4979  ,  -1.5568  ,  -1.6911  ,
        -2.1842  ,   6.0319  ,   0.022573,   2.3824  ,  -1.1002  ,
         0.90216 ,  -1.9113  ,   1.5527  ,   5.7413  ,  -3.1956  ,
         0.68655 ,  -1.6068  ,   1.7404  ,  -3.2142  ,   6.4783  ,
         1.7548  ,  -2.9795  ,   0.97631 ,  -0.018354,  -0.6379  ,
         0.80559 ,   3.1923  ,   3.3335  ,   4.3068  ,  -1.0819  ,
        -1.3839  ,  -4.7626  ,  -4.6637  ,  -1.2201  ,  -3.2741  ,
         1.5204  ,   0.78119 ,   8.7339  ,   1.6009  ,  -0.79332 ,
         5.8416  ,  -1.485   ,   1.5978  ,   2.9746  ,  -0.30759 ,
        -1.8023  ,  -4.8344  ,   1.2817  ,  -2.5469  ,   2.6517  ,
         1.4881  ,   2.1952  ,  -0.12652 ,   1.2223  ,   0.44763 ,
        -3.1445  ,  -2.2051  ,  -4.1785  ,  -3.6539  ,   5.1929  ,
         0.78457 ,  -1.2312  ,   5.5624  ,  -1.8462  ,   6.1262  ,
```

```
       -1.6653  ,  -2.7557  ,  -0.066465,  -3.6362  ,   5.2005  ,
       -1.2865  ,   2.8855  ,   6.1219  ,   1.7824  ,   1.4264  ,
       10.628   ,  -0.36028 ,   1.9268  ,  -7.835   ,   0.57865 ],
      dtype=float32)
```

In [3]:
```
nlp(u'the quick brown fox jumped').vector.shape
```

Out[3]:
```
(300,)
```

In [4]:
```
nlp(u'fox').vector.shape
```

Out[4]:
```
(300,)
```

In [6]:
```
tokens=nlp(u'dog cat pet')
```

In [7]:
```
for token1 in tokens:
    for token2 in tokens:
        print(token1.text, token2.text, token1.similarity(token2))
```

```
dog dog 1.0
dog cat 0.8220816850662231
dog pet 0.7856059074401855
cat dog 0.8220816850662231
cat cat 1.0
cat pet 0.732966423034668
pet dog 0.7856059074401855
pet cat 0.732966423034668
pet pet 1.0
```

In [8]:
```
tokens=nlp(u'love like hate')
```

In [9]:
```
for token1 in tokens:
    for token2 in tokens:
        print(token1.text, token2.text, token1.similarity(token2))
```

```
love love 1.0
love like 0.5212638974189758
love hate 0.5708349943161011
like love 0.5212638974189758
like like 1.0
like hate 0.5065140724182129
hate love 0.5708349943161011
hate like 0.5065140724182129
hate hate 1.0
```

In [10]:
```
len(nlp.vocab.vectors)
```

Out[10]:
```
514157
```

In [11]:
```
nlp.vocab.vectors.shape
```

Out[11]:
```
(514157, 300)
```

In [16]:
```
tokens=nlp(u'dog cat nargle')
```

In [17]:
```
for token in tokens:
    print(token.text, token.has_vector, token.vector_norm, token.is_oov)
```

