



Amrita School of Engineering
Chennai – 601 103, Tamil Nadu, India.

AMRITA SCHOOL OF ENGINEERING

Degree of Bachelor of Technology in
COMPUTER SCIENCE & ENGINEERING

COGNIZANCE TASK-3

[Python –Medicore Lvl]

Submitted by :

GUNDE ELIJAJER

CSE-A

CH.EN.U4CSE20027

Question-1:

File Handling is one of the basic important task when it comes to building machine learning models or neural networks. Building a good model always starts with finding datasets and processing it, for which, file handling acts as a stepping stone.

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.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
ds = pd.read_csv("filename2.csv")
ds.head()
```

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

```

def split_file(filename2):

    header=0
    header_line=""
    file_count=0
    for line in filename2:
        line=line.rstrip()
        a=line.split()
        if header==0:
            header_line=line
            header+=1
        else:
            if a[-1] not in l:
                l.append(a[-1])
                file_count+=1
                if file_count>1:
                    dest.close()
                else:
                    pass
            dest=open(a[-1], 'a')
            dest.write(header_line+"\n"+line+"\n")
        else:
            dest.write(line+"\n")
    source.close()
    dest.close()

```

OUTPUT:-

```

1,Aaa,3.5,Maths
2,Bbb,4.2,Physics
3,Ccc,7.62,Chemistry

```

Question-2

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.

```
import pandas as pd
data = pd.read_csv("dataset.csv")
```

```
[ ] data.describe()
```

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1
count	99.000000	99.000000	85.000000	99.000000	99.000000	99.000000	99.000000	99.000000	99.000000	99.000000
mean	50.000000	52.323232	71.341176	10123.909091	6.000000	5.505051	1971.535354	1984.171717	122.868687	453.404040
std	28.722813	42.638209	21.210146	5156.018489	1.525297	0.983265	29.495871	20.921642	214.285221	464.240897
min	1.000000	20.000000	21.000000	1596.000000	3.000000	2.000000	1910.000000	1950.000000	0.000000	0.000000
25%	25.500000	20.000000	60.000000	7628.500000	5.000000	5.000000	1954.500000	1965.000000	0.000000	0.000000
50%	50.000000	45.000000	70.000000	9600.000000	6.000000	5.000000	1970.000000	1995.000000	0.000000	384.000000
75%	74.500000	60.000000	84.000000	11245.500000	7.000000	6.000000	2000.500000	2003.500000	191.000000	738.000000
max	99.000000	190.000000	122.000000	50271.000000	10.000000	8.000000	2009.000000	2009.000000	1115.000000	1880.000000

```
[ ] data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 36 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Id                    99 non-null    int64
1   MSSubClass            99 non-null    int64
2   MSZoning              99 non-null    object
3   LotFrontage          85 non-null    float64
4   LotArea               99 non-null    int64
5   Street               99 non-null    object
6   Alley                6 non-null     object
7   LotShape              99 non-null    object
8   LandContour           99 non-null    object
9   Utilities             99 non-null    object
10  LotConfig             99 non-null    object
11  LandSlope             99 non-null    object
12  Neighborhood          99 non-null    object
13  Condition1            99 non-null    object
14  Condition2            99 non-null    object
15  BldgType              99 non-null    object
16  HouseStyle            99 non-null    object
17  OverallQual           99 non-null    int64
18  OverallCond           99 non-null    int64
19  YearBuilt             99 non-null    int64
20  YearRemodAdd          99 non-null    int64
21  RoofStyle             99 non-null    object
22  RoofMatl              99 non-null    object
23  Exterior1st           99 non-null    object
24  Exterior2nd           99 non-null    object
```

```
[ ] data.duplicated(keep="first")
```

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

```
[ ] data_without_missing_values = data.dropna(axis=1)
```

```
[ ] cols_with_missing = [col for col in data.columns
                          if data[col].isnull().any()]
reduced_original_data = data.drop(cols_with_missing, axis=1)
reduced_test_data = data.drop(cols_with_missing, axis=1)
```

```
[10] one_hot_encoded_training_predictors = pd.get_dummies(train_predictors)
one_hot_encoded_training_predictors[:5]
```

	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSf1	MSZoning_C (all)	MSZoning_FV	MSZoning_RL	...	ExterQual_Gd	Ex
0	8450	7	5	2003	2003	196	706	0	0	1	...	1	
1	9600	6	8	1976	1976	0	978	0	0	1	...	0	
2	11250	7	5	2001	2002	162	486	0	0	1	...	1	
3	9550	7	5	1915	1970	0	216	0	0	1	...	0	
4	14260	8	5	2000	2000	350	655	0	0	1	...	1	

5 rows × 72 columns

```
cols_with_missing = [col for col in X_train.columns if X_train[col].isnull().any()]
reduced_X_train = X_train.drop(cols_with_missing, axis=1)
reduced_X_test = X_test.drop(cols_with_missing, axis=1)
print("Mean Absolute Error after dropping columns with missing values:")
print(score_dataset(reduced_X_train, reduced_X_test, y_train, y_test))
```

```
Mean Absolute Error after dropping columns with missing values:
26.65
```

Question-3

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.

```
[ ] import pandas as pd
import matplotlib.pyplot as plt

[ ] data = pd.read_csv('about.csv')

[ ] data.describe()
```

	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.
count	0	0	0	0
unique	0	0	0	0
top	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN


```
[1] count = 0;
word = "";
maxCount = 0;
words = [];
file = open("about.csv", "r")
for line in file:
    string = line.lower().replace(',','').replace('.', '').split(" ");
    for s in string:
        if len(s)>=6: #len greater than 6
            words.append(s);
for i in range(0, len(words)):
    count = 1;
    for j in range(i+1, len(words)):
        if(words[i] == words[j]):
            count = count + 1;
    if(count > maxCount):
        maxCount = count; #frequency
        word = words[i];

print("Most repeated word: " + word);
print("Frequency: " ,maxCount);
file.close();
```

Output:-

```
➤ Most repeated word: python
  Frequency: 4
```

-----THE END-----