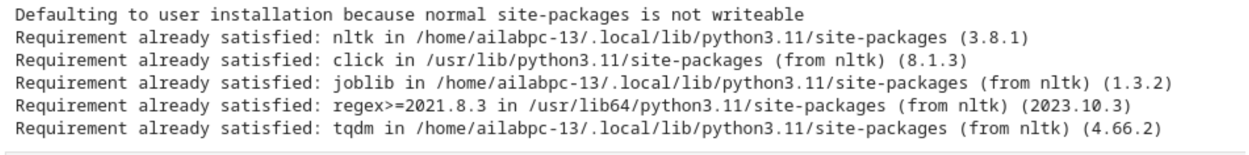
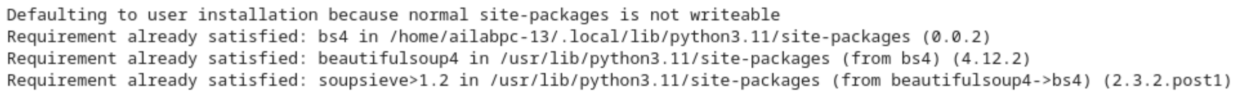
!pip install nltk -U



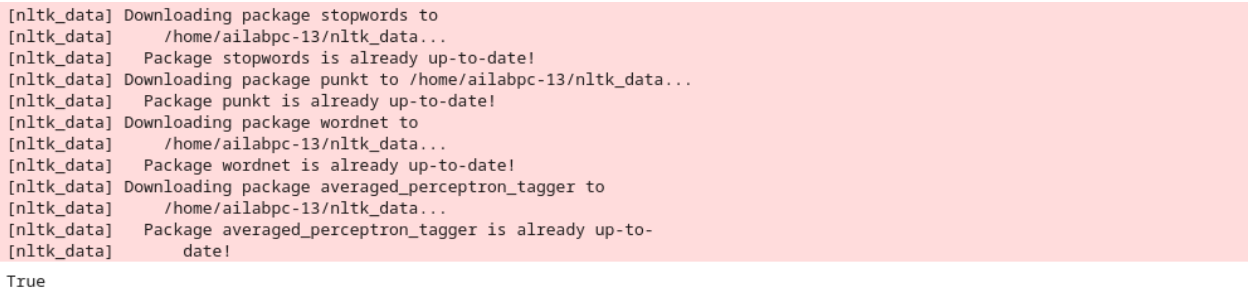
```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: nltk in /home/ailabpc-13/.local/lib/python3.11/site-packages (3.8.1)
Requirement already satisfied: click in /usr/lib/python3.11/site-packages (from nltk) (8.1.3)
Requirement already satisfied: joblib in /home/ailabpc-13/.local/lib/python3.11/site-packages (from nltk) (1.3.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/lib64/python3.11/site-packages (from nltk) (2023.10.3)
Requirement already satisfied: tqdm in /home/ailabpc-13/.local/lib/python3.11/site-packages (from nltk) (4.66.2)
```

!pip install bs4 -U



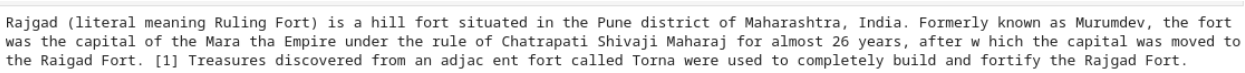
```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: bs4 in /home/ailabpc-13/.local/lib/python3.11/site-packages (0.0.2)
Requirement already satisfied: beautifulsoup4 in /usr/lib/python3.11/site-packages (from bs4) (4.12.2)
Requirement already satisfied: soupsieve>1.2 in /usr/lib/python3.11/site-packages (from beautifulsoup4->bs4) (2.3.2.post1)
```

```
import nltk
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')
```