```
dog True 75.254234 False
cat True 63.188496 False
nargle False 0.0 True
```

In [4]:
```
import nltk
nltk.download('vader_lexicon')
```

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

Out[4]:
```
True
```

In [6]:
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

In [7]:
```
sid=SentimentIntensityAnalyzer()
```

In [8]:
```
a="This is a good movie"
```

In [9]:
```
sid.polarity_scores(a)
```

Out[9]:
```
{'neg': 0.0, 'neu': 0.508, 'pos': 0.492, 'compound': 0.4404}
```

In [10]:
```
a="This is the best, most awesome movie EVER MADE!!!"
```

In [11]:
```
sid.polarity_scores(a)
```

Out[11]:
```
{'neg': 0.0, 'neu': 0.425, 'pos': 0.575, 'compound': 0.8877}
```

In [13]:
```
a="This was the WORST movie that has ever disgraced the screen."
```

In [14]:
```
sid.polarity_scores(a)
```

Out[14]:
```
{'neg': 0.465, 'neu': 0.535, 'pos': 0.0, 'compound': -0.8331}
```

In [15]:
```
import pandas as pd
import numpy as np
```

In [16]:
```
df=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\amazonreviews.t
```

In [17]:
```
df.head()
```

Out[17]:

	label	review
0	pos	Stuning even for the non-gamer: This sound tra...
1	pos	The best soundtrack ever to anything.: I'm rea...
2	pos	Amazing!: This soundtrack is my favorite music...
3	pos	Excellent Soundtrack: I truly like this soundt...
4	pos	Remember, Pull Your Jaw Off The Floor After He...

In [20]:
```
df['label'].value_counts()
```

Out[20]:
```
label
neg    5097
pos    4903
Name: count, dtype: int64
```

In [21]:
```
df.dropna(inplace=True)
```

In [22]:
```
blanks=[]
for i, ib, rv in df.itertuples():
    #index, label, review
    if type(rv)==str:
        if rv.isspace():
            blank.append(i)
```

In [23]:
```
blanks
```

Out[23]:
```
[]
```

In [25]:
```
df.drop(blanks, inplace=True)
```

In [26]:
```
df.iloc[0]['review']
```

Out[26]:
```
'Stuning even for the non-gamer: This sound track was beautiful! It paints th
e senery in your mind so well I would recomend it even to people who hate vi
d. game music! I have played the game Chrono Cross but out of all of the game
s I have ever played it has the best music! It backs away from crude keyboard
ing and takes a fresher step with grate guitars and soulful orchestras. It wo
uld impress anyone who cares to listen! ^_^'
```

In [27]:
```
sid.polarity_scores(df.iloc[0]['review'])
```

Out[27]:
```
{'neg': 0.088, 'neu': 0.669, 'pos': 0.243, 'compound': 0.9454}
```

In [28]:
```
df['scores']=df['review'].apply(lambda review:sid.polarity_scores(review) )
```

In [29]:
```
df.head()
```

Out[29]:

	label	review	scores
0	pos	Stuning even for the non-gamer: This sound tra...	{'neg': 0.088, 'neu': 0.669, 'pos': 0.243, 'co...
1	pos	The best soundtrack ever to anything.: I'm rea...	{'neg': 0.018, 'neu': 0.837, 'pos': 0.145, 'co...
2	pos	Amazing!: This soundtrack is my favorite music...	{'neg': 0.04, 'neu': 0.692, 'pos': 0.268, 'com...
3	pos	Excellent Soundtrack: I truly like this soundt...	{'neg': 0.09, 'neu': 0.615, 'pos': 0.295, 'com...
4	pos	Remember, Pull Your Jaw Off The Floor After He...	{'neg': 0.0, 'neu': 0.746, 'pos': 0.254, 'comp...

In [30]:
```
df['compound']=df['scores'].apply(lambda d:d['compound'])
```

In [31]:
```
df.head()
```

Out[31]:

	label	review	scores	compound
0	pos	Stuning even for the non-gamer: This sound tra...	{'neg': 0.088, 'neu': 0.669, 'pos': 0.243, 'co...	0.9454
1	pos	The best soundtrack ever to anything.: I'm rea...	{'neg': 0.018, 'neu': 0.837, 'pos': 0.145, 'co...	0.8957
2	pos	Amazing!: This soundtrack is my favorite music...	{'neg': 0.04, 'neu': 0.692, 'pos': 0.268, 'com...	0.9858
3	pos	Excellent Soundtrack: I truly like this soundt...	{'neg': 0.09, 'neu': 0.615, 'pos': 0.295, 'com...	0.9814
4	pos	Remember, Pull Your Jaw Off The Floor After He...	{'neg': 0.0, 'neu': 0.746, 'pos': 0.254, 'comp...	0.9781

In [32]:
```
df['comp_score']=df['compound'].apply(lambda score: 'pos'if score>0 else 'neg'
```

In [33]:
```
df.head()
```

Out[33]:

	label	review	scores	compound	comp_score
0	pos	Stuning even for the non-gamer: This sound tra...	{'neg': 0.088, 'neu': 0.669, 'pos': 0.243, 'co...	0.9454	pos
1	pos	The best soundtrack ever to anything.: I'm rea...	{'neg': 0.018, 'neu': 0.837, 'pos': 0.145, 'co...	0.8957	pos
2	pos	Amazing!: This soundtrack is my favorite music...	{'neg': 0.04, 'neu': 0.692, 'pos': 0.268, 'com...	0.9858	pos
3	pos	Excellent Soundtrack: I truly like this soundt...	{'neg': 0.09, 'neu': 0.615, 'pos': 0.295, 'com...	0.9814	pos
4	pos	Remember, Pull Your Jaw Off The Floor After He...	{'neg': 0.0, 'neu': 0.746, 'pos': 0.254, 'comp...	0.9781	pos

In [34]:
```
from sklearn.metrics import accuracy_score, classification_report, confusion_ma
```

In [36]:
```
accuracy_score(df['label'],df['comp_score'])
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[36]:
```
0.713
```

