In [1]:
```
pip list | grep nl
```

```
greenlet                    3.0.3
matplotlib-inline           0.1.6
nltk                        3.8.1
types-greenlet              3.0
Note: you may need to restart the kernel to use updated packages.
```

In [2]:
```
import nltk
```

In [3]:
```
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package stopwords to /home/omkar/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package punkt to /home/omkar/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package wordnet to /home/omkar/nltk_data...
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /home/omkar/nltk_data...
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger.zip.
```

Out[3]:
```
True
```

In [4]:
```
import nltk
```

In [6]:
```
para ="He picked up the burnt end of the branch and made a mark on the stone
```

In [7]:
```
print(para)
```

```
He picked up the burnt end of the branch and made a mark on the stone. Day
52 if the marks on the stone were accurate. He couldn't be sure. Day and ni
ghts had begun to blend together creating confusion, but he knew it was a l
ong time. Much too long.
```

In [8]:
```
para.split()
```

Out[8]:
```
['He',
 'picked',
 'up',
 'the',
 'burnt',
 'end',
 'of',
 'the',
 'branch',
 'and',
 'made',
 'a',
 'mark',
 'on',
 'the',
 'stone.',
 'Day',
 '52',
 'if',
 'the',
 'marks',
 'on',
 'the',
 'stone',
 'were',
 'accurate.',
 'He',
 "couldn't",
 'be',
 'sure.',
 'Day',
 'and',
 'nights',
 'had',
 'begun',
 'to',
 'blend',
 'together',
 'creating',
 'confusion,',
 'but',
 'he',
 'knew',
 'it',
 'was',
 'a',
 'long',
 'time.',
 'Much',
 'too',
 'long.']
```

In [9]:
```
from nltk.tokenize import sent_tokenize
from nltk.tokenize import word_tokenize
```

In [10]:
```
sent = sent_tokenize(para)
```

In [11]:
```
sent[2]
```

Out[11]:
```
"He couldn't be sure."
```

In [12]:
```
words = word_tokenize(para)
```

In [13]:
```
words
```

Out[13]:
```
['He',
 'picked',
 'up',
 'the',
 'burnt',
 'end',
 'of',
 'the',
 'branch',
 'and',
 'made',
 'a',
 'mark',
 'on',
 'the',
 'stone',
 '.',
 'Day',
 '52',
 'if',
 'the',
 'marks',
 'on',
 'the',
 'stone',
 'were',
 'accurate',
 '.',
 'He',
 'could',
 "n't",
 'be',
 'sure',
 '.',
 'Day',
 'and',
 'nights',
 'had',
 'begun',
 'to',
 'blend',
 'together',
 'creating',
 'confusion',
 ',',
 'but',
 'he',
 'knew',
 'it',
 'was',
 'a',
 'long',
 'time',
 '.',
 'Much',
 'too',
 'long',
 '.']
```

In [14]:
```
from nltk.corpus import stopwords
```

In [15]:
```
swords=stopwords.words('english')
```

In [16]:
```
swords
```

Out[16]:
```
['a',
 'about',
 'above',
 'after',
 'again',
 'against',
 'ain',
 'all',
 'am',
 'an',
 'and',
 'any',
 'are',
 'aren',
 "aren't",
 'as',
 'at',
 'be',
 'because',
 'been',
 'before',
 'being',
 'below',
 'between',
 'both',
 'but',
 'by',
 'can',
 'couldn',
 "couldn't",
 'd',
 'did',
 'didn',
 "didn't",
 'do',
 'does',
 'doesn',
 "doesn't",
 'doing',
 'don',
 "don't",
 'down',
 'during',
 'each',
 'few',
 'for',
 'from',
 'further',
 'had',
 'hadn',
 "hadn't",
 'has',
 'hasn',
 "hasn't",
 'have',
 'haven',
 "haven't",
 'having',
 'he',
 "he'd",
 "he'll",
 'her',
 'here',
 'hers',
```

```
'herself',
"he's",
'him',
'himself',
'his',
'how',
'i',
"i'd",
'if',
"i'll",
"i'm",
'in',
'into',
'is',
'isn',
"isn't",
'it',
"it'd",
"it'll",
"it's",
'its',
'itself',
"i've",
'just',
'll',
'm',
'ma',
'me',
'mightn',
"mightn't",
'more',
'most',
'mustn',
"mustn't",
'my',
'myself',
'needn',
"needn't",
'no',
'nor',
'not',
'now',
'o',
'of',
'off',
'on',
'once',
'only',
'or',
'other',
'our',
'ours',
'ourselves',
'out',
'over',
'own',
're',
's',
'same',
'shan',
"shan't",
'she',
"she'd",
"she'll",
```

```
"she's",
'should',
'shouldn',
"shouldn't",
"should've",
'so',
'some',
'such',
't',
'than',
'that',
"that'll",
'the',
'their',
'theirs',
'them',
'themselves',
'then',
'there',
'these',
'they',
"they'd",
"they'll",
"they're",
"they've",
'this',
'those',
'through',
'to',
'too',
'under',
'until',
'up',
've',
'very',
'was',
'wasn',
"wasn't",
'we',
"we'd",
"we'll",
"we're",
'were',
'weren',
"weren't",
"we've",
'what',
'when',
'where',
'which',
'while',
'who',
'whom',
'why',
'will',
'with',
'won',
"won't",
'wouldn',
"wouldn't",
'y',
'you',
"you'd",
"you'll",
```

```
 'your',
 "you're",
 'yours',
 'yourself',
 'yourselves',
 "you've"]
```

