In [36]:
```
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.util import ngrams

potterstem=PorterStemmer()

text = '''
<html>
<head>
<title>150-word Meaningful Text with HTML Tags</title>
</head>
<body>
<h1>Why Learning is Important</h1>

<p>Learning is a lifelong process that plays a crucial role in personal growth and dev

<p>When we embrace learning, we open doors to new opportunities. It enables us to adap

<p>Moreover, continuous learning can enhance our professional prospects. By gaining ex

<p>Learning also has numerous benefits for personal well-being. It promotes personal f

<p>In today's digital age, access to learning resources has become more convenient tha

<p>By embracing a lifelong learning mindset, we can embrace personal growth, professic

</body>
</html>
'''



nonalphanumbericpattern = '[^a-zA-Z0-9\s]'
htmlpattern=r'<.*?>'

nonhtmltext=re.sub(htmlpattern,'',text)
alphanumberictext=re.sub(nonalphanumbericpattern,'',nonhtmltext)

texttoken=nltk.wordpunct_tokenize(alphanumberictext)
lowertexttokens=[w.lower() for w in texttoken]
print(lowertexttokens)
```

```
['150word', 'meaningful', 'text', 'with', 'html', 'tags', 'why', 'learning', 'is', 'i
mportant', 'learning', 'is', 'a', 'lifelong', 'process', 'that', 'plays', 'a', 'cruci
al', 'role', 'in', 'personal', 'growth', 'and', 'development', 'it', 'allows', 'indiv
iduals', 'to', 'acquire', 'knowledge', 'skills', 'and', 'understanding', 'that', 'ca
n', 'empower', 'them', 'to', 'navigate', 'through', 'various', 'aspects', 'of', 'lif
e', 'when', 'we', 'embrace', 'learning', 'we', 'open', 'doors', 'to', 'new', 'opportu
nities', 'it', 'enables', 'us', 'to', 'adapt', 'to', 'changing', 'circumstances', 'de
velop', 'critical', 'thinking', 'abilities', 'and', 'expand', 'our', 'perspectives',
'learning', 'encourages', 'curiosity', 'fostering', 'a', 'sense', 'of', 'exploratio
n', 'and', 'discovery', 'moreover', 'continuous', 'learning', 'can', 'enhance', 'ou
r', 'professional', 'prospects', 'by', 'gaining', 'expertise', 'and', 'staying', 'upd
ated', 'in', 'our', 'respective', 'fields', 'we', 'increase', 'our', 'value', 'in',
'the', 'job', 'market', 'and', 'improve', 'our', 'chances', 'for', 'career', 'advance
ment', 'learning', 'also', 'has', 'numerous', 'benefits', 'for', 'personal', 'wellbei
ng', 'it', 'promotes', 'personal', 'fulfillment', 'boosts', 'selfconfidence', 'and',
'cultivates', 'a', 'sense', 'of', 'accomplishment', 'additionally', 'it', 'provides',
'a', 'platform', 'for', 'personal', 'expression', 'allowing', 'us', 'to', 'pursue',
'hobbies', 'interests', 'and', 'creative', 'endeavors', 'in', 'todays', 'digital', 'a
ge', 'access', 'to', 'learning', 'resources', 'has', 'become', 'more', 'convenient',
'than', 'ever', 'online', 'courses', 'educational', 'platforms', 'and', 'communitie
s', 'offer', 'a', 'wealth', 'of', 'knowledge', 'at', 'our', 'fingertips', 'making',
'it', 'easier', 'to', 'engage', 'in', 'continuous', 'learning', 'by', 'embracing',
'a', 'lifelong', 'learning', 'mindset', 'we', 'can', 'embrace', 'personal', 'growth',
'professional', 'development', 'and', 'a', 'more', 'fulfilling', 'life']
```