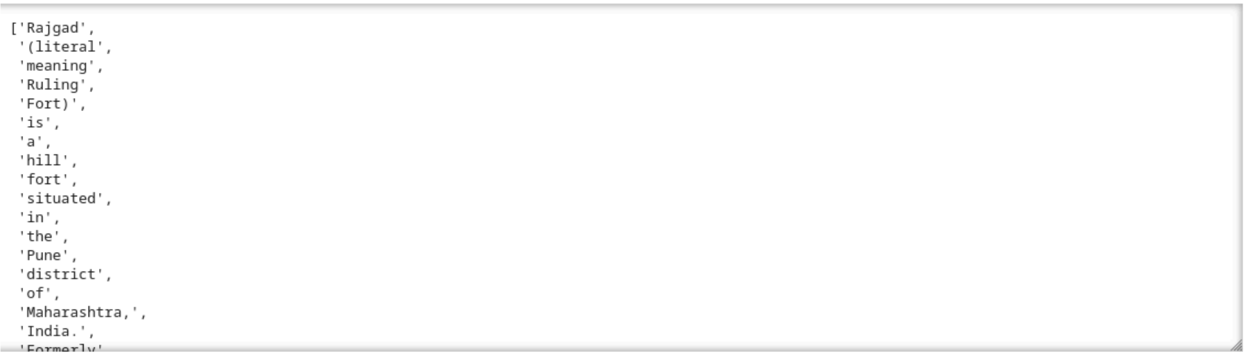
```
[nltk_data] Downloading package stopwords to
[nltk_data]     /home/ailabpc-13/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /home/ailabpc-13/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     /home/ailabpc-13/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /home/ailabpc-13/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
True
```

```
import nltk
para="Rajgad (literal meaning Ruling Fort) is a hill fort situated in the Pune district of Maharashtra, India. Formerly known as Murumdev, the fort was the capital of the Mara tha Empire under the rule of Chatrapati Shivaji Maharaj for almost 26 years, after w hich the capital was moved to the Raigad Fort. [1] Treasures discovered from an adjac ent fort called Torna were used to completely build and fortify the Rajgad Fort."
print(para)
```



```
Rajgad (literal meaning Ruling Fort) is a hill fort situated in the Pune district of Maharashtra, India. Formerly known as Murumdev, the fort
was the capital of the Mara tha Empire under the rule of Chatrapati Shivaji Maharaj for almost 26 years, after w hich the capital was moved to
the Raigad Fort. [1] Treasures discovered from an adjac ent fort called Torna were used to completely build and fortify the Rajgad Fort.
```

```
para.split()
```

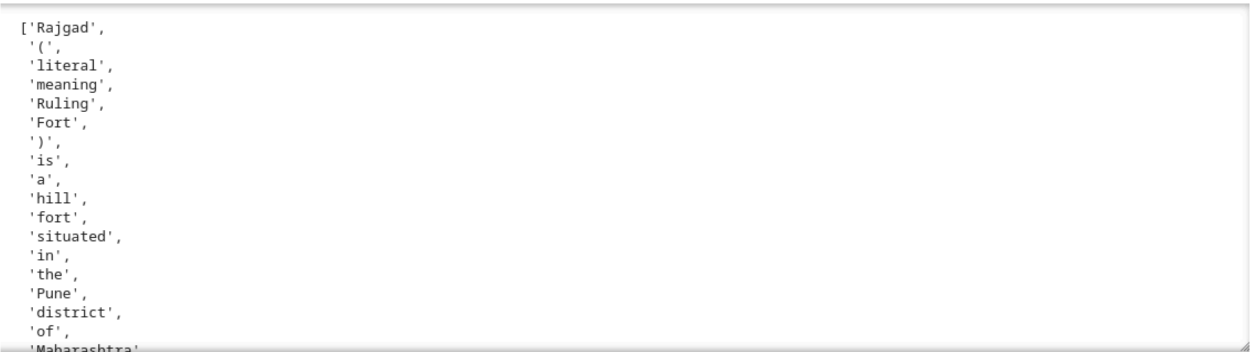


```
['Rajgad',
 '(literal',
 'meaning',
 'Ruling',
 'Fort)',
 'is',
 'a',
 'hill',
 'fort',
 'situated',
 'in',
 'the',
 'Pune',
 'district',
 'of',
 'Maharashtra,',
 'India.',
 'Formerly',
```

```
from nltk.tokenize import sent_tokenize
from nltk.tokenize import word_tokenize
```