In [37]:
```
print(classification_report(df['label'],df['comp_score']))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

	precision	recall	f1-score	support
neg	0.85	0.53	0.65	5097
pos	0.65	0.90	0.75	4903
accuracy			0.71	10000
macro avg	0.75	0.72	0.70	10000
weighted avg	0.75	0.71	0.70	10000

In [38]:
```
print(confusion_matrix(df['label'], df['comp_score']))
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):

[[2716 2381]
 [ 489 4414]]
```

Sentiment Analysis Project

In [39]:
```
import pandas as pd
import numpy as np
```

In [41]:
```
df=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\TextFiles\moviereviews.ts
```

In [42]:
```
df.head()
```

Out[42]:

	label	review
0	neg	how do films like mouse hunt get into theatres...
1	neg	some talented actresses are blessed with a dem...
2	pos	this has been an extraordinary year for austra...
3	pos	according to hollywood movies made in last few...
4	neg	my first press screening of 1998 and already i...

In [43]:
```
df.dropna(inplace=True)
```

In [45]:
```
blanks=[]
for i, lb, rv in df.itertuples():
    #index, label, review
    if type(rv)==str:
        if rv.isspace():
            blanks.append(i)
```

In [46]:
```
blanks
```

Out[46]:
```
[57,
 71,
 147,
 151,
 283,
 307,
 313,
 323,
 343,
 351,
 427,
 501,
 633,
 675,
 815,
 851,
 977,
 1079,
 1299,
 1455,
 1493,
 1525,
 1531,
 1763,
 1851,
 1905,
 1993]
```

In [48]:
```
df.drop(blanks, inplace=True)
```

In [51]:
```
df['label'].value_counts()
```

Out[51]:
```
label
neg    969
pos    969
Name: count, dtype: int64
```

In [52]:
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

In [53]:
```
sid=SentimentIntensityAnalyzer()
```

In [56]:
```
df['score']=df['review'].apply(lambda review:sid.polarity_scores(review))
```

In [57]:
```
df.head()
```

Out[57]:

	label	review	score
0	neg	how do films like mouse hunt get into theatres...	{'neg': 0.121, 'neu': 0.778, 'pos': 0.101, 'co...
1	neg	some talented actresses are blessed with a dem...	{'neg': 0.12, 'neu': 0.775, 'pos': 0.105, 'com...
2	pos	this has been an extraordinary year for austra...	{'neg': 0.068, 'neu': 0.781, 'pos': 0.15, 'com...
3	pos	according to hollywood movies made in last few...	{'neg': 0.071, 'neu': 0.782, 'pos': 0.147, 'co...
4	neg	my first press screening of 1998 and already i...	{'neg': 0.091, 'neu': 0.817, 'pos': 0.093, 'co...

In [58]:
```
df['compound']=df['score'].apply(lambda d:d['compound'] )
```

In [59]:
```
df.head()
```

Out[59]:

	label	review	score	compound
0	neg	how do films like mouse hunt get into theatres...	{'neg': 0.121, 'neu': 0.778, 'pos': 0.101, 'co...	-0.9125
1	neg	some talented actresses are blessed with a dem...	{'neg': 0.12, 'neu': 0.775, 'pos': 0.105, 'com...	-0.8618
2	pos	this has been an extraordinary year for austra...	{'neg': 0.068, 'neu': 0.781, 'pos': 0.15, 'com...	0.9951
3	pos	according to hollywood movies made in last few...	{'neg': 0.071, 'neu': 0.782, 'pos': 0.147, 'co...	0.9972
4	neg	my first press screening of 1998 and already i...	{'neg': 0.091, 'neu': 0.817, 'pos': 0.093, 'co...	-0.2484

In [66]:
```
df['comp_score']=df['compound'].apply(lambda score: 'pos' if score>0 else 'neg
```

In [67]:
```
df.head()
```

Out[67]:

	label	review	score	compound	comp_score
0	neg	how do films like mouse hunt get into theatres...	{'neg': 0.121, 'neu': 0.778, 'pos': 0.101, 'co...	-0.9125	neg
1	neg	some talented actresses are blessed with a dem...	{'neg': 0.12, 'neu': 0.775, 'pos': 0.105, 'com...	-0.8618	neg
2	pos	this has been an extraordinary year for austra...	{'neg': 0.068, 'neu': 0.781, 'pos': 0.15, 'com...	0.9951	pos
3	pos	according to hollywood movies made in last few...	{'neg': 0.071, 'neu': 0.782, 'pos': 0.147, 'co...	0.9972	pos
4	neg	my first press screening of 1998 and already i...	{'neg': 0.091, 'neu': 0.817, 'pos': 0.093, 'co...	-0.2484	neg

In [68]:
```
from sklearn.metrics import accuracy_score, classification_report, confusion_ma
```

In [69]:
```
accuracy_score(df['label'], df['comp_score'])
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