In [25]:

```
unique_tokens=set(lowertexttokens)
print(unique_tokens)
```

```
{'changing', 'market', 'digital', 'also', 'educational', 'personal', 'role', 'accompl
ishment', 'lifelong', 'to', 'pursue', 'wealth', 'it', 'a', 'various', 'life', 'abilit
ies', 'sense', 'convenient', 'value', 'html', 'discovery', 'doors', 'development', 's
elfconfidence', 'become', 'courses', 'wellbeing', 'circumstances', 'of', 'knowledge',
'fostering', 'fulfillment', 'easier', 'allows', 'embrace', 'individuals', 'empower',
'plays', 'open', 'chances', 'opportunities', 'fingertips', 'has', 'todays', 'we', 'on
line', 'promotes', 'adapt', 'making', 'develop', 'skills', 'prospects', 'critical',
'allowing', 'our', 'career', 'interests', 'numerous', 'resources', 'enhance', 'acces
s', 'them', 'moreover', 'benefits', 'learning', 'growth', 'increase', 'exploration',
'embracing', 'through', 'us', 'ever', 'curiosity', 'perspectives', 'the', 'fields',
'provides', 'meaningful', 'offer', 'updated', 'can', 'thinking', 'expand', 'age', 'pr
ocess', 'respective', 'new', 'by', 'platform', 'continuous', 'in', 'advancement', 'ho
bbies', 'acquire', 'aspects', 'endeavors', 'for', 'important', 'job', 'is', 'understa
nding', 'when', 'creative', 'that', 'enables', 'platforms', 'staying', 'gaining', 'ex
pression', 'at', 'crucial', 'expertise', 'more', 'professional', 'tags', 'with', 'wh
y', 'than', '150word', 'and', 'engage', 'mindset', 'encourages', 'fulfilling', 'tex
t', 'communities', 'improve', 'boosts', 'cultivates', 'navigate', 'additionally'}
```

In [23]:

```
stopwords_list = set(stopwords.words('english'))
print(stopwords_list)
```

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

In [26]:
```
filtered_sentence=[]
for w in unique_tokens:
    if(w not in stopwords_list):
        filtered_sentence.append(w)
print(len(filtered_sentence))
```

```
109
```

In [29]:
```
print(filtered_sentence)
```

```
['changing', 'market', 'digital', 'also', 'educational', 'personal', 'role', 'accompl
ishment', 'lifelong', 'pursue', 'wealth', 'various', 'life', 'abilities', 'sense', 'c
onvenient', 'value', 'html', 'discovery', 'doors', 'development', 'selfconfidence',
'become', 'courses', 'wellbeing', 'circumstances', 'knowledge', 'fostering', 'fulfill
ment', 'easier', 'allows', 'embrace', 'individuals', 'empower', 'plays', 'open', 'cha
nces', 'opportunities', 'fingertips', 'todays', 'online', 'promotes', 'adapt', 'makin
g', 'develop', 'skills', 'prospects', 'critical', 'allowing', 'career', 'interests',
'numerous', 'resources', 'enhance', 'access', 'moreover', 'benefits', 'learning', 'gr
owth', 'increase', 'exploration', 'embracing', 'us', 'ever', 'curiosity', 'perspectiv
es', 'fields', 'provides', 'meaningful', 'offer', 'updated', 'thinking', 'expand', 'a
ge', 'process', 'respective', 'new', 'platform', 'continuous', 'advancement', 'hobbie
s', 'acquire', 'aspects', 'endeavors', 'important', 'job', 'understanding', 'creativ
e', 'enables', 'platforms', 'staying', 'gaining', 'expression', 'crucial', 'expertis
e', 'professional', 'tags', '150word', 'engage', 'mindset', 'encourages', 'fulfillin
g', 'text', 'communities', 'improve', 'boosts', 'cultivates', 'navigate', 'additional
ly']
```

In [28]:
```
stemmedtokens=[]
for w in filtered_sentence:
    stemmedtokens.append(potterstem.stem(w))
stemmedtokens
```