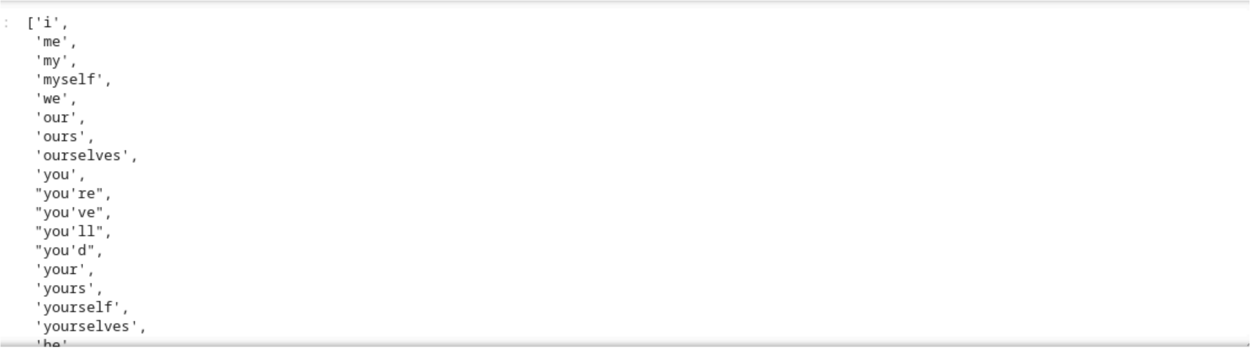
sent=sent_tokenize (para)
sent[2]
'[1] Treasures discovered from an adjac ent fort called Torna were used to completely build and fortify the Rajgad Fort.'

words=word_tokenize(para)
words



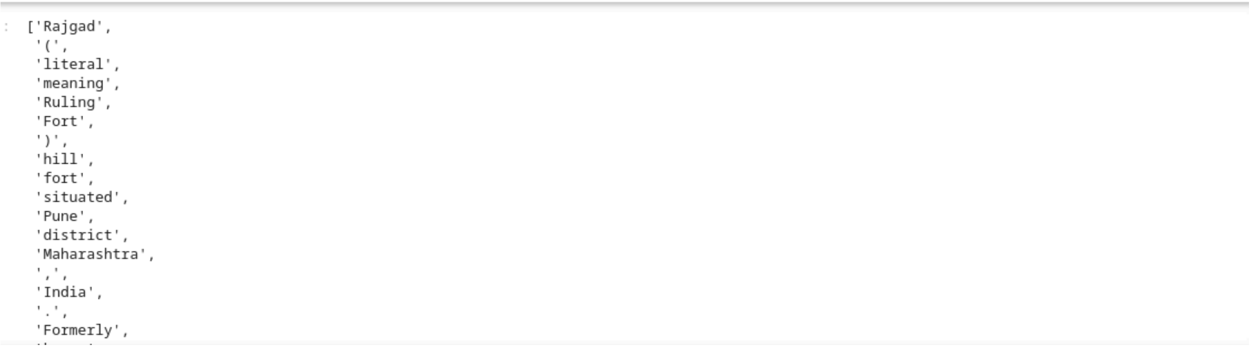
```
['Rajgad',
 '(',
 'literal',
 'meaning',
 'Ruling',
 'Fort',
 ')',
 'is',
 'a',
 'hill',
 'fort',
 'situated',
 'in',
 'the',
 'Pune',
 'district',
 'of',
 'Maharashtra',
```

from nltk.corpus import stopwords
swords=stopwords.words('english')
swords



```
: ['i',
   'me',
   'my',
   'myself',
   'we',
   'our',
   'ours',
   'ourselves',
   'you',
   "you're",
   "you've",
   "you'll",
   "you'd",
   'your',
   'yours',
   'yourself',
   'yourselves',
   'he',
```

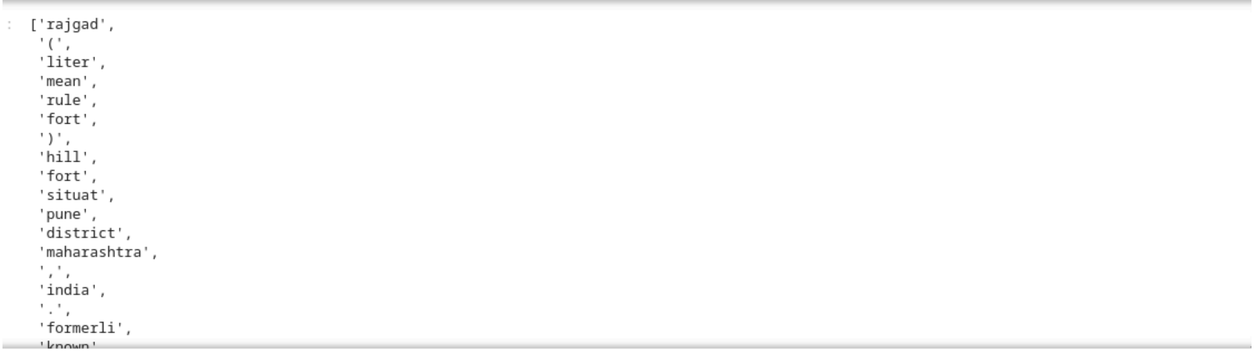
x=[word for word in words if word not in swords]
x



```
: ['Rajgad',
   '(',
   'literal',
   'meaning',
   'Ruling',
   'Fort',
   ')',
   'hill',
   'fort',
   'situated',
   'Pune',
   'district',
   'Maharashtra',
   ',',
   'India',
   '.',
   'Formerly',
```