In [17]:
```
x=[word for word in words if word not in swords]
x
```

Out[17]:
```
['He',
 'picked',
 'burnt',
 'end',
 'branch',
 'made',
 'mark',
 'stone',
 '.',
 'Day',
 '52',
 'marks',
 'stone',
 'accurate',
 '.',
 'He',
 'could',
 "n't",
 'sure',
 '.',
 'Day',
 'nights',
 'begun',
 'blend',
 'together',
 'creating',
 'confusion',
 ',',
 'knew',
 'long',
 'time',
 '.',
 'Much',
 'long',
 '.']
```

In [18]:
```
x = [word for word in words if word.lower() not in swords]
x
```

Out[18]:
```
['picked',
 'burnt',
 'end',
 'branch',
 'made',
 'mark',
 'stone',
 '.',
 'Day',
 '52',
 'marks',
 'stone',
 'accurate',
 '.',
 'could',
 "n't",
 'sure',
 '.',
 'Day',
 'nights',
 'begun',
 'blend',
 'together',
 'creating',
 'confusion',
 ',',
 'knew',
 'long',
 'time',
 '.',
 'Much',
 'long',
 '.']
```

In [19]:
```
from nltk.stem import PorterStemmer
```

In [21]:
```
ps = PorterStemmer()
```

In [22]:
```
ps.stem('working')
```

Out[22]:
```
'work'
```

In [23]:
```
y=[ps.stem(word) for word in x]
y
```

Out[23]:
```
['pick',
 'burnt',
 'end',
 'branch',
 'made',
 'mark',
 'stone',
 '.',
 'day',
 '52',
 'mark',
 'stone',
 'accur',
 '.',
 'could',
 "n't",
 'sure',
 '.',
 'day',
 'night',
 'begun',
 'blend',
 'togeth',
 'creat',
 'confus',
 ',',
 'knew',
 'long',
 'time',
 '.',
 'much',
 'long',
 '.']
```

In [24]:
```
from nltk.stem import WordNetLemmatizer
```

In [25]:
```
wnl=WordNetLemmatizer
```

In [30]:
```
nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package omw-1.4 to /home/omkar/nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
```

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

In [32]:
```
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to /home/omkar/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[32]:
```
True
```

In [33]:
```
wnl=WordNetLemmatizer()
```

In [34]:
```
wnl.lemmatize('working',pos='v')
```

Out[34]:
```
'work'
```

In [35]:
```
print(ps.stem('went'))
print(wnl.lemmatize('went',pos='v'))
```

```
went
go
```

In [37]:
```
z=[wnl.lemmatize(word,pos='v') for word in x]
z
```

Out[37]:
```
['pick',
 'burn',
 'end',
 'branch',
 'make',
 'mark',
 'stone',
 '.',
 'Day',
 '52',
 'mark',
 'stone',
 'accurate',
 '.',
 'could',
 "n't",
 'sure',
 '.',
 'Day',
 'nights',
 'begin',
 'blend',
 'together',
 'create',
 'confusion',
 ',',
 'know',
 'long',
 'time',
 '.',
 'Much',
 'long',
 '.']
```

In [38]:
```
import string
```

In [39]:
```
string.punctuation
```

Out[39]:
```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

In [40]:
```
t=[word for word in words if word not in string.punctuation]
t
```

Out[40]:
```
['He',
 'picked',
 'up',
 'the',
 'burnt',
 'end',
 'of',
 'the',
 'branch',
 'and',
 'made',
 'a',
 'mark',
 'on',
 'the',
 'stone',
 'Day',
 '52',
 'if',
 'the',
 'marks',
 'on',
 'the',
 'stone',
 'were',
 'accurate',
 'He',
 'could',
 "n't",
 'be',
 'sure',
 'Day',
 'and',
 'nights',
 'had',
 'begun',
 'to',
 'blend',
 'together',
 'creating',
 'confusion',
 'but',
 'he',
 'knew',
 'it',
 'was',
 'a',
 'long',
 'time',
 'Much',
 'too',
 'long']
```