Out[69]:
```
0.6357069143446853
```

In [70]:
```
print(classification_report(df['label'],df['comp_score']))
```

```
              precision    recall  f1-score   support

         neg       0.72      0.44      0.55       969
         pos       0.60      0.83      0.70       969

    accuracy                           0.64      1938
   macro avg       0.66      0.64      0.62      1938
weighted avg       0.66      0.64      0.62      1938
```

```
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [71]:
```
print(confusion_matrix(df['label'], df['comp_score']))
```

```
[[427 542]
 [164 805]]

C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:605:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
C:\Users\nilesh\anaconda3\lib\site-packages\sklearn\utils\validation.py:614:
FutureWarning: is_sparse is deprecated and will be removed in a future versio
n. Check `isinstance(dtype, pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

In [2]:
```
#Latent Dirichlet Allocation (LDA) is a popular topic modeling technique to ext
```

In [1]:
```
import pandas as pd
```

In [2]:
```
npr=pd.read_csv(r'F:\daily work\NLP\UPDATED_NLP_COURSE\05-Topic-Modeling\npr.cs
```

In [3]:
```
npr.head()
```

Out[3]:

	Article
0	In the Washington of 2016, even when the polic...
1	Donald Trump has used Twitter — his prefe...
2	Donald Trump is unabashedly praising Russian...
3	Updated at 2:50 p. m. ET, Russian President Vl...
4	From photography, illustration and video, to d...