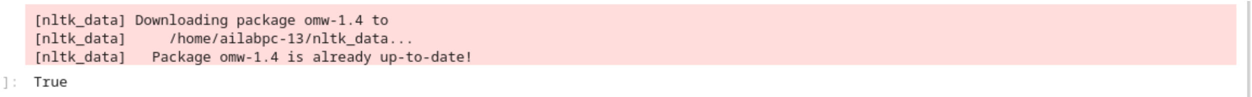
from nltk.stem import PorterStemmer
ps=PorterStemmer()
ps.stem('working')
'work'

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



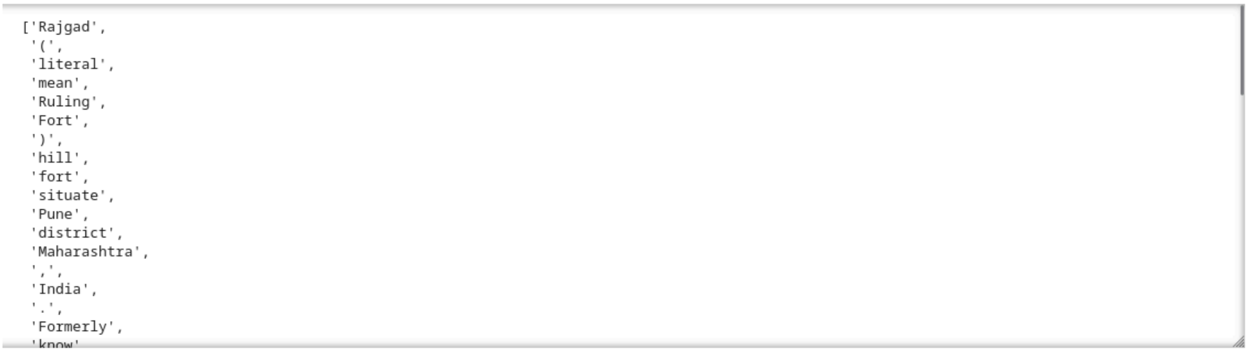
```
:  ['rajgad',
    '(',
    'liter',
    'mean',
    'rule',
    'fort',
    ')',
    'hill',
    'fort',
    'situat',
    'pune',
    'district',
    'maharashtra',
    ',',
    'india',
    '.',
    'formerli',
    'known',
```

```
from nltk.stem import WordNetLemmatizer
wnl=WordNetLemmatizer()
wnl.lemmatize('working', pos='v')
‘work’
nltk.download('omw-1.4')
```



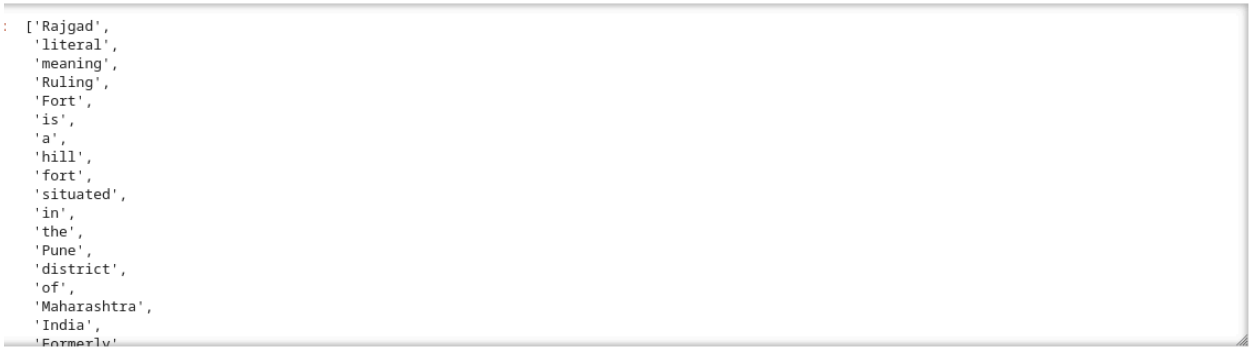
```
   [nltk_data] Downloading package omw-1.4 to
   [nltk_data]     /home/ailabpc-13/nltk_data...
   [nltk_data]   Package omw-1.4 is already up-to-date!
]: True
```

```
print(ps.stem('went'))
print(wnl.lemmatize('went', pos='v'))
went
go
z=[wnl.lemmatize(word, pos='v') for word in x]
z
```