In [41]:
```
from nltk import pos_tag
```

In [42]:
```
pos_tag(t)
```

Out[42]:
```
[('He', 'PRP'),
 ('picked', 'VBD'),
 ('up', 'RP'),
 ('the', 'DT'),
 ('burnt', 'JJ'),
 ('end', 'NN'),
 ('of', 'IN'),
 ('the', 'DT'),
 ('branch', 'NN'),
 ('and', 'CC'),
 ('made', 'VBD'),
 ('a', 'DT'),
 ('mark', 'NN'),
 ('on', 'IN'),
 ('the', 'DT'),
 ('stone', 'NN'),
 ('Day', 'NNP'),
 ('52', 'CD'),
 ('if', 'IN'),
 ('the', 'DT'),
 ('marks', 'NNS'),
 ('on', 'IN'),
 ('the', 'DT'),
 ('stone', 'NN'),
 ('were', 'VBD'),
 ('accurate', 'JJ'),
 ('He', 'PRP'),
 ('could', 'MD'),
 ("n't", 'RB'),
 ('be', 'VB'),
 ('sure', 'JJ'),
 ('Day', 'NNP'),
 ('and', 'CC'),
 ('nights', 'NNS'),
 ('had', 'VBD'),
 ('begun', 'VBN'),
 ('to', 'TO'),
 ('blend', 'VB'),
 ('together', 'RB'),
 ('creating', 'VBG'),
 ('confusion', 'NN'),
 ('but', 'CC'),
 ('he', 'PRP'),
 ('knew', 'VBD'),
 ('it', 'PRP'),
 ('was', 'VBD'),
 ('a', 'DT'),
 ('long', 'JJ'),
 ('time', 'NN'),
 ('Much', 'NNP'),
 ('too', 'RB'),
 ('long', 'RB')]
```

In [44]:

```
error: externally-managed-environment

× This environment is externally managed
╰─> To install Python packages system-wide, try apt install
    python3-xyz, where xyz is the package you are trying to
    install.

    If you wish to install a non-Debian-packaged Python package,
    create a virtual environment using python3 -m venv path/to/venv.
    Then use path/to/venv/bin/python and path/to/venv/bin/pip. Make
    sure you have python3-full installed.

    If you wish to install a non-Debian packaged Python application,
    it may be easiest to use pipx install xyz, which will manage a
    virtual environment for you. Make sure you have pipx installed.

    See /usr/share/doc/python3.12/README.venv for more information.

note: If you believe this is a mistake, please contact your Python installa
tion or OS distribution provider. You can override this, at the risk of bre
aking your Python installation or OS, by passing --break-system-packages.
hint: See PEP 668 for the detailed specification.
```

In [45]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [46]:
```
tfidf = TfidfVectorizer()
```

In [47]:
```
v=tfidf.fit_transform(t)
```

In [48]:
```
v.shape
```

Out[48]:
```
(52, 38)
```

In [49]:
```
import pandas as pd
```

In [50]:
```
pd.DataFrame(v)
```

Out[50]:

	0
0	(0, 15)\t1.0
1	(0, 27)\t1.0
2	(0, 35)\t1.0
3	(0, 30)\t1.0
4	(0, 7)\t1.0
5	(0, 13)\t1.0
6	(0, 25)\t1.0
7	(0, 30)\t1.0
8	(0, 6)\t1.0
9	(0, 2)\t1.0
10	(0, 20)\t1.0
11	
12	(0, 21)\t1.0
13	(0, 26)\t1.0
14	(0, 30)\t1.0
15	(0, 28)\t1.0
16	(0, 12)\t1.0
17	(0, 0)\t1.0
18	(0, 16)\t1.0
19	(0, 30)\t1.0
20	(0, 22)\t1.0
21	(0, 26)\t1.0
22	(0, 30)\t1.0
23	(0, 28)\t1.0
24	(0, 37)\t1.0
25	(0, 1)\t1.0
26	(0, 15)\t1.0
27	(0, 10)\t1.0
28	
29	(0, 3)\t1.0
30	(0, 29)\t1.0
31	(0, 12)\t1.0
32	(0, 2)\t1.0
33	(0, 24)\t1.0
34	(0, 14)\t1.0
35	(0, 4)\t1.0
36	(0, 32)\t1.0
37	(0, 5)\t1.0
38	(0, 33)\t1.0

	0
39	(0, 11)\t1.0
40	(0, 9)\t1.0
41	(0, 8)\t1.0
42	(0, 15)\t1.0
43	(0, 18)\t1.0
44	(0, 17)\t1.0
45	(0, 36)\t1.0
46	
47	(0, 19)\t1.0
48	(0, 31)\t1.0
49	(0, 23)\t1.0
50	(0, 34)\t1.0
51	(0, 19)\t1.0

In [ ]: