PRACTICAL NO 9

Aim: Execute a wordcount problem using Spark and NLTK.

Code:-

```
import nltk
nltk.download('punkt')
from nltk.tokenize import word_tokenize
statment=str(input("enter a sentence:"))
tokens=word_tokenize(statment)
print("the no of words in the given sentence is:",len(tokens))
```

Output:-

```
enter a sentence:this is amruta
the no of words in the given sentence is: 3
PS D:\TYCS 2024\dwdm>
```

Code for the collocations :-

```
from nltk.util import ngrams #ngrams is the pair of  words (collocations)
from nltk.tokenize import word_tokenize,sent_tokenize
#from nltk.collocations import*

statment=['sun','rises','in','the','east','it','sets','in','the','west']
bigrams=ngrams(statment,2)
bigrams_count={}
for b in bigrams:
    if b not in bigrams_count:
        bigrams_count[b]=1
    else:
        bigrams_count[b]+=1

print(statment)
print("Biggrams:",bigrams_count)
```

Output:-

```
['sun', 'rises', 'in', 'the', 'east', 'it', 'sets', 'in', 'the', 'west']
Biggrams: {('sun', 'rises'): 1, ('rises', 'in'): 1, ('in', 'the'): 2, ('the', 'east'): 1, ('east',
'it'): 1, ('it', 'sets'): 1, ('sets', 'in'): 1, ('the', 'west'): 1}
```