In [4]:
```
npr['Article'][0]
```

Out[4]: 'In the Washington of 2016, even when the policy can be bipartisan, the politics cannot. And in that sense, this year shows little sign of ending on Dec. 31. When President Obama moved to sanction Russia over its alleged interference in the U. S. election just concluded, some Republicans who had long called for similar or more severe measures could scarcely bring themselves to approve. House Speaker Paul Ryan called the Obama measures ”appropriate” but also ”overdue” and ”a prime example of this administration’s ineffective foreign policy that has left America weaker in the eyes of the world.” Other GOP leaders sounded much the same theme. ”[We have] been urging President Obama for years to take strong action to deter Russia’s worldwide aggression, including its operations,” wrote Rep. Devin Nunes, . chairman of the House Intelligence Committee. ”Now with just a few weeks left in office, the president has suddenly decided that some stronger measures are indeed warranted.” Appearing on CNN, frequent Obama critic Trent Franks, . called for ”much tougher” actions and said three times that Obama had ”finally found his tongue.” Meanwhile, at and on Fox News, various spokesmen for Trump said Obama’s real target was not the Russians at all but the man poised to take over the White House in less than three weeks. They spoke of Obama trying to ”tie Trump’s hands” or ”box him in,” meaning the would be forced either to keep the sanctions or be at odds with Republicans who want to be tougher still on Moscow. Throughout 2016, Trump has repeatedly called not for sanctions but for closer ties with Russia, including cooperation in the fight against ISIS. Russia has battled ISIS in Syria on behalf of that country’s embattled dictator, Bashar Assad, bombing the besieged city of Aleppo that fell to Assad’s forces this week. During the campaign, Trump even urged Russia to ”find” missing emails from the private server of his opponent, Hillary Clinton. He has exchanged public enco miums with Russian President Vladimir Putin on several occasions and added his doubts about the current U. S. levels of support for NATO — Putin’s longtime nemesis. There have also been suggestions that Trump’s extensive business dealings with various Russians are the reason he refuses to release his tax returns. All those issues have been disquieting to some Republicans for many months. Sens. John McCain, . and Lindsay Graham, . C. prominent senior members of the Armed Services Committee, have accepted the assessment of 17 U. S. intelligence agencies regarding the role of Russia in the hacking of various Democratic committees last year. That includes the FBI and CIA consensus that the Russian goal was not just to discredit American democracy but to defeat Clinton and elect Trump. They say the great majority of their Senate colleagues agree with them, and McCain has slated an Armed Services hearing on cyberthreats for Jan. 5. But the politicizing of the Russian actions — the idea that they helped Trump win — has also made the issue difficult for Republican leaders. It has allowed Trump supporters to push back on the intelligence agencies and say the entire issue is designed to undermine Trump’s legitimacy. Senate Majority Leader Mitch McConnell has so far resisted calls for a select committee to look into the Russian interference in the 2016 campaign. He has said it is enough for Sen. Richard Burr, . C. to look into it as chairman of the Senate Intelligence Committee. Typically, Republican leaders and spokesmen say there is no evidence that the actual voting or tallying on Nov. 8 was compromised, and that is true. But it is also a red herring, as interference in those functions has not been alleged and is not the focus of the U. S. intelligence agencies’ concern. For his part, Trump has shown little interest in delving into what happened. He has cast doubt on the U. S. intelligence reports to date and suggested ”no one really knows what happened.” He also has suggested that computers make it very difficult to know who is using them. This week, Trump said it was time to ”get on with our lives and do more important things.” However, at week’s end he did agree to have an intelligence briefing on the subject next week. The has not wanted the daily intelligence briefings available to him in recent weeks, preferring that they be given to the

```
men he has chosen as his vice president (Mike Pence) and national security ad
viser (Mike Flynn) with Trump taking them only occasionally. The irony of thi
s controversy arising at the eleventh hour of the Obama presidency can scarce
ly be overstated, and it defines the dilemma facing both the outgoing preside
nt and the incoming party in control. Obama appears to have been reluctant to
retaliate against the Russian hacking before the election for fear of seeming
to interfere with the election himself. The Republicans, meanwhile, have for
years called for greater confrontation with the Russians, with Obama usually
resisting. Obama did join with NATO in punishing the Russians with economic s
anctions over the annexation of Crimea. Those sanctions may have been painfu
l, coming as they did alongside falling prices for oil  —  the commodity tha
t keeps the Russian economy afloat. On other occasions, despite Russian provo
cations through surrogates in Syria and elsewhere, Obama did not make overt m
oves to force Russia's hand. That includes occasions when Russia was believed
to be hacking critical computer systems in neighboring Ukraine, Estonia and P
oland. But this week, following a chorus of confirmation from the U. S. intel
ligence community regarding the Russian role in computer hacking in the polit
ical campaign, Obama acted. He imposed a set of mostly diplomatic actions suc
h as sanctioning some Russian officials, closing two diplomatic compounds and
expelling 35 Russian diplomats. There may have been more damaging measures ta
ken covertly, and some Russophobes in Washington held out hope for that. But
the visible portion of the program scarcely amounted to major retribution. An
d Putin saw fit to diminish the Obama sanctions further by declining to respo
nd. Although his government has steadfastly denied any interference in the U.
S. election, Putin rejected his own foreign minister's recommended package of
responses. (He even sent an invitation for U. S. diplomats to send their chil
dren to a holiday party in Moscow.) That allowed Putin to appear for the mome
nt to be "the bigger man," even as he spurned Obama and kept up what has look
ed like a public bromance with Trump, who tweeted: "Great move on delay (by
V. Putin)  I always knew he was very smart!" At the moment it may seem that
the overall Russia question amounts to the first crisis facing the Trump pres
idency. Whether forced by this campaign interference issue or not, Trump must
grasp the nettle of a relationship Mitt Romney once called the greatest threa
t to U. S. security in the world. To be sure, Trump needs to dispel doubts ab
out his ability to stand up to Putin, who has bullied and cajoled his way to
center stage in recent world affairs. But Trump also seems determined to turn
the page on past U. S. commitments, from free trade philosophy to funding of
NATO and the United Nations. And if his Twitter account is any guide, Trump s
hows little concern about the conundrum others perceive to be facing him. Abo
ve all, Trump has shown himself determined to play by his own rules. A year a
go, many were confident that would not work for him in the world of president
ial politics. We are about to find out whether it works for him in the Oval O
ffice.'
```

