Tanusha G
20125

Name: Tanusha Gudise

Roll. No: CH.EN.U4CSE20125

Discord Server: Tanusha Tanu#0880

TASK 6 [PYTHON - MEDICORE LVL]

QUESTION-1 [Code compiled in Python.org idle]

Write a python program that reads the contents from the given file 'onelinefile.txt'. The file contains a single line which is of the format (int)(string)(float)(string) repeatedly. For e.g.

1Aaa3.5Maths2Bbb4.2Physics3Ccc7.62Chemistry

Your main task is to split the contents of the given file based on their format and write it into a .csv file say 'Filename2.csv'. For e.g. the above txt file should be converted into a csv file such that the contents look like this:

1,Aaa,3.5,Maths

2,Bbb,4.2,Physics

3,Ccc,7.62,Chemistry

Contents of 'onelinefile.txt'

1Aaa3.5Maths2Bbb4.2Physics3Ccc7.62Chemistry4Ddd9.55Biology5Eee4.0Social6Fff7.6English7Ggg3.111Maths8Hhh9.99Physics9Iii1.23Civics

Q1.py - C:\Users\Tanusha\Desktop\Cognizance\Task-6\Q1.py (3.10.4)

File Edit Format Run Options Window Help

```python
#Q1
import re,csv
fh = open('onelinefile.txt')
for i in fh:
        r = re.findall(r'[+-]?[0-9]+\.[0-9]+', i)
        q = re.findall(r'[a-zA-Z]+', i)
        j = 0
        for p in range(len(r)):
            with open('onelinefile.csv', 'a', newline='') as file:
                writer = csv.writer(file)
                writer.writerow([str(p+1), q[j],r[p],q[j+1]])
            j += 2

with open('onelinefile.csv', 'r',) as file:
    reader = csv.reader(file)
    for row in reader:
        print(','.join(row))
```

Tanusha G
20125

OUTPUT:

```
1,Aaa,3.5,Maths
2,Bbb,4.2,Physics
3,Ccc,7.62,Chemistry
4,Ddd,9.55,Biology
5,Eee,4.0,Social
6,Fff,7.6,English
7,Ggg,3.111,Maths
8,Hhh,9.99,Physics
9,Iii,1.23,Civics
```

QUESTION-2 [Code compiled in Google Collab]

Data formatting

Python libraries represent missing numbers as nan which is short for "not a number". Most libraries (including scikit-learn) will give you an error if you try to build a model using data with missing values. One of the common solution to get around this issue is to impute or fill in the missing value with a number or value of same format. From the given dataset, find the missing values(Nan/NA/-/Nil) and change those values into an appropriate number.

Tanusha G
20125

Cognizance Task-6

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

```
import pandas as pd
import numpy as np
df = pd.read_csv("https://raw.githubusercontent.com/cognizance-amrita/AI-Tasks/main/Task-1/Q2-Dataset.csv")
df.head(100)

print(df['LotFrontage'].isnull())

print(df.isnull().sum())

df['LotFrontage'].fillna(1, inplace=True)

print(df['LotFrontage'])

print(df['Alley'].isnull())

df['Alley'].fillna('alley not mentioned here', inplace=True)
print(df['Alley'])

print(df['BsmtQual'].isnull())

df[df['BsmtQual'].isnull()]

df['BsmtQual'].fillna('value not mentioned', inplace=True)
```

```
df.tail(10)

df[df['BsmtQual'].isnull()]

print(df['BsmtCond'].isnull())

df[df['BsmtCond'].isnull()]

df['BsmtCond'].fillna('condition not mentioned', inplace=True)
df.tail(10)

df[df['BsmtCond'].isnull()]

print(df['BsmtExposure'].isnull())

df[df['BsmtExposure'].isnull()]

df['BsmtExposure'].fillna('exposure not mentioned', inplace=True)
df.tail(10)

df[df['BsmtExposure'].isnull()]
```

```
df[df['BsmtFinType1'].isnull()]

df['BsmtFinType1'].fillna('value not assigned ', inplace=True)
df.tail(10)

df[df['BsmtFinType1'].isnull()]

print(df['BsmtFinType2'].isnull())

df[df['BsmtFinType2'].isnull()]

df['BsmtFinType2'].fillna('values not found', inplace=True)
df.tail(10)

df[df['BsmtFinType2'].isnull()]

print(df.isnull().sum())
```