Out[28]:
```
['chang',
 'market',
 'digit',
 'also',
 'educ',
 'person',
 'role',
 'accomplish',
 'lifelong',
 'pursu',
 'wealth',
 'variou',
 'life',
 'abil',
 'sens',
 'conveni',
 'valu',
 'html',
 'discoveri',
 'door',
 'develop',
 'selfconfid',
 'becom',
 'cours',
 'wellb',
 'circumst',
 'knowledg',
 'foster',
 'fulfil',
 'easier',
 'allow',
 'embrac',
 'individu',
 'empow',
 'play',
 'open',
 'chanc',
 'opportun',
 'fingertip',
 'today',
 'onlin',
 'promot',
 'adapt',
 'make',
 'develop',
 'skill',
 'prospect',
 'critic',
 'allow',
 'career',
 'interest',
 'numer',
 'resourc',
 'enhanc',
 'access',
 'moreov',
 'benefit',
 'learn',
 'growth',
 'increas',
```

```
 'explor',
 'embrac',
 'us',
 'ever',
 'curios',
 'perspect',
 'field',
 'provid',
 'meaning',
 'offer',
 'updat',
 'think',
 'expand',
 'age',
 'process',
 'respect',
 'new',
 'platform',
 'continu',
 'advanc',
 'hobbi',
 'acquir',
 'aspect',
 'endeavor',
 'import',
 'job',
 'understand',
 'creativ',
 'enabl',
 'platform',
 'stay',
 'gain',
 'express',
 'crucial',
 'expertis',
 'profession',
 'tag',
 '150word',
 'engag',
 'mindset',
 'encourag',
 'fulfil',
 'text',
 'commun',
 'improv',
 'boost',
 'cultiv',
 'navig',
 'addit']
```

In [30]:

```python
print(len(stemmedtokens))
```

```
109
```

In [34]:

```python
tokenizer = nltk.data.load('tokenizers/punkt/english.pickle')
sentences = tokenizer.tokenize(nonhtmltext)
len(sentences)
```