In [5]:
```
len(npr)
```

Out[5]: 11992

In [6]:
```
npr['Article'][4]
```

Out[6]: 'From photography, illustration and video, to data visualizations and immersi
ve experiences, visuals are an important part of our storytelling at NPR. Int
erwoven with the written and the spoken word, images — another visual lang
uage — can create deeper understanding and empathy for the struggles and t
riumphs we face together. We told a lot of stories in 2016 — far more than
we can list here. So, instead, here's a small selection of our favorite piece
s, highlighting some of the work we're most proud of, some of the biggest sto
ries we reported, and some of the stories we had the most fun telling. Transp
ort yourself to Rocky Mountain National Park, with all its sights and sounds,
in an immersive geology lesson with Oregon State University geology professor
Eric Kirby, who discusses the geologic history of the Rockies in video. "To
day, Indians use much less energy per person than Americans or Chinese peopl
e. Many of its 1. 2 population live on roughly $2 a day. But what if all of
those people had electricity at night, a refrigerator, a car? "With ambitious
goals to improve the standard of living, and 400 million people lacking relia
ble electricity, 'This means we need to enhance the energy supply by four to
five times what it is now,' says Ajay Mathur, a climate expert who runs the E
nergy and Resources Institute in New Delhi. He says that no matter how fast I
ndia increases its clean energy, like solar and wind, the country will probab
ly also double its use of coal between now and 2030. "Todd Stern, who served
till last month as the top U. S. envoy on climate change, says India has a st
eeper hill to climb than any other country. 'There is no country, probably, w
ith a bigger challenge — looking at the number of people, the level of the
ir economic growth, the number of people who don't have access to electricit
y,' he says." Can India's Sacred But 'Dead' Yamuna River Be Saved? India's Bi
g Battle: Development Vs. Pollution, In India's Sundarbans, People And Tigers
Try To Coexist In A Shrinking Space, "Trying to understand the Trump Organiza
tion is a daunting task. Donald Trump has not released his tax returns, so
the best clues about his privately held business interests come from a financ
ial disclosure form he released in May. "The document covers scores of pages
with small type, and suggests he is financially involved with hundreds of com
panies, including some that simply license his name. "A dive into that disclo
sure form, submitted to the Office of Government Ethics, shows his largest so
urces of revenue are golf courses and rents. But his interests are far flun
g, and include media, retail, entertainment and much more. "Those business in
terests are affected by government agencies and policies. NPR scoured this do
cument to create an overview of some of his business assets and operations (e
xcluding debts) and the possible areas where conflicts may arise." The protes
ts at the Standing Rock Reservation, which started in early 2016, had small r
oots but grew into the thousands, drawing support from Native Americans from
across the country, as well as activists who joined in solidarity against the
proposed route of the Dakota Access Pipeline just north of the reservation.
In December, those protests won a concession from the federal government: The
Army Corps of Engineers announced it would deny the permit necessary to build
the oil pipeline in that area. In Their Own Words: The 'Water Protectors' Of
Standing Rock, Protesters Mark A Solemn Thanksgiving Day At Standing Rock, Pr
otesters, Police Still Clashing Over Disputed North Dakota Pipeline, N. D. Pi
peline Protester: 'It's About Our Rights As Native People' "Up to 1 in 5 kids
living in the U. S. shows signs or symptoms of a mental health disorder in a
given year. So in a school classroom of 25 students, five of them may be stru
ggling with the same issues many adults deal with: depression, anxiety, subst
ance abuse. And yet most children — nearly 80 percent — who need mental
health services won't get them. "Whether treated or not, the children do go t
o school. And the problems they face can tie into major problems found in sch
ools: chronic absence, low achievement, disruptive behavior and dropping out.
"Experts say schools could play a role in identifying students with problems
and helping them succeed. Yet it's a role many schools are not prepared for."

```
”Grapefruit’s bitterness can make it hard to love. Indeed, people often smoth
er it in sugar just to get it down. And yet Americans were once urged to swee
ten it with salt. ”Ad campaigns from the first and second world wars tried to
convince us that ’Grapefruit Tastes Sweeter With Salt!’ as one 1946 ad for Mo
rton’s in Life magazine put it. The pairing, these ads swore, enhanced the fl
avor. ”In our   world, these curious culinary time capsules raise the questio
n: Does salt really make grapefruit taste sweeter? And if this practice was o
nce common, why do few people seem to eat grapefruit this way today?” Rio de
Janeiro hosted the world’s elite athletes in an Olympics that promised transc
endent moments in sports  —   and potential controversies outside of the comp
etition. The Summer Games began Aug. 5, and more than 10, 000 athletes from 2