Tanusha G
20125

OUTPUT:

```
0     False
1     False
2     False
3     False
4     False
      ...
94    False
95     True
96    False
97    False
98    False
Name: LotFrontage, Length: 99, dtype: bool
Id               0
MSSubClass       0
MSZoning         0
LotFrontage     14
LotArea          0
Street           0
Alley           93
LotShape         0
LandContour      0
Utilities        0
LotConfig        0
LandSlope        0
Neighborhood     0
Condition1       0
Condition2       0
```

```
BldgType         0
HouseStyle       0
OverallQual      0
OverallCond      0
YearBuilt        0
YearRemodAdd     0
RoofStyle        0
RoofMatl         0
Exterior1st      0
Exterior2nd      0
MasVnrType       0
MasVnrArea       0
ExterQual        0
ExterCond        0
Foundation       0
BsmtQual         3
BsmtCond         3
BsmtExposure     3
BsmtFinType1     3
BsmtFinSF1       0
BsmtFinType2     3
dtype: int64
0     65.0
1     80.0
2     68.0
3     60.0
4     84.0
```

Tanusha G
20125

```
      ...
94    69.0
95     1.0
96    78.0
97    73.0
98    85.0
Name: LotFrontage, Length: 99, dtype: float64
0     True
1     True
2     True
3     True
4     True
      ...
94    True
95    True
96    True
97    True
98    True
Name: Alley, Length: 99, dtype: bool
0     alley not mentioned here
1     alley not mentioned here
2     alley not mentioned here
3     alley not mentioned here
4     alley not mentioned here
                ...
94    alley not mentioned here
95    alley not mentioned here
96    alley not mentioned here
97    alley not mentioned here
98    alley not mentioned here
Name: Alley, Length: 99, dtype: object
0     False
1     False
2     False
3     False
4     False
      ...
94    False
95    False
96    False
97    False
98    False
Name: BsmtQual, Length: 99, dtype: bool
0     False
1     False
2     False
3     False
4     False
      ...
94    False
95    False
96    False
97    False
98    False
```

Tanusha G
20125

```
Name: BsmtCond, Length: 99, dtype: bool
0     False
1     False
2     False
3     False
4     False
      ...
94    False
95    False
96    False
97    False
98    False
Name: BsmtExposure, Length: 99, dtype: bool
0     False
1     False
2     False
3     False
4     False
      ...
94    False
95    False
96    False
97    False
98    False
Name: BsmtFinType2, Length: 99, dtype: bool
```

```
Id               0
MSSubClass       0
MSZoning         0
LotFrontage      0
LotArea          0
Street           0
Alley            0
LotShape         0
LandContour      0
Utilities        0
LotConfig        0
LandSlope        0
Neighborhood     0
Condition1       0
Condition2       0
BldgType         0
HouseStyle       0
OverallQual      0
OverallCond      0
YearBuilt        0
YearRemodAdd     0
RoofStyle        0
RoofMatl         0
Exterior1st      0
Exterior2nd      0
MasVnrType       0
MasVnrArea       0
```

```
ExterQual        0
ExterCond        0
Foundation       0
BsmtQual         0
BsmtCond         0
BsmtExposure     0
BsmtFinType1     0
BsmtFinSF1       0
BsmtFinType2     0
dtype: int64
```

Tanusha G
20125

QUESTION-3 [Code compiled in Python.org idle]

Read the file 'about.txt' and find the words with atleast 6 letters and the most frequently used word.

Contents of the file 'about.txt':

Python has tools for almost every aspect of scientific computing. The Bank of America uses Python to crunch its financial data and Facebook looks upon the Python library Pandas for its data analysis. While there are many libraries available to perform data analysis in Python, here are a few: NumPy, SciPy, Pandas and Matplotlib.

Q3.py - C:\Users\Tanusha\Desktop\Cognizance\Task-6\Q3.py (3.10.4)

File Edit Format Run Options Window Help

```
#Q3

import re
with open('about.txt','r') as file:
    contents =file.read()
    string = re.sub('[^a-zA-Z\d\s]', '', contents)
    x=string.split()
    ans = max(x,key=x.count)
    print("Most frequently used word is:",ans)
```

OUTPUT:

```
Most frequently used word is: Python
```