Web mining Assignment 1

19BCE2311 Gaurav Singh

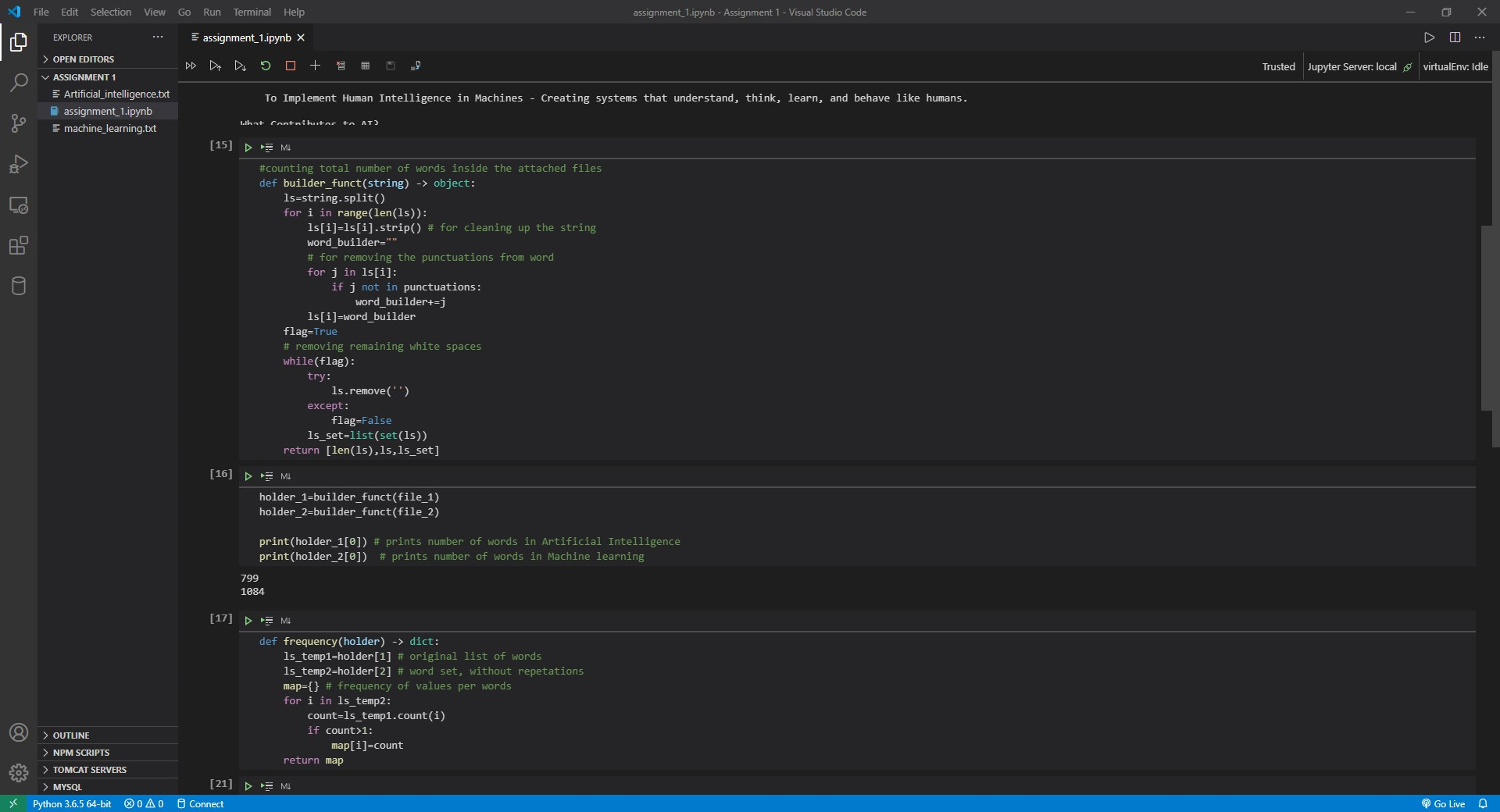
