Assessment 2

Name: Loukik Bhangale

Reg No: 17BCE0961

Solt: L9+L10

Code:

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize



def cleaner(filename):
    filevariablename=open(filename).read()
    stop_words = set(stopwords.words('english'))
    word_tokens = word_tokenize(filevariablename)
    filtered_sent = [w for w in word_tokens if not w in stop_words]
    filtered_sentence = []
    for w in word_tokens:
        if w not in stop_words:
            filtered_sentence.append(w)
    #print(word_tokens)
    #
    #print("stop words removed!")
    punctuations = list('''!()-[]{};:'"\,<>./?@#$%^&*_~''')
    temp=[]
    for char in filtered_sentence:
        if char not in punctuations:
            temp.append(char)
    filtered_sentence=temp
    #print(filtered_sentence)
    return filtered_sentence
```

```
def distinct(doc,li):
    for items in doc:
        if items not in li:
            li.append(items)
doc1=cleaner("Doc 1.txt")
doc2=cleaner("Doc 2.txt")
doc3=cleaner("Doc 3.txt")
doc4=cleaner("Doc 4.txt")
doc5=cleaner("Doc 5.txt")
doc6=cleaner("Doc 6.txt")
doc7=cleaner("Doc 7.txt")
doc8=cleaner("Doc 8.txt")
doc9=cleaner("Doc 9.txt")
doc10=cleaner("Doc 10.txt")
dislist=[]
finaldic={}
posdic={}
def discounter(docnamev,dicvarname):
    temp={}
    temp2={}
    distinct(docnamev,dislist)
    for item in dislist:
        c=docnamev.count(item)
        temp[item]=c
        temp3=[]
        for stuff in range(len(docnamev)):
            if item==docnamev[stuff]:
                temp3.append(stuff)
        temp2[item]=temp3
    finaldic[dicvarname]=temp
```

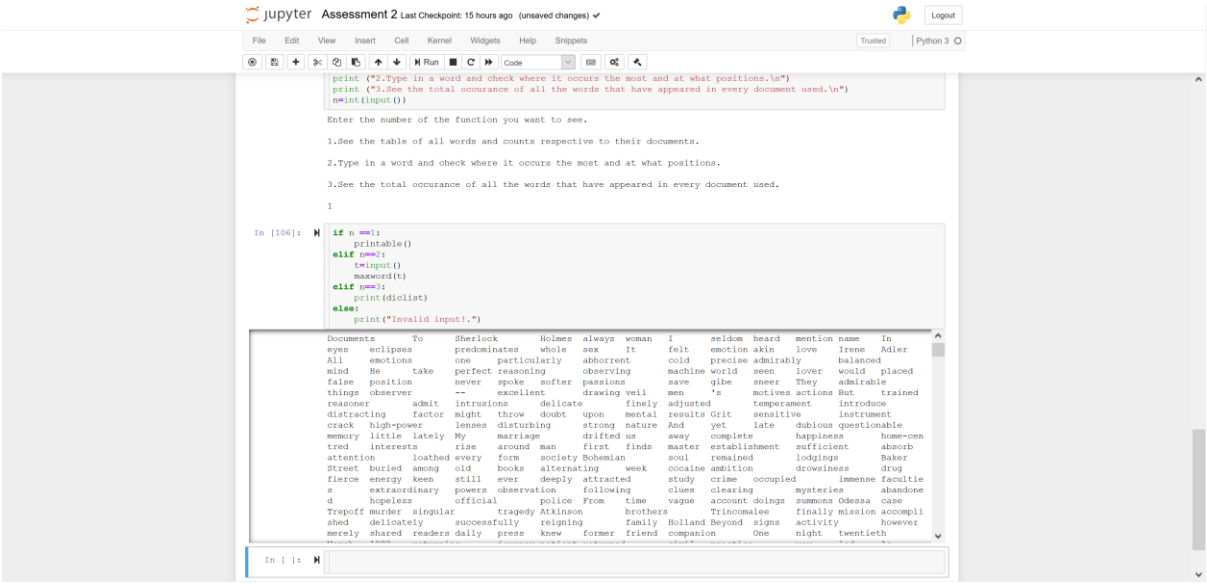
```
    posdic[dicvarname]=temp2
    discounter(doc1,"Document 1")
discounter(doc2,"Document 2")
discounter(doc3,"Document 3")
discounter(doc4,"Document 4")
discounter(doc5,"Document 5")
discounter(doc6,"Document 6")
discounter(doc7,"Document 7")
discounter(doc8,"Document 8")
discounter(doc9,"Document 9")
discounter(doc10,"Document 10")
#run it twice because updation distinct list
discounter(doc1,"Document 1")
discounter(doc2,"Document 2")
discounter(doc3,"Document 3")
discounter(doc4,"Document 4")
discounter(doc5,"Document 5")
discounter(doc6,"Document 6")
discounter(doc7,"Document 7")
discounter(doc8,"Document 8")
discounter(doc9,"Document 9")
discounter(doc10,"Document 10")
def printable():
    print("Documents",end='')
    for item in dislist:
        print("\t"+item,end='')
    for item in finaldic:
        print(item,end='\t')
        for stuff in finaldic[item]:
            print(finaldic[item][stuff],end="\t")
        print("\n")
```