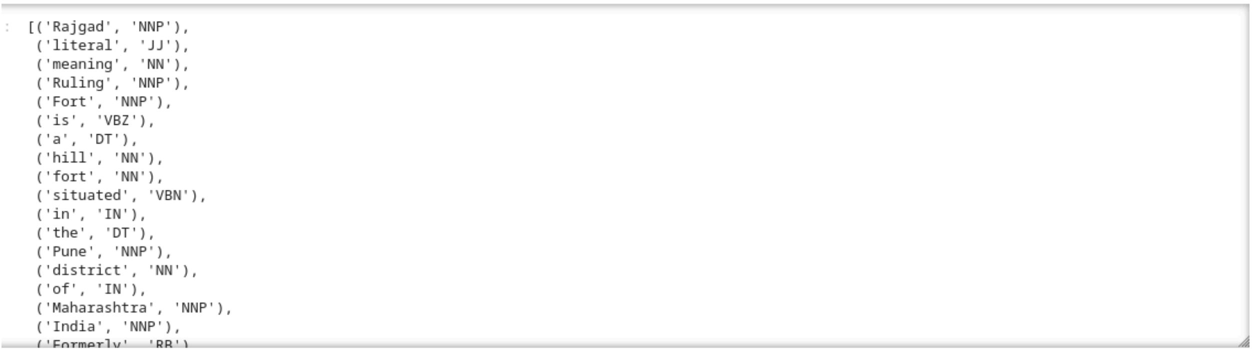
```
['Rajgad',
 '(',
 'literal',
 'mean',
 'Ruling',
 'Fort',
 ')',
 'hill',
 'fort',
 'situate',
 'Pune',
 'district',
 'Maharashtra',
 ',',
 'India',
 '.',
 'Formerly',
 'know',
```

```
import string
string.punctuation
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
t=[word for word in words if word not in string.punctuation]
t
```



```
: ['Rajgad',
   'literal',
   'meaning',
   'Ruling',
   'Fort',
   'is',
   'a',
   'hill',
   'fort',
   'situated',
   'in',
   'the',
   'Pune',
   'district',
   'of',
   'Maharashtra',
   'India',
   'Formerly'
```

from nltk import pos_tag
pos_tag(t)



```
: [('Rajgad', 'NNP'),
   ('literal', 'JJ'),
   ('meaning', 'NN'),
   ('Ruling', 'NNP'),
   ('Fort', 'NNP'),
   ('is', 'VBZ'),
   ('a', 'DT'),
   ('hill', 'NN'),
   ('fort', 'NN'),
   ('situated', 'VBN'),
   ('in', 'IN'),
   ('the', 'DT'),
   ('Pune', 'NNP'),
   ('district', 'NN'),
   ('of', 'IN'),
   ('Maharashtra', 'NNP'),
   ('India', 'NNP'),
   ('Formerly', 'RB')
```

from sklearn.feature_extraction.text import TfidfVectorizer
tfidf = TfidfVectorizer()
v=tfidf.fit_transform(t)
v.shape
(73, 52)
import pandas as pd
pd.DataFrame (v)

	0
0	(0, 37)\t1.0
1	(0, 27)\t1.0
2	(0, 31)\t1.0
3	(0, 39)\t1.0
4	(0, 18)\t1.0
...	...
68	(0, 5)\t1.0
69	(0, 19)\t1.0
70	(0, 43)\t1.0
71	(0, 37)\t1.0
72	(0, 18)\t1.0

73 rows × 1 columns