06 countries participated. From concerns over the Zika virus and Russian athl
etes banned on doping charges to incredible wins by the U. S. women’s gymnast
ics team and sweet moments of   support, the 2016 Olympics was one of the big
gest events  —   and biggest stories  —   of the year. ’A Fantasy Of A Fantas
y’: U. S. Fencer Jason Pryor On Reaching The Olympics, In Rio’s Favelas,   Be
nefits From Olympics Have Yet To Materialize, How The Olympic Medal Tables Ex
plain The World, ”Philando Castile spent his driving career trapped in a seem
ingly endless cycle of traffic stops, fines, court appearances, revocations a
nd reinstatements, raising questions about bias, race and luck. ”Castile’s tr
ouble with traffic stops began when he still had his learner’s permit. He was
stopped a day before his 19th birthday. From there, he descended into a seemi
ngly endless cycle of traffic stops, fines, court appearances, late fees, rev
ocations and reinstatements in various jurisdictions. ”Court records raise bi
g questions: Was Castile targeted by police? Or was he just a careless or unl
ucky driver? ”An NPR analysis of those records shows that the    cafeteria wo
rker who was shot and killed by a police officer during a traffic stop in a S
t. Paul, Minn. suburb, was stopped by police 46 times and racked up more than
$6, 000 in fines. Another curious statistic: Of all of the stops, only six of
them were things a police officer would notice from outside a car  —   things
like speeding or having a broken muffler.” During a week in Cleveland, photog
rapher Gabriella Demczuk explored the ways that people embraced and challenge
d the Republican Party’s mission in this election  —   both from inside and o
utside the party. Then in Philadelphia, Demczuk continued her exploration of
the fractures in America’s political system, examining the Democratic Party’s
attempt to make itself ”stronger together.” True Believers, Protesters And Tr
ump: Scenes From Cleveland, Dissent, Drama And Unity At The Democratic Conven
tion, ” ’With recent events and political environment, these weapons will be
harder to get a hold of.’ ’This is what your   dreams it could be when it gro
ws up.’ ’I can meet . .. near the FL Mall in Orlando or any other time.” ”Cas
h is king.’ ”These classified advertisements for   weapons were listed on Arm
slist, a website where anyone can advertise a firearm they’d like to sell, an
d anyone can contact a seller with an offer to buy. The site is legal. But th
ere’s no way to know whether buyers and sellers who meet through Armslist are
following federal, state or local background check rules. ”We wanted to see h
ow many   firearms  —   defined here as handguns and rifles able to rapidly f
ire a large number of bullets, one shot per trigger pull, without having to r
eload  —   can be currently found on Armslist, and how quickly new listings a
ppear. This provides a window into the difficulty of regulating access to a t
ype of weapon frequently used in mass shootings.” Our favorite albums of the
year draw from all of the genres we cover at NPR Music, from rock, pop and
to classical, jazz, electronic and international artists. These are the recor
ds NPR Music couldn’t stop playing  —   albums that speak to a moment and a l
ifetime, that party, and that exist in their own worlds. Our list of the yea
r’s best songs may begin with Beyoncé and end with Drake, but between those t
wo stars you’ll find a   mix that celebrates all of the music we love. These
are the pop anthems, rallying cries, party jams, riff rockers, perfumed piano
```

```
pieces and emotional exorcisms that we loved to share this year. ”Across the
country, private organizations,  groups and individuals quietly have been wo
rking to ease the plight of Syrian refugees. More than 11, 000 have arrived i
n the U. S. this year, fulfilling a pledge by the Obama administration. That
figure far exceeds the number of Syrian refugees accepted during the previous
four years of the Syrian war, and the White House is calling for a big bump i
n the overall number of refugees next year. ”It had been a long journey for O
sama and Ghada and their four kids, who are among the nearly 5 million Syrian
s who have fled their homeland since the war began in 2011. They survived the
war in Syria and had struggled for three years as refugees in Jordan when the
y were notified by the U. N. refugee agency, UNHCR, that they had been accept
ed for resettlement in the U. S.” ”There are huge gaps in school funding betw
een affluent and  districts. And, with evidence that money matters, especial
ly for disadvantaged kids, something has to change. ”School Money is a  nati
onwide collaboration between NPR’s Ed Team and 20 member station reporters ex
ploring how states pay for their public schools and why many are failing to m
eet the needs of their most vulnerable students.” Is There A Better Way To Pa
y For America’s Schools? Why America’s Schools Have A Money Problem, Presiden
t Obama spoke to NPR as he prepared to leave Washington for the holidays, ref
lecting on the year that was, the 2016 campaign and other news, plus revealin
g what he’s hearing from citizens. In the  exit interview, NPR’s Steve Inske
ep asked Obama about Russian interference in the U. S. election, executive po
wer, the future of the Democratic party and his future role.'
```