```
#suming up the number of occurance of each words
diclist={}#dictionary of sum of each word occurance in all the documents
for item in dislist:
    t=0
    for stuff in finaldic:
        t+=finaldic[stuff][item]
    diclist[item]=t
#here if you print diclist you get the total occurance of all the words that have appeared in every document used.
def maxword(a):#this finds out if the word exists and if it does then it finds where it occurs maximum and also prints out it's positions
    maxc=0
    maxdoc=0
    if a in dislist:
        for item in finaldic:
            if ((finaldic[item][a])> maxc):
                maxc=finaldic[item][a]
                maxdoc=item
        print ("The word {0} has occured the most in {1}, {2} number of times, in positions:".format(a,maxdoc,maxc))
        print(posdic[maxdoc][a])
    else:
        print ("This word does not appear in the documents")
print ("Enter the number of the function you want to see.\n")
print ("1.See the table of all words and counts respective to their documents.\n")
print ("2.Type in a word and check where it occurs the most and at what positions.\n")
print ("3.See the total occurance of all the words that have appeared in every document used.\n")
n=int(input())
if n ==1:
    printable()
elif n==2:
    t=input()
```

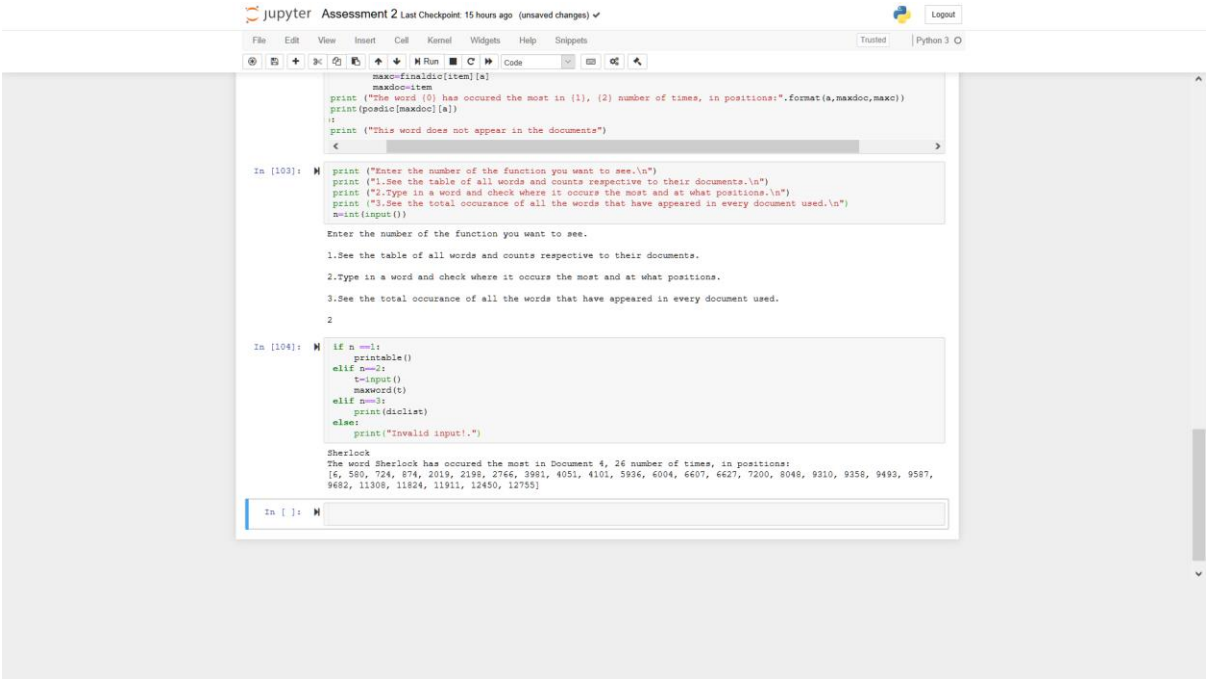
```
    maxword(t)
elif n==3:
    print(diclist)
else:
    print("Invalid input!.")
```

Output:

Table:



Word query:



Sum of occurance: