Practical 34

Write a python program for stemming and remove stop words from a sentence.

```
[2]: import nltk
     nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /home/jayampatel/nltk_data…
[nltk_data]    Unzipping tokenizers/punkt.zip.
```

```
[2]: True
```

```
[4]: from nltk.tokenize import word_tokenize
     from nltk.stem.porter import PorterStemmer
     from sklearn.feature_extraction.text import *
     sentence='sam loves swimming so he swims whole time'
     def tokenize(text):
         tokens=word_tokenize(text)
         l=[]
         ob=PorterStemmer()
         for w in tokens:
             l.append(ob.stem(w))
         return l
     vec=CountVectorizer(tokenizer=tokenize,stop_words='english')
     vec.fit([sentence])
     print(vec.get_feature_names_out())
```

```
['love' 'sam' 'swim' 'time']
```

```
/home/jayampatel/.local/lib/python3.10/site-
packages/sklearn/feature_extraction/text.py:525: UserWarning: The parameter
'token_pattern' will not be used since 'tokenizer' is not None'
  warnings.warn(
/home/jayampatel/.local/lib/python3.10/site-
packages/sklearn/feature_extraction/text.py:408: UserWarning: Your stop_words
may be inconsistent with your preprocessing. Tokenizing the stop words generated
tokens ['abov', 'afterward', 'alon', 'alreadi', 'alway', 'ani', 'anoth',
'anyon', 'anyth', 'anywher', 'becam', 'becaus', 'becom', 'befor', 'besid',
'cri', 'describ', 'dure', 'els', 'elsewher', 'empti', 'everi', 'everyon',
'everyth', 'everywher', 'fifti', 'formerli', 'forti', 'ha', 'henc', 'hereaft',
'herebi', 'hi', 'howev', 'hundr', 'inde', 'latterli', 'mani', 'meanwhil',
'moreov', 'mostli', 'nobodi', 'noon', 'noth', 'nowher', 'onc', 'onli',
'otherwis', 'ourselv', 'perhap', 'pleas', 'seriou', 'sever', 'sinc', 'sincer',
'sixti', 'someon', 'someth', 'sometim', 'somewher', 'themselv', 'thenc',
'thereaft', 'therebi', 'therefor', 'thi', 'thu', 'togeth', 'twelv', 'twenti',
'veri', 'wa', 'whatev', 'whenc', 'whenev', 'wherea', 'whereaft', 'wherebi',
'wherev', 'whi', 'yourselv'] not in stop_words.
  warnings.warn(
```

```
[ ]:
```