Web mining Assignment 1

19BCE2311 Gaurav Singh

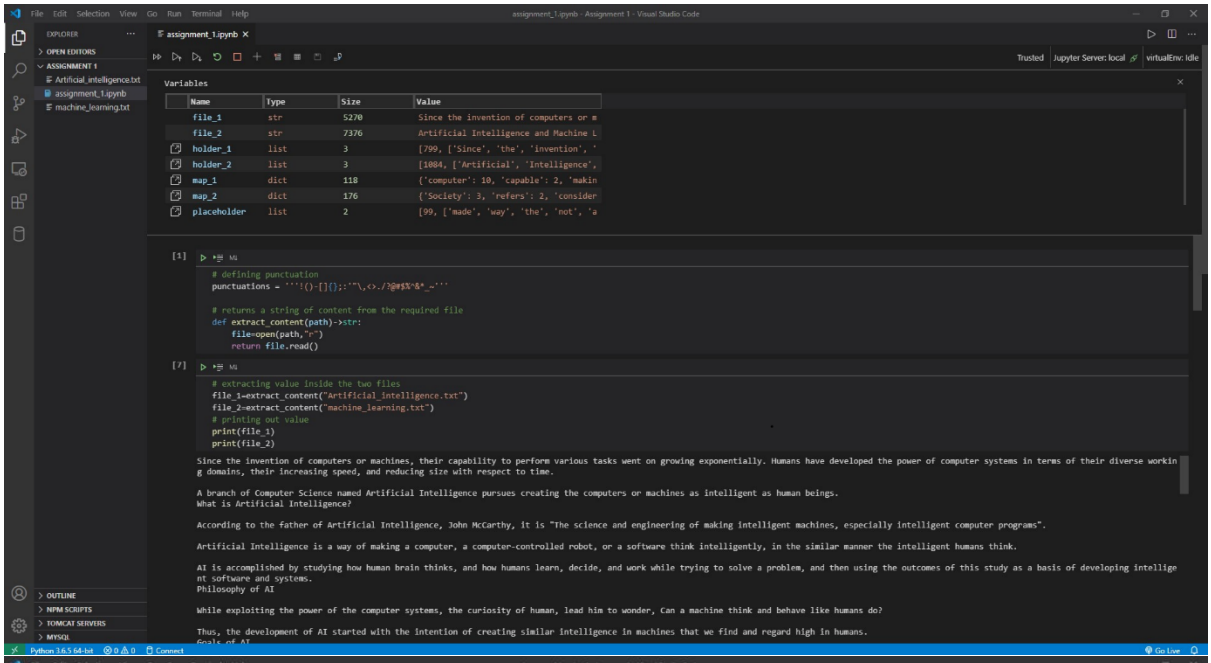
Question:

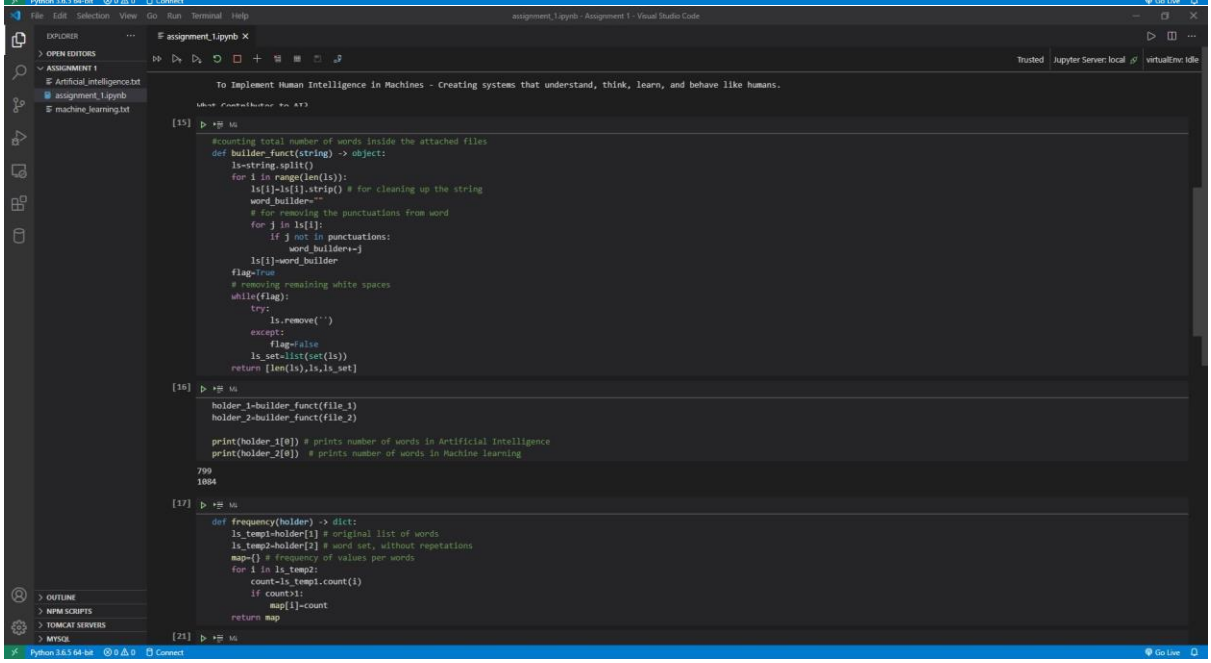
Practice Programming Exercise:

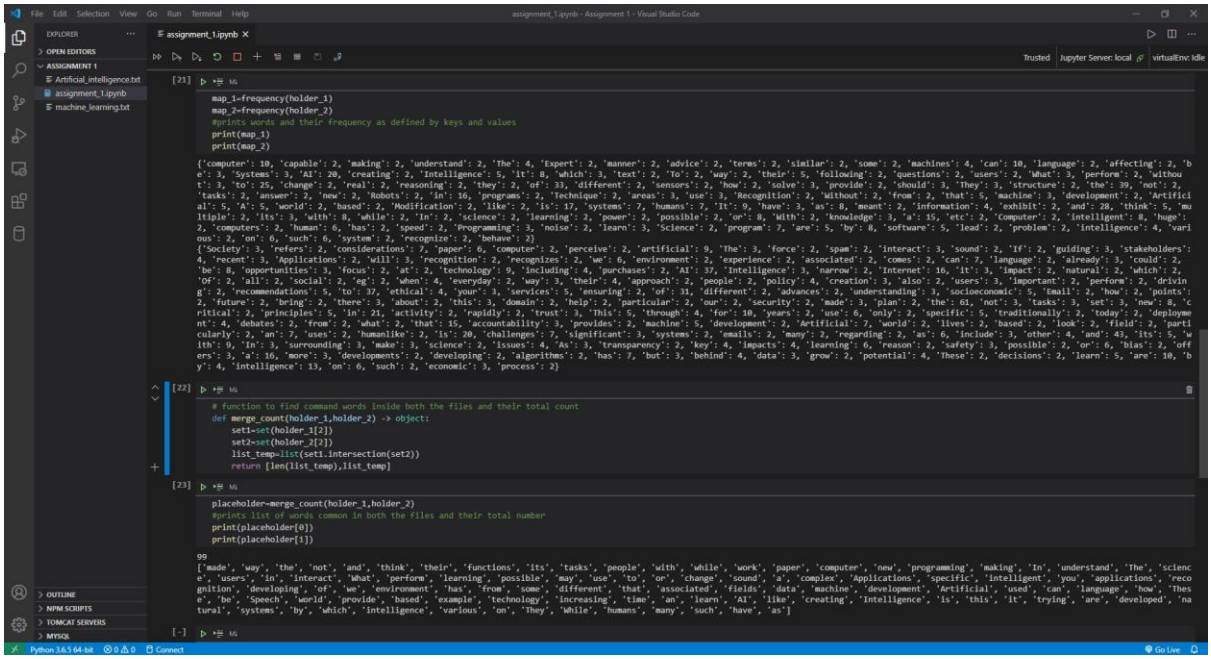
Write a python program to

- 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])
```