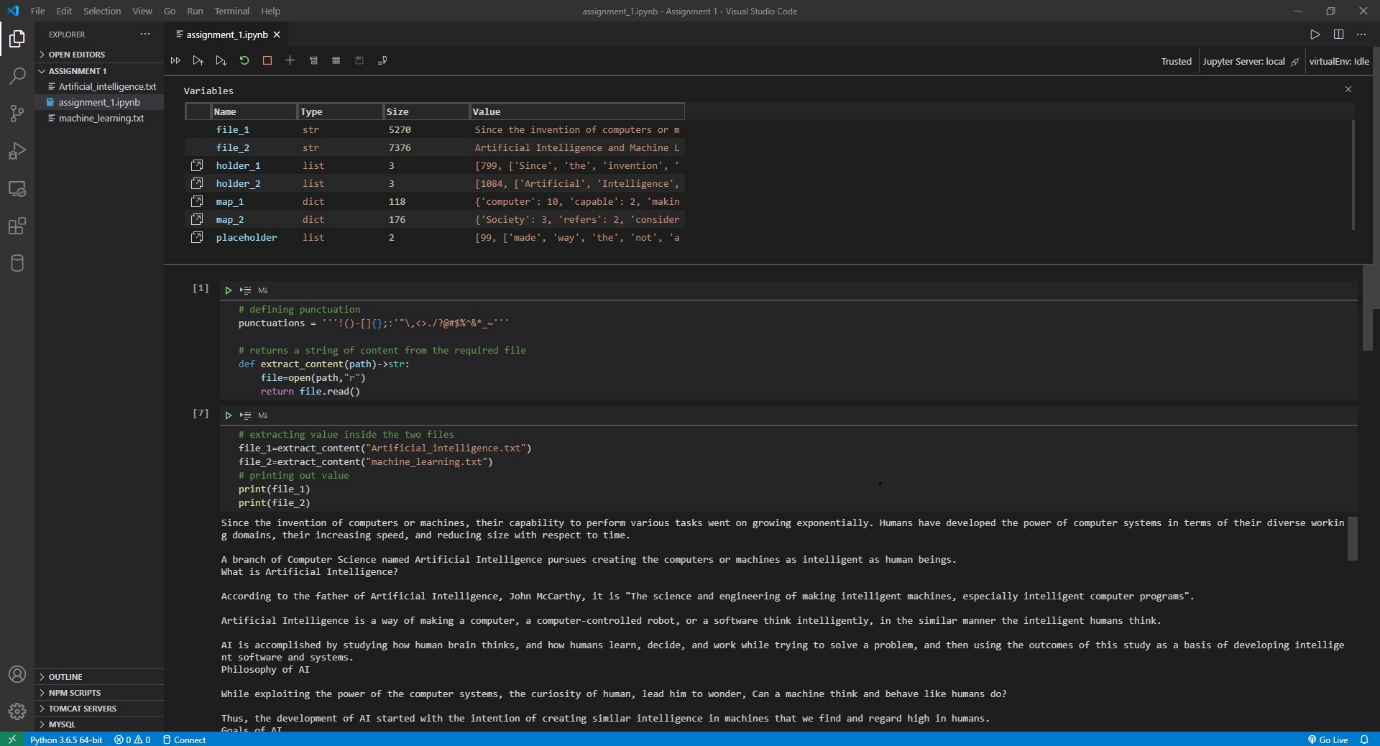
Question:

