Assessment 1

Name : Loukik Bhangale

Reg No: 17BCE0961

Slot : L9+L10

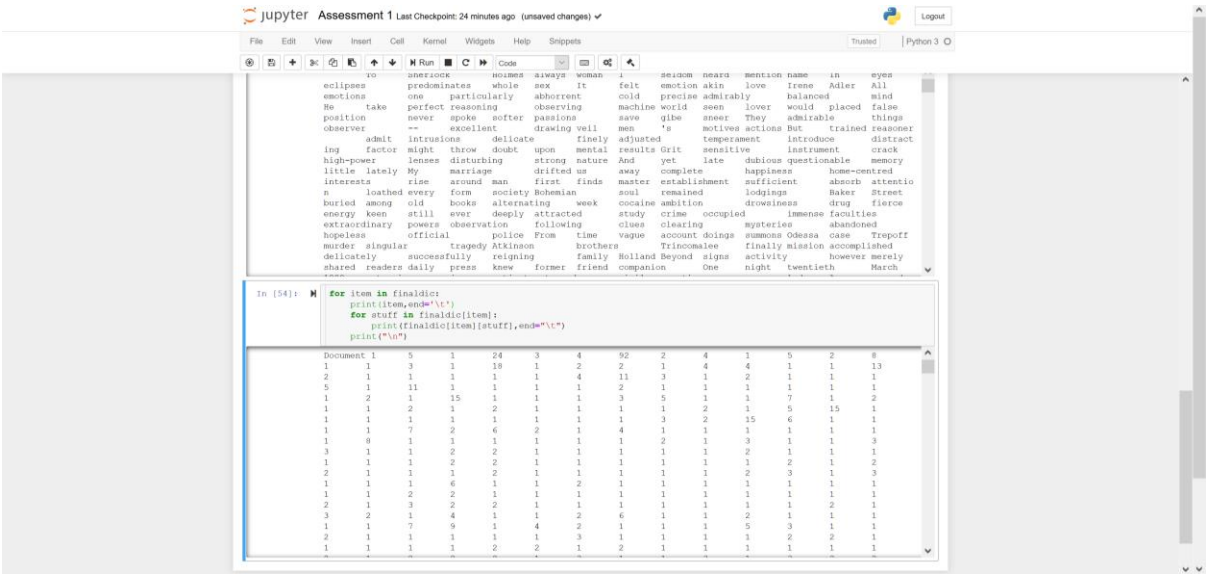
Code:

```
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={}
def discounter(docnamev,dicvarname):
    temp={}
    distinct(docnamev,dislist)
    for item in dislist:
        c=docnamev.count(item)
        temp[item]=c
    finaldic[dicvarname]=temp
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")
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")
```

Output:



This is a table, there are too words to fit in a line hence it is displayed like this.