In [7]:
```
from sklearn.feature_extraction.text import CountVectorizer
```

In [8]:
```
cv=CountVectorizer(max_df=0.9, min_df=2, stop_words='english')
```

In [9]:
```
dtm=cv.fit_transform(npr['Article'])
```

In [10]:
```
dtm
```

Out[10]:
```
<11992x54777 sparse matrix of type '<class 'numpy.int64'>'
        with 3033388 stored elements in Compressed Sparse Row format>
```

In [11]:
```
from sklearn.decomposition import LatentDirichletAllocation
```

In [12]:
```
LDA=LatentDirichletAllocation(n_components=7, random_state=42)
```

In [13]:
```
LDA.fit(dtm)
```

Out[13]:
```
LatentDirichletAllocation(n_components=7, random_state=42)
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [14]:
```
#Grab the vocabulary of words
# Grab the topics
# Grab the highest probability words per topic
```

In [17]:
```
#Grab the vocabulary of words
len(cv.get_feature_names_out())
```

Out[17]: 54777

In [18]:
```
type(cv.get_feature_names_out())
```

Out[18]: numpy.ndarray

In [20]:
```
cv.get_feature_names_out()[41000]
```

Out[20]: 'reproductive'

In [21]:
```
cv.get_feature_names_out()[35000]
```

Out[21]: 'outlet'

In [26]:
```
import random
random_word_id=random.randint(0,54777)
cv.get_feature_names_out()[random_word_id]
```

Out[26]: 'reconvened'

In [27]:
```
#Grab the topics
```

In [28]:
```
len(LDA.components_)
```

Out[28]: 7

In [29]:
```
type(LDA.components_)
```

Out[29]: numpy.ndarray

In [30]:
```
LDA.components_.shape
```

Out[30]: (7, 54777)

In [31]:
```
LDA.components_
```

Out[31]:
```
array([[8.64332806e+00, 2.38014333e+03, 1.42900522e-01, ...,
        1.43006821e-01, 1.42902042e-01, 1.42861626e-01],
       [2.76191749e+01, 5.36394437e+02, 1.42857148e-01, ...,
        1.42861973e-01, 1.42857147e-01, 1.42906875e-01],
       [7.22783888e+00, 8.24033986e+02, 1.42857148e-01, ...,
        6.14236247e+00, 2.14061364e+00, 1.42923753e-01],
       ...,
       [3.11488651e+00, 3.50409655e+02, 1.42857147e-01, ...,
        1.42859912e-01, 1.42857146e-01, 1.42866614e-01],
       [4.61486388e+01, 5.14408600e+01, 3.14281373e+00, ...,
        1.43107628e-01, 1.43902481e-01, 2.14271779e+00],
       [4.93991422e-01, 4.18841042e+02, 1.42857151e-01, ...,
        1.42857146e-01, 1.43760101e-01, 1.42866201e-01]])
```

In [32]:
```
# Grab the highest probability words per topic
for i, topic in enumerate(LDA.components_):
    print(f'THE TOP 15 WORDS FOR TOPIC #{i}')
    print([cv.get_feature_names_out()[index] for index in topic.argsort()[-15:]
    print('\n')
    print('\n')
```

```
THE TOP 15 WORDS FOR TOPIC #0
['companies', 'money', 'year', 'federal', '000', 'new', 'percent', 'governmen
t', 'company', 'million', 'care', 'people', 'health', 'said', 'says']



THE TOP 15 WORDS FOR TOPIC #1
['military', 'house', 'security', 'russia', 'government', 'npr', 'reports',
'says', 'news', 'people', 'told', 'police', 'president', 'trump', 'said']



THE TOP 15 WORDS FOR TOPIC #2
['way', 'world', 'family', 'home', 'day', 'time', 'water', 'city', 'new', 'ye
ars', 'food', 'just', 'people', 'like', 'says']



THE TOP 15 WORDS FOR TOPIC #3
['time', 'new', 'don', 'years', 'medical', 'disease', 'patients', 'just', 'ch
ildren', 'study', 'like', 'women', 'health', 'people', 'says']



THE TOP 15 WORDS FOR TOPIC #4
['voters', 'vote', 'election', 'party', 'new', 'obama', 'court', 'republica
n', 'campaign', 'people', 'state', 'president', 'clinton', 'said', 'trump']



THE TOP 15 WORDS FOR TOPIC #5
['years', 'going', 've', 'life', 'don', 'new', 'way', 'music', 'really', 'tim
e', 'know', 'think', 'people', 'just', 'like']



THE TOP 15 WORDS FOR TOPIC #6
['student', 'years', 'data', 'science', 'university', 'people', 'time', 'scho
ols', 'just', 'education', 'new', 'like', 'students', 'school', 'says']
```

In [33]:
```
dtm
```

Out[33]:
```
<11992x54777 sparse matrix of type '<class 'numpy.int64'>'
        with 3033388 stored elements in Compressed Sparse Row format>
```

In [34]:
```
topic_results=LDA.transform(dtm)
```

In [36]:
```
topic_results[0].round(2)
```

Out[36]:
```
array([0.02, 0.68, 0.  , 0.  , 0.3 , 0.  , 0.  ])
```

In [37]:
```
topic_results[0].argmax()
```

Out[37]:
```
1
```

In [38]:
```
npr['Topic']=topic_results.argmax(axis=1)
```

In [39]:
```
npr
```

Out[39]:

	Article	Topic
0	In the Washington of 2016, even when the polic...	1
1	Donald Trump has used Twitter — his prefe...	1
2	Donald Trump is unabashedly praising Russian...	1
3	Updated at 2:50 p. m. ET, Russian President Vl...	1
4	From photography, illustration and video, to d...	2
...	...	...
11987	The number of law enforcement officers shot an...	1
11988	Trump is busy these days with victory tours,...	4
11989	It’s always interesting for the Goats and Soda...	3
11990	The election of Donald Trump was a surprise to...	4
11991	Voters in the English city of Sunderland did s...	0

11992 rows × 2 columns

In [ ]:
```

```