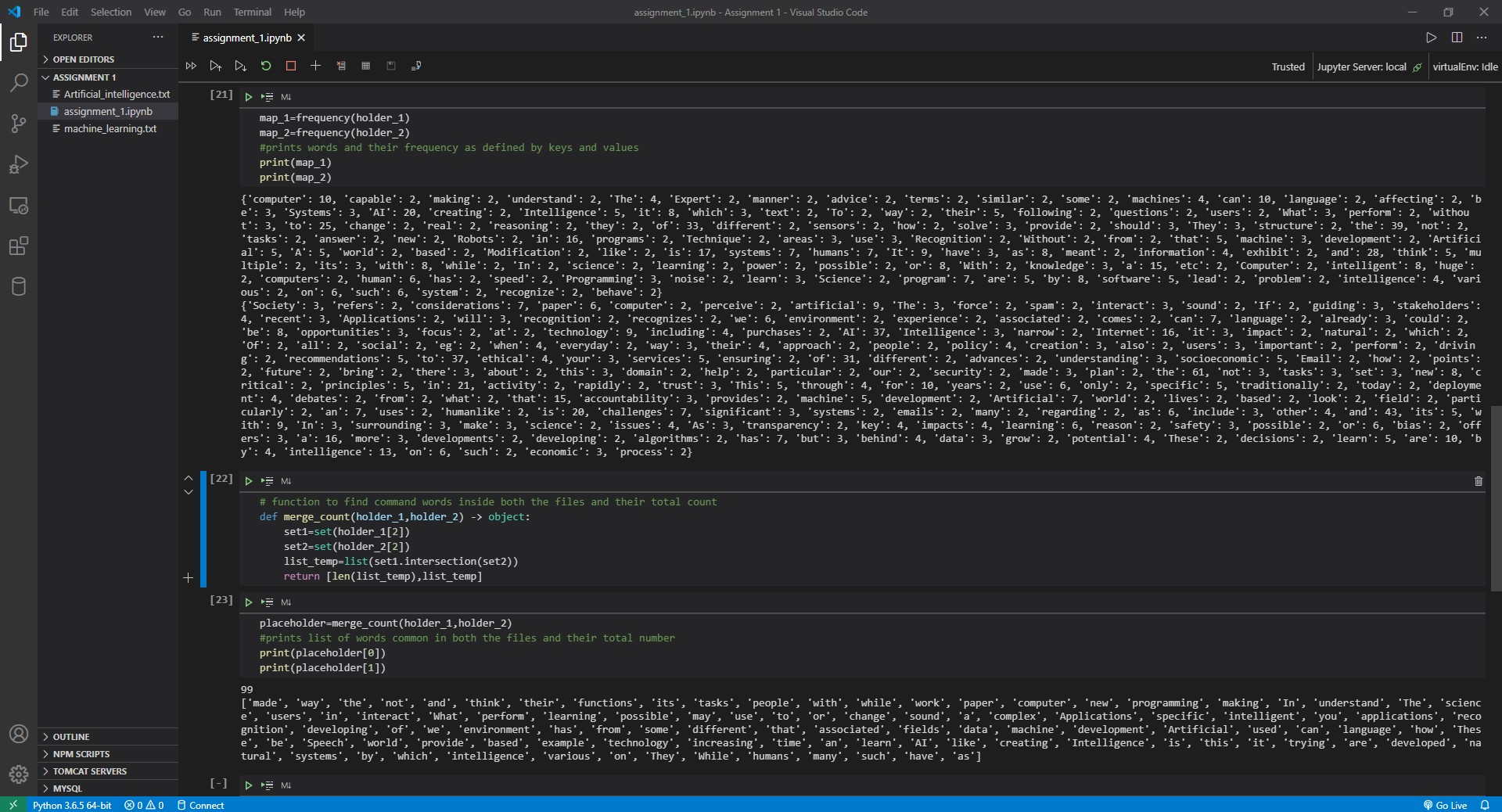
**Practice Programming Exercise:**

Write a python program to

* Extract content from two text files attached
* Count the total number of words in each text file
* Count the frequency of repetition of each word found in each file.
* Group in a list the words common for two text files and show their total count

Screenshots :





Code :

# defining punctuation

punctuations = '''!()-[]{};:'"\,<>./?@#$%^&\*\_~'''

# returns a string of content from the required file

def extract\_content(path)->str:

    file=open(path,"r")

    return file.read()

# extracting value inside the two files

file\_1=extract\_content("Artificial\_intelligence.txt")

file\_2=extract\_content("machine\_learning.txt")

# printing out value

print(file\_1)

print(file\_2)

#counting total number of words inside the attached files

def builder\_funct(string) -> object:

    ls=string.split()

    for i in range(len(ls)):

        ls[i]=ls[i].strip() # for cleaning up the string

        word\_builder=""

        # for removing the punctuations from word

        for j in ls[i]:

            if j not in punctuations:

                word\_builder+=j

        ls[i]=word\_builder

    flag=True

    # removing remaining white spaces

    while(flag):

        try:

            ls.remove('')

        except:

            flag=False

        ls\_set=list(set(ls))

    return [len(ls),ls,ls\_set]

holder\_1=builder\_funct(file\_1)

holder\_2=builder\_funct(file\_2)

print(holder\_1[0]) # prints number of words in Artificial Intelligence

print(holder\_2[0])  # prints number of words in Machine learning

def frequency(holder) -> dict:

    ls\_temp1=holder[1] # original list of words

    ls\_temp2=holder[2] # word set, without repetations

    map={} # frequency of values per words

    for i in ls\_temp2:

        count=ls\_temp1.count(i)

        if count>1:

            map[i]=count

    return map

map\_1=frequency(holder\_1)

map\_2=frequency(holder\_2)

#prints words and their frequency as defined by keys and values

print(map\_1)

print(map\_2)

# function to find command words inside both the files and their total count

def merge\_count(holder\_1,holder\_2) -> object:

    set1=set(holder\_1[2])

    set2=set(holder\_2[2])

    list\_temp=list(set1.intersection(set2))

    return [len(list\_temp),list\_temp]

placeholder=merge\_count(holder\_1,holder\_2)

#prints list of words common in both the files and their total number

print(placeholder[0])

print(placeholder[1])