Out[34]:

```
13
```

In [35]:
```
len(texttoken)/len(sentences)
```

Out[35]:
```
15.76923076923077
```

In [37]:
```
NGRAMS=ngrams(sequence=nltk.word_tokenize(nonhtmltext), n=3)
for grams in NGRAMS:
    print(grams)
```

```
('150-word', 'Meaningful', 'Text')
('Meaningful', 'Text', 'with')
('Text', 'with', 'HTML')
('with', 'HTML', 'Tags')
('HTML', 'Tags', 'Why')
('Tags', 'Why', 'Learning')
('Why', 'Learning', 'is')
('Learning', 'is', 'Important')
('is', 'Important', 'Learning')
('Important', 'Learning', 'is')
('Learning', 'is', 'a')
('is', 'a', 'lifelong')
('a', 'lifelong', 'process')
('lifelong', 'process', 'that')
('process', 'that', 'plays')
('that', 'plays', 'a')
('plays', 'a', 'crucial')
('a', 'crucial', 'role')
('crucial', 'role', 'in')
('role', 'in', 'personal')
('in', 'personal', 'growth')
('personal', 'growth', 'and')
('growth', 'and', 'development')
('and', 'development', '.')
('development', '.', 'It')
('.', 'It', 'allows')
('It', 'allows', 'individuals')
('allows', 'individuals', 'to')
('individuals', 'to', 'acquire')
('to', 'acquire', 'knowledge')
('acquire', 'knowledge', ',')
('knowledge', ',', 'skills')
(',', 'skills', ',')
('skills', ',', 'and')
(',', 'and', 'understanding')
('and', 'understanding', 'that')
('understanding', 'that', 'can')
('that', 'can', 'empower')
('can', 'empower', 'them')
('empower', 'them', 'to')
('them', 'to', 'navigate')
('to', 'navigate', 'through')
('navigate', 'through', 'various')
('through', 'various', 'aspects')
('various', 'aspects', 'of')
('aspects', 'of', 'life')
('of', 'life', '.')
('life', '.', 'When')
('.', 'When', 'we')
('When', 'we', 'embrace')
('we', 'embrace', 'learning')
('embrace', 'learning', ',')
('learning', ',', 'we')
(',', 'we', 'open')
('we', 'open', 'doors')
('open', 'doors', 'to')
('doors', 'to', 'new')
('to', 'new', 'opportunities')
('new', 'opportunities', '.')
('opportunities', '.', 'It')
```

```
('.', 'It', 'enables')
('It', 'enables', 'us')
('enables', 'us', 'to')
('us', 'to', 'adapt')
('to', 'adapt', 'to')
('adapt', 'to', 'changing')
('to', 'changing', 'circumstances')
('changing', 'circumstances', ',')
('circumstances', ',', 'develop')
(',', 'develop', 'critical')
('develop', 'critical', 'thinking')
('critical', 'thinking', 'abilities')
('thinking', 'abilities', ',')
('abilities', ',', 'and')
(',', 'and', 'expand')
('and', 'expand', 'our')
('expand', 'our', 'perspectives')
('our', 'perspectives', '.')
('perspectives', '.', 'Learning')
('.', 'Learning', 'encourages')
('Learning', 'encourages', 'curiosity')
('encourages', 'curiosity', ',')
('curiosity', ',', 'fostering')
(',', 'fostering', 'a')
('fostering', 'a', 'sense')
('a', 'sense', 'of')
('sense', 'of', 'exploration')
('of', 'exploration', 'and')
('exploration', 'and', 'discovery')
('and', 'discovery', '.')
('discovery', '.', 'Moreover')
('.', 'Moreover', ',')
('Moreover', ',', 'continuous')
(',', 'continuous', 'learning')
('continuous', 'learning', 'can')
('learning', 'can', 'enhance')
('can', 'enhance', 'our')
('enhance', 'our', 'professional')
('our', 'professional', 'prospects')
('professional', 'prospects', '.')
('prospects', '.', 'By')
('.', 'By', 'gaining')
('By', 'gaining', 'expertise')
('gaining', 'expertise', 'and')
('expertise', 'and', 'staying')
('and', 'staying', 'updated')
('staying', 'updated', 'in')
('updated', 'in', 'our')
('in', 'our', 'respective')
('our', 'respective', 'fields')
('respective', 'fields', ',')
('fields', ',', 'we')
(',', 'we', 'increase')
('we', 'increase', 'our')
('increase', 'our', 'value')
('our', 'value', 'in')
('value', 'in', 'the')
('in', 'the', 'job')
('the', 'job', 'market')
('job', 'market', 'and')
```

```
('market', 'and', 'improve')
('and', 'improve', 'our')
('improve', 'our', 'chances')
('our', 'chances', 'for')
('chances', 'for', 'career')
('for', 'career', 'advancement')
('career', 'advancement', '.')
('advancement', '.', 'Learning')
('.', 'Learning', 'also')
('Learning', 'also', 'has')
('also', 'has', 'numerous')
('has', 'numerous', 'benefits')
('numerous', 'benefits', 'for')
('benefits', 'for', 'personal')
('for', 'personal', 'well-being')
('personal', 'well-being', '.')
('well-being', '.', 'It')
('.', 'It', 'promotes')
('It', 'promotes', 'personal')
('promotes', 'personal', 'fulfillment')
('personal', 'fulfillment', ',')
('fulfillment', ',', 'boosts')
(',', 'boosts', 'self-confidence')
('boosts', 'self-confidence', ',')
('self-confidence', ',', 'and')
(',', 'and', 'cultivates')
('and', 'cultivates', 'a')
('cultivates', 'a', 'sense')
('a', 'sense', 'of')
('sense', 'of', 'accomplishment')
('of', 'accomplishment', '.')
('accomplishment', '.', 'Additionally')
('.', 'Additionally', ',')
('Additionally', ',', 'it')
(',', 'it', 'provides')
('it', 'provides', 'a')
('provides', 'a', 'platform')
('a', 'platform', 'for')
('platform', 'for', 'personal')
('for', 'personal', 'expression')
('personal', 'expression', ',')
('expression', ',', 'allowing')
(',', 'allowing', 'us')
('allowing', 'us', 'to')
('us', 'to', 'pursue')
('to', 'pursue', 'hobbies')
('pursue', 'hobbies', ',')
('hobbies', ',', 'interests')
(',', 'interests', ',')
('interests', ',', 'and')
(',', 'and', 'creative')
('and', 'creative', 'endeavors')
('creative', 'endeavors', '.')
('endeavors', '.', 'In')
('.', 'In', 'today')
('In', 'today', "'s")
('today', "'s", 'digital')
("'s", 'digital', 'age')
('digital', 'age', ',')
('age', ',', 'access')
```

```
(',', 'access', 'to')
('access', 'to', 'learning')
('to', 'learning', 'resources')
('learning', 'resources', 'has')
('resources', 'has', 'become')
('has', 'become', 'more')
('become', 'more', 'convenient')
('more', 'convenient', 'than')
('convenient', 'than', 'ever')
('than', 'ever', '.')
('ever', '.', 'Online')
('.', 'Online', 'courses')
('Online', 'courses', ',')
('courses', ',', 'educational')
(',', 'educational', 'platforms')
('educational', 'platforms', ',')
('platforms', ',', 'and')
(',', 'and', 'communities')
('and', 'communities', 'offer')
('communities', 'offer', 'a')
('offer', 'a', 'wealth')
('a', 'wealth', 'of')
('wealth', 'of', 'knowledge')
('of', 'knowledge', 'at')
('knowledge', 'at', 'our')
('at', 'our', 'fingertips')
('our', 'fingertips', ',')
('fingertips', ',', 'making')
(',', 'making', 'it')
('making', 'it', 'easier')
('it', 'easier', 'to')
('easier', 'to', 'engage')
('to', 'engage', 'in')
('engage', 'in', 'continuous')
('in', 'continuous', 'learning')
('continuous', 'learning', '.')
('learning', '.', 'By')
('.', 'By', 'embracing')
('By', 'embracing', 'a')
('embracing', 'a', 'lifelong')
('a', 'lifelong', 'learning')
('lifelong', 'learning', 'mindset')
('learning', 'mindset', ',')
('mindset', ',', 'we')
(',', 'we', 'can')
('we', 'can', 'embrace')
('can', 'embrace', 'personal')
('embrace', 'personal', 'growth')
('personal', 'growth', ',')
('growth', ',', 'professional')
(',', 'professional', 'development')
('professional', 'development', ',')
('development', ',', 'and')
(',', 'and', 'a')
('and', 'a', 'more')
('a', 'more', 'fulfilling')
('more', 'fulfilling', 'life')
('fulfilling', 'life', '.')
```

In [42]:
```
def extract_ngrams(data,num):
    n_grams=ngrams(nltk.word_tokenize(data), num)
    return [' '.join(grams) for grams in n_grams]
```

In [45]:
```
my_text='For most of the cases, it is totally fine to use the pre-trained version. So
extract_ngrams(my_text,3)
```

Out[45]:
```
['For most of',
 'most of the',
 'of the cases',
 'the cases ,',
 'cases , it',
 ', it is',
 'it is totally',
 'is totally fine',
 'totally fine to',
 'fine to use',
 'to use the',
 'use the pre-trained',
 'the pre-trained version',
 'pre-trained version .',
 'version . So',
 '. So you',
 'So you can',
 'you can simply',
 'can simply initialize',
 'simply initialize the',
 'initialize the tokenizer',
 'the tokenizer without',
 'tokenizer without providing',
 'without providing any',
 'providing any arguments',
 'any arguments .']
```

In [ ]: